Objective Quality Assessment in Free-Viewpoint Video Production

J. Kilner *, J. Starck, J.Y.Guillemaut, A. Hilton

Centre for Vision, Speech and Signal Processing,
Faculty of Engineering and Physical Sciences,
University of Surrey, Guildford. GU2 7XH. UK

Abstract

This paper addresses the problem of objectively quantifying accuracy in free-viewpoint video production. Free-viewpoint video makes use of geometric scene reconstruction and renders novel views using the appearance sampled in multiple camera images. Previous work typically adopts an objective evaluation of geometric accuracy against ground truth data or a subjective evaluation of visual quality in view synthesis. We consider two production scenarios, human performance capture in a highly constrained studio environment and sports production in a large scale external environment. The accuracy of scene reconstruction is typically limited and absolute geometric accuracy does not necessarily reflect the quality of free-viewpoint rendering. A framework is introduced to quantify error at the point of view synthesis. The approach can be applied as a full-reference metric to measure fidelity to a ground truth image or as a no-reference metric to measure the error in rendering. The framework is applied to a data-set with known geometric accuracy and a comparison is presented for studio-based and sports production scenarios.

Key words: Free-Viewpoint Video, Image-based reconstruction, Image-based rendering, Quality Assessment

1. Introduction

Over the past decade multiple-view capture of events has gained increasing interest as a means to create three-dimensional (3D) video content. Application areas range from on-line visualization for mixed reality environments [1], communications [2], and production or pre-visualization in television [3], games [4] and 3DTV [5]. In 3DTV applications, cameras are typically arranged with a relatively short baseline to synthesise virtual views directly from the camera images [6]. Free-viewpoint video on the other hand is typically based on a relatively sparse set of cameras that surround a scene and makes use of 3D geometry to synthesise arbitrary viewpoints. A novel view is synthesised as shown in Figure 1 by rendering a geometric proxy to a virtual viewpoint using the appearance sampled from the adjacent camera images.

Research in free-viewpoint video currently lacks a consistent framework for quality assessment. Previous work in image-based reconstruction for static scenes evaluates geometric accuracy using ground-truth 3D shape [7]. In image-based rendering relatively little work has addressed the accuracy or quality of view synthesis, relying instead on a subjective visual assessment of performance [8] or pixel-wise error metrics with respect to ground-truth images [9]. We find that geometric accuracy does not necessarily reflect the quality of view synthesis and ground-truth 3D shape is often unavailable for the dynamic scenes used in free viewpoint video. Furthermore, conventional pixel-wise error metrics do not necessarily reflect perceived visual quality and provide no intuition as to the underlying geometric

---

* Corresponding author. Tel: +44 (0)1483 689842. Fax: +44 (0)1483 686031

Email address: `J.Kilner@surrey.ac.uk` (J. Kilner).

Fig. 1. Free-viewpoint video synthesis: a geometric proxy is used to reproject the appearance sampled in camera images to a new viewpoint.

errors that cause artifacts in rendering.

This paper presents a quality assessment framework to quantify the accuracy of free-viewpoint video production. An objective measure of quality is required as a means to optimise the parameters of different algorithms and to benchmark the performance of different production techniques. The key contributions in the paper are as follows:

(i) Two different production scenarios are considered: human performance capture in a highly constrained studio environment and sports production in a large scale external environment. A taxonomy of errors in free-viewpoint video synthesis is presented.

(ii) A framework is proposed to objectively quantify errors. Simple objective measures of fidelity in view synthesis are presented that are designed to reflect perceived visual artifacts and to provide a well-understood measure of accuracy.

(iii) The framework is applied in two cases *full-reference* against ground truth from a novel viewpoint where available; and *no-reference* to measure artifacts in free-viewpoint rendering for the general case where ground-truth is not available. Results of full-reference evaluation are presented for both the studio based and sports production scenarios and a no-reference evaluation is presented for the the studio scenario.

Objective quality assessment is restricted here to the problem of evaluating geometric production for free-viewpoint video synthesis. Production techniques should provide high-quality view synthesis independent of coding, transmission and display. This paper is based on the objective evaluation of production for human performance capture [10] and sporting events [11]. In Section 2 previous work on free-viewpoint video production is presented along with approaches to geometric and image-based quality assessment. Section 3 provides an overview of the framework adopted for quality assessment. Section 4 introduces evaluation in the constrained studio environment and Section 5 large scale external environments for sporting events. An evaluation of different production techniques is then presented in Section 6. Finally, Section 7 draws conclusions and considers future directions.

2. Background

2.1. *Free-Viewpoint Video*

In traditional video and film production, an event is recorded from a specific location using a camera. When the video is rendered it provides the fixed viewing experience dictated by the director. Free-viewpoint video attempts to break this restriction by allowing the specification of the camera location at the point of rendering using multiple video streams recorded from different locations. Research to-date has focused on the multiple camera acquisition systems and the computer vision algorithms required to achieve robust reconstruction and high-quality view synthesis either in real-time or as an offline post-process [8]. Recent advances have exploited image-based reconstruction and image-based rendering to produce free-viewpoint video at a quality comparable to captured video [12].

Image-based reconstruction deals with the problem of deriving scene geometry from the appearance sampled in camera images. Conventional stereo-based techniques [13] reconstruct a 2.5D depth image representation from two or more cameras through a regularized search for image correspondence. Volumetric techniques allow inference of visibility and integration of appearance across all camera views without image correspondence. Shape-from-silhouette (SFS) techniques [14] derive the *visual-hull*, the maximal volume consistent with foreground silhouettes. This is refined in space-carving techniques [15] which provide the *photo-hull*, the maximal volume that has a consistent ap-

pearance across all visible cameras. Multiple shape cues and iterative refinement techniques have been adopted to search for optimal surfaces in a scene [16–18].

Image-based rendering is the process of synthesising novel views from camera images. With no geometric scene information, synthesis is performed directly by treating multiple view images as a set of samples from the light-field in a scene. The light-field is then interpolated and resampled to generate a novel view [19]. In this way, highly realistic view synthesis can be achieved at the cost of a requirement for dense image sampling to avoid interpolation artifacts. Image-based reconstruction and image-based rendering have been combined [20,21] to synthesize novel views from a sparse set of cameras by using scene geometry to provide image correspondence. The advantage of view-dependent rendering is that it can overcome inaccuracies in geometric scene reconstruction by reproducing the change in surface appearance that is sampled in the original camera images.

Free-viewpoint video production within controlled studio environments, which provide fixed illumination, static backgrounds, stationary cameras and a relatively small capture volume, has received considerable interest. The *Virtualized Reality* system [13] first used 51 cameras distributed over a 5m dome to capture the performance of an actor in a studio. Fast and robust reconstruction from silhouettes [22,23,3,1] has allowed real-time systems to be developed for mixed reality environments. Offline systems have the potential for more accurate geometric scene reconstruction. Model-based techniques make use of a prior humanoid model to track the changes in shape, appearance and pose over time [24–27]. Data-driven approaches use image-based reconstruction without restriction to a prior model [28–30,4]. Video-based rendering blends the appearance sampled in camera images using view-dependent rendering [31–33].

In the sports environment, free-viewpoint video has been used to manipulate the viewpoint for playback of sequences within a match. Sports environments such as a soccer stadium require capture of a large volume (50m x 100m x 2m) with severely limited control over the environment. As a result reconstruction must cope with a relatively wide-baseline between cameras, the use of moving broadcast cameras, less accurate calibration and difficulties in player segmentation due to the uncontrolled illumination and backgrounds. Image morphing has

been applied for view interpolation with weak camera calibration [34,35], a simplified geometric representation using planar billboards has been proposed for real-time view synthesis [36] and graph cuts [37] and a deformable model [38] have been proposed for high-quality offline view synthesis.

This field has also been exploited commercially. Eyevision [39] used a bank of robotic cameras and specialist capture equipment to provide free rotation around a fixed instant in time for use during coverage of the Superbowl. Piero [40] use either hardware or software to calibrate broadcast cameras and generate a 3D billboard representation of the players which then allows limited viewpoint mobility. Most recently, Liberovision [41] uses a 3D representation of a soccer match to allow a variety of novel camera viewpoints to be rendered. Most techniques either require a high degree of user input or are limited in the range of viewpoints allowed. High-quality view synthesis from wide-baseline camera positions in an uncontrolled external environment remains an open and challenging problem.

2.2. *Objective Quality Assessment*

In image-based reconstruction, geometric accuracy has been evaluated using ground-truth 3D shape. Seitz et al. [7] present a comprehensive framework to compare reconstruction techniques against 3D geometry acquired from a laser stripe scanner. Recent work [8] in free-viewpoint production has demonstrated that with current camera hardware, geometric accuracy is inherently insufficient to represent the detailed geometry of a scene and that where display resolution reflects camera resolution, image-based rendering is required to achieve sub-pixel accuracy to minimise visual artifacts in view synthesis. An evaluation of free-viewpoint video should therefore target the accuracy, or quality of view synthesis rather than ground truth accuracy in geometric reconstruction.

The problem of defining video quality metrics has received significant interest in the image processing community to assess degradations introduced by video acquisition, processing, coding, transmission and display [42]. There is also active research studying the degradation introduced by watermarking schemes [43]. An overview of the field can be found in Eckert and Bradley [44]. Research into image quality assessment can be broken down into two broad categories, those attempting to model the Hu-

man Visual System (HVS) and those using more direct pixel fidelity criteria.

There has been much work focusing on HVS based measures of the fidelity of an image. Examples include measuring mutual information in the wavelet domain [45], contrast perception modeling [46] and modeling the contrast gain control of the HVS [47] However, HVS techniques do not necessarily reflect the true complexity of the visual system and objective measurement of perception remains an open research problem [48–50].

Pixel-wise fidelity metrics such as Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) remain widely adopted as simple, well-understood measures of fidelity despite a poor correlation with visual quality [51]. An overview of several measures and their performance can be found in Eskigioglu and Fischer [52] while a statistical analysis of various techniques encompassing both pixel metrics and HVS-based metrics can be found in Avcibas et al [53].

Objective evaluation should ideally provide simple, repeatable quality measures that afford a clear physical interpretation tailored to perceived visual quality. However the lack of effective standard measures are testament to the difficulties of achieving this in the general case, and so a more fruitful approach is to use domain-specific measures to target the quality assessment of particular types of images.

3. Quality Assessment Framework

In this section a framework is proposed to quantify accuracy in multiple camera geometric production for free-viewpoint video synthesis. The following definition is provided as the basis for quality assessment.

Free-viewpoint video production should recover a sufficiently accurate 3D scene representation for view synthesis free from visual artifacts. View synthesis should in turn target the resolution of the input camera images such that 3D video provides an acceptable alternative to conventional video.

The human visual system is highly adapted to perceive structural and temporal detail in a scene [51] and errors in the visual assessment of free-viewpoint video become apparent where prominent features or geometric structure is incorrectly reproduced. The accuracy of free-viewpoint video is therefore quantified as the structural error in reproducing the appearance of a scene at the point of view synthesis.
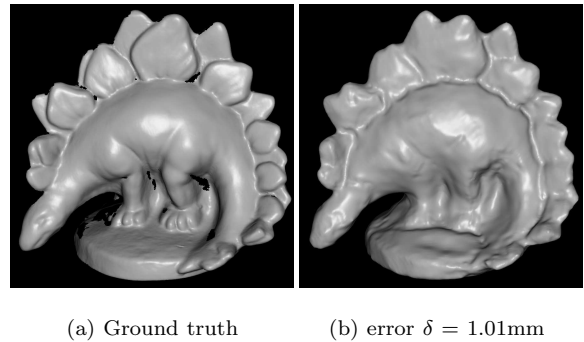

(a) Ground truth   (b) error $\delta = 1.01$mm

Fig. 2. Geometric evaluation of multiple view reconstruction courtesy of the Multi-View Stereo Evaluation Homepage (http://vision.middlebury.edu/mview).

Typical image quality assessment metrics either directly measure image fidelity (i.e. the pixel-wise distance between two images) or image quality (i.e. HVS-based systems). A similar classification can be made in the domain of reconstruction quality assessment between techniques which measures the piecewise accuracy of the reconstruction against some ground truth surface (fidelity measures) versus those that measure the overall suitability of a reconstruction for the task of representing the original surface (quality measures). Typical reconstruction quality assessment measures only reconstruction fidelity[7]. However, in some cases reconstruction fidelity has little or no effect on output image quality, which is itself independent of output image accuracy (for example an image shifted exactly one pixel to the right is preferred to an image which has had shot noise added). Our approach is an attempt to avoid simply measuring reconstruction fidelity and instead measure reconstruction quality by measuring the artifacts in the rendered image.

Seitz et al. [7] present a framework to evaluate geometric error in image-based reconstruction with respect to ground-truth 3D geometry. An error metric δ is defined as the distance such that 90% of the reconstructed geometry lies within the distance δ of the ground truth surface. Figure 2 illustrates the geometry reconstructed using a free-viewpoint video production technique [4] with an accuracy of $\delta = 1.01$mm.

We extend this framework to the image domain to evaluate free-viewpoint video. The accuracy of a synthesised image I is quantified as the registration error with respect to a reference image I'. An image I is represented by a function on a set of pixels where each pixel value (typically a 3 component RGB vec-

4

tor) that makes up the image is obtained by evaluating $I(p)$. We denote those pixels which are in the domain of a specific image with the notation $p \in I$. The error at each pixel is computed as a distance $d(p, I')$ to the reference image. Here for illustration we compute distance using image similarity $S(.)$.

$$d(p, I') = \|p - p'\|_2, \operatorname*{argmax}_{p' \in I'} S(I(p), I'(p')) \quad (1)$$

The maximum error in the distribution can be used to characterise the maximum visual artifact that will be apparent, defined by the *Hausdorff distance* from image I to the reference I'.

$$d(I, I') = \max_{p \in I} d(p, I') \quad (2)$$

In practice the Hausdorff metric is sensitive to outliers in the data and the generalized Hausdorff distance is taken as the kth ranked distance in the distribution, where $Q_{x \in X}^k f(x)$ is the quantile of rank k for $f(x)$ over the set X.

$$d^k(I, I') = Q_{p \in I}^k d(p, I') \quad (3)$$

For consistency with [7] we adopt the 90th percentile measure d^{90}.

The error for a synthesised view I is now defined by a single metric $d^k(I, I')$ that quantifies mis-registration between two images. Intuitively the distance measure is related to the geometric error in the underlying geometry of the scene. With a larger error in the 3D geometry of the scene, there will be a shift in reprojected 2D appearance and greater mis-registration.

This framework provides a single intuitive value in terms of pixel accuracy in view synthesis that can be applied at the resolution of the input video images. The approach can be applied as a full-reference measure of fidelity in aligning structural detail with respect to a ground truth image, or as a no-reference metric where mis-registration of structural detail can be evaluated between the appearance sampled from different camera viewpoints in view-dependent rendering.

4. Studio Production Error Metric

Studio-based production provides a highly constrained environment for multiple view video capture. Studio systems typically consist of between 4 and 51 cameras [8] configured to surround a specific

Fig. 3. A typical image produced in a studio environment.

volume of interest. Figure 3 illustrates one such production studio. In this environment the camera parameters defining the projective transformation to the image plane can be pre-calibrated to sub-pixel accuracy [54,55]. The scene can also be recorded using a single ambient lighting environment with a fixed backdrop such as a blue-screen for accurate foreground matting.

In this section the different visual artifacts are categorised and a metric is presented to evaluate the accuracy of view synthesis for the studio environment. The metric is demonstrated against ground-truth geometric accuracy using the 16-view data-set courtesy of the multi-view evaluation project [4]. In Section 6 the application to multiple view video of people is presented where no ground truth geometry is available.

4.1. A Taxonomy of Errors

With accurate camera parameters, visual artifacts in view synthesis arise principally from an inexact geometric representation of the scene. Geometric error results in an incorrect projected shape or an incorrect sampled appearance for the scene surface. Reconstruction errors can be categorised as global errors in gross shape or local errors in exact surface position. Examples of typical reconstruction errors are shown in Figure 4. Large scale errors are apparent as extraneous or phantom volumes or protrusions that do not correspond to the true surface of the scene. Local errors in surface position give inexact sampling of appearance from the multiple view video images leading to blurring or ghosting. Errors are summarised in Table 1.

Real-time systems [3,1] typically make use of fast and robust reconstruction from image silhouettes. Accurate foreground mattes can be derived in a studio environment, however the resulting visual hull

representation provides the *maximum* volume that is consistent with the silhouettes. These techniques therefore suffer from gross geometric errors where false positive volumes are derived. Such errors often appear as protrusions particularly when only a few cameras are used in reconstruction or there are multiple occlusions in the scene.

Offline systems [4,27] on the other hand optimise the scene geometry and often adopt a minimal surface as a regularisation constraint in reconstruction. These techniques tend to suffer less from gross errors in geometry and visual artifacts arise from local errors in surface extraction.

4.2. Registration Error Metric

Here we focus on evaluating offline scene reconstruction as the basis for high quality free-viewpoint video production. Visual artifacts arise where local errors in surface geometry result in incorrectly sampled surface appearance. These artifacts become apparent at the prominent visual features in the scene. In areas of uniform surface appearance inexact geometry will correctly sample a similar appearance across camera viewpoints.

Structural registration error is computed between images using a public domain optic flow algorithm [56] that accounts for the expected image variance in uniform areas of appearance. Registration is derived as the displacement at each pixel $\underline{d}(p, I)$, providing the motion between two images I, I'.

The 90th percentile Hausdorff distance d^{90} now provides a single measure of the maximum error at distinct image regions, where the effect of geometric error is apparent in view synthesis.

4.3. Full Reference Evaluation

A full-reference (FR) comparison makes use of a ground-truth reference for a frame-by-frame evaluation of the accuracy in view synthesis. This is illustrated using a leave-one-out test where the reconstructed geometry shown in Figure 2(b) is used to synthesise a view that is excluded in synthesis. A comparison is made between the registration error d^{90}, the RMSE registration error as well as PSNR. The comparison is made for varying degrees of geometric error introduced by inflating the surface by 1mm and 2mm beyond the known geometric error $\delta = 1.01$mm [4].

(a) reference   (b) error $\delta = 1.01$mm

(c) +1mm error   (d) +2mm error
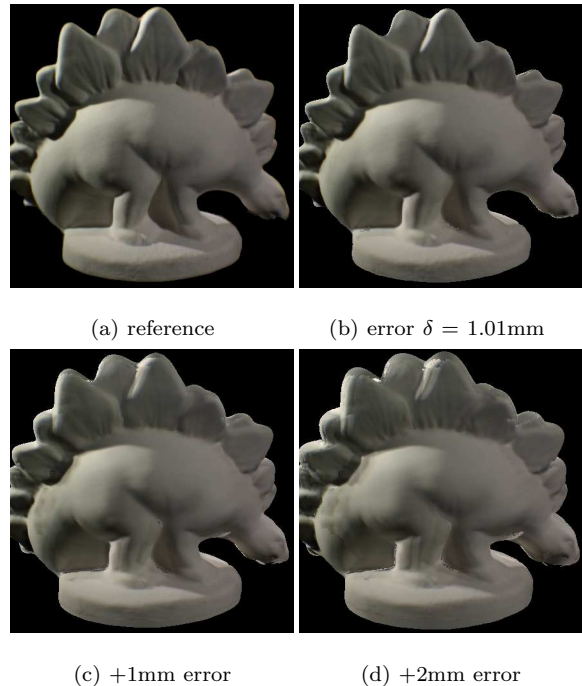
Fig. 5. Synthesised views in a leave-one-out test in comparison to (a) the reference image, with (b) baseline error $\delta = 1.01$mm and additional geometric error (c) +1mm (d) +2mm introduced in order to generate typical mis-registration artifacts such as blurring and double images.

Geometric Error	d^{90}	RMSE	PSNR (dB)
+0mm	0.70	0.65	31.50
+1mm	2.12	1.44	24.40
+2mm	3.81	2.36	21.90

Table 2
Error metrics for FR comparison in a leave-one-out test with varying degrees of additional geometric error. Note that the RMSE presented in these tables is derived from the registration error whilst PSNR is derived from pixel value differences

Figure 5 shows the synthesised views compared to the reference image. The corresponding error metrics for the leave-one-out tests against the ground-truth image are presented in Table 2. As the geometric accuracy is reduced, double exposure effects can be observed in structured image regions such as shadow boundaries. The error metrics follow this subjective decrease in image quality with a reduced PSNR, an increased RMSE and an increase in the generalized maximum error d^{90}. Both the RMSE and d^{90} metrics reflect the change in geometric scene accuracy. The d^{90} measure also provides intuition as to the maximum error that is apparent in Figure 5

Type	Error	Cause	Description
	Phantom volume	Ambiguity due to occlusions	The visual hull contains disconnected components due to the ambiguity in shape caused by occlusions
Gross Errors	Phantom protrusion	Ambiguity due to occlusions	The visual hull is extended into a region where no genuine surface exists. This is often caused by self-occlusion e.g. in the region between the legs.
	Holes	Incorrect matting	Areas incorrectly marked as background will cause the space carving algorithm to generate a visual hull with holes in it where none existed in the original shape
Local Errors	Sunken or raised surface	Stereo Ambiguity	Lack of structure or repeating patterns can cause an ambiguity in the stereo scores for a region. As a result stereo based optimisation cannot determine the true depth of the surface.
	Blurring	Quantisation	Fundamental limits to precision of surface placement and resolution of input images lead to blurring when multiple images are combined

Table 1
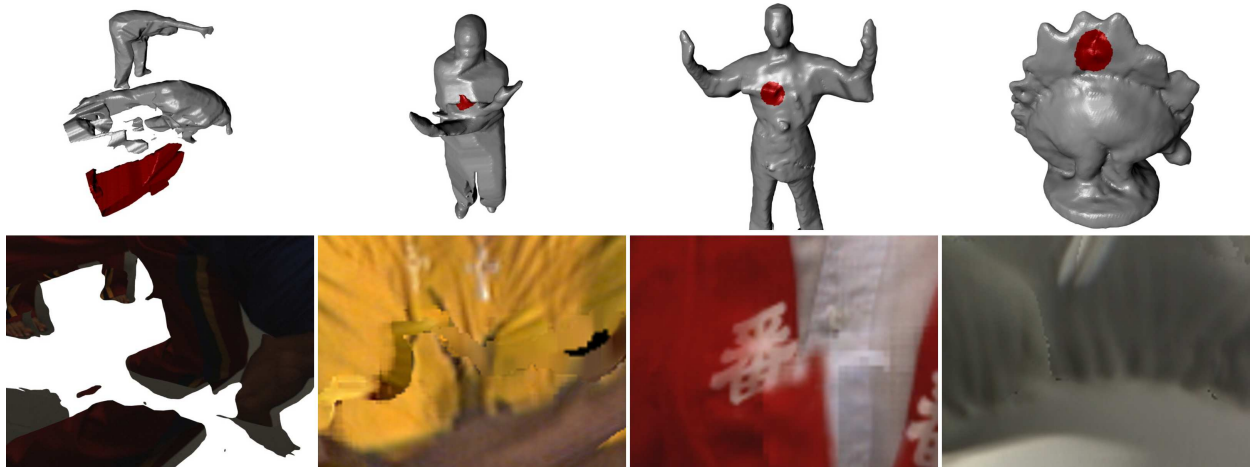Classification of errors in foreground synthesis.

Fig. 4. Gross and local errors. The top row shows the error in reconstruction highlighted in red while the bottom row shows the corresponding artifact in the rendered view. From left to right, phantom volume, phantom protrusion, a raised surface area and blurring. Capoeira and Kimono datasets (columns 1 and 3) courtesy of Matsuyama lab., Kyoto Univ., Japan[29]. Dancer dataset (column 2) courtesy of the public domain multiple camera dataset [4]. Dinosaur model (column 4) courtesy of the Multi-View Stereo Evaluation Homepage (http://vision.middlebury.edu/mview)

as the numerical value indicates the expected scale of mis-registration errors.

4.4. No-Reference (NR)

A no-reference comparison requires no explicit ground-truth. In view synthesis the appearance of a 3D scene is sampled in two or more camera images and reprojected to a new view. In the absence of a ground-truth reference the reprojected appearance from different cameras can be compared directly. This is illustrated using a virtual viewpoint placed at the mid-point between two cameras in the 16-view dataset. From this viewpoint, two images are ren-

dered, each using one of the original cameras. These images are then compared against each other to produce the results shown in Table 3. These results demonstrate that the error metric values obtained in the no-reference case are strongly correlated with those in the full-reference case with known ground-truth.

The NR comparison provides a measure of the potential artifacts in view synthesis without the requirement for a ground truth image. Note that visual artifacts are observed where the reprojected appearance is mis-registered in Figure 5(c),(d). The generalized maximum error d^{90} provides a metric for the maximum apparent error which mirrors the

Geometric Error	d^{90}	$RMSE$	$PSNR$ (dB)
+0mm	1.09	0.76	31.60
+1mm	2.22	1.24	24.90
+2mm	3.74	2.13	21.70

Table 3
Error metrics for NR comparison where the reprojected appearance from two camera images is compared.

Full-Reference metrics shown in Table 2.

4.5. *Summary*

A simple metric has been presented to quantify structural errors in view synthesis. Conventional error metrics such as the *root mean square error* (RMSE) can be adopted. However, a simple mean across an entire image can mask visually distinct errors in highly structured image regions. For example where there are extended areas of uniform appearance the RMSE will naturally tend to zero as the images I, I' are similar. Here the metric targets distinct image regions and features where geometric errors become visually apparent. The technique is relatively simple to implement using public domain software and can be used to compare a synthesised view to a ground truth image for a full-reference evaluation, or to compare the appearance sampled from different camera images in a virtual viewpoint as a no-reference evaluation.

5. Sports Production Error Metric

The sports production environment typically consists of a set of cameras (up to 30) arranged at various mounting points around a stadium. These cameras are either static "locked off" cameras or operator controlled cameras. The cameras are connected to an editing suite in a mobile control unit. The environment is unconstrained in terms of lighting (which may be natural or floodlit), camera calibration (cameras are often hired and cameras are moving constantly during the game) and backgrounds (backgrounds are natural, noisy and temporally varying). In addition, moving broadcast and fixed cameras capture the scene at different resolutions and levels of motion blur. A typical image captured from a sporting event is shown in Figure 6

As such, attempting to use free-viewpoint video techniques developed for studio use results in very

Fig. 6. A typical image captured during a soccer match broadcast.

inaccurate reconstructions [11]. Errors in calibration and matting quickly swamp small ambiguities in shape and stereo-based optimisation becomes unfeasible as the baseline between cameras is widened.

5.1. *A Taxonomy of Errors*

When reconstructing an outdoor sporting event, the magnitude of errors increases and the output resolution decreases dramatically such that the distinction between gross and local errors becomes redundant. All errors, whether they are caused by ambiguity in the input data or by inaccuracies in the calibration and matting, cause significant deviation of the reconstructed shape from the true shape of the scene. As such a different and more fundamental taxonomy of errors is employed to characterise the performance of a reconstruction in this environment.

Free-viewpoint video in the sports environment suffers from input images with a greatly reduced resolution compared to the studio environment. Similarly the output images generated are at a much lower resolution and the foreground elements account for a much lower percentage of the image pixels. This renders the techniques developed for studio use inappropriate for use in the sports environment. Also the lack of overlap between the wide-baseline images coupled with their varying surface sampling rates (some images may zoom in on one or two players while others may cover half the field) and differences in motion blur, render a no-reference metric impractical. However similar principles may be

used to derive a full-reference metric for evaluation in this environment.

The errors inherent in free-viewpoint video synthesis in the sports environment can be classified by considering the ways in which a synthetic video sequence generated for a given viewpoint can differ from the video sequence that would have been captured by a real camera placed at that viewpoint.

5.2. *Errors in Shape and Appearance*

Errors in shape are errors where I is missing a foreground element that is present in I', or I contains an extraneous foreground element that was not present in I'. Examples are missing limbs or double images as shown in Figure 7 (a) and (b). In both cases pixels in I have been incorrectly synthesised as either foreground or background.

Errors in appearance occur when a region $R \in I$ contains different pixel values to $R' \in I'$. This can occur through the rendering of the incorrect surface or due to incorrect sampling of the input image. Examples are the rendering of surfaces in incorrect locations and blurred rendering of surfaces, as shown in Figure 7 (c) and (d).

Table 4 summarises these classifications. Through the distinction between foreground and background (and hence between shape and appearance), a measurement of image fidelity can be composed that conveys more information about the errors in the synthesis pipeline. If the entirety of I' is treated as foreground then this analysis reduces to a comparison of pixel values across the image.

5.3. *Registration Error Metric*

We now derive metrics to measure errors in appearance and shape as described above. We also derive a completeness metric to measure only missing foreground, as this allows us to perform meaningful analysis in the presence of matting errors which may introduce extraneous elements of the scene into the reconstruction.

An r-shuffle is a perturbation of an image such that if I is an r-shuffle of I' then every pixel $p' \in I'$ will be transformed to a pixel $p \in I$ such that $\|p' - p\|_2 < r$ [57]. Due to the accumulation of errors in the reconstruction process the generated image I will be a distortion of I'. By modelling this distortion as an r-shuffle, the accumulation of errors in the view synthesis pipeline can be accounted

for, allowing an assessment of the fidelity of the underlying reconstruction technique. This measure is similar to the measure d used in the studio evaluation, however computation of d requires some image registration technique such as optic flow to be performed reliably on the images, whereas r simply requires an estimate of the mean pixel error in the system which can be inferred directly from known camera calibration and matting errors.

The r-neighbourhood N_r of a pixel p on the image I is defined such that for some other pixel q

$$q \in N_r(p) \iff \|q - p\|_2 < r \qquad (4)$$

We then define the pixel-wise shape matching function s in terms of the function $F(v)$ which returns 1 if pixel value v is foreground and 0 otherwise.

$$s(p, I, I') = F(I(p)) \max_{q \in N_r(p)} (F(I'(q))) \qquad (5)$$

Summing over the entire image and normalising by the number of foreground pixels gives the shape score. When r is taken as 0 this is simply the area of the intersection of both foreground regions divided by the area of their union.

$$shape(I, I') = \frac{\sum_p s(p, I, I')}{\sum_p \max(F(I(p)), F(I'(p)))} \qquad (6)$$

The pixel-wise completeness function is defined similarly to s except that it does not penalise for extraneous foreground.

$$c(p, I, I') = \max_{q \in N_r(p)} (\max(F(I(p))F(I'(q)), 1 - F(I'(q)))) \qquad (7)$$

Summing over the entire image and normalising by the number of foreground pixels gives the completeness score.

$$comp(I, I') = \frac{\sum_p c(p, I, I')}{\sum_p \max(F(I(p)), F(I'(p)))} \qquad (8)$$

The pixel-wise appearance matching function a can be defined.

$$a(p, I, I') = \max_{q \in N_r(p)} (T(\|I(p) - I'(q)\|_2)) \qquad (9)$$

where $T(x)$ returns 1 if $x \geq \tau$ and 0 otherwise, and τ is some chosen small threshold. Summing over the entire image and normalising by the common foreground region gives the appearance score.

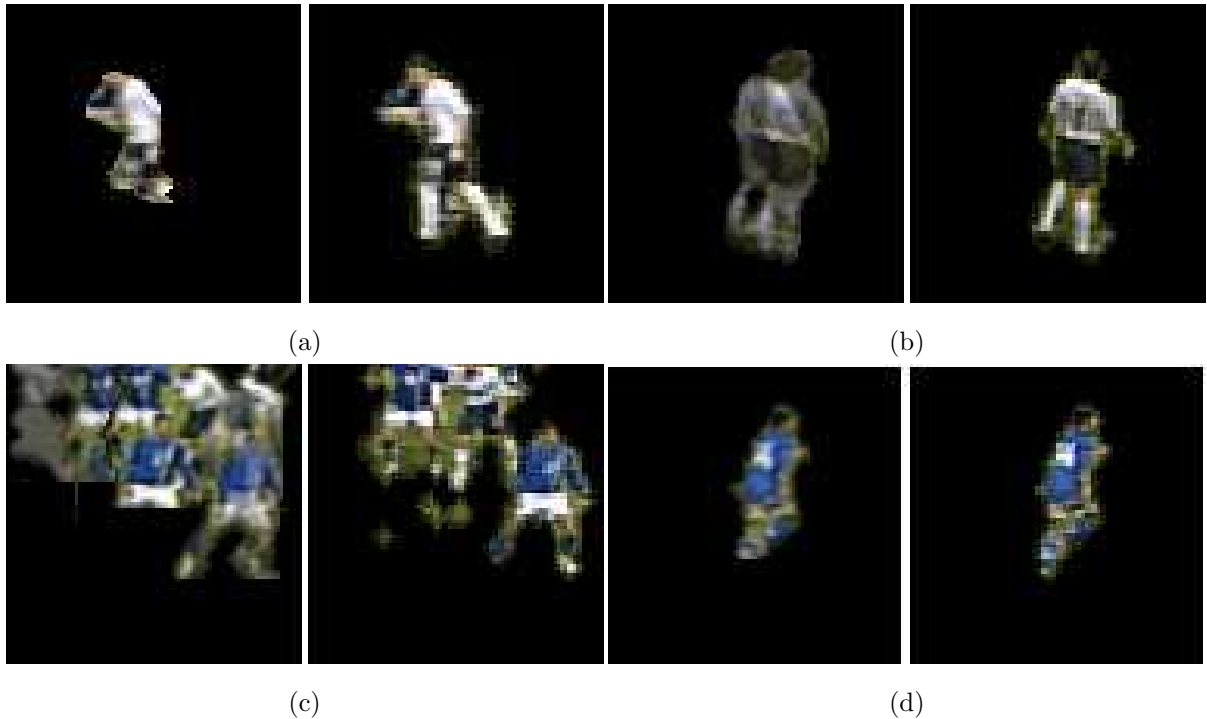$$app(I, I') = \frac{\sum_p a(p, I, I')s(p, I, I')}{\sum_p F(I(p))F(I'(p))} \qquad (10)$$

9

Fig. 7. A comparison of synthetic images to their corresponding ground truths. In each pair the synthetic image is on the left and the ground truth on the right (a) shows an incomplete synthetic image, (b) shows a synthetic image where the player is incorrectly rendered twice. (c) shows a player incorrectly rendered to a foreground region and (d) shows a blurred player.

Error	Image in R'	Image in R	Classification
missing foreground	present	absent	error in shape
extraneous foreground	absent	present	error in shape
none	present	present	correct shape
incorrect image	image of β	image of α	error in appearance
distorted image	distorted image of α	image of α	error in appearance
none	image of α	image of α	correct view synthesis

Table 4
Classification of errors in foreground synthesis comparing a region of the ground truth image R' to a corresponding region of the synthetic image R. α and β denote different elements within the scene

These measures are compared against the PSNR which is given by:

$$\text{PSNR}(I, I') = 20 \log_{10}\left(\frac{K\sqrt{n}}{\sqrt{\sum_p \|I(p) - I'(p)\|_2^2}} \right)$$

(11)

where K is the maximum value that can be given by $\|I(p) - I'(p)\|_2^2$ and n is the number of pixels in the image.

These relationship between r and the d^{90} metric can now be seen. Finding a value of r which yields an appearance score of 0.9 is analogous to the d^{90} score used in the studio evaluation as both identify the distance within which 90% of pixels can be matched to a corresponding pixel in the comparisson image. However, the size and distribution of the errors in images from the sports outdoor broadcast environment render this measure less usefull than it is in the studio environment. Also the wide baseline and small size of the players makes the optic flow algorithm, upon which the d^{90} measure is based, unstable. Plotting the shape, appearance and complete-

Image	*shape*	*comp*	*app*	PSNR	VIF
Original	1	1	1	inf	1
Median	0.99	1	0.98	17.35	0.36
Blur	0.75	0.99	0.96	16.64	0.32
Visual Hull	0.86	0.87	0.99	14.02	0.22
Visual Hull - 1	0.84	0.88	0.95	11.91	0.15
Billboards	0.81	0.89	0.95	11.03	0.17
Billboards - 1	0.70	0.89	0.86	8.94	0.08
VDVH	0.56	0.58	0.98	8.88	0.07
VDVH -1	0.56	0.58	0.95	8.71	0.06
Blank	0	0	0	6.00	0

Table 5
Comparison of evaluation techniques of a single reconstructed or processed image vs. a ground truth image. The suffix "-1" refers to a reconstruction generated from the "leave one out" dataset.

ness scores against the size of the r-shuffle yields more insight as to the nature of the errors present in the reconstruction.

5.4. *Full Reference Evaluation*

Table 5 shows a comparison of these derived measures against the Visual Information Fidelity (VIF) measure [45] of visual quality in an image which was chosen as a baseline full-reference quality metric. The comparison was carried out on several test images, some consisting of filtered versions of an original image and others on reconstructions of the scene using billboards [36], visual hull [58] and the view-dependent visual hull [59] (a fuller description of the experimental setup for reconstruction is provided in Section 6.2). It can be seen that all measures are in broad agreement, justifying the use of the proposed measures in the comparison. However the shape, completeness and appearance measures give more information than VIF or PSNR as to the nature of the reconstruction. For example all scores agree that the blur transformation degrades the image more than the median filter, but the shape and completeness scores correctly indicate that the median filter preserves the shape of the image while the blur does not, and that the distortion caused by blurring is an expansion (*shape* < 1 is but *comp* ≈ 1).

5.5. *Summary*

A methodology has been presented to quantitatively evaluate free-viewpoint video production in the sports environment. A set of metrics are introduced that measure errors in shape, completeness and appearance of view synthesis. The measures have been compared to PSNR and VIF as benchmark no-reference error metrics. The techniques can be used to gain more information as to the nature of errors in foreground view synthesis and are suitable for the low resolution images generated in the sports environment. The techniques are simple to implement and provide an intuitive metric for quality assessment.

6. Results

This section will now evaluate the framework and metrics previously introduced by applying them to two production environments. Firstly the studio based metrics will be applied to capture from a studio environment and the sports based metrics are applied to video from a soccer match.

6.1. *Studio Production*

In free-viewpoint video production sparse camera sets are typical and additional camera views are not necessarily available for a full-reference quality assessment. A no-reference comparison is now presented to evaluate two free-viewpoint video production techniques.

6.1.1. *Capture Environment*
Two data-sets are considered, the first courtesy of [4] consists of a street-dancer performing fast acrobatic motions wearing everyday clothing recorded from 8, 1920 × 1080 resolution cameras, the second courtesy of [29] consists of a Maiko wearing a brightly coloured Kimono performing a slow dance recorded from 16, 1024×768 resolution cameras. Figures 8, 9 illustrate the 3D geometry recovered using a surface optimisation technique [29] with a computational cost of 1 min/frame on an Intel(R) Xeon(TM) 3.6 GHz CPU and a global optimisation technique [4] with a cost of 38 min/frame on an Intel(R) Xeon(TM) 3GHz CPU.

6.1.2. Reconstruction Techniques

Free viewpoint video in the studio environment requires high quality reconstruction techniques that work with well calibrated high quality cameras arranged along a wide baseline. The techniques compared in the studio environment are a deformable 3D mesh model and a graph cut optimised surface. Both approaches use image segmentation techniques to generate silhouette images and then use shape from silhouette to calculate the visual hull.

The deformable 3D mesh model described in Matsuyama et al.[29] then uses a temporo-spatial deformable model to optimise the surface for photoconsistency, silhouette fitting, smoothness, 3D motion flow and inertia.

The graph cut optimised surface described in Starck et al.[4] uses a max-flow min-cut algorithm to find a global optimum surface that maximises stereo correspondence while simultaneously fulfilling silhouette constraints and minimising intraframe distances.

6.1.3. Evaluation

The no-reference comparison is performed using a virtual viewpoint placed at the mid-point between a set of cameras in the studio, a pair of cameras for the planar 8 camera setup and a camera triplet for the non-planar 16 camera setup.

6.1.4. Discussion

The d^{90} metric given in Figures 8, 9 provides an objective comparison of the quality of view synthesis in the virtual viewpoint. The relatively computationally expensive global optimisation approach [4] gives lower d^{90} rendering error. This is expected as the global optimisation, which includes a stereo correspondence term, is expected to increase the geometric accuracy of reconstruction. Note that the errors for the kimono sequence are higher as the appearance in the scene is more highly structured. The metric will be both camera configuration and dataset dependent.

6.2. Sports Production

6.2.1. Capture Environment

The data set chosen for this evaluation was a recording of a soccer match. The recording was made with 15 static cameras arranged around 120 degrees of the stadium [60]. The cameras were interlaced PAL broadcast cameras captured at a resolution of

Fig. 10. Arrangement of cameras in soccer stadium. Camera highlighted in red provided ground truth images and was not used for "leave one out" experiment and cameras highlighted in blue were additionally removed for "leave three out" experiment

702×288 (one deinterlaced widescreen field). The camera configuration is illustrated in Figure 10. This data set was chosen as stable calibration data is available for the cameras, and the arrangement of the cameras allows "leave one out" and "leave three out" comparisons without excessively reducing the quality of the reconstruction. This allows consideration of the behaviour of each technique as information becomes more limited.

6.2.2. Reconstruction Techniques

Different techniques for scene reconstruction have different overheads and trade-offs in terms of quality and fidelity. For the proposed application of sporting event reconstruction, real-time playback capabilities along with the ability to work on data from sparse view-points are important properties. This has led to a number of alternatives considered here: billboards [36], visual hull [58] and the view-dependent visual hull [59].

In billboarding a single polygon is rotated around the Y axis so that it is constantly facing the virtual camera. An image of the original object is then applied to the polygon as a texture map.

A volumetric shape from silhouette approach is adopted [58] that divides space into a voxel grid and back-projects every voxel to each image. By comparing the overlap of the voxel to the silhouettes for each camera you can determine the voxel's occupancy. Overlap can be tested up to a given reprojection error to account for calibration errors in the camera system. The surfaces generated by the vox-
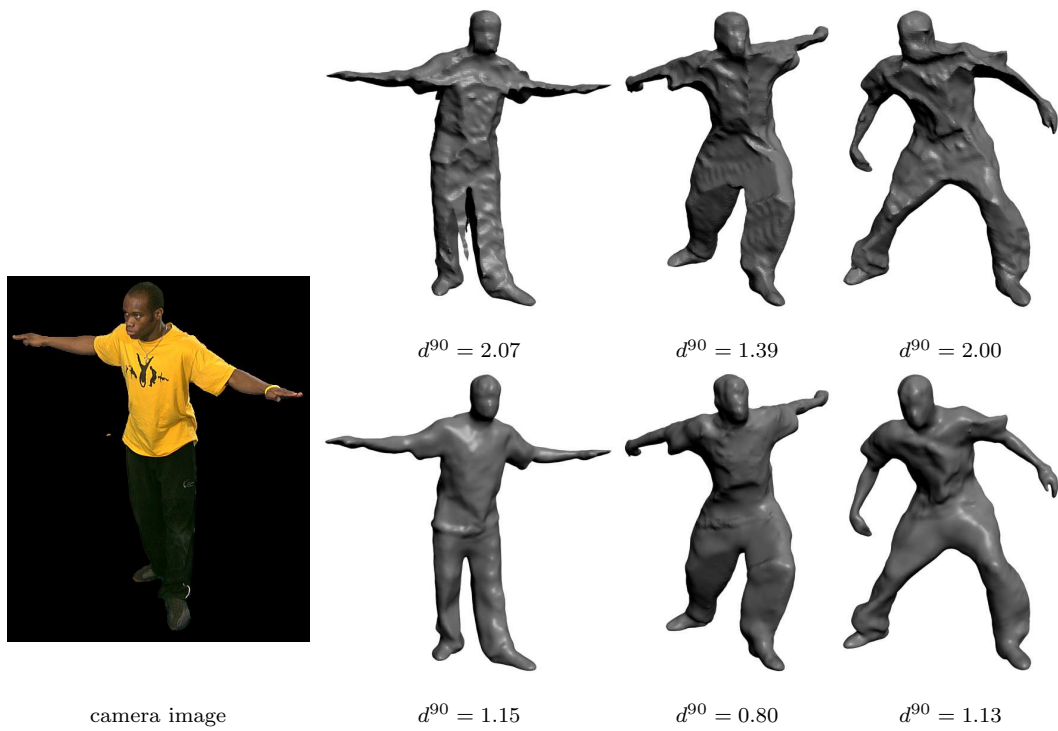
$d^{90} = 2.07$ $d^{90} = 1.39$ $d^{90} = 2.00$

camera image $d^{90} = 1.15$ $d^{90} = 0.80$ $d^{90} = 1.13$

Fig. 8. NR evaluation of two techniques (top) [29], (bottom) [4] for the street dancer sequence courtesy of [4].
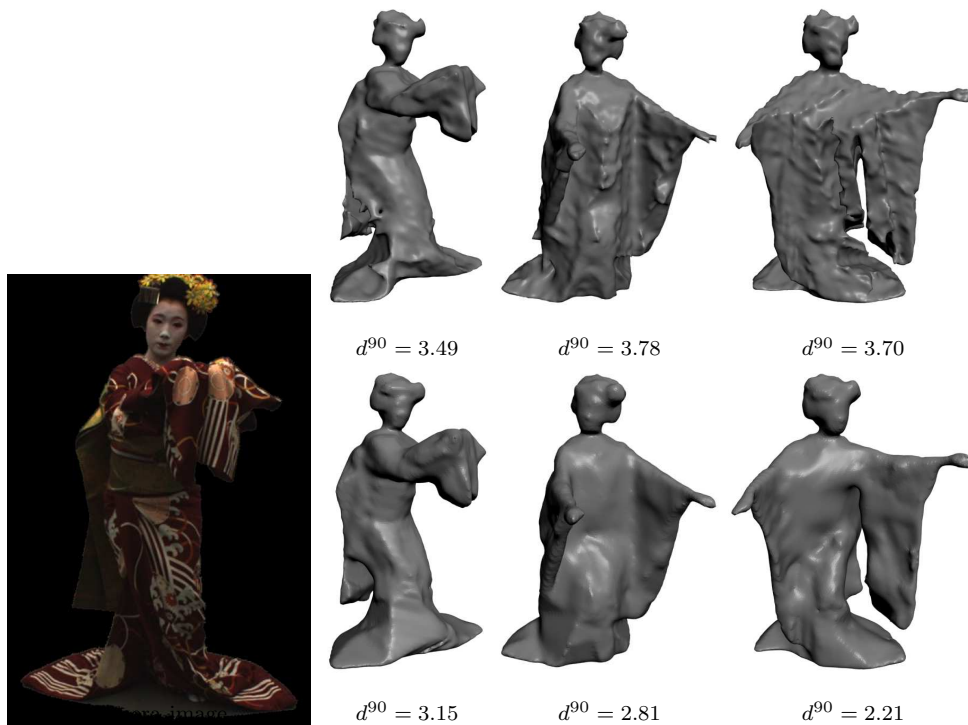

$d^{90} = 3.49$ $d^{90} = 3.78$ $d^{90} = 3.70$

$d^{90} = 3.15$ $d^{90} = 2.81$ $d^{90} = 2.21$

Fig. 9. NR evaluation of two techniques (top) [29], (bottom) [4] for the kimono sequence courtesy of [29].

Fig. 11. Ground truth image from camera 5.
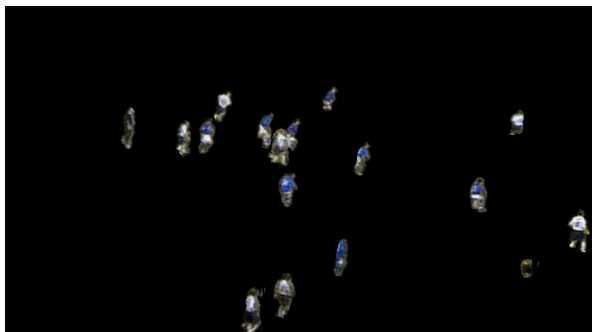
Fig. 12. Visual Hull reconstruction not using camera 5 with compensation for a silhouette error of 3 pixels.

els are then triangulated using the marching cubes algorithm [61] to produce a mesh.

These techniques are compared to the view-dependent visual hull (VDVH) [59] derives a view-dependent 2.5D depth representation for the visual hull with respect to a given viewpoint. Surface geometry is derived in the image domain by reprojecting a ray from this given viewpoint and deriving the exact intersection with each image silhouette to provide a depth-per-pixel.

6.2.3. *Evaluation*

The techniques of billboarding, visual hull, and VDVH were then compared. A single renderer capable of rendering all the scene reconstructions was created. This avoided differences in camera representation or lighting creating significant variation in global error between the techniques. A sequence of 100 frames of video from the 15-camera recording was matted using Bayesian matting [62] and 100 frames of synthetic video produced from the viewpoint of camera 5. In the first experiment, all available data sets were used, in the second, camera 5 was omitted from the data set and in the third, cameras 4, 5 and 6 were omitted (as shown in Figure 10). The

synthetic video streams were then compared with the ground truth video stream using the technique described in Section 5. It should be noted that the ground truth is also segmented using Bayesian matting and as such is not perfect. However any gross errors in the matting were corrected and the remaining errors are small compared to the reconstruction errors.

6.2.4. *Discussion*

The left hand column of Figure 13 shows the behaviour of the shape scores for each of the techniques as the magnitude of the estimated system error is increased. The shape graph for the "leave none out" test shows that none of the techniques achieve a score of 1, even when compensating for large system errors. Some of this is due to errors in the original matting.

The VDVH suffers quite extensively from missing and clipped players. These problems are not exacerbated by the removal of input cameras from the system, hence it does not degrade as much as the billboard or the visual hull. Billboards, being the simplest geometrical representation of the scene, degrade most as cameras are removed from the system as they provide the worst interpolation between views.

The central column of Figure 13 shows the completeness scores for the techniques. This graph clearly shows the difference between both the billboard and visual hull techniques which are generally overcomplete, and the VDVH which is generally incomplete. This can be interpreted in terms of the requirement for agreement between cameras. In the VDVH agreement is required between silhouette projections from all cameras to generate a surface and so errors in calibration and matting erode the shape. The visual hull and billboarding techniques used for these experiments both have some tolerance for disagreement between cameras and so errors do not erode the shape as badly. However the cost for this is an increase in false positives which can adversely affect the shape through the generation of phantom volumes, as seen in the shape scores in the "leave three out" experiment.

The right hand column of Figure 13 shows the behaviour of the appearance scores for each of the techniques as the magnitude of the estimated system error is increased. When a system error of 0 pixels is estimated, the appearance score for all techniques is significantly decreased. This is due to resampling
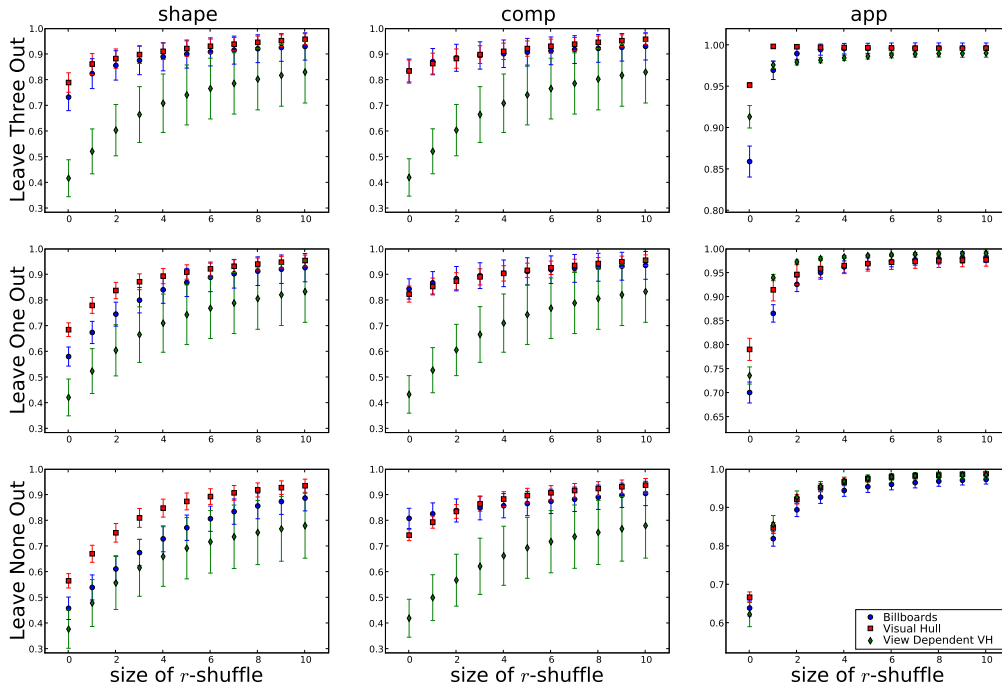
14

Fig. 13. Plots showing the mean and standard deviation over a sequence of 100 frames for shape, completeness and appearance scores vs. the size of the r-shuffle.

errors when the camera images are converted for use in the renderer.

It should be noted that even in this worst case "leave three out" test, appearance is generally synthesised with high fidelity compared to shape. The performance of the billboarding technique, which ignores small scale surface shape, indicates that the pressing problem with all current techniques is one of generating a scene reconstruction that is accurate and complete at the large scale; small-scale parallax having little effect on the appearance.

This conclusion is further supported by the shape of the graphs. The "dog-leg" shape of the appearance graph implies that a large number of pixels are rendered to within a small distance of their correct locations. This is consistent with small displacements of areas of correct view synthesis. However the shape and completeness graphs show a much smoother gradient as the size of the r-shuffle is increased, implying that errors in shape can not be accounted for by a simple displacement and that more serious distortions such as truncations have occurred.

7. Conclusion

An objective error measure for evaluation of free-viewpoint rendering has been presented based on the registration error with respect to either ground-truth reference images or with respect to appearance sampled from multiple input images. This allows the calculation of either full-reference or no-reference image synthesis quality assessment. Results demonstrate that the proposed error measures correlate strongly with the error in scene reconstruction and give an improved estimate of image quality over previously used RMSE and PSNR. Results show that the no-reference metric gives error values in agreement with those obtained from the full-reference metric.

Objective quality evaluation is performed for two free-viewpoint video scenarios: studio capture in a controlled environment; and sports production in an uncontrolled stadium environment with unconstrained illumination, natural backgrounds and moving cameras. Results demonstrate that the proposed error measures are able to quantify perceived visual artifacts and give a good indication of

15

relative algorithm performance.

8. Acknowledgements

References

[1] J. Allard, J.-S. Franco, C. Menier, E. Boyer, B. Raffin, The grimage platform: A mixed reality environment for interactions, IEEE International Conference on Computer Vision Systems (ICVS) (2006) 46.

[2] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. Vande Moere, O. Staadt, blue-c: a spatially immersive display and 3d video portal for telepresence, ACM Transactions on Graphics (SIGGRAPH) 22(3) (2003) 819–827.

[3] O. Grau, T. Pullen, G. Thomas, A combined studio production system for 3D capturing of live action and immersive actor feedback, IEEE Transactions on Circuits and Systems for Video Technology 14(3) (2003) 370–380.

[4] J. Starck, A. Hilton, Surface capture for performance based animation, IEEE Computer Graphics and Applications 27(3) (2007) 21–31.

[5] E. Stoykova, A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, X. Zabulis, 3-d time-varying scene capture technologiesa survey, IEEE Transactions on Circuits and Systems for Video Technology 17(11) (2007) 1568–1586.

[6] W. Matusik, H. Pfister, 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes, ACM Transactions on Graphics (SIGGRAPH) (2004) 814–824.

[7] S. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A comparison and evaluation of multi-view stereo reconstruction algorithms, IEEE International Conference on Computer Vision and Pattern Recognition 1 (2006) 519–526.

[8] J. Starck, A. Maki, S. Nobuhara, A. Hilton, T. Matsuyama, The 3D production studio, Technical Report VSSP-TR-4/2007.

[9] O. Woodford, I. Reid, A. Fitzgibbon, Efficient new view synthesis using pairwise dictionary priors, IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] J. Starck, J. Kilner, A. Hilton, Objective quality assessment in free-viewpoint video production, IEEE Conference on 3DTV (2008) 225–228.

[11] J. Kilner, S. J., A. Hilton, A comparative study of free-viewpoint video techniques for sports events, European Conference on Visual Media Production (2006) 87–96.

[12] C. Zitnick, S. B. Kang, M. Uyttendaele, S. A. J. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, ACM Transactions on Graphics (SIGGRAPH) 23(3) (2004) 600–608.

[13] T. Kanade, P. Rander, P. Narayanan, Virtualized reality: Constructing virtual worlds from real scenes, IEEE Multimedia 4(1) (1997) 34–47.

[14] A. Laurentini, The visual hull concept for silhouette based image understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 16(2) (1994) 150–162.

[15] K. Kutulakos, S. Seitz, A theory of shape by space carving, International Journal of Computer Vision 38(3) (2000) 199–218.

[16] C. Hernandez, F. Schmitt, Silhouette and stereo fusion for 3D object modeling, Computer Vision and Image Understanding 96(3) (2004) 367–392.

[17] Y. Furukawa, J. Ponce, Carved visual hulls for image-based modeling, International Journal of Computer Vision (2008) in press.

[18] G. Vogiatzis, C. Hernandez, P. Torr, R. Cipolla, Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency, IEEE Transactions on Pattern Analysis and Machine Intelligence 29(12) (2007) 2241–2246.

[19] M. Levoy, P. Hanrahan, Light field rendering, ACM Transactions on Graphics (SIGGRAPH) 30 (1996) 31–42.

[20] P. Debevec, Y. Yu, G. Borshukov, Efficient view-dependent image-based rendering with projective texture-mapping, Proceedings of Eurographics Workshop on Rendering, (1998) 105–116.

[21] C. Buehler, M. Bosse, L. McMillan, S. Gortler, M. Cohen, Unstructured lumigraph rendering, ACM Transactions on Graphics (SIGGRAPH) (2001) 425–432.

[22] G. Cheung, T. Kanade, J. Bouguet, M. Holler, A real time system for robust 3D voxel reconstruction of human motions, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2000) 714–720.

[23] W. Matusik, C. Buehler, L. Mcmillan, Polyhedral visual hulls for real-time rendering, Proceedings of Eurographics Workshop on Rendering, (2001) 115–126.

[24] J. Starck, A. Hilton, Model-based multiple view reconstruction of people, IEEE International Conference on Computer Vision (ICCV) (2003) 915–922.

[25] J. Carranza, C. M. Theobalt, M. Magnor, H. Seidel, Free-viewpoint video of human actors, ACM Transactions on Graphics (SIGGRAPH) 22(3) (2003) 569–577.

[26] D. Vlasic, I. Baran, W. Matusik, J. Popovic, Articulated mesh animation from multi-view silhouettes, ACM Transactions on Graphics (SIGGRAPH) (2008) in press.

[27] E. Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, S. Thrun, Performance capture from sparse multi-view

video, ACM Transactions on Graphics (SIGGRAPH) (2008) in press.

[28] B. Goldluecke, M. Magnor, Space-time isosurface evolution for temporally coherent 3D reconstruction, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) I (2004) 350–355.

[29] T. Matsuyama, X. Wu, T. Takai, S. Nobuhara, Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video, Computer Vision and Image Understanding 96(3) (2004) 393–434.

[30] S. Vedula, S. Baker, T. Kanade, Image-based spatio-temporal modeling and view interpolation of dynamic events, ACM Transactions on Graphics 24(2) (2005) 240–261.

[31] S. Yamazaki, R. Sagawa, H. Kawasaki, K. Ikeuchi, M. Sakauchi, Microfacet billboarding, Eurographics workshop on Rendering (EGWR) 97(2) (2002) 169–180.

[32] M. Eisemann, B. Decker, M. Magnor, P. Bekaert, E. Aguiar, N. Ahmed, C. Theobalt, A. Sellent, Floating textures, Computer Graphics Forum (Proc. Eurographics EG'08) 27(2) (2008) 409–418.

[33] M. Waschbusch, S. Wurmlin, M. Gross, 3d video billboard clouds, Computer Graphics Forum 26 (September 2007) 561–569(9).

[34] K. Connor, I. Reid, A multiple view layered representation for dynamic novel view synthesis, Proceedings of the 14th British Machine Vision Conference (BMVC).

[35] N. Inamoto, H. Saito, Arbitrary viewpoint observation for soccer match video, The 1st European Conference on Visual Media Production (CVMP) (2004) 21–30.

[36] T. Koyama, I. Kitahara, Y. Ohta, Live mixed-reality 3d video in soccer stadium, The 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality (2003) 178–186.

[37] J.-Y. Guillemaut, A. Hilton, J. Starck, J. Kilner, O. Grau, A bayesian framework for simultaneous matting and 3d reconstruction, in: 3DIM '07: Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling, 2007, pp. 167–176.

[38] J. Kilner, J. Starck, A. Hilton, O. Grau, Dual-mode deformable models for free-viewpoint video of sports events, in: 3DIM '07: Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling, 2007, pp. 177–184.

[39] Eye-Vision, Carnegie Mellon goes to the Super Bowl, http://www.ri.cmu.edu/events/sb35/tksuperbowl.html.

[40] BBC, The Piero System, BBC Production Magic., http://www.bbc.co.uk/rd/projects/virtual/piero.

[41] LiberoVision, LiberoVision GmBH Website, http://www.liberovision.com/.

[42] T. Pappas, R. Safranek, Perceptual criteria for image quality evaluation (2000).

[43] M. C. Q. Farias, S. K. Mitra, M. Carli, A. Neri, A comparison between an objective quality measure and the mean annoyance values of watermarked videos, in: in Proc. IEEE Int. Conf. Image Proc, 2002, pp. 469–472.

[44] M. P. Eckert, A. P. Bradley, Perceptual quality metrics applied to still image compression, Signal Process. 70 (3) (1998) 177–200.

[45] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444.

[46] A. Pons, J. Malo, J. Artigas, P. Capilla, Image quality metric based on multidimensional contrast perception models, Displays 20 (25 August 1999) 93–110(18).

[47] S. Winkler, A perceptual distortion metric for digital color video, SPIE (1999) 175–184.

[48] K. Seshadrinathan, A. Bovik, A structural similarity metric for video based on motion models, IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP) I (2007) 869—872.

[49] S. Winkler, Video quality and beyond, Proceedings of European Signal Processing Conference (2007).

[50] T. Brandao, P. Queluz, Towards objective metrics for blind assessment of images quality, Image Processing, 2006 IEEE International Conference on (2006) 2933–2936.

[51] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13(4) (2004) 600–612.

[52] A. Eskicioglu, P. Fisher, Image quality measures and their performance, Communications, IEEE Transactions on 43 (12) (1995) 2959–2965.

[53] I. Avcibas, B. Sankur, K. Sayood, Statistical evaluation of image quality measures, Journal of Electronic Imaging 11 (2) (2002) 206–223.

[54] J. Mitchelson, A. Hilton, Wand-based multiple camera studio calibration, CVSSP Technical Report VSSP-TR-2/2003.

[55] T. Svoboda, D. Martinec, T. Pajdla, A convenient multicamera self-calibration for virtual environments, Presence: Teleoper. Virtual Envrion 14(4) (2005) 407–422.

[56] W. Christmas, Filtering requirements for gradient-based optical flow measurement, IEEE Transactions on Image Processing 9(10) (2000) 1817–1820.

[57] K. Kutulakos, Approximate n-view stereo, ECCV 1 (2000) 67–83.

[58] R. Szeliski, Rapid octree construction from image sequences, CVGIP: Image Understanding 58 (1) (1993) 23–32.

[59] G. Miller, A. Hilton, Exact view-dependent visual hulls, Proceedings of the 18th International Conference on Pattern Recognition (ICPR) (2006).

[60] O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, J. Starck, A free-viewpoint video system for visualisation of sport scenes, SMPTE Motion Imaging (2007) 213–219.

[61] W. Lorenson, H. Cline, Marching cubes: A high resolution 3d surface construction algorithm, Computer Graphics 21 (4) (1987) 163–169.

[62] Y. Chuang, B. Curless, D. Salesin, R. Szeliski, A bayesian approach to digital matting, Proceedings of IEEE CVPR 2001 2 (2001) 264–271.