Quantifying Human Reconstruction Accuracy for Voxel Carving in a Sporting Environment *

David Monaghan, Philip Kelly, and Noel E. O'Connor
CLARITY: Centre for Sensor Web Technologies
Dublin City University, Ireland
david.monaghan@dcu.ie

ABSTRACT

Whilst voxel carving approaches exist that allow non-invasive 3D human reconstruction, their performance is heavily dependent on the number of cameras used and the placement of these cameras around the subject. We present a technique to quantify the fall-off in accuracy of spatially carved volumetric representations of humans based on real world constraints. We describe an example of such a quantitative evaluation using a synthetic dataset of typical sports motion in a tennis court scenario, created using computer graphics techniques and motion capture data. Experiments are performed using a baseline voxel carving technique that includes player tracking, background subtraction and player voxel carving. This type of quantitative evaluation could be used by amateur sporting clubs without a sophisticated capture infrastructure to understand how best to instrument a camera network in order to obtain a good trade-off between reconstruction accuracy and installation cost.

Categories and Subject Descriptors

I.4.5 [**Computing Methodologies**]: Image Processing and Computer Vision—*Reconstruction*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Image processing, 3D Reconstruction, Space Carving

1. INTRODUCTION

In order for coaches to both technically and tactically improve a player's performance they must be able to ascertain the deficiencies in an athlete's abilities and effectively communicate to the player how to correct these. In our on-going work, we aim to maximise the impact of coaching feedback by developing tools for both performance analysis and visualisation of player performance. In tennis, for example, 3D visualisation could be used to provide detailed feedback on an athlete's biomechanical movement during specific actions crucial to overall performance. Such a visualisation would be useful for coaches for identifying faults in a player's stroke mechanics during serves, drop shots, etc as well a useful means of conveying this information back to the athlete.

Clearly, 3D human reconstruction is a key component in achieving this, as it forms a core element of both performance analysis and visualisation. For example, Zhao et.al. [10] fit a human skeletal poses to a temporal series of 3D volumetric representations, resulting in the acquisition of the movement of a human's limbs and body over time. This technique could be employed as the basis for the evaluation of sporting performance, for example determining how parameters, such as fatigue, affect an athlete technique over time. However, the accuracy of such skeletal fitting techniques are intrinsically linked to the precision of the 3D human representations. Similarly, an accurate reconstruction of the player at any point in time would allow the coach flexibility when clarifying a tactical or technical issue by allowing the game to be viewed from any angle; or enabling highly beneficial tactical information to be obtained by an athlete from reliving a match from their own viewpoint, a spectator's viewpoint, and from their opponent's viewpoint. The usefulness of this depends on the accuracy of the visualisation.

In an ideal scenario, the 3D human volumetric models would be acquired from voxel carving techniques using a camera network consisting of numerous high quality cameras, with the cameras in the network positioned on the surface of a virtual sphere around the athlete. Real world constraints, however, do not normally allow for this. In a low cost amateur set-up, the quality and number of cameras available is limited. Furthermore, the physical constraints of retro-fitting a camera network to an existing environment will often mean the non-linear placement of cameras that do not provide full 360 degree encirclement. For such a scenario, deciding the optimum camera placement positions can lead to significant gains in the accuracy of generated volumetric models. To the best of our knowledge, however, no tools exist to help inform how best to instrument the camera network in order to ensure maximum quality given a limited budget (number of cameras) and real world practical mounting constraints.

In this work, we describe a technique that, starting from

*Area chair: Qi Tian

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

an idealized capture scenario, can quantify the fall off in accuracy of a 3D reconstruction as the number of cameras is reduced. Using this as a tool, the number, placement and cost of the camera network could be optimised, allowing the highest quality 3D volumetric models to be acquired given a defined budget and the physical constraints of the capture environment (note that we do not claim to address this complex optimization problem here, it is targeted for future work). The paper is organised as follows: Section 2 provides a brief overview on previous work conducted in the area 3D reconstruction techniques. Section 3 gives an overview of the proposed approach. Section 5 provides quantitative and subjective evaluation of the approach and conclusions and future work are outlined in section 6.

2. RELATED WORK

The 3D reconstruction techniques described in the literature range from emerging techniques based on wearable cheap sensors [8] to well-known and broadly used light scanning and infra-red optical motion capture systems. Although systems such as Vicon [2] are generally regarded as being the *gold standard* in accurate motion reconstruction, they can be impractical and cumbersome in some scenarios. Athletic motion, in particular, is challenging because subjects often move very quickly through large spatial volumes that may be difficult to densely and safely populate with enough cameras to ensure adequate coverage and resolution. The practicalities of installing a large number of expensive cameras in such spaces may also be impractical due to the inherent cost of the hardware. Within the context of real world sporting environments, where athletes tend not to be permitted to be instrumented or wear restrictive motion capture suits, a non-invasive technique that does not disturb or distract the athlete is the most desirable solution.

Space carving [3, 5, 6, 7], or voxel (a 3D pixel) carving, techniques can provide such a non-invasive solution and can be implemented on a low-cost camera network allowing the approach to be feasibly employed by both amateur and elite level sports clubs. Space carving techniques create a 3D representation of a subject from a temporally calibrated sequence of 2D images captured from a number of cameras at multiple viewing angles. In each image, a silhouette of the region of interest (i.e. the athlete) is identified and segmented. A virtual cube is drawn around the volume a subject occupies in 3D space and this volume is subsequently populated with voxels. Each voxel is analogous to a virtual lump of clay that is carved by iterating through each of the cameras and eliminating inconsistent voxels from this predefined initial volume using the extracted silhouettes from each image [6]. The resultant voxels occupy the space that corresponds to the most probable 3D spatial location of the subject. However, the accuracy of the resultant 3D representation is heavily dependent on the number of cameras used and their placement.

3. GENERATING A GROUND TRUTH

In this paper, we present a technique to quantify the fall-off in accuracy of a spatially carved volumetric representation based on real world constraints. In this case study, these constraints are derived from (i) a modest budget that will support a limited number of our preferred cameras, and (ii) based on the capture environment that imposes a physical restriction on the placement of these cameras (for example, cameras cannot be placed in areas where they will be damaged, or located where no physical structures exist on which they can be mounted). However, in order to quantitatively calculate the expected error of a space carving from a given camera set-up, a *groundtruth* 3D representation is needed. The groundtruth represents the best possible space carving that could be achieved in the environment, given a near unlimited budget (i.e. a large number of cameras). This groundtruth data can be used to quantify the fall-off in accuracy resulting from the reduction in the number of cameras from an ideal scenario to a scenario governed by real world constraints. However, the generation of this dataset in a real-world scenario is highly complicated as the groundtruth volume of the region of interest is generally unknown to a high degree of accuracy, and acquiring and positioning a very high number of cameras is unrealistic due to budgetary and engineering constraints (it is what we are trying to avoid in the first place!).

We therefore propose the use of synthetic datasets. We first accurately model the capture area in an OpenGl graphical environment. Using this graphical model, a high number of plausible camera positions (and camera intrinsic parameters) can be identified. This number is set to be sufficiently high to ensure saturation in coverage within the required area of capture. By using a hypothetical camera's intrinsic and extrinsic parameters, a camera viewpoint in a real world scenario can be mimicked to a high degree of accuracy in the virtual graphical environment.

In order to generate a groundtruth human, we captured a typical sample set of the types of motions that the subject is expected to perform at capture time. For example, in this work, the area of interest is a tennis court, so we used typical human tennis movements captured in a Vicon motion capture studio. However, a variety of publicly available motions [1] could have been used for different scenarios. The data allows synthetic human motion to be rendered within the virtual reconstruction. Using this virtual environment, synthetic video streams of expected human animation from each of the virtual cameras can be generated. The voxel carving of the synthetic player is then be performed (as described in section 4) from all of the virtual camera viewpoints for any individual time instant or complete video sequence. The resulting volumetric representations are taken as a groundtruth dataset, as they can be seen as the highest quality 3D reconstruction that could be achieved given an unlimited number of cameras.

Using this setup, the carvings from any subset of the camera positions can also be acquired. We quantify the resulting error by calculating the Normalised Mean Square Error (NMSE) of the groundtruth volumetric reconstruction against any reconstruction from an inferior camera setup. The 3D NMSE between the two volumes is possible as the space carvings are written in real world Euclidean coordinates. In this work, the MSE is normalised with respect to the number of voxels used in a carving, typically of the order of 10^5 voxels, to allow for a direct comparison. The NMSE cost function will be 0 if the two volumes are identical, rising to 1 if there is no spatial volume in common between the two volumes.

4. VOXEL CARVING APPROACH

We use our own approach to the acquisition of the 3D

Figure 1: Real-world reconstruction; (Row 1) 5 camera views; (Row 2) Model from 5 camera views.

human volumetric representations that is tailored towards a tennis scenario [4]. The approach is completely autonomous and requires no human interaction. There are several components of this system including player tracking, background subtraction, silhouette extraction and voxel carving. An overhead camera is used for player tracking. Player tracking is required as it automatically approximates player position on the tennis court groundplane, allowing for a choice of the initial virtual cube for voxel carving – note that a good choice can have a considerable impact on the results of the reconstruction process [6]. Using the tracked players bounding box, a cube is then drawn around the player and this represents the most probable spatial volume that is occupied by the player. The cube height is set to a standard player height, in this study 2.4 metres. The *number* of voxels dynamically changes with each frame so as to maintain a constant voxel spatial resolution.

Using intrinsic and extrinsic camera calibration parameters the cameras in which the player-cube appears can be easily identified. The real-world coordinates of this spatial volume can then be transformed into camera coordinates for any of the cameras and specific region where the player appears in each image can be isolated from the rest of the image. By restricting the silhouette extraction algorithm to only search within the bounds of the player-cube in each image the computational effort of the silhouette extraction is significantly reduced. The silhouette extraction algorithm is based on an approximate median background subtraction model [4] and the entire process is autonomous.

After the player silhouettes are constructed, the space carving technique, which follows from the technique of Yang *at el* [9], is performed. To reduce the strain on available computer memory that traditional voxel carving techniques can have, an initial carving is carried out at a low voxel spatial resolution. The spatial volume surrounding the player can therefore be optimised and a second, more accurate, player-cube is calculated. A second space carving is then carried out at a higher voxel spatial resolution. An example carving for a single time instant in a real-world setup with 5 cameras using this approach can be seen in figure 1.

5. QUANTITATIVE EVALUATION

For this study, we evaluate the trade-off between cost (i.e. number of cameras) and reconstruction accuracy using the NMSE measure in a specific application scenario, namely our real-world a tennis court facility. For this analysis, we focused 50 virtual cameras on a single side of a tennis court – cameras were positioned according to the physical constraints of the environment (for example, cameras were not placed on the court, or located in areas where there are no physical structures – as such, cameras could only be placed on a 220 degree arch around the court due to the presence of other tennis courts). Taking several still frames from a video sequence (see figures 3(a), (b) and (c) for example), we implemented the voxel carving algorithm using all 50 cameras to perform the carving operation. The resultant volumetric representation of the tennis player represents the best possible, or groundtruth, space carving that could be obtained. Three such acquired models can be seen in figure 3, rows (i), (iii) and (v), in each of these figures the reconstruction is a volumetric shape and is shown at 8 different angles, each separated by 45 degrees.

We then randomly remove a virtual camera and perform the voxel carving again. This new volumetric representation is slightly different to the previous one as it has been cut with fewer cameras and the decrease in accuracy from the groundtruth carving is quantitatively evaluated using the NMSE measure, as outlined in section 3. Figure 2 shows how the NMSE error indicates the fall off in accuracy of the space carving as the number of cameras is reduced. It can be seen that by using only 5 cameras a space carving with an NMSE of 0.4729 with respect to the groundtruth can be obtained. Some example carvings for synthetic data from our 5 real-world cameras positions, as shown in figure 1, are illustrated in figure 3, rows (ii), (iv) and (vi). It should be noted that the results in figure 3(vi) are noisier than those in rows (ii) and (iv) due the increased complexity of the pose, as the player is leaning forwards and his arms are folded inward generating occlusions. From figure 2 we could thus infer that better quality representations may be achievable by adding another camera (to bring the NMSE < 0.4) but that no further significant improvements beyond this would be obtained until we reached 9 cameras.

6. CONCLUSIONS

In this study we presented a technique to quantify the accuracy achievable via spatially carved volumetric representations based on real world constraints. We used a 3D reconstruction technique developed for an amateur sporting environment and illustrated how we could quantitatively evaluate its performance by simulating realistic human motion and virtual placement of a large number of cameras.

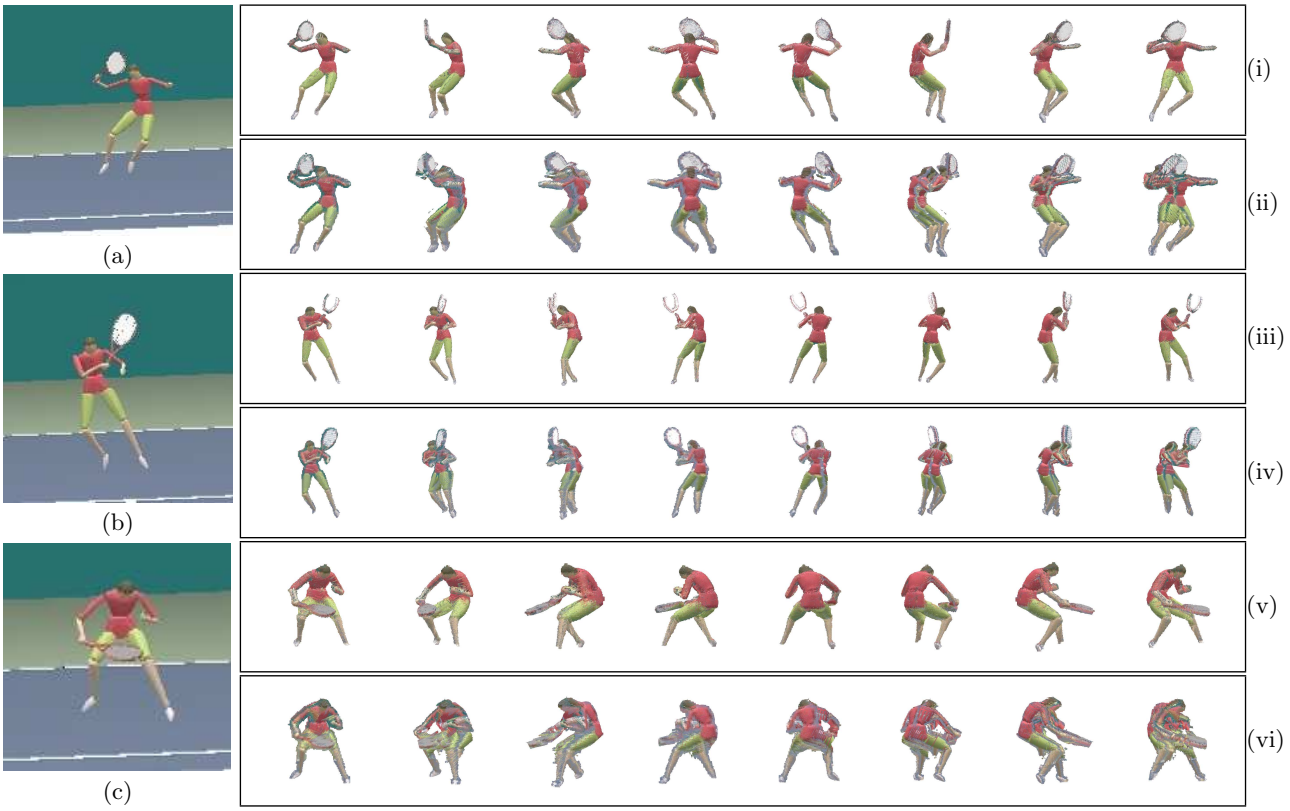

Figure 3: Synthetic data volumetric reconstruction; (a)/(b)/(c) Example viewpoint of model pose; (i)/(iii)/(v) Reconstruction produced from 50 cameras; (ii)/(iv)/(vi) Reconstruction produced from 5 cameras.

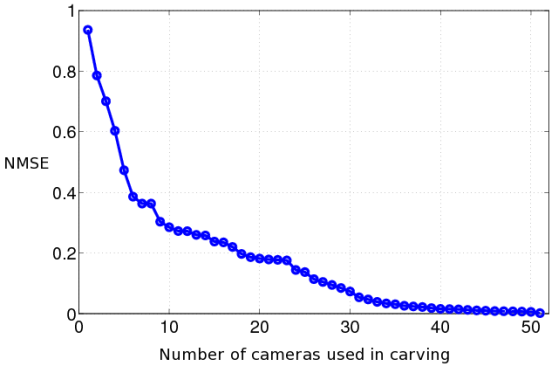

Figure 2: Quantitative evaluation of space carvings with regard to the number of cameras used.

7. ACKNOWLEDGMENTS

This research was partially supported by the European Commission under FP7-247688 3DLife. This work is supported by Science Foundation Ireland under grant 07/CE/I1147

8. REFERENCES

[1] Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2011.

[2] Vicon. http://www.vicon.com, 2011.

[3] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *Conf on Computer Vision*, volume 1, page 388, 2001.

[4] C. O. Conaire, P. Kelly, C. Kim, and N. E. O'Connor. Automatic camera selection for activity monitoring in a multi-camera system for tennis. In *ACM/IEEE Conference on Distributed Smart Cameras*, 2009.

[5] W. B. Culbertson, T. Malzbender, and G. Slabaugh. Generalized voxel coloring. In *Intern. Workshop on Vision Algorithms: Theory and Practice*, pages 100–115, 1999.

[6] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *Intern. Journal of Computer Vision*, 38:199–218, July 2000.

[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.

[8] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.-P. Seidel, and B. Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Trans on Graphics*, 2011.

[9] Y.-K. Yang, J. Lee, S.-K. Kim, and C.-H. Kim. Adaptive space carving with texture mapping. In *ICCSA (3)*, pages 1129–1138, 2005.

[10] X. Zhao and Y. Liu. Generative tracking of 3d human motion by hierarchical annealed genetic algorithm. *Pattern Recognition*, 41:2470–2483, August 2008.