DYNAMIC VOXEL CARVING IN TENNIS BASED ON PLAYER LOCALISATION USING A LOW COST CAMERA NETWORK

David Monaghan, Philip Kelly and Noel E. O'Connor

CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland

ABSTRACT

In this paper, we address the problem of reconstructing 3D volumetric models, illustrating human sporting performance for use in coaching scenarios. We advocate the use of low cost camera networks for acquiring such data, allowing the approach to be feasibly adopted by both amateur and elite level sports athletes. A dynamic voxel carving approach is described, coupled with over-head player tracking and autonomous background subtraction, to automatically produce a 3D reconstruction technique that intelligently uses memory resources. We demonstrate the efficacy of our approach in the context of tennis as a challenging application scenario.

***Index Terms*—** Image processing, Reconstruction algorithms, Voxel carving

1. INTRODUCTION

In order for a coach to improve an athlete's performance, both technically and tactically, deficiencies in the student's abilities must be both *identified* and *eliminated* – starting with larger faults in technique and tactical awareness, and finally fine tuning to remove smaller issues. However, these two complementary objectives of identification and elimination are difficult to achieve in practice. Identifying the factors needing improvement, requires a solid understanding of the many variables critical to optimal performance. In addition, the coach must be able to effectively convey this information to the player for the benefits of the coach's knowledge to be efficiently leveraged.

In this paper, we describe a non-invasive approach for acquiring a 3D reconstruction of a human's performance using a low cost camera network. An accurate 3D rendering of an athlete's motion can be used to maximise the impact of coaching feedback, allowing a coach further flexibility when clarifying a tactical issue, by allowing the game to be viewed from any angle. In addition, the coach has the ability to replay the match from the point of view of either player. This can provide the athlete with highly beneficial tactical information obtained from reliving the match not only from their own viewpoint, or a spectator's viewpoint, but also from their opponent's point of view.

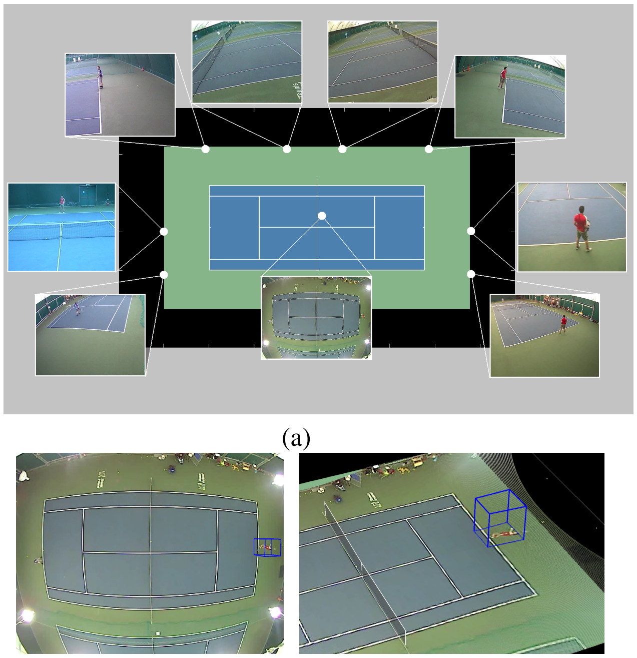
This paper is organized as follows: Section 2 details previous work in the area. Section 3 provides a description of the 3D reconstruction of an athlete's movements from a number of low cost cameras. In section 4 we quantitatively evaluate our approach in a real world test scenario, namely for obtaining rapid motions from a tennis player in competitive action. Tennis is a challenging test scenario, given both the speed and explosive nature of the actions performed by athletes. Finally, section 5 provides conclusions and future work.

2. PRIOR WORK

The main motivation in this work is to extract 3D motion from athletes using low cost techniques, which can be feasibly acquired by most amateur sporting clubs. This data can then be applied in performance evaluation from both training and match scenarios. Although sensor-based techniques could be employed in training, such as the strapping of wireless wearable sensors [1] to players, the use of such technologies in competitive action is currently not a feasible solution, due to sporting body regulations. In addition, despite the relatively small size of such sensors, discussions with elite athletes and sports professionals lead us to believe that their size could still impair the performance of athletes. As such, in this work we focus on the use of non-invasive, image processing techniques from camera networks.

Space carving, or voxel (three dimensional pixel), techniques [2, 3, 4, 5] have received much attention in the literature in recent years. These approaches create a 3D representation of a subject from multiple 2D images, synchronously captured from a number of calibrated cameras at different viewing angles. In each camera view, the subject of interest is identified, and its silhouette is segmented from the image background. The approach of space carving, as described in the seminal work of Kutulakos and Seitz [4], is to progressively remove inconsistent voxels from a pre-defined initial volume, using the extracted silhouettes from each image. The output of the technique is the maximum volume in 3D space which may have produced the extracted silhouettes. However, the choice of this initial volume can have a considerable impact on the outcome of the reconstruction process [4].

Many others have expanded on the Kutulakos and Seitz algorithm. Broadhurst and colleagues present a probabilistic approach to space carving, accessing the likelihoods of a voxel existing or not existing [2]. This approach benefits from assuring no holes are carved in the final model. A more accurate carving, in a moving environment, is produced by introducing a ’hexel’ and performing motion carving in six dimensions in [6]. Yang at el. presents a progressive scheme to improve the reconstruction of surfaces that lack sufficient texture details coupled with robustness to variations in lighting conditions [7]. Kolmogorov and Zabih take a slightly different approach, and employ an energy minimization technique via graph cuts [8]. Bhatia at el. designed a bottom-up 3D limb detection system that utilises a multi-view Eigen model to provide approximate 3D pose information that combines the concepts of space carving and Eigen recognition [9]. Seitz and colleagues presented a quantitative comparison of several multi-view stereo reconstruction algorithms and try to address the challenges in comparing and evaluating multi-view stereo algorithms [5]. Real time space carving implementations have been proposed that use graphics hardware [10] or several computers [11]. Cheung at el. [11] present a near-real time system for voxel reconstruction that uses multiple computers, one computer is used for each camera in the system and a further computer to compute the final voxel carving.


(a)

(b)

(c)

Fig. 1. (a) Camera Setup; (b/c) Player tracking.

3. APPROACH

In general, there are a number aspects required to achieve high quality voxel carvings; (a) a high number of images, captured from multiple calibrated cameras at varying viewing angles; (b) good camera calibration; (c) quality object silhouette segmentation; and (d) the initial volume selected for carving. Within these constraints, the number of images available is generally predetermined by the number of available cameras in a particular physical setup, and good manual calibration can ensure the second.

In this work, we focus on the last two aspects, extending the work of Yang at el. [12] to cope with the difficulties in object segmentation in relatively unconstrained real world environments. In addition, we detail a technique for providing a good choice for the 3D initial volume to be used for carving. The camera setup used in this study, shown in figure 1(a), consists of nine low cost IP-network cameras, (seven cameras with 640×480 pixels and two cameras with 704×576 pixels), forming a 220 degree arch around a tennis court and one overhead camera. This overhead camera was a specific requirement of the tennis coaches, in order to visualise tactical shots and movement during matches. It should be noted that in this scenario, 360 degree ring of cameras was not possible in this case due the presence of three additional tennis courts beside the test court. However this represents a realistic court setup that could be obtained in most amateur sports clubs.

This camera setup tends to differ significantly when compared to those adopted in the voxel carving, both in terms of camera pixel resolution and the spatial volume that reconstruction can occur (the nine cameras cover the 78×36 feet size of a standard tennis court, in addition the overhead camera is 45 feet above the ground). In our experimental setup, the spatial region covered by a bounding box of a human subject can range from maximum of 92×357 pixels (width $\times$ height) when the athlete is close to a given camera, to a minimum of 48×135 pixels when the player is at the opposing end of the court to the camera. In addition, the spatial region covered by the player from the overhead viewpoint is an average of 42×21 pixels. As we are focusing on low cost camera networks, the camera pixel resolution and number of cameras should be kept low, in order to minimise expense. As such, it is imperative that the segmentation quality of the silhouette must be high. For each camera viewpoint, we employ a layered background modelling approach, which includes robust shadow removal techniques [1], to learn the background scene. However, even with such a technique, accurate segmentation is difficult to achieve in real world scenarios, where moving backgrounds, lighting conditions and lack of variability in colour intensity between players and background objects in the scene are unconstrained. In order to maximise the accuracy of the segmentation technique, we leverage the benefits of having multiple calibrated cameras.

An overhead camera is used to perform player tracking [13], which identifies the groundplane player location on the court in real-world Euclidean coordinates (see figure 2(b), where the location of the base of the cube indicates the player position). A player-cube is then placed around the spatial vol-

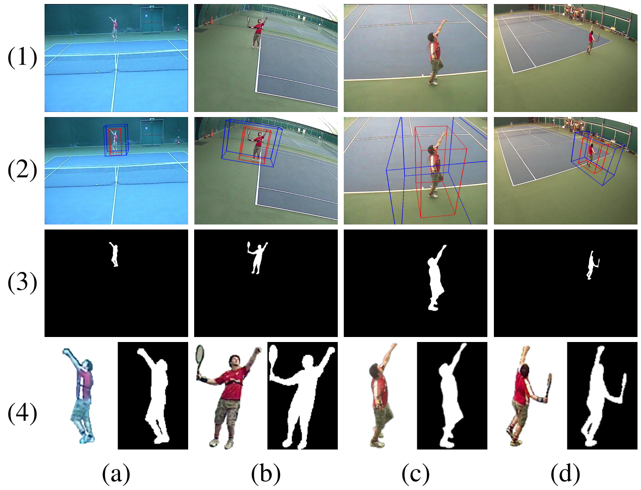


Fig. 2. Player localisation and silhouette extraction.

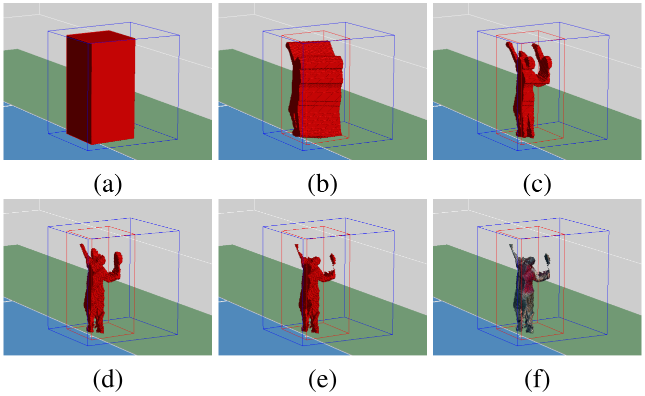


Fig. 3. Volumetric reconstructions using (a) 0, (b) 1, (c) 2, (d) 3 and (e) 4 camera viewpoints. Figure (f) shows the resultant volumetric model with surface colouring.

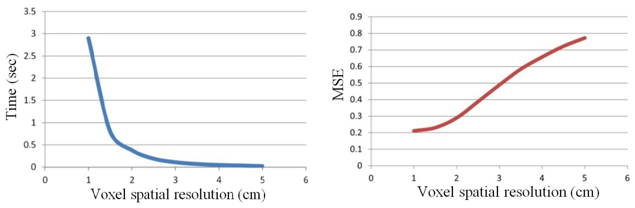
ume occupied by the player. The cube's groundplane dimensions are determined using the bounding box of the player tracker, the height of the cube is set to 2.4 metres (see figure 1(c)). As the intrinsic and extrinsic camera calibration parameters of each the other eight cameras in the network are known, the player-cube location can be obtained in each one of the other camera's local coordinate systems – see figure 2 row (2), for the projection of a 3D player-cube onto four different camera image planes. Using the bounds of the player-cube on each image plane, the player location can be constrained to a specific section of the image. As such, in order to segment the players silhouette from each viewpoint, foreground pixels are obtained (using the layered background modelling technique of Ó Conaire and colleagues[1]) from within the bounds of the 2D player-cube *only*. These pixels are then filtered to remove noise, via connected component analysis and regions under a predefined threshold are removed. This technique of silhouette extraction can cut down on computational complexity, while simultaneously increasing precision on silhouette extraction – see figure 2 rows (2) and (3) for example results of this process.

Once player silhouettes are extracted, space carving as per the technique of Yang at el. [12] can be performed. However, the amount of computer memory utilised in traditional space carving techniques can be both computationally and memory intensive. In order to reduce this complexity, the number of voxels used in the initial volume selected for carving can be decreased. This technique will also reduce the spatial resolution of the 3D volumetric reconstruction. In this work, we use the available computer memory as efficiently as possible by adopting a dynamic space carving methodology, which is controlled by player localisation on the court, again provided by the overhead camera tracker. The initial volume used for carving is centred, as before, on the area occupied by the player. Only this specific volume is filled with voxels. As such, we can keep the spatial resolution of the voxels constant, but the *number* of voxels can be dynamically changed with respect the size of the player bounding box supplied by the overhead tracker. This intelligent use of memory resources means that the approach can run comfortably on a standard desktop computer, producing high quality reconstructions, despite the low cost, low resolution cameras.

4. EXPERIMENTAL RESULTS

All experiments in this section were performed a standard desktop computer; (Windows 32Bit, 3GHz CPU, 4GB RAM). Figure 3(a) shows 3D volumetric reconstruction results, obtained from our approach, on a sample frame using between one and four cameras. In this example, the voxel resolutions in the carved volume are set to 1cm^3. It should be noted, that if the initial volume extended to the whole tennis court, a voxel resolution of approximately 1000cm^3 (one voxel has a resolution of $10\text{cm}\times10\text{cm}\times10\text{cm}$) would have to be set in order for the same amount of memory to be utilised, resulting in a significant decrease in reconstruction accuracy. In figure 3, a maximum of four cameras is used, although there are nine cameras in the network, due to limited camera overlaps, at most the player will appear in four fields of view at a given moment in time. The graph in figure 4(a) illustrates the trade-off between reconstruction time and voxel resolution. For example, the process runs at 38 frames per second when $x = 3$, i.e. a voxel is $3\text{cm}\times3\text{cm}\times3\text{cm}$, or 27cm^3.

A difficulty in evaluating a space carving technique that covers a large area, such as a tennis court, is the lack of available ground truth data for comparison. In an attempt to provide quantitative evaluation figures, the Mean Squared Error (MSE) of a single frames space carving and its associated silhouettes are compared in the graph of figure 4(b). In this graph, a space carving was automatically generated from a



(a) (b)

Fig. 4. Quantitative evaluation results; (a) Reconstruction time versus voxel resolution (where x is the length of each of a voxel's faces in cm); (b) Mean-square error (MSE) versus voxel resolution.

number of sets of input images. The resultant carving was then back-projected into each camera plane, forming a silhouette of the final volumetric model. For each image, this back-projected silhouette was compared to a manually segmented ground truth silhouette using a Mean Square Error metric. This qualitative measure was then averaged across all cameras. This process was repeated and graphed for differing voxel spatial resolutions. As expected, when the voxel spatial resolution is decreased, the MSE decreases accordingly. This highlights the importance of using memory efficiently, to acquire the highest possible voxel spatial resolution.

5. DISCUSSION

In this work, we implemented a viable 3D visualisation technique, by means of dynamic voxel carving, and using a low cost, low resolution, camera network. We demonstrated this approach in the context of tennis, however a variety other real world applications are possible. With the exception of initial camera synchronisation and calibration the entire process is completely autonomous and does not require any manual interaction. In future work, we will investigate the possibility of fitting a 3D human skeleton to the volumetric model and acquiring dynamic motion data associated with an athlete's body. The acquisition of such data from athletes on the field of play would allow to quantify performance data, such as velocities, ground reaction forces and torques, amongst other data to be obtained for evaluation of sporting performance.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] C. Ó Conaire, D. Connaghan, P. Kelly, N. E. OConnor, M. Gaffney, and J. Buckley, "Combining inertial and visual sensing for human action recognition in tennis," in *ACM ARTEMIS*, 2010, p. 51.

[2] A. Broadhurst, T.W. Drummond, and R. Cipolla, "A probabilistic framework for space carving," in *Intern. Conference on Computer Vision*, 2001, vol. 1, p. 388.

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

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

[5] 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*, 2006, vol. 1, pp. 519–528.

[6] S. Vedula, S. Baker, S. Seitz, and T. Kanade, "Shape and motion carving in 6d," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2000, p. 592.

[7] R. Yang, M. Pollefeys, and G. Welch, "Dealing with textureless regions and specular highlights-a progressive space carving scheme using a novel photo-consistency measure," in *IEEE Conference on Computer Vision*, 2003, vol. 2, p. 576.

[8] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *European Conference on Computer Vision-Part III*, 2002, pp. 82–96.

[9] S. Bhatia, L. Sigal, M. Isard, and M. J. Black, "3d human limb detection using space carving and multi-view eigen models," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2004, p. 17.

[10] C. Nitschke, A. Nakazawa, and H. Takemura, "Real-time space carving using graphics hardware," *IEICE - Trans. Inf. Syst.*, vol. E90-D, pp. 1175–1184, 2007.

[11] G. K.M. Cheung, T. Kanade, J-Y Bouguet, and M. Holler, "A real time system for robust 3d voxel reconstruction of human motions," in *Computer Vision and Pattern Recognition*, 2000, vol. 2, p. 714.

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

[13] Ciaran O Conaire, Philip Kelly, Chanyul Kim, and Noel O'Connor., "Automatic camera selection for activity monitoring in a multi-camera system for tennis," in *ACM/IEEE Intern. Conference on Distributed Smart Cameras*, 2009.