2010 International Symposium on Ubiquitous Virtual Reality

Image-based modelling for augmenting reality

(Invited Paper)

Anton van den Hengel
Australian Centre for Visual Technologies
School of Computer Science
University of Adelaide, Australia
Anton.vandenHengel@adelaide.edu.au

***Abstract*—The interaction between real and synthetic geometry is fundamental to Augmented Reality. Portraying such interactions in a manner that is convincing to the user requires 3D models of the shape of the real objects involved. The large-scale application of Augmented Reality technologies will thus require practical methods for generating 3D models of real objects. These methods will need to be fast, flexible, and capable of operating in-situ in order to generate models in unforeseen environments. Image-based modelling offers a means of creating such models by direct analysis of an image set. This paper describes two approaches to image-based modelling for Augmented Reality, and argues that technologies of this type are critical to the future of the domain.**

***Keywords*-component; formatting; style; styling;**

I. Introduction

Methods for estimating the shape of an environment have become an important component of practical Augmented Reality (AR) systems. The ability of any AR system to operate in an unknown environment critically depends on the ability to recover this information. A number of methods and modalities are possible, from live laser range finding to pre-calculated models. The challenge for Ubiquitous Augmented Reality (UAR) however, is to make AR both flexible and practical.

Image-based Modelling (IBM) offers the prospect of building a 3D model of an environment by analysing an image set. In the AR case this may well be the set of images captured by a head mounted camera, or a more diverse set, possibly to accessed over the network.

This paper describes 2 approaches to the problem of IBM for UAR, both developed within The Australian Centre for Visual Technologies. These technologies exploit IBM as an in-situ modelling tool, allowing the user to generate 3D models of real objects on the basis of images taken by the AR system itself. The fact that this is in-situ IBM is important to it's application, as UAR implies the deployment of AR within unknown environments. Pre-existing models cannot be relied up in such circumstances, and a separate modelling process is unlikely to represent a practical solution. The modelling must take place within the AR system itself.

The first IBM system for AR we describe is labelled Jiim, and has an interface aimed at a tablet-style device.

Figure 1. The gnome desktop, showing the original (modelled) object on the left and two synthetic copies inserted into the live video.

The modelling process is interactive, and suitable for a wide variety of objects, from buildings to bagels.

The second system is still in development, but the design goal has been to reduce the amount of interaction required in order to generate a model. The method is thus more suitable to a head-mounted display, or similar. One of the applications of particular interest is that of inserting special effects into video as it is taken. This would allow location scouts to visualise synthetic geometry in-situ as they assess the suitability of a location for a particular scene. More interestingly, however, it might also allow more general users of video cameras to perform advanced 3D video editing operations while the video is captured. Figure 1, for instance, shows a frame from a sequence demonstrating the process of modelling geometry in-situ and inserting it back into the same live video stream. This is may thus be seen as a copy and paste of real objects from and to live video.

II. Image-based modelling

Image-based modelling is the process of creating a 3D model of an object on the basis of a set of images or a video sequence. Many approaches have been developed, from the fully automatic to the largely manual. For example, Photomodelerand Facaderequire the user to interactively specify shapes in the scene by marking corresponding primitives in

978-0-7695-4124-2/10 $26.00 © 2010 IEEE
DOI 10.1109/ISUVR.2010.11

Figure 2. A screen capture from Jiim showing a synthetic car model leaving the end a real ramp, and shadowing both the ramp and a toy ambulance. The geometry of the table, ramp, and ambulance required were modelled within Jiim in under a minute.

multiple images. Based on image markup, they estimate the parameters of the cameras that took the images, and thence the 3D shape and texture of the scene.

Jiim uses an image based modelling approach based on VideoTrace [1], an interactive system by which a user can generate a 3D model of an object by simply tracing over it in an image. VideoTrace thus takes as input a previously captured video sequence depicting the object to be modelled. This video sequence is passed to a camera tracker (such and Boujou [2] or PFTrack [3]) in order to extract such as the path of the camera and a set of reconstructed 3D point locations. On the basis of this information, and the user interaction, a texture-mapped 3D model of the object is generated. VideoTrace thus allows the user to specify the model they require by tracing out the edges of the polygons from which it is to be constructed. This interaction is carried out over a frame or frames of an input video sequence, and involves purely 2D interactions (see Figure 4). One of the key ideas behind VideoTrace is that the image *is* the interface. The interactions are all 2-dimensional, and thus well suited to being carried out using a mouse or similar. The 3D is implied by analysing combination of the image set and the interactions.

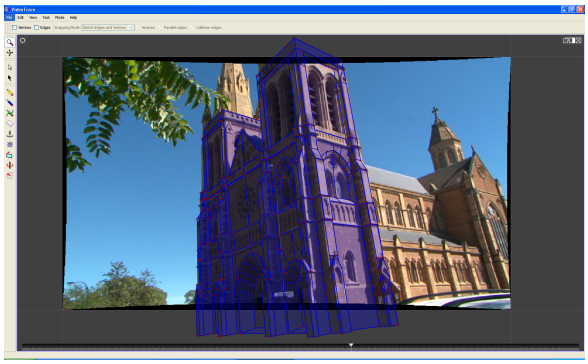

Figure 3. Modelling a building using VideoTrace

The primary advantage of live image-based modelling over capture-then-process approaches is the fact that it gives the user the ability to see the model as it evolves and capture further data as required to complete the model. It is typically the case when using a capture then-process-approach that multiple attempts at capturing the images are required to complete the modelling process.

A. Real time camera tracking

One way to improve the integration of data capture and modelling is to use camera tracking software that operates in real time. Methods for real-time camera tracking have evolved from fiducial marker-based approaches, to those based on simultaneous localisation and mapping (SLAM) which are marker-less and require neither additional hardware nor a-priori knowledge of the camera or environment.

The MonoSLAM system [4] showed that SLAM can recover, in real-time, the path of a single camera and a set of sparse 3D points (called a map) which describes the shape of the scene. A key limitation of MonoSLAM and similar systems is the sparsity of the 3D map typically maintained by these approaches, whose primary goal is to estimate camera pose relative to selected keypoints in the scene rather than to obtain a complete model of the scene's shape. The PTAM system of Klein and Murray [5], which combines real-time camera tracking with incremental bundle-adjustment, is able to build far denser maps containing over 10,000 point features. It does this by decoupling the camera tracking from map estimation, updating the map estimate using bundle adjustment while in parallel updating camera state using a faster, frame-rate process. By applying PTAM to the live video, we obtain an estimate of the current camera, relative to a fixed world coordinate system, and a map of 3D scene point locations which is dense enough to form the basis for interactive image based modelling software.

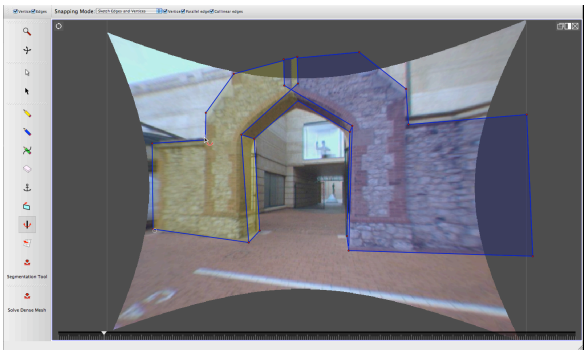

Figure 4. Part of the process of modelling an architectural feature using Jiim. A polygonal mesh is traced out over the corresponding structure in the image, and its 3D location is estimated using available image data. Each image is automatically undistorted.

III. In-situ modelling

In-situ modelling is a process whereby models are constructed using the Virtual Reality (VR) or AR system within

which they will be used [6]. Many such modelling systems exist, and the advantages of the in-situ approach have been well documented (see [6] for a survey). However, the modelling facilities in these systems are typically not designed to create models that accurately represent objects in the world, and using them for this purpose can be somewhat laborious. Examples include Piekarski et al. [7] which proposes a 3D constructive solid geometry approach to the construction of models within the Tinmith AR system, and Baillot et al. [8] a more CAD-like interface. Of note also is [9] which uses a contact probe to model a surface within an AR system. None of these systems perform any analysis of the image data, meaning that the modeller must fully specify all aspects of each object. Bunnun et al. in [10] propose a SLAM based and hence image assisted modelling process using a camera attached to a mouse, but this requires that each vertex in the model is individually specified in multiple images. Kim et al. in [11] describe an 'online 3D modelling' approach which uses satellite images to model outdoor structures for their AR system.

The distinction between these systems and the approach we propose here is that the latter facilitates user-assisted generation of accurate synthetic models of real objects based on analysis of video. Using PTAM and image based modelling methods, information is extracted automatically from the video which reduces the number of interactions required to construct a model which accurately reflects the shape of the real geometry.

A. *In-situ image based modelling*

The fact that the frame of reference for the modelling process and the use of the model are the same eliminates the potential for misalignment. The approach we propose here thus uses no position information other than that recovered through the SLAM process. In-situ modelling also allows the user to make direct comparisons between the real geometry and its synthetic counterpart, greatly simplifying the verification process.

Another advantage of integrating the video capture and modelling processes is that it allows the user to identify instantly any of the image data required to generate the model which is missing, and capture it. This has the effect of ensuring that the user returns from the scene with a complete model, rather than a subset of the data required to create one.

IV. Jiim

Jiim is In-situ Image-based Modelling (JIIM), and was first presented in [12]. The method uses information gained through automated analysis of video to empower an interactive 3D modelling process. The result is a flexible and efficient method for creating accurate 3D models of real objects in the scene. These models can be used within the AR system itself to enable real and synthetic objects to interact convincingly, or for non-AR purposes such as importing into Google Earth. The creation of the models and their use in AR can be interleaved, allowing "on demand" creation of the minimal 3D structure that is necessary for a particular application.

V. Minimal interaction modelling

The level of interaction required to generate a model by the Jiim approach renders it unsuitable for devices and applications where this is impractical. This includes phones and to a lesser extent head-mounted displays, but also video cameras.

The goal of the minimal interaction approach to modelling has thus been to reduce the interaction requirement to that of simply identifying the object to be modelled. The goal is to reach the point whereby the interaction is semantic, rather than geometric. Modelling using the silhouette-based system we have developed does not quite achieve this ambition, as the user is required to drag a crosshair across the object to be modelled. The crosshair being fixed within the field of view of the camera. This is, however, a one button operation, the user pushes the button to indicate the beginning of the stroke, and releases it at the end of the stroke. The object indicated in this manner will then be modelled without further interaction.

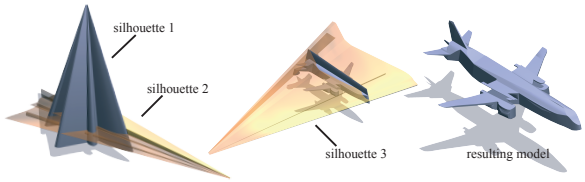

Figure 5. Silhouette intersection modelling.

The silhouette, or outline, of an object conveys a lot of information about its shape. The modelling process we propose combines the shape information from successive silhouettes in order to generate a 3D model. The set of shapes which may be modelled using silhouettes is very broad, which is important given the proposed application. The silhouettes of an object may also be identified using the segmentation process described above, which allows the majority of the modelling process to be performed in-camera.

Because the projection of every point on an object's surface must lie within its silhouette in all views, the 3D surface of an object must lie within the *visual cone* formed by back-projecting its segmentation into scene-space . The visual cone for one image is defined by the surface silhouette, which represents the cone's cross-section, and the camera's optical centre, which is its apex. Each new view (and segmentation) increasingly constrains the shape, leading to the *visual hull* constructed by the intersection of each camera's visual cone. Although the visual hull is

the tightest bound on the reference surface given only its silhouette, it cannot recover holes. This problem is somewhat compensated for by the view-dependent texturing process described below, however.

A. Visual hulls

The silhouette-based method generates high fidelity visual hulls with an associated projective texture. These models are suitable for a wide class of objects and applications, but less so for objects which exhibit significant transparency or holes. Figure1 shows a copy and paste application within live video. The system has been used to copy the gnome and duplicate it back into the same video sequence. This operation in live, and in-situ, no other programs are used and the process of modelling directly precedes the insertion back into the video. The insertion process is again a one-button interaction, although it does require a second button in order to distinguish it from the modelling interaction.

Figure 6. A desktop monster and a synthetic copy added into the live video.

Figure 6 shows the result of a similar copy and paste within live video. The fidelity of the models is suitable for the application, to the extent that it can be difficult to distinguish between the real and the synthetic objects. In the live video this distinction is clear as inaccuracies in the SLAM process cause the synthetic object to move subtly relative to its surroundings.

VI. Conclusion

The quality of the results produced by the methods we have described depend upon the quality of the camera tracking and point location information provide by PTAM. Although effective, SLAM is not perfect, and often loses track, particularly in outdoor environments. SLAM systems typically have a number of other limitations also, such as requiring a wide–angle camera and losing track under certain circumstances [5]. SLAM technology is developing, however, and our approach is independent of the particular SLAM system used.

Despite current limitations, however, it seems inevitable that IBM will form an important part of the future of practical AR.

Acknowledgment

This research was supported under Australian Research Council's Discovery Projects funding scheme (project DP0988439).

References

[1] A. van den Hengel, A. R. Dick, T. Thormählen, B. Ward, and P. H. S. Torr, "VideoTrace: Rapid interactive scene modelling from video," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[2] T. Thormählen and H. Broszio, "Voodoo camera tracker," free download at www.digilab.uni-hannover.de.

[3] The Pixel Farm, "PFTRACK: A commercial camera tracking and image based modelling producthttp://www.thepixelfarm.co.uk." [Online]. Available: http://www.thepixelfarm.co.uk

[4] A. J. Davison, I. D. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[5] G. Klein and D. W. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. International Symposium on Mixed and Augmented Reality (ISMAR'07, Nara)*, 2007.

[6] G. A. Lee, G. J. Kim, and M. Billinghurst, "Immersive authoring: What you experience is what you get (wyxiwyg)," *Commun. ACM*, vol. 48, no. 7, pp. 76–81, 2005.

[7] W. Piekarski and B. H. Thomas, "Tinmith-metro: New outdoor techniques for creating city models with an augmented reality wearable computer," in *Proc. 5th IEEE International Symposium on Wearable Computers*, 2001.

[8] Y. Baillot, D. Brown, and S. Julier, "Authoring of physical models using mobile computers," in *Proc. 5th IEEE International Symposium on Wearable Computers*, 2001.

[9] J. Lee, G. Hirota, and A. State, "Modeling real objects using video see-through augmented reality," *Presence: Teleoper. Virtual Environ.*, vol. 11, no. 2, pp. 144–157, 2002.

[10] P. Bunnun and W. Mayol-Cuevas, "Outlinar: an assisted interactive model building system with reduced computational effort," in *7th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE, September 2008. [Online]. Available: http://www.cs.bris.ac.uk/Publications/Papers/2000883.pdf

[11] S. Kim, S. DiVerdi, J. S. Chang, T. Kang, R. Iltis, and T. Höllerer, "Implicit 3d modeling and tracking for anywhere augmentation," in *Proc. ACM symposium on Virtual reality software and technology*, 2007.

[12] A. van den Hengel, R. Hill, B. Ward, and A. Dick, "In situ image-based modeling," in *ISMAR '09: Proceedings of the 2009 8th IEEE International Symposium on Mixed and Augmented Reality*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 107–110.