OBJECTIVE QUALITY ASSESSMENT IN FREE-VIEWPOINT VIDEO PRODUCTION

J. Starck, J. Kilner and A. Hilton

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

ABSTRACT

This paper addresses the problem of objectively measuring quality in free-viewpoint video production. The accuracy of scene reconstruction is typically limited and an evaluation of free-viewpoint video should explicitly consider the quality of image production. A simple objective measure of accuracy is presented in terms of structural registration error in view synthesis. This technique can be applied as a full-reference metric to measure the fidelity of view synthesis to a ground truth image or as a no-reference metric to measure the error in registering scene appearance in image-based rendering. The metric is applied to a data-set with known geometric accuracy and a comparison is also demonstrated between two free-viewpoint video techniques across two prototype production studios.

Index Terms— Free-Viewpoint Video, Image-based reconstruction, Image-based rendering

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 and communications, as well as production or pre-visualization in television, games and 3DTV. In 3DTV applications, cameras are typically arranged with a relatively short baseline to synthesise virtual views directly from camera images [1]. Free-viewpoint video is based on a relatively sparse set of cameras that surround a scene and typically makes use of 3D geometry to synthesise arbitrary viewpoints.

This paper presents a technique to objectively measure quality in free-viewpoint video production independent of coding, transmission and display. A quality assessment framework is required to benchmark the performance of production techniques as well as to provide a means to optimise the parameters of different algorithms. The paper contributes a simple objective measure of fidelity in view synthesis that is designed to reflect perceived visual artefacts and to provide a well-understood measure of accuracy.

2. BACKGROUND

Research to-date in free-viewpoint video 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 off-line post-process [2]. Recent advances have exploited image-based reconstruction and image-based rendering to produce free-viewpoint video at a quality comparable to captured video [3].

In image-based reconstruction, geometric accuracy has been evaluated using ground-truth 3D shape. Seitz et al. [4] present a comprehensive framework to compare reconstruction techniques against 3D geometry acquired from a laser stripe scanner. 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. Objective evaluation has been performed using pixel-wise error metrics with respect to a ground-truth view, for example by using a "leave-one-out" test [5].

Recent work [2] in free-viewpoint production of people has demonstrated that with current camera hardware, geometric accuracy is 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 artefacts 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. Recent research has focused on modelling the Human Visual System (HVS) to evaluate perceived image quality, however techniques do not necessarily reflect the true complexity of the visual system and objective measurement of perception remains an open research problem [6, 7]. In contrast pixel-wise 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 [8]. Objective evaluation should ideally provide simple, repeatable quality measures that afford a clear physical interpretation tailored to perceived visual quality.

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

3. EVALUATING FREE-VIEWPOINT VIDEO

Free-viewpoint video should be evaluated in terms of view synthesis. The geometric accuracy of the 3D scene representation does not necessarily reflect visual fidelity or visual quality in view 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 artefacts. View synthesis should in turn target the resolution of the input camera images such that 3D video provides an acceptable alternative to conventional video.*

View synthesis is performed through image-based rendering by sampling the appearance of the 3D scene geometry in a set of camera images and reprojecting the appearance to a new view. Visual artefacts arise either from inaccurate sampling of appearance in the imaging system or through an inexact 3D scene representation. For example, where the colour or geometric calibration of the imaging system is incorrect the appearance of a 3D surface point will not be sampled correctly in a camera image. Geometric error in turn results in an incorrect projected shape for the scene and an incorrect sampled appearance for the scene surface.

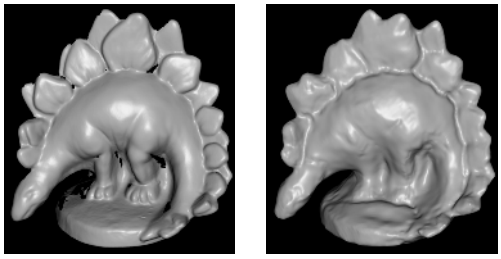
The human visual system is highly adapted to perceive structural detail in a scene [8] and errors in the visual assessment of free-viewpoint video become apparent where prominent features are incorrectly reproduced. A metric is therefore presented to measure structural error in view synthesis. The metric can either 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 causes visual artefacts in view synthesis. The measure 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.

4. A METRIC TO EVALUATE VIEW SYNTHESIS

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 set of pixels $p \in I$ and the registration error at each pixel is computed as the minimum distance to a similar pixel in I'. The error in view synthesis is now characterised by the distribution of pixel-wise error distances $d(p, I')$. Here the function $S(.)$ defines image similarity where the distance between similar pixels is minimised. We make use of a public domain optic flow algorithm [9] which performs a patch-based image registration and accounts for the expected image variance in uniform areas of appearance.

$$d(p, I') = \|p - p'\|_2, \max_{p' \in I'} S(I(p), I'(p')) \qquad (1)$$

A single error metric can be defined using the *root mean square error* (RMSE) across the entire image. However, a simple mean 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. Any errors at distinct image features will not be measured. We therefore seek to measure the maximum error in the distribution where the image is sufficiently structured to define the registration error $d(p, I')$.



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

Fig. 1. Geometric evaluation of a free-viewpoint video production technique [10] courtesy of the Multi-View Stereo Evaluation Homepage (http://vision.middlebury.edu/mview).

The *Hausdorff distance* is adopted to measure the maximum distance from image I to the reference I'.

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

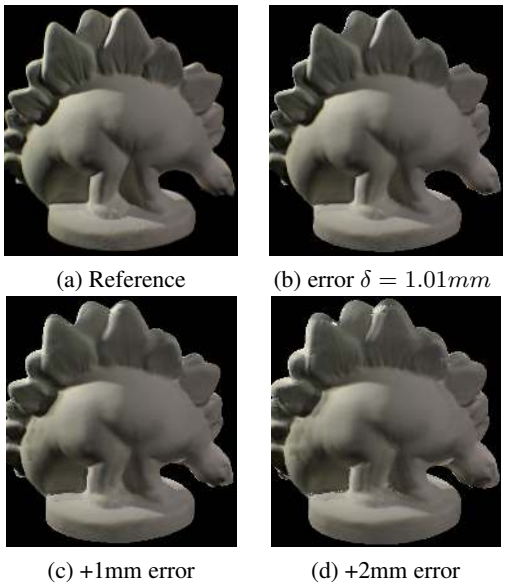
In practise the Hausdorff metric is sensitive to outliers in the data and the generalized Hausdorff distance is taken as the kth ranked distance in the distribution, where $Q^k_{x \in X} f(x)$ is the quantile of rank k for $f(x)$ over the set X.

$$d^k(I, I') = Q^k_{p \in I} d(p, I') \qquad (3)$$

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 misregistration. The metric is however specifically tailored to measure the registration of distinct image regions where the effect of geometric error is most apparent to an observer.

5. RESULTS

Seitz et al. [4] 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 1 illustrates the geometry reconstructed using a free-viewpoint video production technique [10] with an accuracy of $\delta = 1.01mm$. A full-reference and no-reference evaluation is now demonstrated



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

(c) +1mm error (d) +2mm error

Fig. 2. Synthesised views in a leave-one-out test in comparison to (a) the reference image, with (b) baseline error $\delta = 1.01mm$ and additional geometric error (c) +1mm (d) +2mm.

using the 16-view data-set courtesy of the multi-view evaluation project. We adopt a 90th percentile measure $k = 90\%$ to reflect the geometric evaluation.

5.1. Full-Reference (FR)

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 1(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 $\delta = 1.01mm$.

	+0mm	+1mm	+2mm
d^{90} (pixels)	0.70	2.12	3.81
$RMSE$	0.65	1.44	2.36
$PSNR$ (dB)	31.5	24.4	21.9

Table 1. Error metrics for FR comparison in a leave-one-out test with varying degrees of additional geometric error.

Figure 2 shows the synthesised views compared to the reference image. 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}. Note that relatively little difference is observed in the PSNR whereas the registration errors follow the marked change in apparent visual quality in the views. Both the RMSE and d^{90} metrics reflect the change in geometric scene accuracy and the d^{90} measure provides intuition as to the maximum error that is apparent in Figure 2.

5.2. 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. Table 2 now shows the error measured for the reprojected appearance between two cameras used in view synthesis.

	+0mm	+1mm	+2mm
d^{90} (pixels)	1.09	2.22	3.74
$RMSE$	0.76	1.24	2.13
$PSNR$ (dB)	31.6	24.9	21.7

Table 2. Error metrics for NR comparison where the reprojected appearance from two camera images is compared.

The NR comparison provides a measure of the potential artefacts in view synthesis without the requirement for a ground truth image. Note that visual artefacts are observed where the reprojected appearance is misregistered in Figure 2(c),(d). The generalized maximum error d^{90} provides a metric for the maximum apparent error which mirrors the Full-Reference metrics shown in Table 1.

5.3. Free-Viewpoint Video Evalution

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. Two data-sets are considered, the first courtesy of [10] consists of a street-dancer performing fast acrobatic motions wearing everyday clothing recorded from 8, 1920×1080 resolution cameras, the second courtesy of [11] consists of a Maiko wearing a brightly coloured Kimono performing a slow dance recorded from 16, 1024×768 resolution cameras. Figures 3, 4 illustrate the 3D geometry recovered using a surface optimisation technique [11] with a computational cost of 1 min/frame on an Intel(R) Xeon(TM) 3.6 GHz CPU and a global optimisation technique [10] with a cost of 38 min/frame on an Intel(R) Xeon(TM) 3GHz CPU.

The no-reference comparison is performed using a virtual viewpoint placed at the mid-point between a set of cameras in

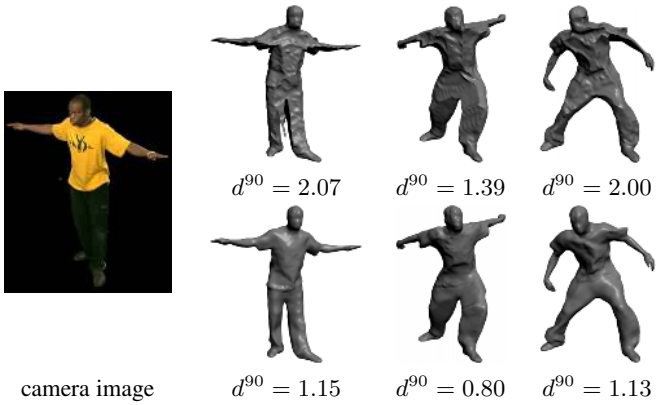


Fig. 3. NR evaluation of two techniques (top) [11], (bottom) [10] for the street dancer sequence courtesy of [10].

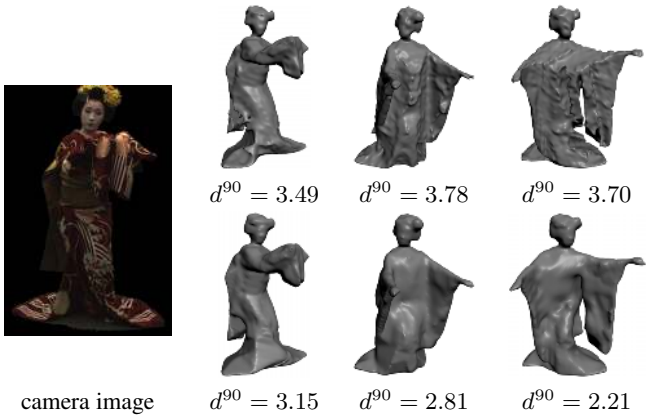


Fig. 4. NR evaluation of two techniques (top) [11], (bottom) [10] for the kimono sequence courtesy of [11].

the studio, a pair of cameras for the planar 8 camera setup and a camera triplet for the non-planar 16 camera setup. The d^{90} metric given in Figures 3, 4 provides an objective comparison of the quality of view synthesis in the virtual viewpoint. Subjectively the computationally expensive global optimisation technique provides a more accurate geometric representation of the scene, objectively the registration error also demonstrates a reduced distortion in view synthesis. 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 data-set dependent.

6. CONCLUSIONS

A methodology has been presented to quantitatively evaluate free-viewpoint video production. The goal of production is defined as accurate view synthesis in the presence of approximate scene geometry. A simple metric is introduced that measures errors in structural registration in view synthesis. Structural registration provides an objective quality measure that is analogous to geometric error where ground truth geometry is not available. 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. The technique can be applied to benchmark the performance of production techniques as well as to provide a means to optimise the parameters of different algorithms. Ultimately such objective measures should be verified by testing subjective perceived quality for typical interactive user behaviour in free-viewpoint video rendering.

7. REFERENCES

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

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

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

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

[5] J. Kilner, Starck J., and A. Hilton, "A comparative study of free-viewpoint video techniques for sports events," *European Conference on Visual Media Production*, pp. 87–96, 2006.

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

[7] S. Winkler, "Video quality and beyond," *Proceedings of European Signal Processing Conference*, 2007.

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

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

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

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