Full-body Performance Capture of Sports from Multi-view Video

Lewis Bridgeman
l.bridgeman@surrey.ac.uk
Jean-Yves Guillemaut
j.guillemaut@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk

Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey

Full-body human performance capture has been extensively explored in constrained environments, but less attention has been applied to the task of performance capture in sports scenes. Sports datasets provide a wealth of challenges: player contact and occlusion; fast motion; and low resolution and wide-baseline cameras. We present a method that uses multi-view pose to disambiguate players and overcome some of these issues. The applications are broad, including player performance analysis, sports broadcast and immersive experiences.

We present a method for the full-body performance capture of multiple people in a sports environment from multi-view video. This paper leverages previous work on multi-person 3D pose estimation and tracking in sports [1], and model-based human shape and pose estimation. These techniques are combined to produce an estimate of the body shape and pose for all players in a sports scene. We demonstrate results on a soccer dataset comprising over 20 subjects.

Related Work: A scene reconstruction of a soccer match is generated in [3] using an optimization-based approach that assigns a depth label to the pixels in each camera image. The recent method [7] employs CNN-based depth estimation to generate 3D billboard representations of players from monocular video. Both methods produce plausible free-viewpoint renderings, but the scene representations contain no information on the underlying human motion. Estimation of human shape and pose has been explored in constrained environments in [4]; the SMPL body model [5] is aligned to multi-view video of a single subject according to pose estimations in an energy minimisation. A recent trend has seen the use of CNNs to estimate the model parameters directly from images, as in [6]. Results are promising, but are currently restricted to single images.

Methodology: The input to our method is multi-view video plus calibration. Each video is passed through a pose detector [2], producing independent pose estimations per view and per frame. The multi-person pose detector under-performs on football data due to the small size of the players (averaging 80 pixels in height). To overcome this, we automatically crop bounding boxes of the players and run the pose estimator on these individually. Following the method outlined in [1], we are able to identify and correct a variety of erroneous pose estimations, such as flipped limbs, and find correspondences between the poses in different camera views. A secondary stage uses the correspondences between poses to generate 3D skeletons; these are subsequently tracked throughout the sequence. The method outputs tracks of 3D skeletons, however, we adapt it to output tracked 2D poses with correspondences between camera views, which we use as input to our model-fitting stage.

The model-fitting stage is based upon the method in [4], in which the SMPL body model is aligned to multi-view video of a single person. We extend this method to multiple people by using the sorted pose estimations from the previous stage. We minimise the energy function given in equation 1.

$$E_M(\beta,\theta) = E_J(\beta,\theta) + \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta) \quad (1)$$

E_J is the joint fitting term, E_θ and E_β are pose and shape priors, and θ and β are the pose and shape parameters under optimisation. The joint term E_J aligns the joints within the SMPL model to the pose estimations. The built-in SMPL joints differ from the joint locations identified by the pose estimator. We build a regressor that produces a second set of joints directly from the vertices of the SMPL model; these joints are equivalent to that of the pose estimator. Thus our joint fitting term aligns the newly regressed joints with the pose estimations in the input images. The prior terms E_θ and E_β constrain the model to realistic poses and body shapes. We introduce an additional constraint to ensure that the SMPL body shape parameters are constant throughout the sequence. Although foreground segmentation is available, we do not include a silhouette fitting term so that our method remains applicable to sports sequences without segmentation.

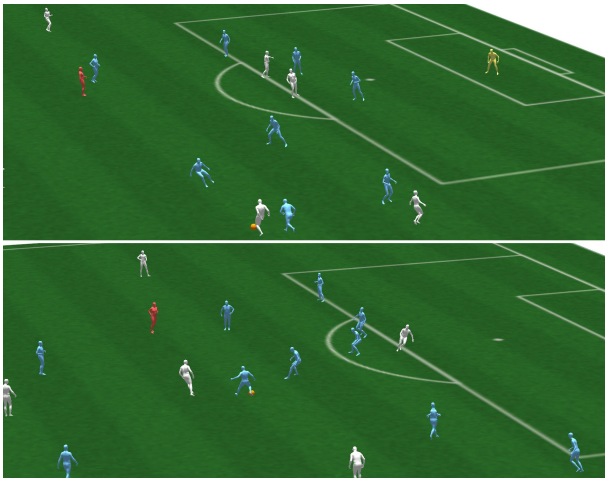

Figure 1: Results of the first and last frames in the sequence. Team colours and ball added manually.

We employ a DCT-based temporal smoothing term as defined in [4]. We ensure that the trajectory of each joint can be approximated by the first 6 basis functions of a 30-frame window. We use 30-frame term as a moving window with 50% overlap on the entire sequence. We allow the DCT term to fill in gaps in the sequence where players lose tracking.

Results: We demonstrate results on a short soccer sequence of 120 frames. The sequence comprises 6 HD cameras, both fixed and moving, which are an average of 48m from the pitch centre. The dataset proves challenging due to motion blur caused by moving cameras, and the small size of the players in the images. The pose-sorting stage produces 21 tracked skeletons (including one ID switch) with occasional missing frames. We apply the model-based reconstruction process individually on each tracked skeleton. Select frames from the resultant reconstruction can be seen in figure 1. The method successfully captures highly dynamic motion of the players with minimal jitter, although artefacts caused by loss in tracking remain apparent. Future work could include generating a motion prior from video, specifically tailored to soccer players, that encompasses more dynamic poses.

Acknowledgements: *This work was funded by EPSRC Grant EP/N50977/1.*

[1] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *CVPR Workshop on Computer Vision in Sports*, 2019.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Jean-Yves Guillemaut and Adrian Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *IJCV*, 93:73–100, 2010.

[4] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 34(6):248:1–248:16, 2015.

[6] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018.

[7] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *CVPR*, 2018.