3D Action Matching with Key-Pose Detection

J.Kilner, J-Y.Guillemaut, A.Hilton
Center for Vision, Speech and Signal Processing
University of Surrey, Guildford, United Kingdom
{J.Kilner,J.Guillemaut,A.Hilton}@surrey.ac.uk

Abstract

This paper addresses the problem of human action matching in outdoor sports broadcast environments, by analysing 3D data from a recorded human activity and retrieving the most appropriate proxy action from a motion capture library. Typically pose recognition is carried out using images from a single camera, however this approach is sensitive to occlusions and restricted fields of view, both of which are common in the outdoor sports environment. This paper presents a novel technique for the automatic matching of human activities which operates on the 3D data available in a multi-camera broadcast environment. Shape is retrieved using multi-camera techniques to generate a 3D representation of the scene. Use of 3D data renders the system camera-pose-invariant and allows it to work while cameras are moving and zooming. By comparing the reconstructions to an appropriate 3D library, action matching can be achieved in the presence of significant calibration and matting errors which cause traditional pose detection schemes to fail. An appropriate feature descriptor and distance metric are presented as well as a technique to use these features for key-pose detection and action matching. The technique is then applied to real footage captured at an outdoor sporting event.

1. Introduction

Multi-camera techniques simultaneously record a single event from multiple angles. These techniques have recently been used in the field of sports media production. Examples include the use of EyeVision[7] and LiberoVision[20] to render novel views, and HawkEye[12] for analysis in a virtual environment.

Most multi-camera techniques are developed for use in special-purpose studios with static cameras where calibration and a known background are readily available, *e.g*. the Virtualized Reality system developed by Kanade *et al*. [16]. For multi-camera techniques to be used in unconstrained outdoor environments, errors in calibration and matting must be explicitly dealt with alongside other domain-specific challenges such as manually-operated cameras, unconstrained illumination, multi-body occlusion, rapid motion and low resolution images.

Figure 1. A split image depicting four stages of processing. Clockwise from the top-left - the original image, a segmented "key" separating the players from the background, a 3D reconstruction of the players and the final synthetic proxies of the players.

In this paper we propose a technique to automatically generate synthetic proxies representing players on a pitch. Proxies performing an appropriate action are selected from a motion capture library and synchronised to the original action. These synthetic proxies can be used directly for analysis, as avatars in a virtual view of the event, or as priors for further scene reconstruction.

Multi-camera video is recorded and a robust shape-from-silhouette technique is used to generate a 3D scene representation. This allows the fusion of multiple moving, zooming cameras, each of which may have been unsuitable for use on its own. Direct pose retrieval in this environment is severely challenged by the ambiguity in the relationships between the cameras. Therefore a Markov model is used to match both the pose and the action dynamics to a predefined library of human motion capture. This process is augmented with a scissor-pose detector to aid the synchronisation of the synthetic action to the original motion. The full process is illustrated in Figure 1

The background to this work is presented in Section 2. Section 3 describes the techniques used to generate the shape-from-silhouette data and to track the players. An appropriate feature descriptor and distance metric are presented in Section 4 and a technique to use these features for action matching is presented in Section 5. Section 6 contains the results obtained from real footage captured at an outdoor sporting event, and Section 7 concludes the paper with a discussion of the work.

2. Background

Pose detection is typically employed with the aim of generating high level descriptions of the detected motion for applications such as action recognition for surveillance[2]. Detailed pose recovery is generally confined to studio systems such as markerless motion capture systems[3]. Pose recovery in crowded scenes such as video of outdoor team sports remains an open and challenging problem.

2.1. Pose Recognition

There is a substantial body of work on 2D pose recognition. Recently Dimitrijevic *et al*. used Bayesian templates to recognise walking poses in natural scenes[5], Ferrari *et al*. used progressive search space reduction to estimate body pose in TV and film data[8], and Gammeter *et al*. used a statistical model of human pose to refine a pedestrian tracking system[9]. For sports applications, Efros *et al*. used optic flow to recognise low resolution video sequences of soccer players [6] and Lu and Little used histograms of orientated gradients to perform action recognition on low resolution soccer and hockey video sequences[22].

Less work has been done on 3D human pose matching. Weinland *et al*. used a volumetric exemplar representation to perform camera-pose-invariant pose recognition, however the matching was performed in 2D[25]. Huang *et al*. performed 3D pose matching for 3D video summarisation of studio data [23] and Balan *et al*. optimised the re-projection of a 3D model into silhouettes obtained in a studio environment to recover pose estimates from multi-camera video[3]. Although highly successful in the studio, such techniques fail in the presence of the calibration and matting errors typical of an outdoor broadcast scenario.

2.2. 3D in Sports

There has been considerable interest in application of 3D computer vision techniques to the field of sports since Kanade *et al*. developed EyeVision for use at the Super Bowl[7]. Connor and Reid demonstrated 3D reconstruction of a soccer game from multiple cameras [4], Innamoto and Saito demonstrated a system for 3D playback of a recorded soccer game [14] and Loy *et al*. reconstructed 3D motion from monocular video using pose templates[21]. Multi-camera systems are used in recent work by Guillemaut *et al*. where a graph-cut optimisation generates high-quality 3D scene reconstructions and matting refinement[11].

While progress is being made in this field, most techniques either require specialist equipment or manual intervention. A general solution has yet to be found.

3. Data Capture

The action matching algorithm relies on per-player 3D meshes as input. These are generated using multi-camera shape-from-silhouette techniques followed by 3D player tracking. The steps in this process are outlined in this section

3.1. Capture, Calibration and Matting

For a sports broadcast of soccer or rugby there are approximately 15 manually-operated cameras in the stadium. Of these, 6-8 cameras are typically following the action of interest and capture footage suitable for calibration. The other cameras include slow-motion cameras, cameras tightly focused on a single player and cameras giving an overall view of the stadium.

Calibration and matting is as described by Grau *et al*. [10]. The cameras are calibrated by detecting pitch markings in the image and comparing them to a pre-generated model. It is important that an on-line calibration technique such as this is used as it allows for calibration of moving cameras, does not require any additional equipment, does not require prior knowledge of, or access to, the equipment (which is typically hired for the match) and does not require prior access to the ground, which may be difficult to obtain. While this technique produces relatively accurate results, the resulting calibration contains significant errors of the order of 1-2 pixels.

The images captured during the match are segmented using a hybrid chroma-key and difference keying technique to separate the image into background and foreground elements. This technique is chosen for its relative accuracy and suitability for running on a video sequence with minimal human intervention. An example segmentation is shown in the top right of Figure 1. This technique typically segments players with an error of 1-2 pixels but also generates an appreciable amount of background clutter.

3.2. Robust Multi-View Reconstruction

The calibration data and silhouettes are used as input to a robust shape-from-silhouette technique to calculate the visual hull[19] of the foreground region of the scene. As the calibration and matting contain combined errors of up to 3 or 4 pixels (which is of similar magnitude to the resolution of players' limbs), the volume generated by apply-

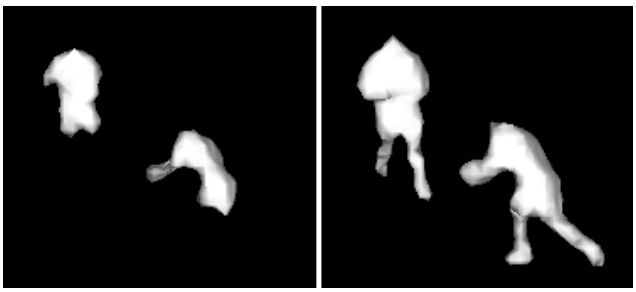

Figure 2. A typical player representation generated by the traditional (left) and conservative (right) shape-from-silhouette techniques.

ing a straightforward shape-from-silhouette technique will be severely truncated. To compensate for these errors the conservative shape-from-silhouette technique introduced by Kilner *et al*. [18] is used to generate a Conservative Visual Hull (CVH) as shown in Figure 2. This technique provides a complete reconstruction of the scene at the cost of some loss of accuracy.

3.3. Player Tracking

A triangle mesh is generated from the CVH on a per-frame basis. This represents all players on the pitch, the referee and the ball with a single mesh. The mesh is divided up into discrete sub-meshes through a connected component analysis.

A greedy algorithm is used to concatenate these sub-meshes into coherent sequences minimising a cost function which penalises the distance between mesh centroids and changes in volume. By ensuring that tracks span the entire sequence, a set of per-player mesh sequences are generated for use as input to the action matching algorithm.

4. Shape Similarity

Pose estimation attempts to recover the configuration of a human body from a set of observations. Our technique determines the similarity between the 3D shape obtained from a multi-view reconstruction and a set of exemplar poses contained within a library. The best match is used as an estimate of the observed pose.

To determine the similarity between a candidate pose and the real data requires a measure of similarity between two different 3D shapes. To measure similarity both a feature vector and a distance metric are required. These must be chosen such that they maximise the distance between shapes derived from differing poses, but minimise the distance between shapes derived from the same pose. The measure should be robust to the amount of reconstruction error expected in the system. In this section we evaluate a shape similarity metric which is robust to the errors in large-scale, multi-view reconstruction.

4.1. Feature Vector

Various shape descriptors have been proposed for 3D shape matching, including spin images[15], spherical harmonics[17] and shape histograms[1]. Huang *et al*. [13] investigated the use of these measures for matching sequences of people in the studio environment and proposed a volumetric shape histogram which is robust to small differences in shape while retaining rotational information (unlike spin images which will match a shape but will not indicate the orientation that produces the optimal match).

For this work, volumetric shape histograms using a spherical co-ordinate system are used. A surface S is isotropically scaled such that it lies within a sphere of radius 1 located at the origin and re-oriented per frame such that the direction of motion of the centroid is always along the Z axis. S is then represented by an implicit function V where $V(x) = 1$ when x lies within S and $V(x) = 0$ when x is outside of S.

The shape histogram H is then obtained as:

$$H_{i,j,k} = \sum_x V(x)B(i,j,k,x), \tag{1}$$

where B is the bin-membership function:

$$B(i,j,k,x) = \begin{cases} 1 & \begin{pmatrix} i\delta r < r_x < (i+1)\delta r \\ j\delta\theta < \theta_x < (j+1)\delta\theta \\ k\delta\phi < \phi_x < (k+1)\delta\phi \end{pmatrix} \\ 0 & otherwise \end{cases} \tag{2}$$

with x expressed in spherical co-ordinates (r_x, θ_x, ϕ_x) and with histogram quantisation steps $(\delta r, \delta\theta, \delta\phi)$. A 5 by 6 by 12 histogram was used in this work as that was empirically determined to capture pose information in sufficient detail while minimising computational costs.

4.2. Similarity Metric

Several distance measures were evaluated against ground truth to determine the most suitable matching score. We considered the histogram intersection I, Euclidean distance E, quadratic distance Q[1], Mahalanobis distance M, chi-squared measures $\chi^2 1$ and $\chi^2 2$ and the Kullback Leibler Divergence (KLD) K. These measures are given by the following equations:

$$I(h,e) = \sum_i \min(h_i, e_i), \tag{3}$$

$$E(h,e) = \sqrt{(h-e)(h-e)^T}, \tag{4}$$

$$Q(h,e) = \sqrt{(h-e)Q(h-e)^T}, \tag{5}$$

$$M(h,e) = \sqrt{(h-e)M(h-e)^T}, \tag{6}$$

$$\chi^2 1(h,e) = \sum_i (h_i - e_i)^2/e_i, \tag{7}$$

$$\chi^2 2(h,e) = \sum_i (h_i - e_i)^2/(h_i + e_i), \quad (8)$$

$$K(h,e) = \sum_i h_i \log(h_i/e_i). \quad (9)$$

Where h and e are histograms of similar dimensions and i is an index over all histogram bins, Q is the quadratic distance matrix which encodes the Euclidean distance between the centres of the histogram bins and M is the covariance matrix of all the data. Where a measure is asymmetric (such as K, $\chi^2 1$ and $\chi^2 2$), a symmetric measure was used $X_{sym}(a,b) = 0.5(X(a,b) + X(b,a))$.

4.3. Evaluation

Several tests were carried out to determine the most suitable measure for use in the sports scenario. A library of motion-captured animations were used to generate animated meshes from which a library of shape histograms $L = \{L_0 \dots L_{|L|}\}$ was generated. From this library, a sub-sequence $S = \{S_0 \dots S_{|S|}\}$ was chosen to use for comparison. A noisy sequence of shape histograms $N = \{N_0 \dots N_{|S|}\}$ was then generated from the same source meshes as S. Noise was added by translating all vertices along the vertex normals by 10cm and adding shot noise to 40% of the voxels during voxelisation. These operations simulate the expansion of shape caused by using the CVH and the additional effects of poor matting and phantom volumes.

A qualitative evaluation was then performed by generating a set of similarity matrices (a matrix Ψ where $\Psi_{A,B,C}(i,j) = A(B_i, C_j)$ for a given measure A and sequences B and C). $\Psi_{X,S,N}$ was generated for each measure X and was then compared to a ground truth self-similarity matrix $\Psi_{E,S,S}$, (i.e. comparing S to itself using the Euclidean distance E defined in equation 4). As both S and N are derived from the same source meshes, the ideal measure should generate a matrix displaying similar structure to the ground truth, with particular importance being attached to the main diagonal feature which corresponds to the distance of each frame from itself. These results allowed a comparison of the different measures as shown in Figure 3. This diagram shows that the only measure which maintains the strong diagonal feature is the Kullback Leibler distance.

A quantitative evaluation was also performed. A similarity matrix Ψ can be converted into a Boolean classification function $C^{\tau}_{\Psi}(i,j) = \Psi(i,j) \leq \tau$. A ground truth classification $G = C^{0.2\sigma}_{\Psi_{E,S,L}}$ is then generated where σ is the standard deviation of $\Psi_{E,S,L}$. In this way, G contains the matches between frames in S and L which fall within an acceptable threshold. Each measure X was then used to compare N with L and a set of classification functions $F^{\kappa}_X = C^{\kappa}_{\Psi_{X,N,L}}$ were generated. Comparing F^{κ}_X with G gives a set of true positives (where $F^{\kappa}_X(i,j) = true = G(i,j)$) and false positives (where $F^{\kappa}_X(i,j) = true \neq G(i,j)$). By varying κ from $min(\Psi_{X,N,L})$ to $max(\Psi_{X,N,L})$ a Receiver Operating Characteristic (ROC) curve for each technique X can be generated as shown in Figure 4.

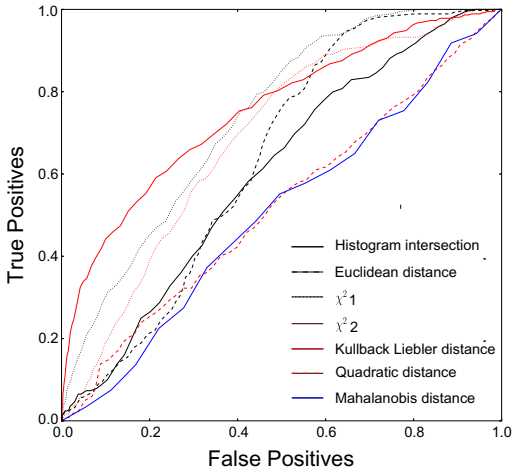

Figure 4. ROC curves for various 3D shape similarity measures in the presence of multi-view reconstruction errors. More discriminative techniques produce a curve which tends towards the upper left corner, less discriminative techniques tend towards a straight line from the bottom left to the upper right corner.

The ROC curves agree with the qualitative analysis that, in the presence of multi-view reconstruction errors, the KLD provides the most appropriate measure of shape histogram similarity.

5. Action Matching

The KLD allows shape to be matched between noisy data and a clean exemplar. However, ambiguities in the recovered signal and variation between the detail of the recorded action and the library motions mean that frame-by-frame matching generates poor results. In order to resolve these ambiguities a two-step process is used. Firstly a key-pose detector identifies each key-pose in the target sequence. Secondly, a Markov model is used to match the library to the target sequence, taking into account the shape, dynamics and detector hits from the first step. The resulting synthetic sequence is then an animation from the library which best matches the shape and dynamics, and is synchronised to the target sequence.

5.1. Key-Pose Detection

One important feature of action matching is the synchronisation of the synthesised action to the original activity. This is mainly a problem for cyclic activities which can drift out of phase over time. The majority of cyclic activities in the outdoor sports environment are motions such as walking, jogging and running. Dimitrijevic *et al.* [5] showed

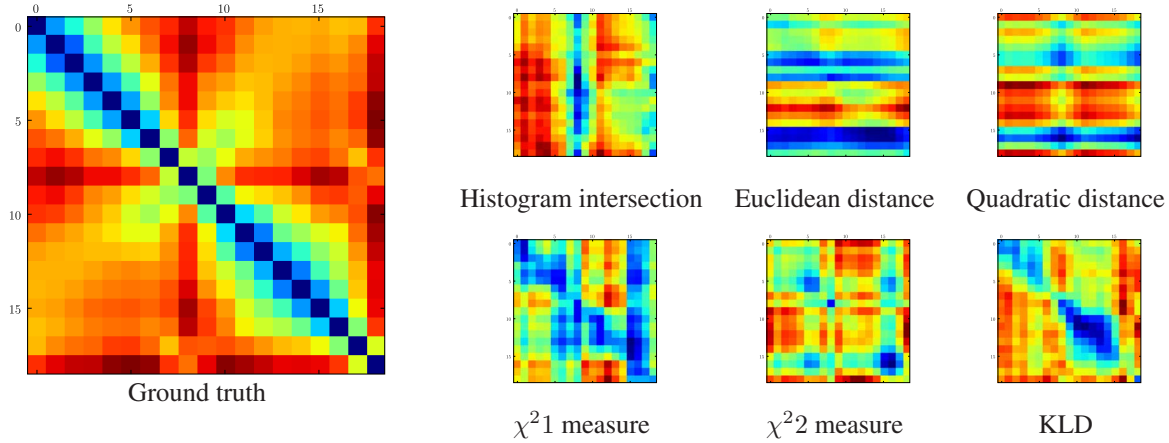

Figure 3. Similarity matrices demonstrating the performance of various similarity measures when comparing a noisy sequence to the original clean sequence. A good measure should preserve the structure shown in the ground truth, particularly the strong diagonal feature.

that identification of the "scissor poses" in 2D data of people walking is sufficient to accurately specify the motion of a tracked subject, as this key-pose is the most distinctive feature in the walking cycle. By detecting these distinctive poses and tying them to the scissor poses in the library actions, synchronisation between the synthetic and target sequences can be improved.

The pose detector compares shape histograms of the left scissor pose ϵ_l, the right scissor pose ϵ_r and the mean pose of the library $\epsilon_m = \sum_i \frac{L_i}{|L|}$ to the input sequence of histograms $D = \{D_0 \dots D_{|D|}\}$. The matching score λ is given by

$$\lambda(t) = 2K(D_t, \epsilon_m) - K(D_t, \epsilon_l) - K(D_t, \epsilon_r) \qquad (10)$$

ϵ_m provides a correction for changes in the matching score due to non-pose-related factors (for example changes in the volume of the CVH or gross errors such as the truncation of a limb). This combined signal is then smoothed and the local maxima are taken as the detector hits Γ. An example graph showing the peaks in the signal can be seen in Figure 5. By testing $K(D_t, \epsilon_l) > K(D_t, \epsilon_r)$ each hit can also be classified as 'left' or 'right'.

These detector hits are then incorporated as soft constraints in the second step of the process to aid synchronisation of the synthetic sequence to the target activity.

5.2. Markov Model

Pose interpolation is achieved by modelling the evolution of the motion as a Markov process represented in the standard way using a state transition matrix T and an emission matrix E[24]. Each frame of each animation in the library L is a state in the model. The original structure of the animations is encoded into the state transition matrix T:

$$T(i,j) = \eta(i,j)\rho(i,j) + \alpha(1 - \eta(i,j))K(L_i, L_j). \quad (11)$$

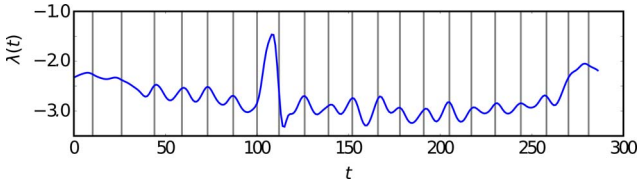

Figure 5. Example matching scores from real data. The gray vertical lines are the frames manually identified as scissor poses. The error near frame 100 is due to a large phantom volume which distorts the CVH. Other scissor poses are all correctly detected.

ρ is a function that returns a maximal value when $j = i + 1$ and hence favours the natural playback of a library animation. However, non-zero values for $j = i$ and $j = i + 2$ allow the animation to repeat or skip frames to match animations at different rates. η is 1 if i and j are in the same animation, zero otherwise, and α is a small value that controls the rate of switching between animations. $\alpha = 0.1$ was empirically determined to be a suitable value - higher values tend to make the model choose a single animation when it would be better to switch, while lower values tend to encourage the model to switch animations excessively. The emission matrix is then defined by:

$$E_{i,j} = -b(i,j)\sqrt{K(D_i, L_j)} \qquad (12)$$

where b is given by:

$$b(i,j) = \begin{cases} 1.2 & (i \pm 1 \in \Gamma, j \pm 1 \in \kappa) \\ 1 & (otherwise). \end{cases} \qquad (13)$$

b boosts scores where a scissor pose was detected to promote matching of these frames to scissor poses in the library. In this way, the detector hits Γ constrain the model to fit to the detected poses and to improve the synchronisation of running and walking actions. T and E are normalised

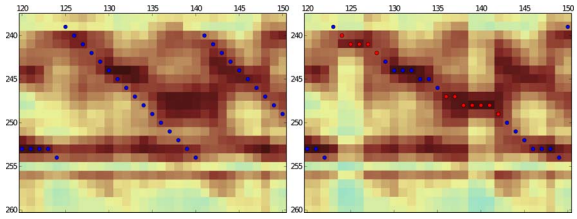

Figure 6. A crop from the emission matrix overlaid with the optimal path obtained by the Viterbi algorithm using the unmodified Markov model (left) and the Markov model boosted with the key-pose detector (right). Red circles show library states that correspond to scissor poses. The effect of the key-pose detector can be seen at frames 125 and 140. Without boosting, the model will tend to favour playback of the animation at its natural rate, key-pose detection can overcome this and improves the synchronisation to the original action.

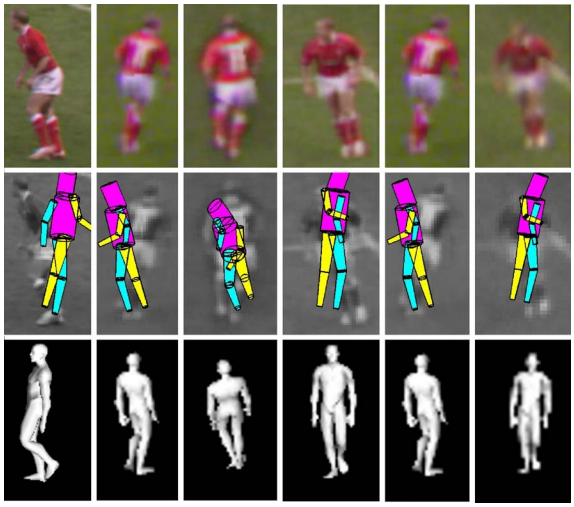

Figure 7. Pose estimation for a single player. The top row shows the original images, the middle row shows a traditional 2D pose estimation technique[2] and the bottom row shows the proposed 3D action matching technique. Due to calibration and matting errors, the 2D technique is unable to recover the pose from the images.

appropriately and the optimal state sequence is calculated using the Viterbi algorithm, maximising the product of the transition probability and the emission probability for each frame in the sequence. As shown in Figure 6, the use of key-pose data improves the synchronisation of activities, but the addition of repeated or skipped frames can be uneven. If the technique is being used to generate models for direct viewing, a post-processing step of smoothing the path through the output animations could be applied to remove these artefacts.

6. Results

The technique was evaluated with footage of professional rugby soccer matches. Data was recorded using standard broadcast equipment. The recording equipment consisted entirely of standard HD broadcast cameras. 12 cameras were recorded for the rugby data - 6 static cameras used only for reconstruction and 6 operator-controlled cameras used for broadcast. At any given time roughly half of these cameras were suitable for reconstruction, the others were unsuitable due to inappropriate framing or excessive motion blur. The cameras were distributed around the pitch as shown in Figure 8. A set of close-up images showing the low quality of per-player imaging is shown in Figure 7. Two sequences were captured - one of 287 frames showing 29 players and the other of 50 frames showing 31 players. 16 cameras recorded the football footage - 3 were static and the rest were operator-controlled. Similar restrictions on usability meant that approximately 6 cameras were usable for reconstruction. Two sequences were captured - one of 100 frames showing 26 players and the other of 120 frames showing 22 players.

Figure 8. Arrangement of cameras used for the rugby data set. Static cameras are coloured blue while moving cameras are coloured red. Note that the static cameras are aimed at either one end of the pitch or the other, so that at any one time only 3 static cameras and up to 3 moving cameras are usable for reconstruction.

The synthetic library for comparison consisted of motion capture data skinned to a human model. Motion capture consisted of various running, jogging, walking, standing, jumping and skipping animations. Neither the motion capture nor the model were specific to this domain. It should be noted that domain-specific motion capture would improve the quality of the matching, and use of properly rigged football/rugby player models would greatly improve the visual quality of the results.

An example frame is shown in Figure 9. Further results are presented in the accompanying video and shown in Figure 10. For comparison, a result generated with a state-of-the-art pose tracking system[2] is shown in 7. The calibration and matting errors in the data mean that the technique does not converge on a solution and quickly drifts away from a reasonable pose estimate.

A quantitative analysis was performed on the results from the longest and most complex sequence (from the rugby data). Due to matting errors and occlusions in the

Figure 9. Left to right, the original input image, the resultant CVH and the final matched pose. Some synthetic players are mirror images of the original player, as left/right symmetry introduces an ambiguity that is difficult to resolve completely.

	Matches				
Sequence	Exact	Near	Mirror	Hit	Miss
Shape	191	88	85	364	392
Action	363	95	80	538	218
Action & Pose	**458**	**94**	**49**	**601**	**155**

Table 1. Evaluation of the generated pose estimates for shape matching, action matching and key-pose augmented action matching. Exact matches are matches where it is reasonable to conclude that no closer match exists in the library and the action is appropriate. Near matches are matches where the action is appropriate but the pose is not exactly the same, and mirror matches are where the left/right ambiguity has been resolved incorrectly. The proposed technique clearly performs far better than simple matching or a Markov model alone.

original images, standard recall-precision measurements comparing the pixels in a virtual view against the pixels in the original mattes are not very meaningful. The lack of a ground truth also precludes a direct quantitative analysis of the technique. Instead, every tenth frame of each sequence was examined and each player assessed as either a 'hit' or a 'miss'. Players were considered an exact match if the pose was of an appropriate action and was as accurate as was possible given the limitations of the library. Close matches and matches where left/right ambiguity resulted in the selection of a mirror image of the correct pose were also considered hits. Anything else was considered a miss, including poses which were close to correct but came from an inappropriate action. The results are shown in Table 1 which compares simple pose matching, action matching and the proposed key-pose augmented action matching technique.

It should be noted that many of the failures common to all techniques are due to players performing actions which are not in the pose library. In many such cases the system selects a pose close to the correct pose, suggesting that if the library contained more appropriate motions they would be selected. Examples of actions that caused failure in this way include walking sideways, kicking and walking backwards. Another cause of failure was excessive clutter. The system can perform well in the presence of occlusions in several cameras, but in cases where multiple players were occluded for long periods of time (creating a single 3D volume for multiple players), the system failed. Improved pre-processing could separate these volumes into individual players and hence improve results. Roughly 1/3 of failures were due to actions not represented in the library and 1/3 were due to excessive clutter in the scene. Finally, left/right ambiguity was an issue in many cases (and indeed is often hard to resolve visually from the original images). The system will occasionally switch modes as it changes from incorrect to correct representation of the left/rightness. Changing in this way introduces either an additional cycle or doubles the length of a cycle. Key-pose matching helps to resolve this by reducing mode-switching but does not eliminate the error entirely.

7. Discussion

7.1. Conclusions

The presented technique is effective in this challenging environment, making use of multiple noisy and disjoint input images to extract player pose information. The technique allows manually-operated broadcast cameras to be used for pose estimation in the presence of calibration and matting errors and is independent of camera pose and player appearance, allowing a single library to be used in all scenarios. The technique provides good pose estimation for up to 70% of the footage, overcoming many of the limitations of single-view pose estimation due to occlusion and projective ambiguity.

7.2. Further Work

Players in close proximity to each other such that individual volumes are not separable are not handled well by the system. More sophisticated player tracking that can segment merged player tracks would enable the pose matching system to function better in these cases.

A larger action library targeted at the specific domain would improve the matching performance of the system. Similarly, post-processing of the synthetic sequence and use of higher quality domain-specific models would improve the quality of the output from the technique.

8. Acknowledgements

This work was supported by the DTI Technology programme under Free-viewpoint video for interactive entertainment production TP/3/DSM/6/I/ 15515 and EPSRC Grant EP/D033926. For further details visit the iview project (http://www.bbc.co.uk/rd/iview).

Figure 10. Crop of players showing every 20th frame over a rugby sequence (top two rows) and every 10th frame over a football sequence (bottom two rows). The system fails due to excessive clutter in the second row, but quickly recovers despite continued occlusions.

References

[1] M. Ankerst, G. Kastenmller, H. peter Kriegel, and T. Seidl. 3d shape histograms for similarity search and classification in spatial databases. *International Symposium In Advances in Spatial Databases*, pages 207–226, 1999.

[2] A. Balan, L. Sigal, and M. Black. A quantitative evaluation of video-based 3d person tracking. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 349–356, 2005.

[3] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

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

[5] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose detection using bayesian spatio-temporal templates. *Computer Vision and Image Understanding*, 104(2):127–139, 2006.

[6] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *IEEE International Conference on Computer Vision*, pages 726–733, 2003.

[7] Eye-Vision. Carnegie Mellon goes to the Super Bowl. *http://www.ri.cmu.edu/events/sb35/tksuperbowl.html*, 1996.

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] S. Gammeter, A. Ess, T. Jaeggli, K. Schindler, B. Leibe, and L. J. V. Gool. Articulated multi-body tracking under egomotion. *European Conference on COmputer Vision*, pages 816–830, 2008.

[10] O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, and J. Starck. A free-viewpoint video system for visualisation of sport scenes. *SMPTE Motion Imaging*, pages 213–219, 2007.

[11] J.-Y. Guillemaut, J. Kilner, and A. Hilton. Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. *IEEE International Conference on Computer Vision*, 2009.

[12] HawkEye. HawkEye Technology Website. *http://www.hawkeyetechnology.co.uk/*, 2008.

[13] P. Huang, J. Starck, and A. Hilton. A study of shape similarity for temporal surface sequences of people. *International Conference on 3-D Digital Imaging and Modeling*, pages 408–418, Aug. 2007.

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

[15] A. Johnson. *Spin-Images: A Representation for 3-D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, 1997.

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

[17] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–164, 2003.

[18] J. Kilner, J. Starck, A. Hilton, and O. Grau. Dual-mode deformable models for free-viewpoint video of sports events. *International Conference on 3-D Digital Imaging and Modeling*, pages 177–184, 2007.

[19] A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[20] LiberoVision. LiberoVision GmBH Website. *http://www.liberovision.com/*, 2008.

[21] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson. Monocular 3d reconstruction of human motion in long action sequences. *European Conference on Computer Vision*, pages 442–455, 2004.

[22] W.-L. Lu and J. Little. Simultaneous tracking and action recognition using the pca-hog descriptor. *Canadian Conference on Computer and Robot Vision*, pages 6–6, June 2006.

[23] P.Huang, A.Hilton, and J.Starck. Automatic 3d video summarization: Key frame extraction from self-similarity. *International Symposium on 3D Data Processing, Visualization and Transmission*, June 2008.

[24] L. Rabiner. A tutorial on hmm and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[25] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. *IEEE International Conference on Computer Vision*, pages 1–7, 2007.