Pacific Graphics 2012
C. Bregler, P. Sander, and M. Wimmer
(Guest Editors)

Volume 31 (2012), Number 7

Performance Capture of High-Speed Motion Using Staggered Multi-View Recording

Di Wu[1,2], Yebin Liu[1], Ivo Ihrke[3,4], Qionghai Dai[1], and Christian Theobalt[3]

[1]Deptartment of Automation, Tsinghua University, [2]Graduate School at Shenzhen, Tsinghua University , [3]MPI Informatik, [4]Saarland University

Abstract

We present a markerless performance capture system that can acquire the motion and the texture of human actors performing fast movements using only commodity hardware. To this end we introduce two novel concepts: First, a staggered surround multi-view recording setup that enables us to perform model-based motion capture on motion-blurred images, and second, a model-based deblurring algorithm which is able to handle disocclusion, self-occlusion and complex object motions. We show that the model-based approach is not only a powerful strategy for tracking but also for deblurring highly complex blur patterns.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Computer Graphics]: Scene Analysis—Time-Varying Imagery

1 Introduction

In recent years, professional movie and game productions have shown increasing interest in non-intrusive markerless technology to capture human performances that has been developed in the research community. Performance capture systems record an actor with a camera array or a combination of cameras and controlled lighting, and use computer vision algorithms to reconstruct detailed models of dynamic geometry, texture or reflectance, e.g., [dAST*08, VPB*09, ECJ*06]. The captured models can serve as basis for high-quality animation content.

Performance capture approaches, however, are still reaching their limits when the captured scene is moving extremely fast, such as during a martial arts kick. The reason is that they typically rely on standard video cameras that operate at much lower frame rate than, for instance, specialized marker cameras in marker-based motion estimation. When capturing rapid motion, the performance is thus often temporally under-sampled and the images are motion blurred, such that neither geometry nor texture or reflectance can be recovered.

Several options exist to overcome the frame rate bottleneck. Special high-speed video cameras could be used [WGT*05]. However, they are very expensive, have tremendous bandwidth requirements, often only capturing to RAM, and require extremely strong illumination due to their very short exposure times. Using only a handful of them may already be infeasible. Arrays of closely-spaced coaxial standard video cameras could also be used. As shown by Wilburn et al. [WJV*04] such systems, triggered in a staggered sequence, can yield temporally highly resolved video footage. Alternatively, in order to overcome the strong illumination constraint, they can be run in a coded sampling fashion to collect more light [AGVN10]. However, in this arrangement, every viewpoint has to be observed by a multitude of cameras to achieve the desired effect, resulting in considerable expense of the resulting system. These approaches are thus, so far, limited to record a single viewpoint.

In this paper, we propose a new 360^o high-speed multi-view performance capture approach that reconstructs rapidly moving actors with standard lighting and multiple standard video cameras placed around the scene. Despite being based on off-the-shelf low frame rate cameras, it allows us to reconstruct both the shape and the dynamic surface texture at a much higher effective frame rate than the physical frame rate of each individual camera, Fig. 1. We purposefully accept and even exploit the fact that captured video frames are blurred, and use clever exposure sequencing, model-based scene reconstruction, and 3D model-based deblurring to recover the scene content at a very high effective temporal sampling rate. In particular, we make the following major contributions.

First, the standard video cameras that are placed around the scene are capturing in a time-shifted exposure sequence, Sec. 3. Each individual camera's exposure time is relatively long.

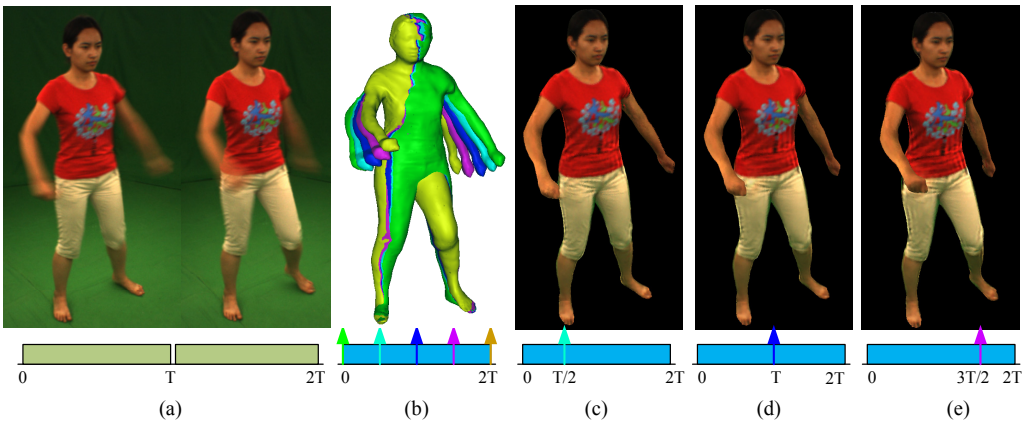


Figure 1: *When recording fast motion with standard video cameras, successive frames recorded with exposure time T are blurred (a). We propose a method to capture shape and unblurred multi-view textures of rapidly moving scenes from blurry multi-view images recorded in a temporally staggered order. We can capture 3D scene geometry at a much higher frame rate than the recording rate of the cameras, i.e., at sub-frames within an original long exposure interval - an overlay of such sub-frame geometries is shown in (b). Our algorithm also recovers deblurred textures for the sub-frame models (c)-(e).*

But the exposure intervals of different cameras do not start at the same time but in a staggered sequence. We show that by using 20 cameras recording at 20 fps, we can sample time and space densely, and capture shape and multi-view texture at an effective rate of 400 fps.

Second, we use a model-based marker-less approach to capture the dynamic scene geometry of the human actor at the high effective frame rate, Sec. 4. This algorithm expects as input multi-view silhouette images at the effective frame rate of the system. These effective silhouette images can be computed from the blurred silhouette images of each video camera through spatio-temporal intersection based on the staggered exposure sequence.

Finally, the ability to follow scene motion with a 3D shape template enables us to properly model the image formation process that leads to blurred video frames. Since we know both, the scene shape and motion, we can form a hypothesis on the individual blur kernel for each 3D scene point as it projects into any of the available camera views. We exploit this feature to develop a 3D model-based deblurring algorithm which can handle important effects like occlusion, self-occlusion and out-of-plane rotation. A robust regularization framework enables the reconstruction of deblurred textures despite potential inaccuracies in our scene model, Sec. 5.

The final result of our method is a three-dimensional, dynamic, completely textured shape model of a human actor at high-speed frame rates which can be displayed from synthetic view points, at arbitrary speeds.

2 Related work

Our work touches on a number of important subjects with a rich number of publications each. We therefore restrict ourselves to an exemplary overview. The three main approaches to capture fast motions are high-speed imaging using specialized hardware, capture using high-speed illumination in controlled environments, and deconvolution on motion-blurred imagery, see Wetzstein et al. [WILH11] for a survey.

Specialized hardware solutions suffer from high band-width requirements, forcing a capture-to-RAM scheme that limits the capture ability to only a few seconds. Another restriction in multi-view applications is that synchronization of these cameras is not easily possible. Alternatively, spatial sensor resolution can be traded for temporal resolution. A DMD-based implementation [BTH*10] of the assorted pixels framework [BEZN05] has recently been demonstrated. In addition to single camera high-speed imaging, multi-camera alternatives with staggered exposure capturing strategy have been explored, e.g. by Shechtman et al. [SCI02, SCI05], Wilburn et al. [WJV*04, WJV*05] and Li et al. [LDXY12].

High-Speed Illumination to picture high-speed events has been pioneered by Harold Edgerton. Theobalt et al. [TAH*04] use photo-cameras and strobe illumination to capture high-speed motion of an arm and a ball in a single exposure, but do not recover texture or complex geometry. Recently, it has been demonstrated that arrays of consumer video cameras can be synchronized to a virtual frame rate by using arrays of strobe lights [BAIH09]. However, the virtual frame rate cannot surpass that of the cameras. DLP projectors have been used to code temporal events for computer vision tasks such as dynamic structured light [NKY08]. Veeraraghavan et al. [VRR11] show that using coded strobe illumination combined with sparse recovery techniques can yield high frame rate videos of periodic events.

Image deblurring methods critically depend on the quality

of the PSF estimates. Hardware-based blur kernel estimation can be performed using inertial measurement sensors [JKZS10] or a dual camera setup [BEN04, THBL10], where one camera records high spatial resolution/low frame rate video while the second one provides high frame rate/low resolution video. In contrast, designed PSFs can simplify and stabilize the deblurring problem considerably [RAT06]. Levin et al. [LSC*08] implement a parabolic camera shake and demonstrate that this way the PSF can be made invariant to the speed of the object motion.

Purely software-based methods, on the other hand, are referred to as blind deconvolution methods. They come in different flavors: single image vs. multiple image methods, spatially invariant PSFs vs. spatially varying ones, and methods that assume some underlying scene model vs. methods that do without. An example for spatially invariant blind deconvolution for single images is the work by Fergus et al. [FSH*06]. In subsequent work this was improved by considering errors on the estimated PSFs in a Bayesian framework [SJA08]. A recent example of multiple image based motion deblurring is [CML07].

Alternatively, the PSF can be estimated directly by detecting blurred versions of sharp edges. Using differently oriented edges throughout the image, a spatially invariant blur kernel can be estimated [JSK08]. Recently, it has been observed that essentially low-parametric spatially varying PSFs can be recovered by including model assumptions about the blur origin. It has been demonstrated that, in the case of a static scene observed by a shaking camera, a set of homographies can model the scene sufficiently well for high-quality deblur results [WSZP10, TTB11]. Ding et.al [DMY10] describe model-based deconvolution in two dimensions by selecting a PSF family (linear, parabolic, or oscillating).

Performance Capture, 3D Video and Dynamic Scene Reconstruction methods aim at recovering a virtual spatio-temporal representation of a given scene. Several types of approaches for dynamic scene reconstruction from multi-view video were proposed (see e.g. [TWdAN07] for an overview). Geometry-based 3D video methods reconstruct the scene using some form of shape-from-silhouette or stereo method [ZKU*04, WWC*05, TNM09, FP09] or fit a template model to the data [CTMS03]. Performance capture methods extend these ideas, and enable further refined model reconstruction from video in a spatio-temporally coherent way, either by a variant of shape-from-silhouette and correspondence finding, through stereo, or by fitting a deformable template to the images [dAST*08, VBMP08, BPS*08, VPB*09, CBI10]. In the geometry community, animation reconstruction approaches [TBW*12, LLV*12, BHLW12] mainly consider point data from dynamic 3D scanners as input, sometimes, however, exploiting texture information for stabilization purposes [LLV*12]. All of these approaches are challenged by extremely fast motions.

One of the first performance capture approaches to employ high speed cameras was the method of Wenger et al. [WGT*05]. The cameras record the images of a performer under changing illumination from a lighting dome at several hundred fps, such that - effectively - a complete reflectance field is captured for successive frames at normal video rate. The video can then be relit. Later, they extended the approach to full-body and several cameras, enabling viewpoint change through image-based warping [ECJ*06]. The goal in this line of research was to achieve higher temporal sampling of controlled illumination to obtain relightable video at a normal effective frame rate. In our work, we pursue a different goal.

Contributions:

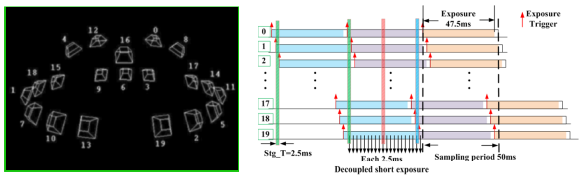
Temporally staggered recording strategies similar to ours have been proposed previously (e.g. [WJV*05, WJV*04] and [SCI02, SCI05]). Li et al. [LDXY12] used staggered recording from a multi-view camera system for capturing motions at high speed with low frame rate cameras. Their strategy differs from ours in several ways. N cameras are clustered into M groups of synchronized cameras, and a combination of shape-from-silhouette reconstruction and image-warping is used to synthesize input images at time instants that have not been sampled. However, their approach does not produce spatio-temporally coherent scene geometry, and thus they can and do not address the deblurring problem for fast motion as we do. Also, their approach struggles with strong occlusions in the scene, and the effective frame rate gain is limited to a factor of N/M as opposed to N in our approach. We use reconstructed approximate geometry to fit the motion of a template model to obtain more accurate high-speed reconstructions of the actual geometry. The temporal motion tracking extracted such can be used as a high-level prior (human body template model) that can predict blur kernels for a much more general class of motions than previously employed models. Previous model-based approaches like e.g. Whyte et al. [WSZP10] are using rather low-level models (homographies in this case) which implies a scene that is essentially planar. Correspondingly, the motions are restricted to rigid body motions of a plane. In contrast, we use a fully articulated model for estimating fully three-dimensional blur paths, respecting occlusion and disocclusion.

3 Acquisition

Our acquisition system consists of $N = 20$ off-the-shelf Point Grey Flea 2 cameras running at a resolution of 1024×768 pixels with a frame rate of 20 fps. The cameras are calibrated, arranged in two complete rings, and are mounted on a dome of roughly 6 m diameter, Fig. 2 (a). To ease backgrund subtraction, we use a green screen background.

The key to high-speed performance capture and model-based deblurring is our spatial and temporal sampling strategy. Camera exposures are triggered in a specific staggered timing sequence as illustrated in Fig. 2 (b). We explain the main design features of this sampling strategy as follows.

Temporal staggering: Each camera is capturing at $f_r = N = 20$ fps, and thus integrates for what we henceforth refer to as a *long exposure interval* $T_l = 1/N\ ms = 50ms$. Here we account for an additional readout and processing



(a) (b)

Figure 2: *Recording Strategy: (a) spatial camera arrangement (for more details please refer to the animation in the supplementary video), (b) temporal triggering strategy.*

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

Figure 3: *Computing long exposure silhouette images: (a) captured blurred long exposure image $I_{c,i}$. (b) Silhouette segmentation $S_{c,i}$ of (a). (c) Overlap image of (a) and (b). (d) 19 short exposure silhouettes $S'_{c,j}$ in a long exposure frame with a color coding of different high speed silhouette indices j.*

time of $2.5ms$ at the end of each exposure. Naturally, each of the corresponding long exposed images $I_{c,i}$ of camera c at frame index i will show motion blur on rapidly moving scenes. However, the cameras are not triggered synchronously. Their exposures are temporally staggered, such that the next camera in the staggered time sequence starts its long exposure $T_s = T_l/N = 2.5ms$ after the previous one, see Fig. 2 (b).

Spatial staggering: It is important that the temporal order in which cameras are triggered covers as much viewpoint diversity as possible. We therefore make sure that two successively exposed cameras have a certain distance in space and are not directly adjacent on the dome. Thus we ensure effective acquisition of directional motion by avoiding spending too much samples on views with little spatial variation.

In the following sections, we explain how we can transform this set of temporally and spatially staggered frames $I_{c,i}$ that were captured at f_r fps into a set of multi-view silhouette images that correspond to a much higher effective frame rate $f_e = f_r \cdot N$. Through model-based marker-less performance capture, the dynamic scene geometry can then be reconstructed at the effective frame rate f_e. Based on this temporally densely reconstructed scene model, we can recover deblurred textures for the captured performance at f_e.

4 Capturing Geometry at High Frame Rate

4.1 Silhouette Extraction from Long Exposure Images

Markerless performance capture algorithms [GSA*09] are based on silhouette information at the effective reconstruction frame rate f_e. In order to compute these images we are processing the motion blurred images in a two-step process: First, we compute silhouette images $S_{c,i}$ of the long exposure frames $I_{c,i}$, these silhouettes include all potentially foreground regions, including motion blur, see Fig. 3 (a)-(c). The long exposure silhouettes can be thought of as the union of all short exposure silhouettes $S'_{c,j}$ that are needed for the motion capture algorithm: $S_{c,i} = \cup_j S'_{c,j}$, see Fig. 3 (d). In a second step, we therefore decompose the long exposure silhouettes into their constituent parts $S'_{c,j}$ by computing silhouettes at the high effective frame rate f_e (see Sec. 4.2). Henceforth index i is used to denote a long exposure cycle number, index j refers to an exposure interval or frame index at the higher effective frame rate f_e. Once the $S'_{c,j}$ are computed, they are fed into a template-based performance capture approach.

Each original long exposure image $I_{c,i}$ exhibits significant blur due to the rapid motion of the person in the foreground, Fig. 3. A long exposure silhoutte image $S_{c,i}$ is a binary image whose value is 1 at every pixel of the potentially blurred foreground, and 0 otherwise, as Fig. 3(b), an overlay image is as Fig. 3(c). We compute such silhouette images by subtracting from each $I_{c,i}$ a background image B_c that was captured for each camera prior to recording. Background subtraction is performed in HSV space to simplify thresholding.

4.2 Computing Silhouettes at High Frame Rate

Silhouette images are reprojections of the 3D scene geometry's visual hull into each camera view [Lau94]. If we can reconstruct the visual hulls $\mathcal{V}_j$ of the moving scene at f_e fps we can thus generate the corresponding $S'_{c,j}$ through projection. Exact reconstruction of the visual hulls from our given input images is of course infeasible. However, due to the spatial and temporal staggering of the long exposure images, we are able to compute a very close approximation to the true visual hull sequence, $\mathcal{V}'_j$.

Let's assume we want to reconstruct $\mathcal{V}'_t$ for short exposure time frame t. As illustrated in Fig. 2(b), such a short exposure time frame t overlaps with exactly one staggered long exposure interval of each camera. Let $\mathcal{C}(t) = (i_{c_1}, \ldots, i_{c_N})$ be the N-tuple of long exposure cycle numbers for each of the cameras c_k, $k = 1..N$ with which t overlaps. The approximate visual hull $\mathcal{V}'_t$ can now be computed by back-projecting the long exposure silhouettes $S_{c_k,i_{c_k}}$ from all cameras c_k at respective cycle index i_{c_k} from $\mathcal{C}(t)$:

$$\mathcal{V}'_t = \bigcap_{i_{c_k} \in \mathcal{C}(t), k=1..N} H(S_{c_k,i_{c_k}}), \qquad (1)$$

where $H(\cdot)$ reprojects a silhouette into a generalized cone in space based on the given camera parameters. These generalized cones are then intersected to obtain the approximate visual hull. An illustration of the concept is shown in Fig. 4 (a). The high frame rate silhouette images $S'_{c,t}$ are now trivially obtained by

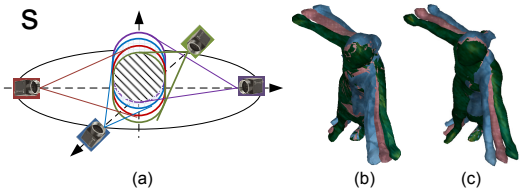


Figure 4: *Computing high-speed visual hulls: (a) intersecting the cones resulting from overlapping long-exposure silhouettes results in an approximation of the high-speed object volume (dashed region). (b) real high-speed visual hulls, (c) template model fitted into the visual hulls of (b).*

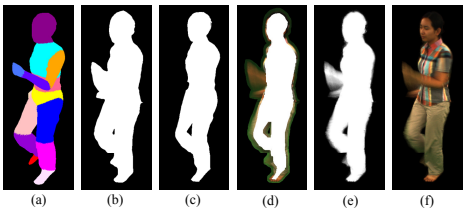


Figure 5: *Illustration of blur matting algorithm: (a) Segmentation into regions moving at different speeds used in markerless performance capture, different colors correspond to different body parts. Sec. 4.3. (b) Long exposure silhouette $S_{c,i}$. (c) Foreground regions $\mathcal{F}$. (d) Trimap. (e) Matte image. (f) Foreground image rendering after matting.*

reprojecting $\mathcal{V}'_t$ into each camera view. Please note that only through joint temporal and spatial staggering, such intersection will result in an approximation of the visual hull at a short time slice. In most circumstances, this approximation is close to the true visual hull and thus the true high frame rate silhouettes, see Fig. 4 (b). Only in some cases, this assumption is violated, which we discuss in Sec. 7.

4.3 Performance Capture

The images $S'_{c,j}$ serve as input to a template-based marker-less performance capture approach. In particular, we adapt the approach by Gall et al. [GSA*09] to our setting. The template model for a human actor comprises of a surface mesh $\mathcal{M}$ with a fitted skeleton and skinning weights. This static surface mesh is created from multi-view images of the actor standing still using the reconstruction approach of Wu et al. [WLDW10]. From the same images, we also create a static surface texture for the template, C_s. Skeleton and skinning weights for this mesh are semi-automatically created using the same procedure as Gall et al. [GSA*09]. For every time step of multi-view high-frame rate silhouette images $S'_{c,j}$, the performance capture approach first determines the model pose by finding optimal skeleton pose parameters. Subsequently, a silhouette refinement step is performed, in which the surface mesh is non-rigidly deformed to align with all silhouette images. Please note that, as opposed to the original paper [GSA*09], due to motion blur we are only able to employ silhouette constraints for tracking and not additional feature points. Since the approximation of $\mathcal{V}'_j$ and thus $S'_{c,j}$ may suffer from artifacts in regions of fast motion, we decrease the influence of silhouette regions in the silhouette adaptation step that were found to move fast after skeleton pose estimation, see Fig. 5 (a). The end result is a sequence of configurations of the template surface mesh $\mathcal{M}_j$, such that the scene geometry at each effective time frame is properly reconstructed, Fig. 4 (c) and Fig. 9.

5 Reconstruction of Deblurred Textures

In the following, we exploit the spatially and temporally dense scene description constructed as in Sec. 4 to estimate spatially varying point-spread functions and to reconstruct deblurred surface textures for the capture geometry from the blurry long exposure images.

5.1 Long Exposure Image Matting

In contrast to a binary segmentation, we need to extract for every pixel in the foreground region the contribution of its color due to the moving foreground and static background. The process is known as alpha matting [DW08, TKLS10].

In order to apply alpha matting, we exploit the model information obtained after high-speed motion capture to generate a trimap, see Fig. 5. The main idea is to use the intersection of all high frame rate silhouettes $S'_{c,j}$ that make up a long exposure silhouette $S_{c,i}$ as sure foreground, i.e. $\mathcal{F} = \cap_j S'_{c,j}$ where j runs over the high frame rate indices that are contained in the long exposure silhouette. Known background regions are obtained by the inverse of the long exposure silhouette $\mathcal{B} = \overline{S_{c,i}}$. The region that has to be matted is $\overline{\mathcal{F} \cup \mathcal{B}}$. For increased robustness we erode $\mathcal{F}$ and $\mathcal{B}$ prior to computing the matting region. With the trimap such defined we run the matting approach by Levin et al. [LLW07].

5.2 Model-based Deblurring

Standard image-base deblurring methods assume that the deblurred original (or latent) image can be obtained by deconvolving the blurred image with the (potentially spatially-varying) blur kernel or PSF [Ric72]. However, in case the scene motion is not simple, and there are occlusions and disocclusions of 3D points occurring during the exposure interval, this image formation model of the convolution of a single latent image with a spatially-varying PSF is not valid anymore. Fortunately, we can exploit the densely sampled 3D scene geometry $\mathcal{M}_j$ and instead perform a direct reconstruction of the *latent surface texture* as previosuly proposed for texture super-resolution [GC09]. By this means, we are able to properly handle occlusions and disocclusions of surface points even for complex motions. We refer to this approach as model-based deblurring.

In model-based deblurring, we are not solving for pixel colors of a deblurred latent image, but for colors of a set of infinitesimally

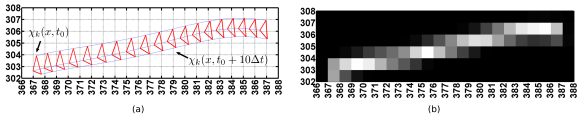


Figure 6: *From patch to PSF: (a) The 3D motion of a patch in one long exposure interval is projected into one camera view from which we compute (b), an approximation of the patch PSF.*

small surface patches that cover the 3D surface of the human body. The colors of these patches are assumed to be constant during the exposure interval of each blurred image $I_{c,i}$. Our image formation model now assumes that the color of each pixel in an image $I_{c,i}$ is obtained by suitable temporal integration of the colors of all patches whose projected motion paths ever pass through that pixel during the exposure interval, taking into account patch visibility.

The motion path of each patch, when projected into the camera view while taking occlusion into account, can be thought of as the PSF assigned to that particular patch, Fig. 6. Based on these considerations, we can formulate image formation as a simple linear model of the form

where $\mathbf{x}$ is the stacked vector of n_p colors of patches that are visible from the camera for a non-zero time interval during the exposure, while $\mathbf{b}$ is the stacked vector of pixel colors in the (matted) blurred image $I_{c,i}$. $\mathbf{A}$ is a matrix with n_p columns and n_i rows, with n_i being the number of pixels in $I_{c,i}$. In case the PSF of patch k has an influence on pixel l, i.e., the projected motion path of k passes through l, $\mathbf{A}_{l,k}$ is the contribution factor of patch k with respect to l. This contribution factor describes the percentage with which patch k's color contributes to l in $I_{c,i}$. In particular, $\mathbf{A}_{l,k}$ can be computed by considering the integral over the pixel area over the full exposure period:

$$\mathbf{A}_{l,k} = \int_{A_l} \int_{t_0}^{t_1} \chi_k(x,t) dt dx, \tag{2}$$

where $\chi_k(x,t)$ is the characteristic function of the projected area of patch k, Fig. 6 and x is the spatial image coordinate. Since the patch is moving, the characteristic function depends on t. The area of pixel l is described by A_l, and $[t_0,t_1]$ is the exposure interval of the long exposure image $I_{c,i}$. We normalize the contribution factors, so for each pixel, the factors of all contributing patches sum to 1.

5.3 Model-based Deblur Algorithm

Given the captured high frame rate geometry and the blurred long exposure images, we can invert the image formation model, and compute a separate set of surface patch colors $C_{c,i}$ for every camera c and for every long exposure frame i. In practice, we use a finite number of discrete surface patches, and in the following we describe how they are initialized. Since we have N subsequent discrete mesh poses $\mathcal{M}_j$ within each long exposure interval i, we can reconstruct the individual PSF of each surface patch [†] through proper interpolation of the discrete mesh positions. Given these PSF estimates, we can formulate deblurred texture reconstruction as a linear least squares problem including robust regularization terms to account for possible inaccuracies in the images or reconstructed 3D models.

We solve for the visible patch colors $\mathbf{x} = C_{c,i}$ as follows:

$$\begin{aligned} \min_{\mathbf{x}} & \left[\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda_1 f_1 + \lambda_2 f_2 + \lambda_3 f_3 \right] \\ f_1 &= \sum_i \|\mathbf{x}_i\|_2^2 \\ f_2 &= \sum_i \sum_{j \in N(i)} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \\ f_3 &= \sum_i v_i (\|\mathbf{x}_i - \mathbf{x}_i^0\|_2^2), \end{aligned} \tag{3}$$

We add three regularization terms to the linear system to overcome inaccuracies in the estimated PSFs and the patch visibility. The first term f_1 is a Thikonov regularization, penalizing large norm solutions. The second term f_2 enforces smoothness in the final solution by enforcing color similarity between neighboring patches in 3D. The third term is a model-based regularizer that encourages the final color of each patch to be similar to the color of the patch in the static texture $x^0 = C_s$. Since tracking errors will have a most deteriorating effect in those regions of the blurred image $I_{c,i}$ where fast motion was observed, we adaptively weight the static texture regularizer f_3 depending on the speed of the moving patch through the factor $v_i \in \{0,1\}$ which is 1 for patches moving at the maximum observed velocity in 3D and $\varepsilon = 10^{-6}$ for the slowest moving patches. The influence of each regularization term is shown in Fig. 7. The necessity of a dynamic texture is shown by the comparison between Fig. 7(b) and (c)-(f) in highlighted blue rectangle regions around the object's abdomen and right leg. The varying geometric details due to motion, e.g. wrinkles on clothes, can not be revealed in a static texture, which greatly decreases the realism of the results. On the other hand, as the misalignment errors (see limitations in Sec.7) from motion capture artifacts increase, the dominant regularization term turns out to be the static texture constraint, which suppresses the small geometric details. The combination the the three regularizers is essential to achieve good results.

[†] When creating the set of surface patches two aspects need to be considered. First, the number of visible patches should roughly correspond to the number of foreground pixels to make sure the solution to the linear system is stable. Second, since we are approximating the model texture by constant-colored patches, the footprint of a projected surface patch in an image should be smaller than the size of the pixel. Our blurred images are of size 1024×768, the number of foreground pixels is in the range of $60000 - 100000$. As a result, we use a tessellated surface mesh $\mathcal{M}$ of around 200000 triangles for performance capture, which should fulfill the constraints above at all time steps and in all camera views.

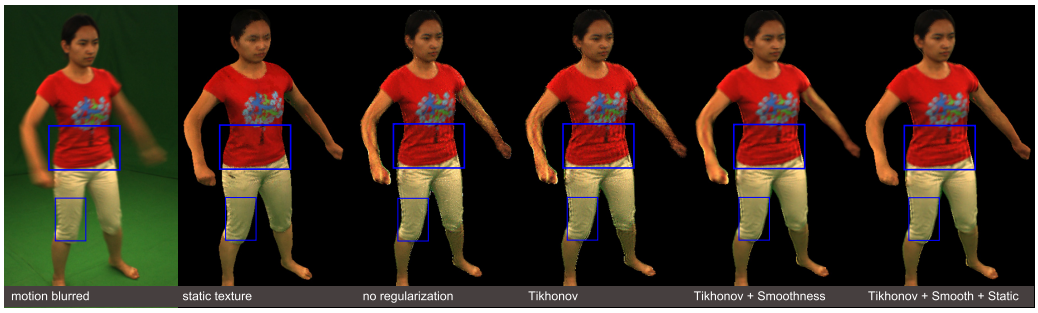


Figure 7: *Effect of regularization terms: (a) Blur image, (b) static texture, (c) no regularization, (d) Tikhonov, (e) Tikhonov with spatial smoothness, (f) Using all three regularizers. Note that the face is recovered well even though the static texture is quite different.*

6 Rendering Captured Performances

The outcome of the deblur algorithm is a separate set of colors for the visible surface patches in each long exposure image $I_{c,i}$. In other words, these colors are constant for each viewpoint c and long exposure time interval i. We chose this representation to avoid coupling of the variables across different views due to the staggered recording scheme. Our final goal, however, is the rendering of the performance captured surface meshes $\mathcal{M}_j$ at the high frame rate f_e from arbitrary viewpoints, including texture detail at the same frame rate exhibiting the true temporal and spatial variation. We accomplish this goal in two steps: First, we generate virtual high frame rate videos from the recording cameras' points of view. In a second step these are used in a standard multi-view projective texturing algorithm.

For high-speed video generation, we propose a spatio-temporal image warping scheme. Let us consider the case of rendering high-speed video for a specific camera view c'. To synthesize the output view at some high speed frame index $j' \in [i', i'+1]$ that falls in a time interval between two long-exposure indices i' and $i'+1$, we render the mesh $\mathcal{M}_{j'}$ into the camera viewpoint c' with two different textures, namely $C_{c',i'}$ and $C_{c',i'+1}$ where i' is the long exposure frame index of camera view c' that is temporally closest to high-speed frame time j', but earlier in the sequence. This procedure yields two video frames of the model in the same pose, albeit with a different texture. We apply optical flow to these two frames and interpolate the final image by a warped blend.

Fig. 8 shows examples of $\mathcal{M}_{j'}$ rendered with textures $C_{c',i'}$ and $C_{c',i'+1}$ (subfigures (a) and (b)) and the difference between these images in the camera view (subfigure (c)). Warped blending yields the final rendered performance at each high-speed frame. After generating high-speed video for each recording camera, the resulting images may be used for projective texturing of the high-speed model in a free-viewpoint video style rendering technique [CTMS03].

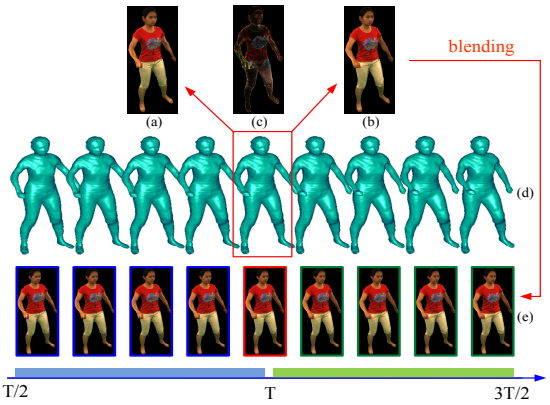


Figure 8: *Temporal interpolation of the deblur results to achieve high frame-rate textures: (a) and (b) are the deblur results, (c) being the 2 times differences between these two images. (d) is the temporally super-resolved model geometry $\mathcal{M}_{j'}$ at different high-speed frame indices j'. (e) shows the results of warped blending using (a) and (b) on the intermediate model.*

7 Results

In this section we verify our algorithms on 3 real-world test sets. In addition, we present validation experiments for some of the sub-steps involved in motion deblurring.

Capture of High-Speed Motion: To demonstrate the effectiveness of our high speed motion capture algorithm as well as the motion oriented deblur framework, we captured different subjects performing high speed motions dressed with different clothes, see Fig. 9.

Our data sets consist of three sequences of 20-30 captured long exposure frames each. The three motions include "marching", "running" and "spinning". All sequences were recorded with 20 cameras running at 20 fps using the staggered exposure framework. The resulting sequences have an effective frame

rate f_e of $20 \times 20 = 400$ fps. They consist of between 400 and 600 high-speed multi-view frames each. Some of the captured performance models, and rendered versions of them with deblurred textures, can be seen in Fig 9. As one can see, we are able to reliably capture detailed high-frame rate performance geometry, as well as plausible deblurred surface textures, despite very rapid motions and strong motion blur in the original camera images. The choice of regularization parameters is quite stable. We produced all results with one common set of parameters for each sequence. For all sequences, we only used two settings for each of the parameters: $\lambda_1 \in \{0, 0.02\}$, $\lambda_2 \in \{0.05, 0.1\}$, and $\lambda_3 \in \{0.06, 1.0\}$, the static texture constraint showing the most variation. The low parameter was used for the "running" and "spinning" sequences, whereas the higher setting was necessary for "marching". The main reason for using a higher value in the static texture constraint are inaccuracies in the tracking phase and resulting imprecise PSF estimates.

Comparison of Image-Based and Model-Based Deblurring: To compare the performance of image-based and model-based deblurring for the case of a complex motion with spatially-varying PSF, occlusions and disocclusions, we synthesized a blur image according to the captured motion using a static texture, see Fig. 10. We then generated PSFs. For the image-based deblur case, we computed the 2D PSF for each pixel in the image from the 3D motion of the geometry model. In particular, we back-projected each pixel to the model surface, marked the affected surface patches, and used their motion paths to generate a spatially varying two-dimensional PSF. The PSF is the sum of all patch-motions projected back to the image plane. Occlusion is not taken into account in this case. We then applied a version of Levin et al.'s code [LFDF07] that was modified to accept spatially varying PSFs. The results are shown in the middle sub-images of Figs. 10(b) and (c). As expected, the results show major artifacts in occlusion regions.

In comparison, our model-based motion deblur method can achieve high-quality deblurring results given the accurate motion information (right sub-images of Fig. 10(b)(c)). Given this analysis, the remaining artifacts in our results, Fig. 9, can be attributed to inaccuracies in the PSF estimates due to errors occurring during motion estimation, Sec. 4.

Limitations Our approach is subject to a few limitations.

First, our high speed visual hulls $\mathcal{V}'_j$ are computed from a spatial intersection of the back-projected cones of the temporally staggered blurry silhouettes, whose accuracy depends on the number of cameras, the motion speed, and matting accuracy, etc. Since there are no texture constraints to the performance capture algorithm, the template model is required to be close to the input. Furthermore, we can not handle wide garments.

Second, there are ringing and swimming artifacts. The ringing artifacts are mainly due to the geometries from performance capture are off and thus there are misalignment between blurry image pixels and the corresponding PSFs from high-speed geometry. The misalignment is usually severe in the boundary of the foreground object. In our deblurring algorithm, high frequency errors from the above misalignment are dealt with by 3 regularization terms that are equivalent to "low frequency filter", and then it will produce unwanted ringing artifacts. On the other hand, the swimming artifacts are caused by the instability of high-speed geometry from performance capture. There could be weird shaking, jittering, etc. that change randomly over time, and make different regularization terms to dominant the deblurring results. Thus, we observe the swimming artifacts.

Third, The regularization in texture deblurring suppresses unwanted artifacts, but also suppresses certain detail. However, our results show that high-speed textures can be recovered at a sufficient amount of detail for most applications. The current rendering may introduce additional loss of detail through blending.

Discussion: Our work explores the area of inexpensive high-speed motion capture using setups that are readily available in a number of research labs today. While more expensive solutions, like the use of high-speed cameras, could avoid some of the problems assoiated with this setting, additional problems (except for the price difference: PtGrey Flea2 – $300, Vision Research Miro eX1 – $9900), like difficulty of synchronization, very strong illumination requirements, and short capture times due to capture to RAM cast a doubt on the practical utility of these imagined systems. Simply using a static texture on the model, is inadequate since the texture is temporally changing, not only due to folds and wrinkles, but also due to illumination changes, facial expressions, etc., all very important cues for making a performance capture believable and convincing. Alternative capture strategies like using short exposure times and a low frame rate will introduce temporal aliasing such that high-speed, high-frequency motion cannot be resolved. In our proposed strategy, on the other hand, these motions leave a trace of motion blur which might facilitate their recovery.

8 Conclusion and Future Work

We presented an approach to capture shape and multi-view texture of rapidly moving human actors from multi-view video footage recorded with normal off-the-shelf cameras. Even though individual camera images are blurred, we can apply template-based performance capture and model-based texture deblurring by recording the video frames in a spatially and temporally staggered sequence. In the end, we obtain spatio-temporally coherent scene geometry and spatially and temporally varying surface textures at an effective frame rate that is an order of magnitude higher than the physical frame rate of the camera. The reconstructed dynamic scene models can be used for rendering at very high temporal resolution.

Our results show that the deblurring of complex motion trajectories involving occlusion and disocclusion, for which traditional 2D PSF-based descriptions are insufficient, is possible. We expect that further research into this problem will lead to improved reconstruction algorithms. As the most limiting factor, we plan to improve the robustness of our

Figure 9: *Results for three different reconstructed performances (see also accompanying video). Each row shows two results from two different viewpoints. From left to right: captured blurred image, one of the reconstructed high-frame rate models with the skeleton, model with deblurred texture.*

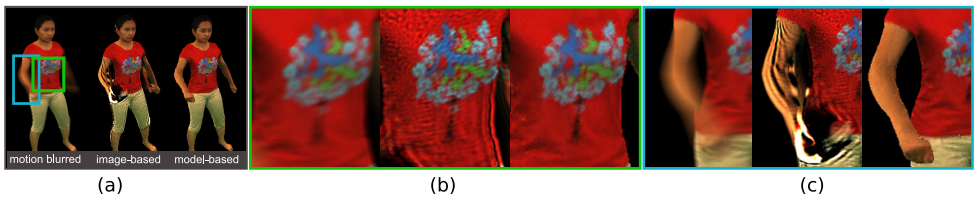


Figure 10: *Image-based vs. Model-based Deblurring: (a) from left to right: synthetically generated blur image based on motion capture data, deblur results of image-based and model-based deconvolution. Zoom-in on regions in (a), (b) shows deblur results for constant scene visibility whereas (c) shows a region affected by occlusions. Model-based deblurring can handle complex motion scenarios where image-based deblurring fails.*

approach with respect to the inaccuracies of motion capture, e.g. by using soft silhouette representations for more accurate high-speed visual hull computation and by treating the motion not as fixed, but as an initial guess for the unknown PSFs. Also, we plan to perform integrated multi-view texture deblurring, instead of per-camera reconstruction and view interpolation.

Acknowledgements

This work was supported by the German Research Foundation (DFG) through the Emmy-Noether fellowship IH 114/1-1, as well as China National Basic Research Project (No.2010CB731800) and the Key Project of NSFC (No. 61120106003 and 61021063).

References

[AGVN10] AGRAWAL A., GUPTA M., VEERARAGHAVAN A., NARASIMHAN S.: Optimal coded sampling for temporal super-resolution. In *Proc. IEEE CVPR* (2010), pp. 599–606. 1

[BAIH09] BRADLEY D., ATCHESON B., IHRKE I., HEIDRICH W.: Synchronization and Rolling Shutter Compensation for Consumer Video Camera Arrays. In *Proc. ProCams* (2009). 2

[BEN04] BEN-EZRA M., NAYAR S.: Motion-based Motion Deblurring. *IEEE Trans. PAMI 26*, 6 (2004), 689–698. 3

[BEZN05] BEN-EZRA M., ZOMET A., NAYAR S.: Video Superresolution using Controlled Subpixel Detector Shifts. *IEEE Trans. PAMI 27*, 6 (2005), 977–987. 2

[BHLW12] BOJSEN-HANSEN M., LI H., WOJTAN C.: Tracking surfaces with evolving topology. *ACM Trans. Graph. (SIGGRAPH) 31*, 4 (2012). 3

[BPS*08] BRADLEY D., POPA T., SHEFFER A., HEIDRICH W., BOUBEKEUR T.: Markerless garment capture. *ACM Trans. Graph. (SIGGRAPH) 27*, 3 (2008), 99. 3

[BTH*10] BUB G., TECZA M., HELMES M., LEE P., KOHL P.: Temporal Pixel Multiplexing for Simultaneous High-Speed, High-Resolution Imaging. *Nature Methods 7* (2010), 209–211. 2

[CBI10] CAGNIART C., BOYER E., ILIC S.: Free-form mesh tracking: a patch-based approach. In *Proc. IEEE CVPR* (2010). 3

[CML07] CHO S., MATSUSHITA Y., LEE S.: Removing non-uniform motion blur from images. In *Proc. of ICCV* (2007). 3

[CTMS03] CARRANZA J., THEOBALT C., MAGNOR M., SEIDEL

H.-P.: Free-viewpoint video of human actors. In *ACM Trans. Graph. (SIGGRAPH)* (2003). 3, 7

[dAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H.-P., THRUN S.: Performance capture from sparse multi-view video. *ACM Trans. Graph. (SIGGRAPH) 27* (2008). 1, 3

[DMY10] DING Y., MCCLOSKEY S., YU J.: Analysis of motion blur with a flutter shutter camera for non-linear motion. In *Proc. ECCV* (2010), pp. 15–30. 3

[DW08] DAI S., WU Y.: Motion from blur. In *Proc. IEEE CVPR* (2008). 5

[ECJ*06] EINARSSON P., CHABERT C.-F., JONES A., MA W.-C., LAMOND B., HAWKINS T., BOLAS M. T., SYLWAN S., DEBEVEC P. E.: Relighting human locomotion with flowed reflectance fields. In *Rendering Techniques* (2006), pp. 183–194. 1, 3

[FP09] FURUKAWA Y., PONCE J.: Carved visual hulls for image-based modeling. *International journal of computer vision 81*, 1 (2009), 53–67. 3

[FSH*06] FERGUS R., SINGH B., HERTZMANN A., ROWEIS S. T., FREEMAN. W. T.: Removing camera shake from a single photograph. *ACM Trans. Graph. (SIGGRAPH)* (2006). 3

[GC09] GOLDLUECKE B., CREMERS D.: Superresolution texture maps for multiview reconstruction. In *Proc. ICCV* (2009), pp. 1–8. 5

[GSA*09] GALL J., STOLL C., AGUIAR E., THEOBALT C., ROSENHAHN B., SEIDEL H.-P.: Motion capture using joint skeleton tracking and surface estimation. In *Proc. IEEE CVPR* (2009), pp. 1746–1753. 4, 5

[JKZS10] JOSHI N., KANG S.-B., ZITNICK L., SZELISKI R.: Image Deblurring using Inertial Measurement Sensors. *ACM Trans. Graph. (SIGGRAPH) 24*, 3 (2010), 15–22. 3

[JSK08] JOSHI N., SZELISKI R., KRIEGMAN D. J.: PSF Estimation using Sharp Edge Prediction. In *Proc. IEEE CVPR* (2008). 3

[Lau94] LAURENTINI A.: The Visual Hull Concept for Silhouette-Based Image Understanding. *IEEE Trans. PAMI* (1994), 150–162. 4

[LDXY12] LI K., DAI Q., XU W., YANG J.: Temporal-dense dynamic 3d reconstruction with low frame rate cameras. *IEEE Journal of Selected Topics in Signal Processing* (2012). 2, 3

[LFDF07] LEVIN A., FERGUS R., DURAND F., FREEMAN W.: Image and Depth from a Conventional Camera with a Coded Aperture. *ACM Trans. Graph. (SIGGRAPH) 26*, 3 (2007), 70. 8

[LLV*12] LI H., LUO L., VLASIC D., PEERS P., POPOVIĆ J., PAULY M., RUSINKIEWICZ S.: Temporally coherent completion of dynamic shapes. *ACM Trans. Graph. 31*, 1 (2012), 2:1–2:11. 3

[LLW07] LEVIN A., LISCHINSKI D., WEISS Y.: A closed-form solution to natural image matting. *IEEE Trans. PAMI* (2007), 228–242. 5

[LSC*08] LEVIN A., SAND P., CHO T. S., DURAND F., FREEMAN W. T.: Motion-invariant photography. *ACM Trans. Graph. (SIGGRAPH) 27*, 3 (2008). 3

[NKY08] NARASIMHAN S. G., KOPPAL S. J., YAMAZAKI S.: Temporal Dithering of Illumination for Fast Active Vision. In *Proc. ECCV* (2008), pp. 830–844. 2

[RAT06] RASKAR R., AGRAWAL A., TUMBLIN J.: Coded exposure photography: Motion deblurring via fluttered shutter. *ACM Trans. Graph. (SIGGRAPH) 25*, 3 (2006), 795–804. 3

[Ric72] RICHARDSON W.: Bayesian-based iterative method of image restoration. *JOSA 62*, 1 (1972), 55–59. 5

[SCI02] SHECHTMAN E., CASPI Y., IRANI M.: Increasing Space-Time Resolution in Video. In *Proc. ECCV* (2002), pp. 753–768. 2, 3

[SCI05] SHECHTMAN E., CASPI Y., IRANI M.: Space-Time Super-Resolution. *IEEE Trans. PAMI 27*, 4 (2005), 531–545. 2, 3

[SJA08] SHAN Q., JIA J., AGARWALA A.: High-quality motion deblurring from a single image. *ACM ToG (SIGGRAPH)* (2008). 3

[TAH*04] THEOBALT C., ALBRECHT I., HABER J., MAGNOR M., SEIDEL H.-P.: Pitching a baseball: tracking high-speed motion with multi-exposure images. *ACM Trans. Graph. (SIGGRAPH) 23* (2004), 540–547. 2

[TBW*12] TEVS A., BERNER A., WAND M., IHRKE I., BOKELOH M., KERBER J., SEIDEL H.-P.: Animation cartographyintrinsic reconstruction of shape and motion. *ACM Trans. Graph. 31*, 2 (Apr. 2012), 12:1–12:15. 3

[THBL10] TAI Y.-W., HAO D., BROWN M. S., LIN S.: Correction of Spatially Varying Image and Video Motion Blur using a Hybrid Camera. *IEEE Trans. PAMI 32*, 6 (2010), 1012–1028. 3

[TKLS10] TAI Y.-W., KONG N., LIN S., SHIN S. Y.: Coded exposure imaging for projective motion deblurring. In *Proc. IEEE CVPR* (2010). 5

[TNM09] TUNG T., NOBUHARA S., MATSUYAMA T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *Proc. of ICCV* (2009), pp. 1709 –1716. 3

[TTB11] TAI Y.-W., TAN P., BROWN M. S.: Richardson-lucy deblurring for scenes under projective motion path. *IEEE Trans. PAMI 33*, 8 (2011), 1603 –1618. 3

[TWdAN07] THEOBALT C., WUERMLIN S., DE AGUIAR E., NIEDERBERGER C.: New trends in 3d video. In *Eurographics Courses* (2007). 3

[VBMP08] VLASIC D., BARAN I., MATUSIK W., POPOVIĆ J.: Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph. (SIGGRAPH)* (2008). 3

[VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P., POPOVIĆ J., RUSINKIEWICZ S., MATUSIK W.: Dynamic shape capture using multi-view photometric stereo. In *ACM Trans. Graph. (SIGGRAPH Asia)* (2009). 1, 3

[VRR11] VEERARAGHAVAN A., REDDY D., RASKAR R.: Coded strobing photography: Compressive sensing of high speed periodic videos. *IEEE Trans. PAMI 33*, 4 (2011), 671 –686. 2

[WGT*05] WENGER A., GARDNER A., TCHOU C., UNGER J., HAWKINS T., DEBEVEC P.: Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Trans. Graph. (SIGGRAPH) 24* (2005), 756–764. 1, 3

[WILH11] WETZSTEIN G., IHRKE I., LANMAN D., HEIDRICH W.: State of the Art in Computational Plenoptic Imaging. In *Proc. Eurographics (STAR)* (2011), pp. 25–48. 2

[WJV*04] WILBURN B., JOSHI N., VAISH V., LEVOY M., HOROWITZ M.: High Speed Video using a Dense Array of Cameras. In *Proc. IEEE CVPR* (2004), pp. 294–301. 1, 2, 3

[WJV*05] WILBURN B., JOSHI N., VAISH V., TALVALA E.-V., ANTUNEZ E., BARTH A., ADAMS A., HOROWITZ M., LEVOY M.: High Performance Imaging using Large Camera Arrays. *ACM Trans. Graph. (SIGGRAPH) 24*, 3 (2005), 765–776. 2, 3

[WLDW10] WU C., LIU Y., DAI Q., WILBURN B.: Fusing multi-view and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE Trans. Vis. Comput. Graph.* (2010). 5

[WSZP10] WHYTE O., SIVIC J., ZISSERMAN A., PONCE J.: Non-uniform deblurring for shaken images. In *Proc. IEEE CVPR* (2010). 3

[WWC*05] WASCHBÜSCH M., WÜRMLIN S., COTTING D., SADLO F., GROSS M.: Scalable 3D video of dynamic scenes. In *Proc. Pacific Graphics* (2005), pp. 629–638. 3

[ZKU*04] ZITNICK C. L., KANG S. B., UYTTENDAELE M., WINDER S. A. J., SZELISKI R.: High-quality video view interpolation using a layered representation. *ACM Trans. Graph. (SIGGRAPH) 23*, 3 (2004), 600–608. 3