2011 15th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications

Camera image synchronisation in multiple camera real-time 3D reconstruction of moving humans

Tobias Duckworth, David J. Roberts
Centre for Virtual Environments and Future Media
University of Salford
Manchester, UK
Email: t.w.duckworth@edu.salford.ac.uk

***Abstract*—We present an analysis of the requirement for input synchronisation in a multi-camera 3D reconstruction system for real-time applications such as telepresence. Synchronisation of the cameras at the acquisition stage is universally used to ensure the images feeding the reconstruction algorithm were taken at the same time. However, this requirement adds delays to the reconstruction pipeline, therefore increasing the end to end latency of the system. While this has not been a significant problem for many of the applications of 3D reconstruction, it is for its application to tele-presence. Furthermore, synchronising the firing of cameras adds much financial cost to the system. Using real camera images of moving humans, we study the effect removing synchronisation has on the output reconstructed model over a range of camera configurations and relative frame delays. From this we determine the synchronisation requirements for a 3D reconstruction telepresence system in terms of the maximum time between camera frames that gives rise to acceptable results.**

***Keywords*-Tele-immersion; telepresence; 3D reconstruction; synchronization**

I. Introduction

Real-time 3D reconstruction from multiple images is an area of research that could yield great advances in applications such as telepresence. While 2D video can communicate what someone looks like and 3D virtual avatars what they look at, 3D reconstruction from video could do both. That is, if only it could be done at sufficient temporal and image quality. Reconstruction of 3D forms from multiple images is a popular area of research in the field of computer vision. There are numerous applications for systems capable of determining the 3D shape of an object, and the specific requirements of such a system depend on the application. With this approach a 3D shape is derived from a set of camera images taken from various sides of a real object. Where the object to be modelled is a static rigid body, there is no requirement to ensure that the pictures are taken at the same time. This may be achieved by either by moving the camera around the object, or by moving the object relative to the camera. In the case of dynamic objects, synchronised images from several cameras surrounding the object must be used to reconstruct the form and position of the object at a particular moment in time. Achieving high levels of synchronisation greatly increases the cost of system, the complexity of time management and potential for delay. There has been very little research into the use of unsynchronised cameras when modelling dynamic objects, and in this paper we attempt to develop further understanding on this subject.

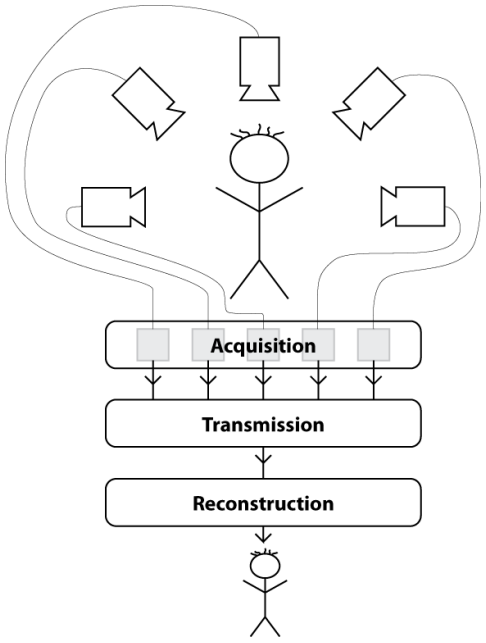

Figure 1: The 3D reconstruction process from camera image acquisition to the reconstruction algorithm.

Synchronisation of frames from many cameras is typically employed at two levels: capturing; and delivery to an algorithm on a different machine. Lack of capture synchronisation would result in slight time offsets in the frame acquisition of each camera. Synchronisation of capture is typically achieved by a common hardware trigger. Various triggering approaches have distinct levels of synchronisation, a lack of which produces a relative temporal drift between frames. Whilst high quality hardware triggering is a good and robust solution, it requires that cameras are equipped

1550-6525/11 $26.00 © 2011 IEEE
DOI 10.1109/DS-RT.2011.15

with a suitable interface for such a signal, which rules out most commodity cameras. This and the sending unit greatly increase the cost of the system. Delivery delay comes from the a chain of components that depends on the system design. A typical system will capture images from each camera using a separate computer and then send these across a network to the computer that runs the reconstruction algorithm. Various forms of time management could be used to ensure this synchronisation. Before embarking on the design of a higher performance time management scheme, this paper aims to better understand the temporal requirements such should meet.

II. Related Work

The field of multiple camera image based modelling relies on two fundamental concepts: Shape from Silhouette, and the visual hull. The concept of shape from silhouette was first introduced in [1] This work proposed that a 3D form could be roughly geometrically approximated from the intersection of a finite number of silhouettes of images taken around the object. Laurentini [2] later proposed the concept of the visual hull as a theoretical entity which could be constructed from the intersection of an infinite number of silhouettes of the object to be modelled. The visual hull is the maximal surface enclosing the form of the object, it is unable to represent any surface concavities. Since these important contributions, the visual hull and methods for constructing it have been the focus of much study in the field of computer vision. These methods largely fall into two categories: Volumetric approaches reconstruct the volume of the object using voxels to represent a region of space. Surface based approaches reconstruct the surface of the object, often using polyhedral geometry to do so. There exists a third category that is able to generate an arbitrary viewpoint of the object from images, but does not create a 3D model, this is known as image based rendering.

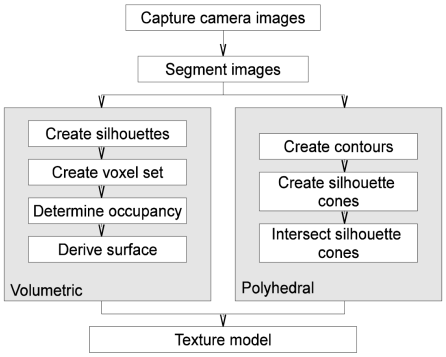

Figure 2: Overview of two common shape from silhouette techniques

Most work describing 3D reconstruction systems claim to use synchronised frames, although the level or method of synchronisation is rarely mentioned. We have found none that claim not to do so. Few give specific details of the exact synchronisation system in place, but some mention use of a hardware trigger. This is a signal shared by all cameras in the system, and usually driven by a host computer. The cameras can be configured to acquire images on the rising or falling edge of the signal. Use of such a hardware triggering mechanism requires that cameras are equipped with a suitable interface for the signal. This rules out use of commodity cameras such as USB webcams if the system requires synchronised images. Origami [3] and Blue-C [4], both use a hardware triggering system to ensure simultaneous image acquisition from all cameras. Wu and Matsuyama [5], claim to use a software triggering system in which cameras connected to individual computers are triggered by sending a command over the network. Matsuyama et al [6], also study the effect synchronization has on the performance of a 3D reconstruction system. Towles et al [7], used a hardware acquisition trigger, and frames were sent across the network synchronized, but arrived unsynchronized. Some camera streams were one or more frames behind the others. They found that this was due to the TCP protocol competition among the streams.

We previously published work [8] in which the synchronization requirements of a real-time 3D reconstruction system was studied using simulation. We analyzed reconstructed model deformation using a shape from silhouette reconstruction algorithm for a variety of camera synchronization scenarios using simulated cameras and images. The images where taken of a virtual human using virtual cameras. It was shown that unsynchronized camera images resulted in negligible distortion in the reconstructed 3D model. The shortcoming of this work were:

- Images were not of a real person
- Images did not come from real cameras
- Only studied the distortion arising from the rotation of a human head
- The reconstructed models were not textured, and as such only the effect on the form of the model was studied.

In this research we aim to improve upon the previous study by using images of the entire human body moving, taken from real cameras. This will provide a better understanding of the effect of synchronization, or lack of it, in terms of the whole human body and the typical movements of each part of it.

III. Approach

We aim to determine whether synchronised cameras are necessary for a 3D reconstruction system, since removal of the synchronisation requirement could significantly increase performance. Camera synchronisation for the pur-

poses of 3D reconstruction falls into two distinct categories: Synchronisation of the acquisition of camera images, and synchronisation of the delivery of these images to the reconstructing algorithm. For the purposes of this paper, we will be focussing on the effects of delivery synchronisation. In order to achieve this we use pre-recorded datasets from synchronised cameras, which are fed to a 3D reconstruction algorithm, and the effect on the reconstructed model of delaying one or more frames to the algorithm analysed.

We have re-implemented a state of the art polyhedral 3D reconstruction algorithm, Exact Polyhedral Visual Hulls (EPVH) [9] [10], which is based on the shape from silhouette concept. Details of our accelerated implementation of the algorithm can be found in [11]. The algorithm produces a manifold and watertight mesh in the form of a polyhedral surface representation of the visual hull. For texturing the resulting model, we calculate the surface normal of each polygon (Figure 3), then determine which camera's principle ray is pointing in the opposite direction, and closest to parallel to it. This is a fairly crude method which can lead to some artefacts, for example no occlusion detection is employed, and therefore a polygon can be textured with a camera which cannot observe that region. More advanced texturing methods, such as blending the best three matching camera images have been reported by other researchers, but add additional processing cost to the reconstruction system.

Figure 3: Texture camera determination using surface normals

For the purpose of this experiment we use images from synchronized cameras, stored on disk, we delay the frame for certain cameras one at a time, so that the images used by the reconstruction algorithm are no longer synchronized. This provides a perfectly synchronized set of images to use as a control against which to compare the results of the unsynchronized reconstructions. For example, if four cameras were used during capture of the sequence, we might choose frame 4 as the reference frame, and begin by providing frames [4, 4, 4, 4] to the reconstruction algorithm to produce the perfectly synchronized control model. Following this, to simulate unsynchronized cameras, for example as a result of network transmission delays, we could choose to delay one or more cameras by providing the frame previous to the reference frame, or even earlier frames to the reconstruction algorithm. For example, providing frames [3, 4, 3, 4] would result in a reconstructed model where two of the camera frames were delayed by a single frame. Visual inspection of the reconstructed model can then be used to determine how the unsynchronized cameras have affected the resulting model.

The reconstruction algorithm was run on an Apple Mac Pro with 2 x 2.8GHz quad core CPUs and 18GB of RAM. Images were loaded from a pre-captured dataset stored on disk. Rendering was achieved using an nVidia GTX 285 graphics card with 1GB of video memory. Three datasets were used in the study:

- Dancer [12] was from Inria's 4D Repository, shot at 780 x 582 pixels using 8 cameras at 30 frames per second.
- Nikos [13] was from Surrey's i3DPost Multi-view human action dataset, shot at 1920 x 1080 pixels, using 8 cameras at 25 frames per second.
- Juggler was from our own facility, shot at 1004 x 1004 pixels, using 6 cameras at 10 frames per second. For this dataset, cameras were synchronised with a software trigger, reported in a sister paper as achieving 9ms between frame starts.
- Martial [14] was from Inria's 4D Repository, shot at 1624 x 1224 pixels, using 16 cameras at 30 frames per second.

For each dataset we first reconstruct the correctly synchronised model from the corresponding frames. Then we reconstruct models with a new camera delayed by a frame each time. When half the number of cameras used by the dataset are delayed, we consider this the maximum single frame delay possible for that dataset.

IV. Results

We present the results of the effect on reconstruction of the frame delays as textured reconstructed models for visual inspection. For each dataset, the correctly synchronised reconstructed model is first presented, followed by the models from the reconstructions with frame delays. The delayed models are presented in order of increasing number of cameras being delayed.

A. Dancer

The sequence is filmed at 30 frames per second using 8 cameras at a resolution of 780 x 582 pixels. There are therefore 33.3 milliseconds between frames.

Figure 4: Dancer Correctly synchronised reference reconstruction

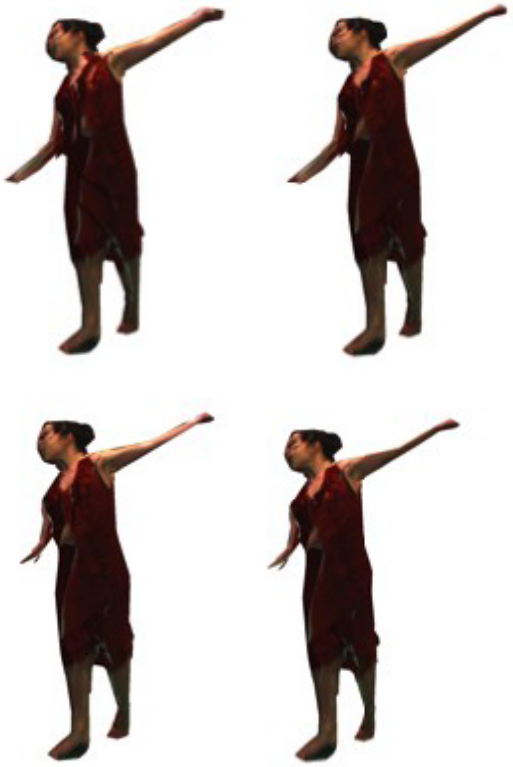

Figure 5: Dancer Top left 1 camera delayed by a frame (33ms), top right 2 cameras delayed by a frame, bottom left 3 cameras delayed by a frame, bottom right 4 cameras delayed by a frame.

The only noticeable quality degradation in the reconstruction with unsynchronized camera images (Figure 5) is for the limbs, particularly the arms, although some distortion is visible in the rear leg as more cameras are delayed. The dancers right arm is almost entirely lost when four camera images are delayed. Thinning can be observed on her left arm. There are no obvious texturing errors. The corresponding frames from the video sequence have been analysed, and it can be observed that there is a fair amount of displacement in the dancer's right arm, and some movement in the left arm too. The thinning of limbs in the reconstructed model corresponds to the parts of the body that move the most between frames.

B. Juggler

The sequence is filmed at 10 frames per second using 6 cameras at a resolution of 1004 x 1004 pixels. There are therefore 100ms between frames.

Figure 6: Juggler Correctly synchronised reference reconstruction

These images were taken from cameras with a software networked synch with recorded synchronisation of between 0 and 9 msecs (As reported in a sister paper submitted in parallel). With all frames synchronised to within this accuracy, it is hard to attribute any related errors in the reconstruction. With one camera delayed by a frame (Figure 7a) there is a noticeable texturing artifact, the juggling ball now appears as part of the texture on the front of the jugglers body. With two delayed cameras, the real juggling ball disappears altogether, and there are further texturing artifacts visible in the hand region. With 3 delayed cameras there is visible distortion in the reconstruction around the hand and lower arm area, as well as obvious texturing errors.

Figure 7: Juggler Top left 1 camera delayed by a frame (100ms), top right 2 cameras delayed by a frame, bottom 3 cameras delayed by a frame.

C. Nikos

The sequence was filmed at 25 frames per second using 8 cameras at a resolution of 1920 x 1080 pixels. There are therefore 40ms between frames.

The synchronised model (Figure 8) is of fairly high quality, due largely to the use of full HD cameras in the dataset. Delaying a single camera by a frame (Figure 9a) drastically reduces the faithfulness of the reconstructed model. The face is immediately affected, and the left forearm is shortened. Delaying further cameras results in a progressive thinning of the limbs in motion, a shortening of the arms, and degradation of the quality of head reconstruction. Texturing also suffers from the unsynchronised cameras, the left forearm can be seen projected onto the side of the body.

D. Martial

The sequence is filmed at 30 frames per second using 16 cameras at a resolution of 1624 x 1224 pixels. There are

Figure 8: Nikos Correctly synchronised reference reconstruction

Figure 9: Nikos Top left 1 camera delayed by a frame (40ms), top right 2 cameras delayed by a frame, bottom left 3 cameras delayed by a frame, bottom right 4 cameras delayed by a frame.

therefore 33ms between frames.

Figure 10: Martial Correctly synchronised reference reconstruction

Due to the high camera count in this dataset, it was expected that by delaying a single camera few visible results would be observable. However, it can be seen in the first reconstructed image (Figure 11a) that the assailant's right foot is lost. It should be mentioned that a different choice of which camera to delay could have resulted in different observable loss in the reconstructed model since the camera silhouette containing the foot segment may not have differed. As the number of cameras being delayed increases, other losses can be observed in the reconstructed model. The assailant's right leg shortens, and his left arms progressively thins and truncates. Fewer differences can be seen in the reconstructed defendant, probably because his body is moving more slowly than the assailant's. However, there is noticeable shrinking of the head and upper body. Whilst some texturing problems arise, this are reduced compared to the other datasets tested. It is likely that this is because there are more cameras from which to select textures, and therefore more suitable candidate cameras for each polygon.

V. Discussion

From the datasets studied, we have clearly shown that when one or more frames are out of sequence, the reconstruction of the human form is compromised to a generally unacceptable level. Conversely, synchronisation errors that do not exceed a frame did not cause noticeable degradation. We have focussed on whole frame delays for up to half the cameras used in a dataset Since we found that for all datasets loss occurred within a single frame we did not test delays of greater than one frame, as this would certainly lead to greater loss in the resulting reconstruction. Delays of a frame or more are likely to be caused by unsynchronised

Figure 11: Martial Top left 1 camera delayed by a frame (33ms), top right 2 cameras delayed by a frame, second row left 3 cameras delayed by a frame, second row right 4 cameras delayed by a frame, third row 5 left 5 cameras delayed by a frame, third row right 6 cameras delayed by a frame, bottom left 7 cameras delayed by a frame, bottom right 8 cameras delayed by a frame.

delivery in a real 3D reconstruction system, whereas delays of less than a frame are likely to be cause by unsynchronised acquisition. Therefore we have demonstrated that delivery synchronisation is important for faithful model reconstruction, but have not studied acquisition synchronisation. It is not possible to study acquisition synchronisation using pre-recorded datasets from synchronized cameras, as the minimum granularity available for simulating the desynchronisation is the time between frames. Our previous work on this subject studied the effect unsynchronised acquisition had on reconstruction of virtual objects, and we are undertaking further research on the subject using real cameras. It is likely that unsynchronised acquisition will result in similar reconstruction errors to those observed in this research, but probably on a smaller scale, since the amount of time cameras will be out of sync with each other will be up to an entire frame period and not more.

VI. Conclusions and future work

We have shown that for dynamic sequences of human movement, synchronisation of camera frames to within the period of one frame is necessary. We have not conclusively shown that sub frame synchronisation has a visually identifiable effect. Camera synchronization is important in faithfully reproducing the moving human form. This is particularly true for the limbs, which can move significant distances in the time between camera frames. Our previous work on the subject found that negligible model deformation occurred for unsynchronised cameras, but that research only looked at the rotation of the human head, used much shorter desynchronisation between cameras, and did not use textured models. We have not looked specifically at the human head in this work, but instead the entire human body. Since our goal is to develop a telepresence system for human communication, it would be useful to repeat this study looking specifically at the human head, and the effect desynchronised cameras have on lip movement, eye gaze, and facial expression, especially with the texturing of the resulting models. This forms the focus for future work.

Acknowledgments

This research was supported by EPSRC, OMG Vicon and Electrosonic Ltd.

References

[1] B. Baumgart, “A polyhedron representation for computer vision,” *AFIPS ’75: Proceedings of the May 19-22, 1975, national computer conference and exposition*, May 1975.

[2] A. Laurentini, “The visual hull concept for silhouette-based image understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[3] O. Grau, T. Pullen, and G. Thomas, “A combined studio production system for 3-D capturing of live action and immersive actor feedback,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 370–380, 2004.

[4] M. Gross, S. Wurmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. van Gool, and S. Lang, “blue-c: A spatially immersive display and 3D video portal for telepresence,” *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 819–827, 2003.

[5] T. Wu and T. Matsuyama, “Real-time active 3D shape reconstruction for 3D video,” *Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on*, vol. 1, pp. 186–191 Vol.1, 2003.

[6] T. Matsuyama, X. Wu, T. Takai, and T. Wada, “Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 357–369, 2004.

[7] H. Towles, S. Kum, T. Sparks, S. Sinha, S. Larsen, and N. Beddes, “Transport and rendering challenges of multi-stream, 3d tele-immersion data,” *NSF Lake Tahoe Workshop on Collaborative Virtual Reality and Visualization (CVRV’03)*, 2003.

[8] C. Moore, T. Duckworth, R. Aspin. D. Roberts *Synchronization of Images from Multiple Cameras to Reconstruct a Moving Human*. , DS-RT ’10 IEEE Computer Society, Oct. 2010.

[9] J. Franco and E. Boyer, “Exact polyhedral visual hulls,” *British Machine Vision Conference*, Jan. 2003.

[10] Franco, “Efficient Polyhedral Modeling from Silhouettes,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 414–427, 2009.

[11] T. Duckworth and D. Roberts, “Accelerated polyhedral visual hulls using OpenCL,” in *Virtual Reality Conference (VR), 2011 IEEE*, 2011, pp. 203–204.

[12] http://4drepository.inrialpes.fr/public/viewgroup/1

[13] http://kahlan.eps.surrey.ac.uk/i3dpost_action/data

[14] http://4drepository.inrialpes.fr/public/viewgroup/4