Research White Paper

WHP 154

September 2007

Dual-Mode Deformable Models for Free-Viewpoint Video of Outdoor Sports Events

J. Kilner, J. Starck, A. Hilton, O. Grau

BRITISH BROADCASTING CORPORATION

White Paper WHP 154

Dual-Mode Deformable Models for Free-Viewpoint Video of Outdoor Sports Events

Abstract

Generating free-viewpoint video in outdoor sports environments is currently an unsolved problem due to difficulties in obtaining accurate background segmentation and camera calibration. This paper introduces a technique for the reconstruction of a scene in the presence of these errors. We tackle the issues of reconstruction completeness, and accuracy of surface shape and appearance. We introduce the concept of the conservative visual hull as a technique to improve reconstruction completeness. We then present a view-dependent surface optimisation technique using deformable models to improve surface shape and appearance. We contribute a novel dual-mode snake algorithm that is robust to noise and demonstrates reduced dependence on parameterisation by separating the search of the solution space from the data fitting. We conclude by presenting results of this technique along with a quantitative evaluation against other reconstruction techniques using a leave-one out data set.

This document was originally published in Proc. of The 6th International Conference on 3-D Digital Imaging and Modeling (3DIM'07), August 21-23, 2007, Montréal, Québec, Canada.

Additional key words: 3D reconstruction, sport visualization, image-based rendering

White Papers are distributed freely on request.

Authorisation of the Head of Research is required for publication.

Dual-Mode Deformable Models for Free-Viewpoint Video of Outdoor Sports Events

J Kilner, J Starck, A Hilton
University of Surrey, U.K.
{J.Kilner, J.Starck, A.Hilton}@surrey.ac.uk

O Grau
BBC Research, U.K.
O.Grau@bbc.co.uk

Abstract

Generating free-viewpoint video in outdoor sports environments is currently an unsolved problem due to difficulties in obtaining accurate background segmentation and camera calibration. This paper introduces a technique for the reconstruction of a scene in the presence of these errors. We tackle the issues of reconstruction completeness, and accuracy of surface shape and appearance. We introduce the concept of the conservative visual hull as a technique to improve reconstruction completeness. We then present a view-dependent surface optimisation technique using deformable models to improve surface shape and appearance. We contribute a novel dual-mode snake algorithm that is robust to noise and demonstrates reduced dependence on parameterisation by separating the search of the solution space from the data fitting. We conclude by presenting results of this technique along with a quantitative evaluation against other reconstruction techniques using a leave-one-out data set.

1. Introduction

When traditional fixed-viewpoint video of an event is rendered, the only viewpoint available for playback is that of the camera that recorded the event. Free-viewpoint video (FVV) attempts to break this restriction by allowing the specification of the viewpoint at the point of rendering rather than the point of recording. This ability is of interest to television companies producing sports coverage as it allows them to generate virtual replays of key events showing them from angles that give greater insight in to the match. This paper addresses some of the issues encountered while attempting to apply FVV techniques in this way to video recorded at outdoor sporting events, specifically the problems of generating output of a sufficiently high quality for use in a broadcast.

Most current FVV techniques are designed around the multi-camera studio environment with controlled lighting and well-calibrated static cameras. These techniques do not perform at an acceptable quality in the context of outdoor sports coverage with unconstrained illumination and poor calibration[10]. As a result there is a requirement for novel reconstruction techniques that can produce output of an acceptable quality in these environments.

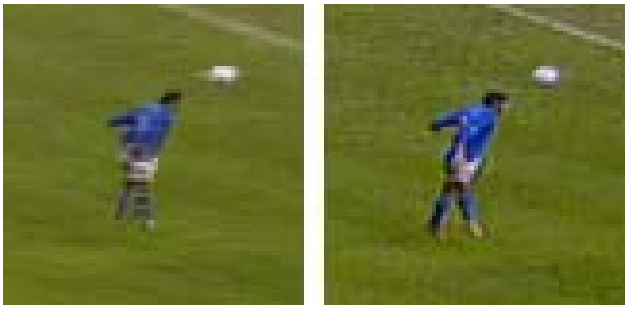

Figure 1. Left: detail from a free-viewpoint video generated from the leave-one-out data set, Right: ground truth image from the unused camera

One major issue that must be overcome is the lack of accurate camera calibration. Due to environmental constraints (moving cameras, lack of access to the pitch etc.), cameras may not be calibrated by traditional methods. As such the only viable techniques for calibration are techniques that extract calibration data from natural features in the scene such as those developed by Thomas et al.[18]. These techniques can result in calibration error of the order of 2 pixels, rather than the sub pixel accuracy commonly achieved in the studio environment. The presence of this calibration error means that a global reconstruction which satisfies all input cameras does not necessarily exist.

A second important factor is the complexity and low resolution of the objects in the scene. This, combined with the unconstrained lighting and natural backgrounds, makes image segmentation a difficult process and introduces a large amount of additional error into the system. This problem is exacerbated by the requirement in free-viewpoint video for all techniques, such as segmentation, to be automatic, as manual initialisation across multiple cameras for each

frame of a segment of video is unfeasible.

In this paper we propose that techniques that combine shape-from-silhouette with refinement using snakes are particularly suitable for this type of reconstruction. We propose a novel technique that uses a deformable model to combine multi-view segmentation with shape and stereo optimisation across the 3D reconstruction of a scene. Our technique uses only video from a set of standard calibrated cameras, can handle multiple self-occluding objects and is error tolerant with regards to initial scene segmentation and poor camera calibration. The novel elements of the technique are the formulation of the silhouette term, which avoids the need for an initial high quality image segmentation, and the two-phase use of the deformable model which reduces the dependence on parameterisation.

In Section 2 we provide relevant background to this paper and in Section 3 present a description of our technique using a dual-mode deformable model. Some results of reconstructions using this technique (including a quantitative comparison with other techniques) are presented in Section 4 and discussed in Section 5. Section 6 provides some conclusions and ongoing work.

2. Background

2.1. Free-viewpoint video and sports

Free-viewpoint video is the technique of combining multiple video sources to generate a novel video from a virtual viewpoint. Techniques have so far focused on the studio environment, beginning with the *Virtualized Reality* system developed by Kanade et al.[8] which used 51 cameras distributed over a 5m dome. Since then techniques including shape-from-silhouette [1] and shape from photo-consistency [19] have been used to generate 3d scene reconstructions from reduced numbers of cameras.

The problem of generating free-viewpoint video in external environments, such as football stadia, have received much less attention than the study of reconstruction in studio environments. Kanade et al. have developed the eye-vision system that was demonstrated at the Superbowl[4]. Koyama et al. have produced a real-time system using billboards[11] and Inamoto et al. have demonstrated a system using image morphing [7]. However these systems are limited either in the quality or the freedom of the virtual viewpoint, and some also require specialist equipment.

2.2. Reconstruction techniques

Most current work on free-viewpoint video for sports uses variations on the billboarding technique [5]. In billboarding a single polygon is placed co-incident with the object that it represents. This polygon is then rotated around an axis or point (typically the Y axis) so that it retains its original position, but is constantly facing the virtual camera. An image of the original object is then applied to the polygon as a texture map. This technique can often give good results with very little overhead, however the lack of correct shape with which to allign the images from multiple cameras quickly becomes apparent once the virtual camera moves any great distance from the original camera location, as can been seen in Figure 6. Techniques such as image morphing [2] have been proposed to improve the quality of view interpolation, but the problem of view extrapolation has not been solved for this representation.

2.3. Deformable models

Snakes were introduced by Kass et al.[9] as an algorithm for extracting a contour from an image. They have two key properties: 1) a physical simulation combines an internal regularisation force with an external data force to fit a smooth contour to the data, 2) the shape of the snake determines the region of the image that influences the snake. An iterative approach then allows the image to update the shape of the snake and then uses the shape of that snake to determine the region of the image that is examined. In 3D, snakes can be implemented with an implicit geometric representation such as level sets[14], or with an explicit geometric representation such as elastic deformable models[17]. Deformable models are particularly attractive when the surface to be reconstructed is small compared to the reconstruction volume and a high resolution result is desired, as the computation cost is proportional only to the number of surface elements, whereas in level sets it is proportional to the number of volume elements. Another discriminant between the two types of technique is that level sets can change their topology whereas deformable models cannot.

Snakes using deformable models are an attempt to use physical modelling to combine multiple data cues in a smooth way. The deformable model is initialised with a shape known to be close to the final solution. This surface is then modeled as an elastic object acted on by many springs, fields or other physical constraints. The simulation of the object's elasticity provides the internal regularisation force and the physical constraints are used to generate the external forces. A physical simulation is then used to determine the movement of the model and, at some termination point, the final shape of the model is recorded. Deformable models have recently been used as a successful framework for combining stereo and silhouette constraints in order to refine 3D geometry [6, 15] and these techniques are amongst the highest quality techniques currently in use[13].

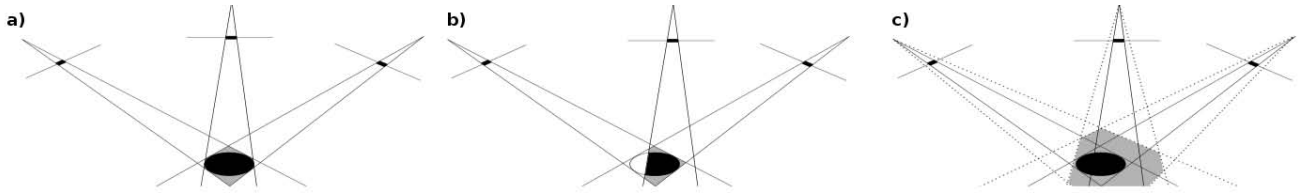

Figure 2. Reconstruction of an object (solid ellipse) using shape-from-silhouette techniques: a) with no calibration error, b) with calibration error and c) conservative shape-from-silhouette with calibration error

3. Methodology

3.1. Overview

In this work we improve the quality of reconstruction by: 1) improving the completeness of the reconstruction, and 2) improving the accuracy of the shape and appearance of the final reconstructed surface. We address the issue of completeness with a modified shape-from-silhouette technique, and improve the shape and appearance with a dual-mode snake to optimise the final reconstruction.

As was previously mentioned, the presence of errors in calibration and matting mean that there is no globally optimal solution to the reconstruction of the scene. However, by optimising solely for one view we can guarantee that there exists a solution that is correct for that view. In this way we can improve the accuracy of shape and appearance without sacrificing completeness. By optimising for multiple shapes from the same global initialisation we can generate a set of view-dependent meshes with constant topology, which aids blending between them for intermediate views.

Our technique uses a deformable model that allows the integration of various cues to reconstruct the surface shape in one consistent framework. Level sets were not used due to their computational complexity and our requirement for consistent topology between the per-view reconstructions.

Snakes suffer from two problems. Firstly there is the question of locality - how far from the snake's current position do you look for data to drive the snake? Too small a band, and the model will be too constrained by the initialisation, possibly unable to consider a deformation large enough to bring the surface into agreement with the data, or trapped in a local minima that happens to be closer to the initialisation than the true surface is. Too large a band, and the contextual information from the initialisation is lost and the model may evolve to undesirable solutions.

Secondly there is the question of parameterisation - how do you balance data forces against regularisation and how do you balance various data forces against each other? Although both data and regularisation costs are united in the deformable model framework, there is no pre-defined relationship between these costs. Therefore solving for the data forces or the internal forces in isolation is a simple mathematical exercise, combining these terms requires arbitrary scaling parameters to be applied.

These two problems are of particular relevance in our application. As the initialisation error is of similar scale to the features in our data, a band large enough to take a poorly initialised surface from its initial position to the correct position is also large enough to take a well-initialised surface and move it to an undesirable position. The scaling of costs is even more problematic. Due to noise in the image and the variable lighting conditions, a set of values which balance to find the correct surface in one section of the image will not balance in another section, meaning it is not possible to use one set of parameters for the entire image.

To avoid these problems our approach is to give each set of forces equal space within the framework, but not to balance them directly against each other. To this end we have developed our dual-mode deformable model. This algorithm uses the same deformable model in each of two modes, differing only in the parameterisation of the model. The first mode is a "search" mode which addresses the issue of banding by performing a search through configuration space, seeking out a consistent set of data points for the reconstruction. We do not have to worry about fitting to weak data points or over-smoothing as we are simply searching for data in this mode. The second mode is a "fitting" mode where the deformable model is used a more conventional manner albeit with a much weaker regularisation force as we have already discarded the major outliers in the data.

3.2. Initialisation

The visual hull of an object, determined by shape-from-silhouette, is often used as the initialisation for deformable model techniques[6, 15]. As shown in figure 2b), errors in calibration will yield a result from shape-from-silhouette techniques that is truncated (shown in light grey), especially if the objects being reconstructed are small compared to the scene size - this can also be seen in the truncation of the player's legs in the shape-from-silhouette results shown in Figure 5. Figure 2c) shows how a conservative shape-from-silhouette technique, where we simply expand the silhou-

ettes by n pixels, will yield a more complete shape. We refer to the shape generated by this technique as the Conservative Visual Hull (CVH - shown in light grey). The benefit of using a CVH in this way is that it provides a more complete reconstruction. The disadvantage of a CVH is that it is only weakly related to the true surface: if a large enough value of n is used the CVH will contain the true surface, but no other guarantees are given.

As a technique for accurate surface reconstruction, calculation of the CVH is not suitable, but properly configured it provides sufficient guarantees of completeness to be used to initialise a refinement technique. Due to the weaker constraints for agreement between the original input images, the conservative shape-from silhouette technique is more susceptible to phantom volumes and to errors from pixel noise in the silhouettes. However with some simple domain knowledge (no surfaces exist below pitch level, no very small objects etc.) clean-up is a trivial task.

3.3. Search

3.4. Deformable model

Taking a "key" view with index γ, the deformable model is optimised to reduce the energy for that view E_γ over the surface S of the model. S is represented by a simple triangle mesh composed of vertices v and edges e.

E_γ is composed of a data driven energy term D_γ and an internal elastic energy term I. These are combined using a weighting term β such that:

$$E_\gamma(S) = (1-\beta)D_\gamma(S) + \beta I(S). \quad (1)$$

$I(S)$ is the elastic energy of the mesh:

$$I(S) = \sum_{e \in S} (L_e)^2 k \quad (2)$$

where L_e is the length of edge e and k is a stiffness constant.

$D_\gamma(S)$ is a term expressing the data fitness of the surface. It is expressed as a per-vertex energy in terms of a vertex v's most desireable local position v':

$$D_\gamma(S) = \sum_{v \in S} \parallel v - v' \parallel^2 \quad (3)$$

v' is calculated by maximising a per vertex data score. This score is a silhouette fitness score for edge generating vertices, and a stereo score for other vertices. First we determine if a vertex is an edge generator when viewed from a camera with index σ by calculating the value $\mu_\sigma(v)$ which is a measure of how perpendicular the unit vertex normal n is to a unit vector along the camera viewing direction a_σ:

$$\mu_\sigma(v) = (1 - |n.a_\sigma|)^2 \quad (4)$$

If we then consider a set of cameras Ω which contains σ, we can classify each vertex as being in κ_σ, the group of edge generators for the view σ:

$$v \in \kappa_\sigma \; if \; \sigma = \arg\max_{x \in \Omega} \mu_x(v) \; and \; \mu_\sigma(v) \geq \lambda \quad (5)$$

By taking $\Omega = \{\gamma - 1, \gamma, \gamma + 1\}$ we only consider edge generators from a set of cameras near to the key view. λ controls the thickness of the strip of edge generators that is considered, and a value of $\lambda = 0.8$ was used. v' can now be given as the location which maximises the data fitness term ϵ which is expressed in terms of a silhouette fitness score G and a stereo score C:

$$v' = v + \delta n \quad (6)$$

$$\epsilon(v') = \begin{cases} G(v') & v \in \kappa_\sigma \\ C(v') & v \notin \kappa_\sigma \end{cases} \quad \sigma \in \Omega \quad (7)$$

$$\delta = \arg\max_\delta \epsilon(v + \delta n), \quad r \geq \delta \geq -r \quad (8)$$

δ is determined by sampling along n within some range r to maximise $\epsilon(v')$. v' is therefore the projection of the strongest local data cue on to the line $v + \delta n$. If no local data cue exists $\delta = 0$ and $v = v'$.

G is a term representing the silhouette matching score for the projection of v' into image σ and C is a term representing the stereo matching score for regions around the projection of v' in to the most appropriate cameras in Ω. Similar terms are used in other deformable model based work such as that by Hernandez et al.[6]. However, although we can use a standard stereo cross-correlation for C, we do not have accurate silhouettes to use for the formulation of G.

This problem is addressed by formulating the silhouette energy in terms of image gradients. It can be noted that silhouette shape can be determined from just the gradient of a matte ($\nabla\alpha$) as the silhouette boundary occurs where $\nabla\alpha$ is maximised. This allows us to formulate the silhouette energy in terms of the image gradient ($\nabla\mathcal{I}$) using the following approximation[16]:

$$\mathcal{I} = \alpha\mathcal{F} + (1-\alpha)\mathcal{B} \quad (9)$$

$$\nabla\mathcal{I} = (\mathcal{F} - \mathcal{B})\nabla\alpha + \alpha\nabla\mathcal{F} + (1-\alpha)\nabla\mathcal{B} \quad (10)$$

$$\nabla\alpha \propto \frac{1}{\mathcal{F} - \mathcal{B}} \nabla\mathcal{I}, \nabla\mathcal{F} \approx C_1, \nabla\mathcal{B} \approx C_2 \quad (11)$$

Equation 9 is the classical matting equation in terms of an image $\mathcal{I}$, its foreground $\mathcal{F}$ and background $\mathcal{B}$ and Equation 10 is its first derivative. Equation 11 then states that where the foreground and background are smooth the rate of change of alpha is proportional to the image gradient. As high $\nabla\alpha$ occurs at silhouette boundaries it can be seen that in regions where $\nabla\mathcal{F}$ and $\nabla\mathcal{B}$ are constant, silhouette boundaries coincide with regions of high $\nabla\mathcal{I}$ and hence silhouette shape can be determined without explicitly calculating α.

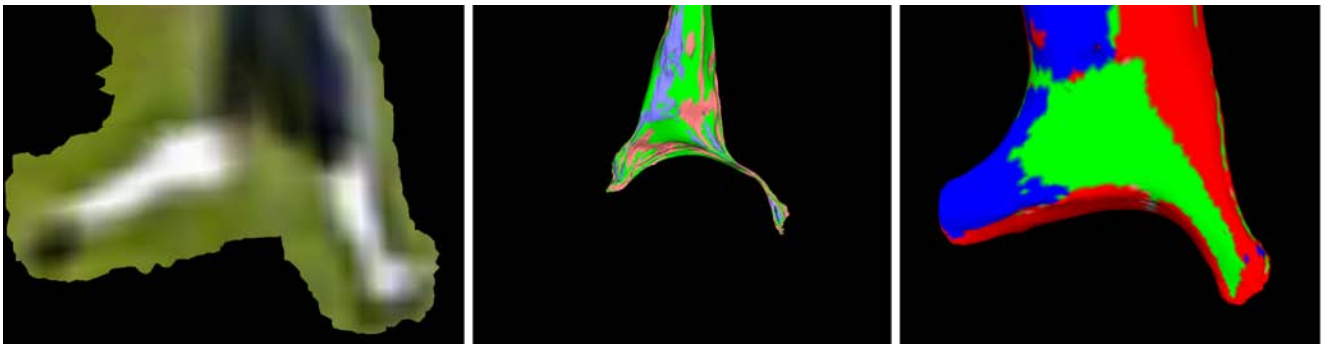

Figure 3. Left: Original image, Centre: Standard deformable model, Right: dual mode deformable model

Both G and C are thresholded to avoid noise. Due to the nature of the snake algorithm used, these thresholds are not particularly sensitive and can be set by inspection from the average edge intensity and standard deviation in the source images.

An optimisation over the mesh to minimise $E_\gamma(S)$ is performed using conjugate gradient descent. The step length in conjugate gradient descent is defined by performing a line search using the back-tracking algorithm.

We initially use the deformable model to act as a banded search through the solution space. Due to the conservative nature of the initialisation, the desired solution is taken to lie within or near to the initialisation surface. The surface is therefore allowed to collapse under the internal elastic energy of the mesh. As it collapses it searches for regions that agree with the input data.

As initialisation is poor, a value for β is chosen (typically 0.1) so that regularisation effects dominate during this phase. This allows the model to evolve in a smooth fashion, collapsing from its initial state to a smaller shape. As $D_\gamma(S)$ is not zero (even though it is small) the strongest data terms affect this evolution so that the collapsed mesh represents the strongest features of the object. If the mesh was simply treated as a normal snake with strong regularisation this would pull the surface away from the relevant image regions as shown in Figure 3. In order to avoid this we do not re-evaluate v' on every iteration.

If v' was never re-evaluated the model would fix onto any edges that happened to fall near the initialisation, and this is equally undesirable. Hence a technique of selective updates is used to determine when to update v' for a given vertex. If N_v^a is the a-neighbourhood of v then the local support $L(v)$ is measured as:

$$L(v) = \frac{|\mathcal{K}_v^a|}{|N_v^a|} \tag{12}$$

where $\mathcal{K}_v^a$ is the the subset of N_v^a only containing those vertices with a valid current data cue. For these experiments $a = 3$ was used. If $L_v \geq$ some threshold then v' is not updated, but if not then a new value of v' is calculated. If the vertex has not moved far from its previous position then recalculation will yield the same value of v'. If however v is pulled away from its previous position then the previous value of v' will fall out of the considered range and a new value will be calculated.

In this way the local support of the data fitness of the model is used to determine the update. If a section of the model is driven by some data cues but cannot be integrated continuously with the rest of the model then the model will update itself to discard the anomalous data cues and to seek new cues that are more consistent with the rest of the model. This does not compromise the model's ability to jump gaps in the data, as it is only where the regularisation is attempting to move vertices far from their previous positions that cues are discarded - if the cues form a smooth whole, then the model will not discard them. Thus only data cues that are inconsistent with a smooth shape incorporating the majority of data cues are discarded. The process is terminated once variation between iterations falls below a certain level.

3.5. Fitting

Having selected the most appropriate values of v' to use for the model we now allow the data term to dominate by relaxing the regularisation ($\beta = 0.01$). The deformable model is re-initialised with the original vertex positions, but maintains the values of v' discovered through the first phase. Returning the vertices to their original positions allows the model to better fit to detail that may have been passed over during the search phase.

Another known problem of snake-like techniques is that they will not move into concavities that are larger than the search band. In order to determine the correct final shape an "exploratory" or "ballooning" force must be used. Our exploratory force operates on those vertices that are part of κ_Ω but for which $G_\sigma(v')$ is less than the specified threshold. This force is modelled by simply moving the relevant vertices inwards along the normal direction. This allows the

Table 1. Analysis of foreground reconstruction on a variety of techniques. BB = billboards, SFS = Shape-from-silhouette, SFS w BM = Shape-from-silhouette with Bayesian mattes, CSFS = conservative shape-from-silhouette and DMS = dual mode snakes (our technique). Scores are shape, completeness, appearance and combined appearance with completeness

Tech.	Shape	Compl.	Appear.	Comb.
BB	0.71	0.86	0.78	0.67
SFS	0.77	0.83	0.94	0.78
SFS w BM	0.75	0.80	0.94	0.75
CSFS	0.27	0.98	0.84	0.82
DMS	0.56	0.95	0.91	0.85

deformable model a chance to find edges further inside the shape if they exist, but allows the deformable model's internal energy to pull the vertices back out to the surface if no such edge exists.

4. Results

The deformable model technique was applied to a data set featuring video from 6 cameras arranged around one quarter of a football pitch. Camera calibration was obtained using natural image features[18] and an initial segmentation of the images was performed using a chroma-based technique. Both calibration and initial segmentation were fully automatic and representative of the kind of input data expected in a real world scenario and contained significant errors. Footage from a seventh camera was used as the ground truth against which comparisons were made. Figure 4 shows the relationship between the two nearest cameras and the virtual viewpoint, showing the wide baseline (approaching 45 degrees) used in this reconstruction.

The technique was compared against a number of alternative techniques: billboards[7], standard shape-from-silhouette[12], shape-from-silhouette using refined mattes, and the conservative shape-from-silhouette technique described earlier in this paper. The refined mattes used in the third technique were generated by applying a dilation operation to the original segmentation to generate a tri-map, and then using Bayesian matting techniques[16] to refine the matting. It should be noted that the ground truth segmentation used in this comaprison was generated by hand, and due to the nature of the video used (compression artefacts, YUV encoding effects) this produced imperfect results. In addition the camera calibration for the ground truth is itself inaccurate as it was obtained via the same calibration techniques used for the other cameras, hence an amount of reprojection error has been accounted for in the comparison.

View-dependent texturing was introduced by Debevec et al. [3] and refers to the technique of choosing between multiple texture maps to apply to a surface based on the orientation of the virtual camera relative to the surfaces of the mesh. All of the mesh-based techniques used were rendered using view-dependent texturing. The two closest textures are chosen and blended together based on the angular distance between the viewing rays of the virtual camera and those of the original cameras that generated the images being used as texture maps.

Images from a set of frames from the input sequence were rendered and Table 1 shows the results of comparison against the ground truth images. This was carried out using an implementation of the technique described by Kilner et al.[10] which provides a framework for measuring the shape (pixels correctly classified as foreground) and appearance (foreground pixels with correct values) for the reconstruction of foreground elements in a scene, as well as techniques for accounting for re-projection error in the ground truth. For this comparison we used an estimated reprojection error of 1 pixel. We introduce an additional "completeness" measure to this analysis:

$$c(p) = \max(f(p)f(p'), (1 - f(p'))) \quad (13)$$

$$C(I) = \frac{\sum_{i=1}^{n} c(p_i)}{\sum_{i=1}^{n} \max(f(p_i), f(p_i'))} \quad (14)$$

where p is a pixel in image I, p' is the corresponding pixel in ground truth image I' and $f(p) = 1$ if p is a foreground pixel, $f(p) = 0$ if p is a background pixel. The completeness score is similar to the shape score except that no penalty occurs if a pixel appears in the synthetic image but not in the ground truth. Thus only missing pixels are penalised. This score is introduced as we are primarily interested in the reconstruction of players, hence they are all that is included in the ground truth. We do not want to penalise the reconstruction of pitch markers and static items such as goal posts, as these static background elements could be removed automatically.

Figure 5 shows the results of the various reconstruction techniques including detail of the reconstruction of an individual player.

5. Discussion

It is clear from the results that the segmentation and calibration errors are great enough to render simple shape-from-silhouette techniques un-usable. Even with the improved mattes provided using the Bayesian refinement, the reconstruction is still poor. This shows that it is not just poor

Figure 4. Left and right: Original images used in reconstruction, Centre: reconstructed view

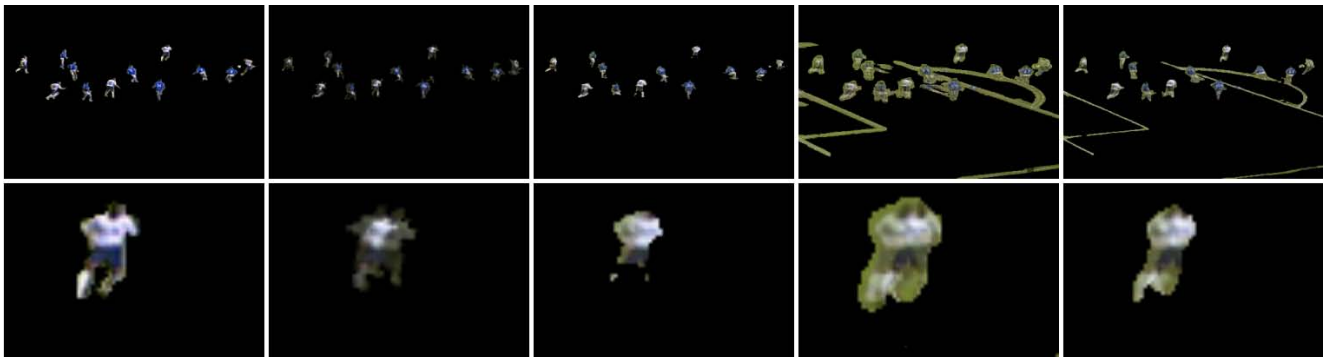

Figure 5. Reconstructions of the view from a camera between cameras 3 and 4, Top: crop from full render, Bottom: detail. Left to right: ground truth, billboards, shape-from-silhouette, conservative shape-from-silhouette, dual-mode snakes.

segmentation that causes the truncated reconstructions, but that errors in camera calibration are dominating the result.

The results from the conservative shape-from-silhouette technique show poor correspondence between view dependent textures. The difference in volume between the standard visual hull (which can be considered the intersection of the reconstruction for each individual view) and the conservative visual hull (which can be considered the union of the reconstructions for each individual view) shows that the ambiguity in the system due to calibration errors is large.

The results therefore show that restriction to a smaller set of inputs reduces ambiguity in the reconstruction from camera calibration error. The deformable model technique provides a solution that is optimal for a certain range of virtual viewing positions.

The use of the CVH for initialisation gives the technique much greater completeness than the normal shape-from-silhouette-based techniques. The initialisation from the CVH also means that information from all input cameras is still incorporated in the final shape.

Finally it can be seen that the dual-mode snakes technique improves the appearance of the reconstruction compared to directly rendering the CVH. The original shape-from-silhouette reconstruction also has good appearance scores, but as only the core of each player is rendered this can partly be accounted for by the fact that the areas most prone to incorrect appearance (the player boundaries) are not rendered at all and hence not considered for comparison - the high completness score for our technique demonstrates that the improvement in appearance is genuine.

Despite the dual mode approach, the final shape is still strongly affected by the initialisation. The model can cope well with inaccurate placement of a roughly correct shape, but is unable to recover from gross errors in initialisation which are consistent with the source data (such as the inclusion of pitch lines in the initial mattes - as shown in Figure 5), however these problems can be easily avoided through applying domain specific cleanup to either the input or the processed data.

Concavities remain a problem for this technique as can be seen in Figure 5. This is because the low resolution of the images and the presence of shadows and image bleeding can provide a strong edge across the concavity that prevents the surface from evolving in to the concavity. Improvement of the initialisation to ensure that some remnant of the concavity is present in the initial surface would give better results.

6. Conclusions and further work

We have demonstrated a technique that combines simultaneous multi-view shape extraction with stereo refinement to generate a view-dependent optimisation of an initial scene reconstruction. The dual-mode snake technique presented in this paper shows clear improvements in the completeness of the reconstruction compared to standard shape-from-silhouette techniques, and improves both the shape

Figure 6. Extreme view extrapolation giving view from pitch level inside playing area. Left: Dual mode deformable model, Right: billboards. Player-player occlusions in the original images lead to a double image of player 15 with the billboarding technique

and the appearance of the reconstruction compared to the conservative shape-from-silhouette technique.

The technique is still susceptible to poor initial segmentation and to clutter in particularly noisy parts of the image (such as viewing the goalkeeper through the goal net) which can lead to poor reconstruction even when clear views of the objects exist from other directions. Further work is required to improve the performance of the technique in these situations.

For the results presented in this paper only a single view-optimised reconstruction was used, further work will investigate techniques for blending multiple optimised surfaces together to improve the quality of synthesis of intermediate views.

The data used for this work was a set of Standard Definition (SD) images. As High Definition (HD) technology is becoming prevalent in the broadcast industry, future work will concentrate on the use of HD images.

Acknowledgements

This work was supported by the DTI Technology programme project iview: Free-viewpoint video for interactive entertainment production TP/3/DSM/6/I/15515 and EPSRC Grant EP/D033926, lead by BBC Research and Development (http://www.bbc.co.uk/rd/iview). The authors gratefully acknowledge the project partners for providing the sports footage and discussion of the research reported in this paper.

References

[1] J. Carranza, C. Theobalt, M. Magnor, and H. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, 2003.

[2] K. Connor and I. Reid. A multiple view layered representation for dynamic novel view synthesis. In *British Machine Vision COnference*, 2003.

[3] P. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. *Eurographics Rendering Workshop*, pages 105–116, 1998.

[4] EyeVision. *www.ri.cmu.edu/events/sb35/tksuperbowl.html*.

[5] K Hayashi and H Saito. Synthesizing free-viewpoint images from multiple view videos in soccer stadium. *International Conference on Computer Graphics, Imaging and Visualisation*, 0:220–225, 2006.

[6] C. Hernandez and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[7] N. Inamoto and H. Saito. Arbitrary viewpoint observation for soccer match video. *European Conference on Visual Media Production*, pages 21–30, 2004.

[8] T. Kanade, P. Rander, and P. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.

[9] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[10] J. Kilner, J. Starck, and A. Hilton. A comparative study of free-viewpoint video techniques for sports events. *European Conference on Visual Media Production*, 2006.

[11] T. Koyama, I. Kitahara, and Y. Ohta. Live mixed-reality 3d video in soccer stadium. *International Symposium on Mixed and Augmented Reality*, pages 178–186, 2003.

[12] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2):150–162, 1994.

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

[14] J. Sethian. *Level Set Methods*. Cambridge University Press, ISBN: 0521572029, 1996.

[15] J. Starck and A. Hilton. Virtual view synthesis of people from multiple view video sequences. *Graphical Models*, 67(6):600–620, 2005.

[16] J. Sun, J. Jia, C. Tang, and H. Shum. Poisson matting. *ACM Trans. Graph.*, 23(3):315–321, 2004.

[17] D. Terzopoulos, A. Witkin, and M. Kass. Symmetry-seeking models and 3D object reconstruction. *International Journal of Computer Vision*, 1(3):211–221, 1987.

[18] G. Thomas. Real-time camera pose estimation for augmenting sports scenes. *European Conference on Visual Media Production*, 2006.

[19] S. Vedula, S. Baker, and T. Kanade. Image-based spatio-temporal modeling and view interpolation of dynamic events. *ACM Trans. Graph.*, 24(2):240–261, 2005.