Computer Vision in the Post Production Industry

Andrew J. Stoddart
2d3 Ltd, 14 Minns Business Park
Wesy Way, Oxford, OX2 0JB, U.K.
a.stoddart@2d3.com
www.2d3.com

Abstract

The field of special effects in post-production has undergone an exciting period of innovation in recent years. There has been a surge in demand for camera tracking or "matchmoving". A number of companies have built manual matchmoving products, and more recently automatic matchmovers have come to market. We will discuss the nature of the problems that the industry requires solved, and the products that have attempted to deliver solutions. Complex vision algorithms have inevitable failure modes, and this raises several issues including the relationship between automatic and manual solutions and the performance of published algorithms.

1 Introduction

Post production is today a huge global enterprise. Its most prominent public manifestations in recent times have been the Hollywood blockbusters with ever more impressive special effects.

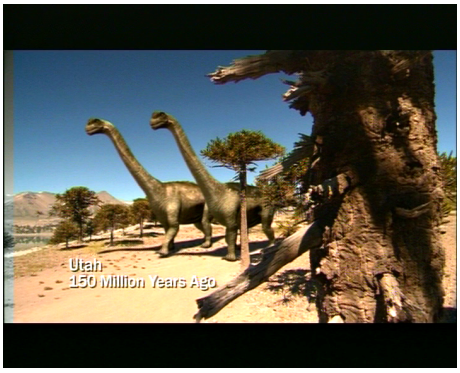
One of the pillars of modern special effects is the field of computer graphics, a forward imaging problem that neatly contrasts the inverse problems of computer vision. The amount of overlap between the fields has accelerated in recent years.

One of the pioneers to open up a research topic that could be published in both the computer vision and computer graphics literature was Paul Debevec. He developed the Facade system for user assisted reconstruction of scene models from still images [1, 2], and has gone on to look at other issues such as image-based modeling and illumination.

There are now many vibrant topics of common research interest and recent SIGGRAPH conferences have had a growing computer vision presence in areas such as reality-based modeling (scanning), animating faces and bodies, texture synthesis and image-based modeling.

Some of the most striking movie landmarks in the late 90's were fully CG movies. In 1995 the first fully computer animated feature film hit the screens. *Toy Story* by Pixar was popular and a huge financial success, it went on to gross $358 million. *Toy Story* used stylised humans and the animation requirements

BMVC 2002 doi:10.5244/C.16.65



(a)

(b)

Figure 1: Examples of movie augmentation

of plastic toys avoided a number of difficulties with human clothing and skin. The year 2001 saw the first attempt at a fully CG animated feature film with photo realistic human characters. *Final Fantasy*, by Square Pictures/Columbia, grossed $30 million in North America, but cost $137 million to make. Square went bankrupt. Nevertheless it was a landmark achievement.

It is important to realise that these films do not represent the mainstream. Human actors are not about to be replaced, despite their huge costs. Shooting live action city and country landscapes is still cheaper than CG.

In practice the majority of films are currently made with a pragmatic blend of live action and animated imagery (but model miniatures, matte paintings and animatronics still have life in them yet, see [10]!). Augmenting empty landscapes with dinosaurs is a well known use of CG, see figure 1 (a). A more every day example is a pop video with CG flowers stuck on a live action studio hedge, see 1 (b).

To augment a moving camera live action shot with CG requires knowledge of the camera path. In post production this task is called camera tracking or matchmoving. Determining this information from the images alone is none other than the structure from motion (sfm) problem that has received so much attention in the computer vision community, dating back to Longuet-Higgins [4] in 1981. Today the subject has matured considerably, for a comprehensive textbook see Hartley and Zisserman [5].

Since most shots in contemporary movies and video have a moving camera there will inevitably be a demand for matchmoving for some time to come. As one moves down the cost spectrum to budget television commercials and prosumer digital video the demand for augmentation still maintains a presence.

2 Camera Tracking

Whenever real images or "live action" footage are to be combined with computer generated (CG) images an alignment operation is necessary.

In the case of a static scene shot from a static camera this need only be done for a single frame. The CG material must be translated, rotated and scaled in 3D to be consistent with the scene in the live action background and then rendered with an appropriate focal length into an image sequence. Finally the two sources of material must be composited. At least one of the two source materials must have an alpha (opacity) channel associated with it. If a CG background is to be placed behind human actors the outline must be established to subpixel precision. This is called rotoscoping and can be done automatically by means of a "blue screen" background or by manually drawing the necessary outline contour in each frame.

Augmentation need not imply a simple CG + live action composite. A process widely used in *Attack of the Clones* [10] was to shoot foreground live action characters in front of a blue screen. The camera move was transferred to a motion control rig to shoot a live action miniature for the midground. Finally a digital matte painting was added for the far background.

A purely rotating camera is called a nodal pan shot. Strictly, a nodal pan will require the camera setup to be specially engineered so that the tripod rotation axis is coincident with camera's optical center, but if the scene is sufficiently far away this step can be ignored.

The live action to CG alignment in this case requires a 3 dof rotation (although roll is very rare in cinema) to be applied to the CG material after a global alignment operation. In the past post production houses have used point tracking software combined with small inhouse software programs to recover the rotation angles, but proper tracking software gives better results in the presence of occlusion.

Complicating factors are zoom and radial distortion. In the vast majority of cases radial distortion is ignored, whereas zoom can be accommodated in the simple point tracking scheme.

If a camera is moving freely the position and orientation must be recovered for each frame. In the earliest period effects shots would typically be filmed using a crane or dolly to obtain a smooth camera move. In this case the CG move may be matched to the live action by manual alignment at key frames followed by interpolation. This was termed "matchmoving". This is extremely difficult for material filmed with a handheld camera, use of which has become more common as cameras become smaller.

This process takes experience and is tedious, so it is natural to move to a computer based technique. The earliest techniques were based on tracking either natural features or markers placed in the scene using 2D image correlation techniques. In general, this has usually proved to be a process that demands modest to quite significant user supervision and/or editing before the 2D track quality is sufficient. Such techniques can also include manual input of the surveyed 3D coordinates corresponding to the 2D tracks.

In their mature form manual camera tracking products have evolved to provide a rich interactive feature set for feature tracking and editing. The solving software is based on the assumption that all the 2D tracks generated by the user are correct and correspond to real points in 3D. Survey knowledge is an optional extra. Arbitrary numbers of 2D tracks, starting and ending at arbitrary points, are solved to produce the camera pose and focal length. It has also to be said that, in the hands of a typical user, this manual procedure will fail on a significant

percentage of shots for a wide variety of reasons.

2.1 Automatic Camera Tracking

Dinotopia, a six hour mini-series from Framestore/Hallmark Entertainment will hit TV screens in 2002. There are 1700 effects shots in the series. Productions like this have driven demand for automatic camera tracking. Here, instead of users choosing features to track, features are detected and tracked automatically. Using techniques such as RANSAC [5] incorrect tracks can be eliminated while computing the structure and camera path. The final stage of this process is usually bundle adjustment to redistribute errors evenly throughout the sequence.

The user workflow is as follows. The image sequence is loaded, a camera is specified (the pixel aspect ratio), and automatic camera tracking is initiated. In the boujou camera tracking product shown in figure 2 the tracking in broken up into two phases. Firstly a set of features is extracted and tracked. The feature tracks are shown in figure 2 (a). Secondly these 2D tracks are used to solve for the camera path and scene points as shown in figure 4. The 3D points are shown in figure 2 (b) projected back into the image.

On completion the quality of the solve is visually evaluated by means of rendering a small test object over the background. Figure 3 shows a test object placed on the wall in 2D and 3D. The 3D points have been rotated to show the planar quality of 3D points recovered on the pillar and foreground box. This shot is an easy one, but does have some non-trivial features. It has "T-junctions" that are not real features, repeated structure, moving people, and transparent substances.

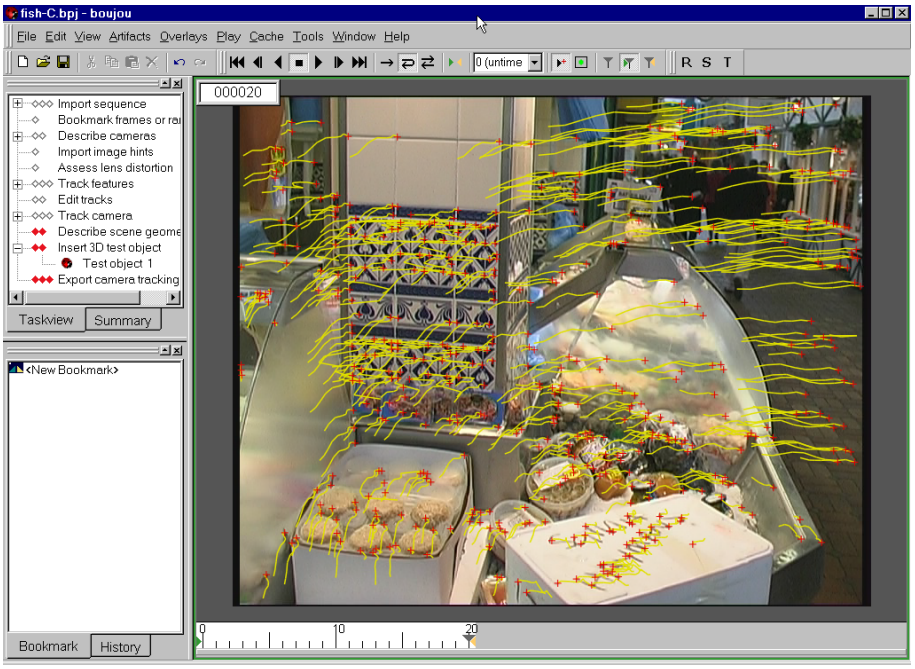
The solve will contain 3D points and these may need to be aligned and scaled to be consistent with 3D models held in the animation software. The resulting camera data is exported to one of the many 3D (or even 2D) animation packages used in the industry so that the CG images may be rendered with the right camera path.

The whole process is now at the stage when many companies are considering performing this step in "batch" mode as the film material comes off the telecine machine that digitizes the images.

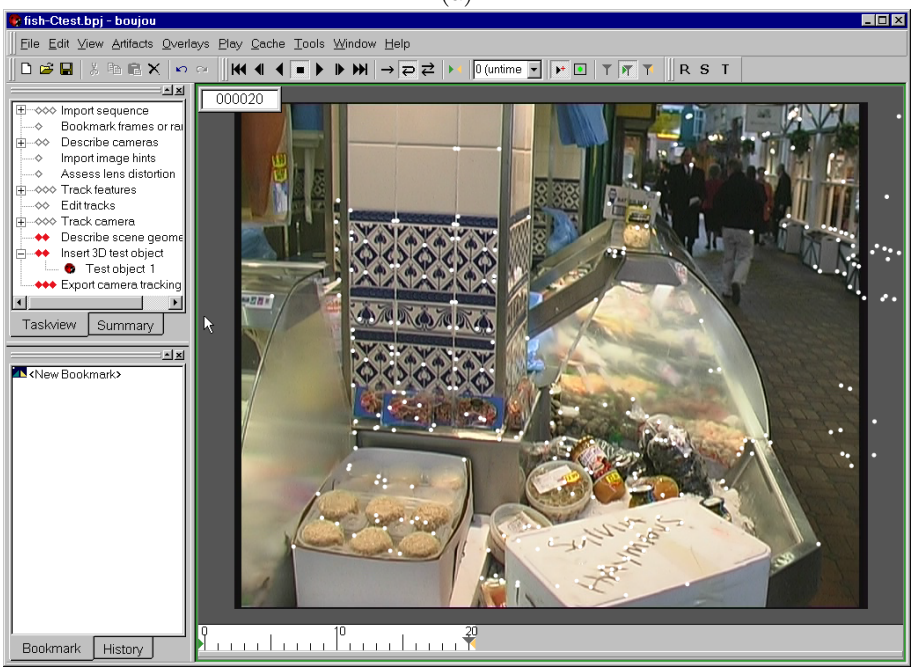
The vast majority of moving camera effects shots are currently filmed with a constant focal length. The camera is assumed to have zero skew, a centered principal point, and a known (pixel) aspect ratio. The focal length could in principle be recorded from the lens and supplied as an input to the matchmoving software. In practice allowing the software to solve for an unknown and constant focal length is preferred. It is often the case that allowing the focal length to vary while shooting results in unstable values being recovered by the camera tracking. For this reason vfx supervisors will usually resist zooming shots, despite pressure from the director. A practical compromise has been to allow the user to specify the focal length curve as an input.

2.2 Product history

The earliest significant work in camera tracking for visual effects was the ras_track package developed by Thaddeus Beier of Hammerhead [3]. ras_track is now a

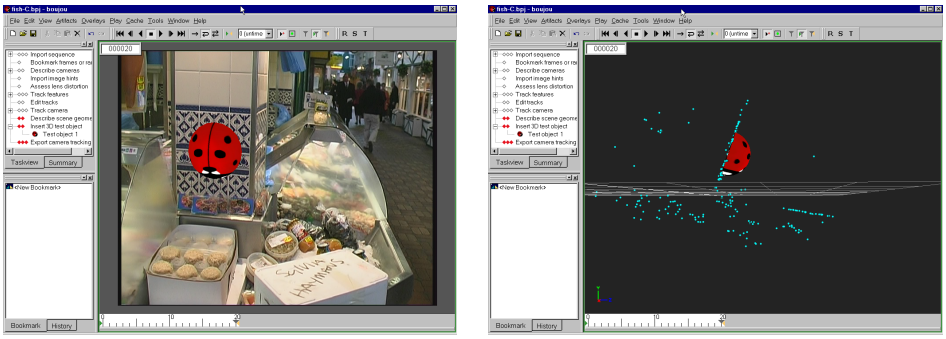


(a)



(b)

Figure 2: Feature tracks and recovered 3D points



(a) (b)

Figure 3: A test object in 2D and 3D

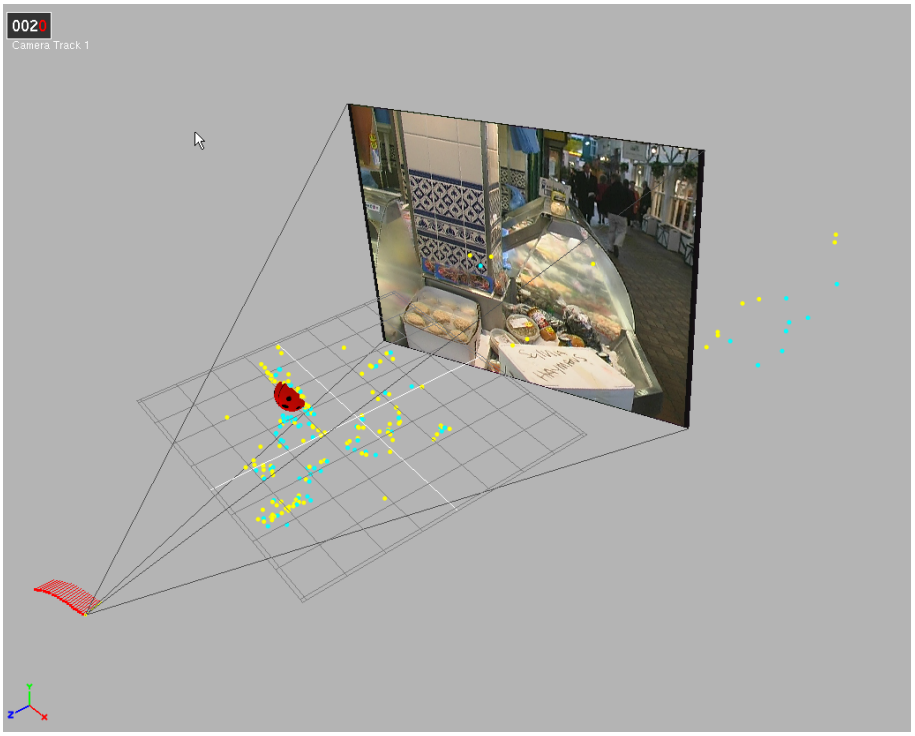


Figure 4: Camera path and 3D structure

public domain package.

In 1995 the first commercial manual tracking product was released, "3D Equalizer" from Sci-D-Visions. This was later followed by "Matchmover" from RealViz. Manual trackers with more limited functionality have been offerred as plug-ins to industry standard 3D packages, e.g. Maya Live for Maya and SceneGenie for 3D Studio Max. The industry has also had its first company failure with Synapix, which offered the SynaMatch camera tracking product.

In 2001 the first fully automatic camera tracker was released, "boujou" from 2d3.

The contribution of camera tracking technology to filmmaking has been acknowledged by the Academy of Motion Pictures Arts and Sciences (who award the Oscars). Since 1930 they have run a program of Scientific and Technical Awards.

In 1998 they made two technical achievement awards to Douglas R. Roble for TRACK and Thaddeus Beier for ras_track.

In 2001 they made a further two awards. The first went to Steve Sullivan and Eric Schafer of Industrial Light and Magic (ILM) for the Motion and Structure Recovery System (MARS). This is in house software at ILM, and not available for purchase. The second award went to Uwe Sassenberg and Rolf Schneider for the development of the "3D Equalizer" camera tracking product.

2.3 Camera tracking: Problem definition

The simplest problem definition is as follows. The director wants to shoot without any contraints and have the technical guys do their stuff afterwards.

In practice the VFX supervisor will have discussed the shot and argued for shooting to take account of a number of technical issues that can be forseen. Even today there are still significant technical and cost limitations on what can be achieved.

During the filming of a big budget special effects blockbuster there may be several pages of notes taken for each shot. These would record lens details, f stop and details of the set geometry.

A typical camera tracking problem is as follows. The camera has known filmback width and height, known image width and height. From this the (pixel) aspect ratio can be computed. The skew may be assumed zero. The principal point is usually assumed centered, but see [9]. The focal length is usually constant, but unknown. There is some mild amount of radial distortion present.

The sequence length is between 2 and 20 seconds, i.e. less than a thousand frames. The image size varies between the standard broadcast sizes of PAL and NTSC to film resolution. 35mm Full Aperture film is typically digitized at 4096x3112. Effects work is usually done at half resolution, and camera tracking can sometimes be done safely at quarter res, i.e. 1024x778.

The quality requirement is easy to state. There must be no visible slip or misalignment to a discerning viewer in a cinema. An rms reprojection error of less than half a pixel is usually a necessary but not sufficient requirement. A camera track based on a small number of tracks may have visible 'kicks' when a track exits the view, this is a track handover problem. Another common problem is

(a) (b) (c)

Figure 5: Radial lens distortion: (a) original material, (b) undistorted (c) original material plus distorted CG flowers and tree

when the augmentation volume is much larger than the volume containing the tracked points. In this case there is no data to lock down the camera parameters. Examples include augmenting a near view scene with a horizon, or an empty scene with characters in close up.

2.4 Camera tracking in practice

Attendees at computer vision conferences will be familiar with the extreme care taken in selecting appropriate images or image sequences for a paper presenting a computer vision algorithm. Indeed, computer vision algorithms are notorious for working on as little as one dataset!

The standard data set chosen for automatic camera calibration in computer vision papers is a richly textured empty rigid outdoor scene under diffuse daylight conditions shot with a camcorder in progressive scan mode with an exposure time of 1/125th second.

Almost all of these conditions will be different in professional footage. Richly textured outdoor scenes give way to burnt out desert or studios with a blue screen backdrop. Empty scenes are filled with people moving about. Rigid objects are replaced with water, moving plants, fire and smoke. The lighting might be good studio lighting, but may also contain flashes or be very dark. Film footage has quite different local noise characteristics. Interlaced footage is sometimes used in video work. Exposure times can increase to the level where motion blur is a serious problem. It is often unacceptable to film e.g. wheels in such a way that the image is not blurred.

Possibly the only way in which professional footage is ever easier than computer vision footage is in the camera move. Amateur hand held footage invariably has more high frequency rotational noise than professional footage!

2.5 Radial lens distortion

Much of the time radial lens distortion is ignored in the post production industry. The vfx supervisor may ask in advance of filming for a grid to be shot, but by the time the shot is in post production it may not be possible to trace which camera was actually used for the shot!

Inevitably, cases arise where radial lens distortion cannot be ignored. Set extension where the set fills the field of view requires consistency over the entire image. Wide angle shots can have extreme lens distortion, for an example see the Microsoft X-box *The Prophecy* advertisement [8]. The camera, in extreme wide angle close up, moves slowly up a sick patient's arm that was then augmented with CG mosquitos.

Properly dealing with radial distortion causes considerable inconvenience in the post production workflow. This is mainly down to the fact that 3D packages don't support this parameter with anything like the ease that they support parameters such as focal length or pixel aspect ratio. (Pure CG never needed radial distortion!)

The user firstly needs to establish the radial lens distortion either by eyeballing, manually matching an image of a grid/ straight line in the scene, or by bundle adjustment (see also [7]).

After tracking they need to create an undistorted copy of the original sequence so that they can set up and verify the scene to be rendered in the 3D software. This is an inconvenient step, it is best to avoid maintaining copies of image sequences due to disk space considerations. The scene may then be previewed using a low resolution render in this software.

The final high quality render must then be made and the distortion applied to rendered images. Typically this is done in software provided by the camera tracking software vendor as CG renderers offer limited support for radial and conventions for expressing the radial distortion differ widely. The CG may then be composited with the original footage. One could imagine using a final result based on undistorted footage, however using the original distorted footage is preferred so as to avoid cropping and resampling the source material.

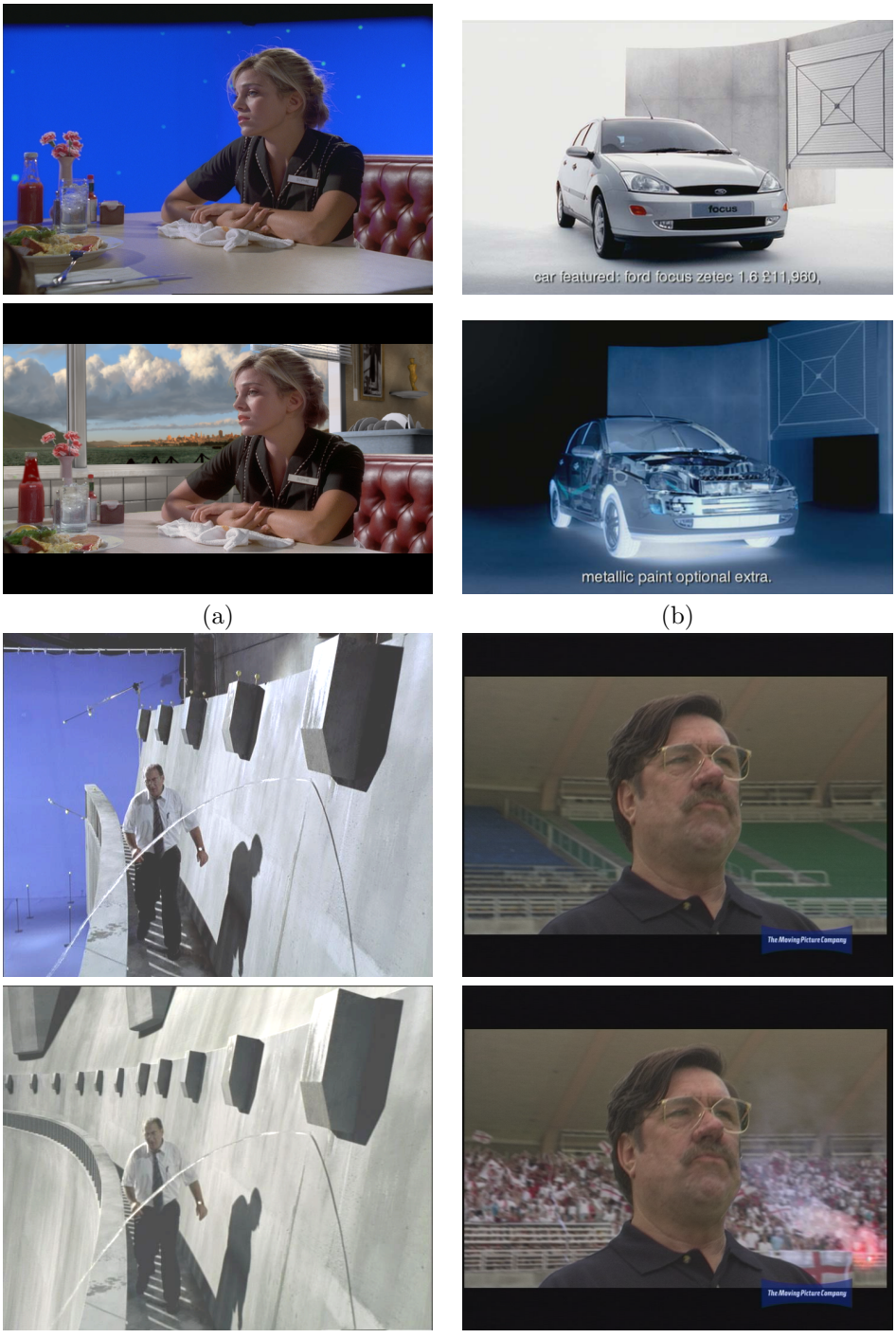
An example is shown in figure 5. The material is shot with a camcorder and a wide angle adaptor. The undistorted source material is shown in (b) and the final augmented scene is shown in (c).

Finally we note that CG packages generally do not support non-centered principal points. This means that a non-trivial principal point has all the same difficulties associated with radial lens distortion.

3 Examples

In this section we review further examples. Figure (6) shows some before and after stills from shots which were camera tracked and augmented in some way. (a) shows a shot filmed in a studio in front of a blue screen and augmented with a background. The blue screen has markers placed on it to facilitate tracking. The ketchup bottle picked up some blue light and was replaced in the final shot.

(b) shows a shot in which the live action object and background were entirely replaced halfway through the shot to give the impression of X-ray vision. This required an accurate 3D model of the car structure to be built. The tracking shot was difficult due to smooth rounded glossy surfaces with high reflectivity as well as significant cut glass lenses. This shot was taken using a motion control rig. In anticipation of tracking difficulties, and to avoid having to remove manual markers, an alternate shot with the same programmed move was taken with marked



(a) (b)

(c) (d)

Figure 6: Before and after shots

materials in shot. However the original was successfully tracked automatically.

(c) shows a classic set extension shot. The near field dam wall and actor are real and shot in a studio with a blue screen, the far field dam wall is CG and seamlessly blended into the real wall.

(d) shows a crowd background added to a stadium scene. This is an interesting shot insofar as it could be attempted as a 2D or 3D augmentation. The camera is hand held and moves around the foreground character, however the background is sufficiently far that the background may be replaced by modeling the shot as a nodal pan.

4 Conclusion

Camera tracking is an example of a solution to a difficult problem in computer vision that has been commercialized and is in daily use around the world. It is a near perfect match between a real world need and a long standing research problem. Another mature commercial technology is human motion capture by use of optical markers [**?**]. This is now beginning to attract more attention in the computer vision community [**?**].

Many other vision applications are in various stages of commercialization, although the challenge of being more cost effective than hard working animators with manual tools should not be underestimated. Especially when you need to allow for the fact the vision algorithms never work 100% of the time, and manual escape routes are often needed.

An obvious challenge is rotoscoping which maps onto the general segmentation problem in computer vision. There is no general rotoscoping product on the horizon at present although sub problems have been addressed, see Mokey for nodal pans [12] and recent research by Szeliski *et al* [11].

5 Acknowledgements

We are grateful for material supplied by: Condor Post Production, Zomba records, *Here and Now*; Meteor Studios, courtesy of Discovery Channel, *When dinosaurs roamed America*; Skippingstone Productions, *Doppelganger*; Marmalade Animation and Effects, *Ford - See Through*; Rushes/Flying Fish, *Infineon - Ultimate Convergence*; Moving Picture Company, Hallmark Entertainment, Gruber Films, *Mike Bassett - England Manager* .

References

[1] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik, Modeling and Rendering Architecture from Photographs, In SIGGRAPH '96, August 1996.

[2] C. Taylor, P. Debevec, and J. Malik, Reconstructing polyhedral models of architectural scenes from photographs, European Conference on Computer Vision, Cambridge, (1996).

[3] www.vicon.com.

[4] Low Density Feature Point Matching for Articulated Pose Identification, Horst Holstein and Baihua Li, British Machine Vision Conference, 2002.

[5] www.hammerhead.com/ras_track/ras_track.html.

[6] H. C. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, Nature, pages 133-135, volume 293 (1981).

[7] Multiple View Geometry in Computer Vision, R. Hartley and A Zisserman, Cambridge University Press, (2000).

[8] C.G. Harris, Determination of ego-motion from matched points, In Third Alvey Vision Conference, pages 189-192, (1987).

[9] A. W. Fitzgibbon, Simultaneous Linear Estimation of Multiple View Geometry and Lens Distortion, IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, U.S.A, pages 125-132, (2001).

[10] A. Daffy, *The Prophecy*, Computer Generated Imagery Magazine, pages 30-34, (2002).

[11] R. I. Hartley, R. Kaucic, Sensitivity of calibration to principal point position European Conference on Computer Vision, Copenhagen, vol II, pages 433-446, (2002).

[12] J. Duncan, "Love and War", a feature on Star Wars Episode II: Attack of the Clones, Cinefex 90, pages 60-70, (2002).

[13] R. Szeliski, S. Avidan, and P. Anandan, Layer extraction from muliple images containing reflections and transparency, IEEE Conference on Computer Vision and Pattern Recognition, U.S.A., pages 246-253, (2000).

[14] Mokey, a mosaicing and matte extraction product, www.is-distribution.com.