VIEWPOINT-DEPENDENT 3D HUMAN BODY POSING FOR SPORTS LEGACY RECOVERY FROM IMAGES AND VIDEO

Luis Unzueta, Jon Goenetxea, Mikel Rodriguez and Maria Teresa Linaza

Vicomtech-IK4, Paseo Mikeletegi, 57, Parque Tecnológico, 20009, Donostia, Spain

ABSTRACT

In this paper we present a method for 3D human body pose reconstruction from images and video, in the context of sports legacy recovery. The video and image legacy content can include camera motion, several players, considerable partial occlusions, motion blur and image noise, recorded with non-calibrated cameras, which increases even more the difficulty of solving the problem of 3D reconstruction from 2D data. Therefore, we propose a semi-automatic approach in which a set of 2D key-points are manually marked in key-frames and then an automatic process estimates the camera calibration parameters, the positions and poses of the players and their body part dimensions. In-between frames are automatically estimated taking into account constraints related to human kinematics and collisions with the environment. Experimental results show that this approach obtains reconstructions that can help to analyze playing techniques and the evolution of sports through time.

Index Terms— Motion capture, human body posing, multibody mechanism fitting, sports preservation and promotion

1. INTRODUCTION

Human motion can be tracked by means of many different devices. Currently, marker-based systems are the most accurate motion capture systems, while markerless solutions that rely on 3D sensing devices, such as the Microsoft Kinect camera, are the most popular choices when accuracy is not so important due to their lower cost. However, all these options require a specific hardware to be used, with their corresponding installation constraints (maximum allowed workspace, controlled illumination conditions, dry weather, etc), which can limit considerably their usage for the motion capture of sports players in action. In such cases, video-based motion capture can be an alternative to be considered as it only requires video images of players, which can be obtained directly from TV footage of matches.

Additionally, video-based motion capture is of special interest for Traditional Sports and Games, as an expression of Intangible Cultural Heritage [1] from different world regions under the threat of disappearing. More specifically, it can help to recover in 3D playing techniques from legacy video content. Thus, it can help to analyze the evolution of sports through time and promote them to broader audience all around the world. For example, the extracted 3D data can help to generate virtual content to be shown in virtual museums, virtual immersive systems or videogames, in which the motions captured from users can be compared to those of past top players.

Monocular motion capture is substantially more challenging than multi-view systems [2], as no depth can be directly measured from image data. Some approaches, like [3], rely on motion databases in order to relate 2D data with 3D poses, obtaining reliable results in specific applications with few and simple human motions, such as walking. However, these cannot be extended to more complex cases such as sports. Others, like [4], do not need any prior learning on motion capture/image annotation data and can generalize better, however they need additional constrains such as good image quality and static cameras, as they rely on background subtraction. There are also some approaches focused on sport player pose reconstruction, such as [5] and [6], which can obtain visually acceptable results for more general cases. The main drawback in [5] is that at least five key-frames in the same sequence are to be manually set in order to estimate the 3D body parameters, while [6] focuses on human pose reconstruction in 2D.

In this paper we propose an approach for 3D human body pose reconstruction from uncalibrated monocular cameras, which can be applied for the recovery of sport player motions from TV footage, without pose limitations, not limited to video sequences but also applicable to single snapshots.

The paper is organized as follows. Section 2 gives insight on our video-based motion capture approach. Section 3 explains the method we propose to locate and pose the 3D human body on images and videos. Section 4 shows the experimental results we obtain with monocular TV sports

The work presented in this paper was supported by the European Commission under contract FP7-601170 RePlay.

footage. Finally, in section 5 we discuss the obtained results and the future work.

2. VIDEO-BASED MOTION CAPTURE

The proposed general procedure for video-based motion capture is shown in Figure 1. It consists of two main stages: (1) the manual setup of key-frames and (2) the automatic estimation of in-between frames. For the first stage, the problem to be solved is the viewpoint-dependent 3D model posing. For the second stage, the players, the ball and the camera parameters are tracked.

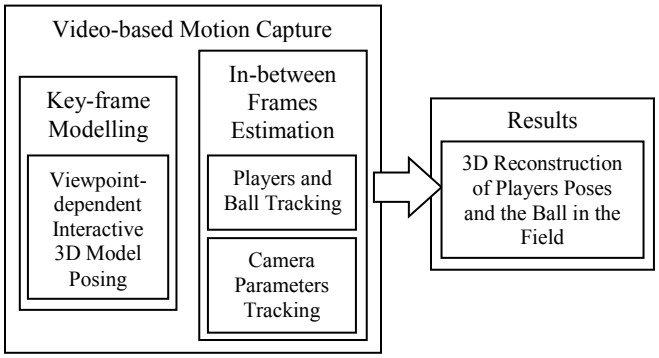

Fig. 1. Video-based motion capture general procedure.

In this work we simplify the second stage by interpolating linearly the positions and orientations of the camera, players and the ball. In the case of the orientations we rely on SLERP (Spherical Linear Interpolation) [7]. The interpolated body joint orientations, correspond to their local orientations, which are corrected by the kinematic constraints where required (see subsection 3.1).

3. VIEWPOINT-DEPENDENT INTERACTIVE 3D HUMAN BODY POSING

In order to solve the viewpoint-dependent 3D human body posing, we propose the approach shown in Figure 2. Here the input data are the image and 2D posing features corresponding to the player bodies, the sports objects and field. Additionally, a set of kinematic constraints will allow for inferring plausible configurations to the ambiguities derived from the perspective projection.

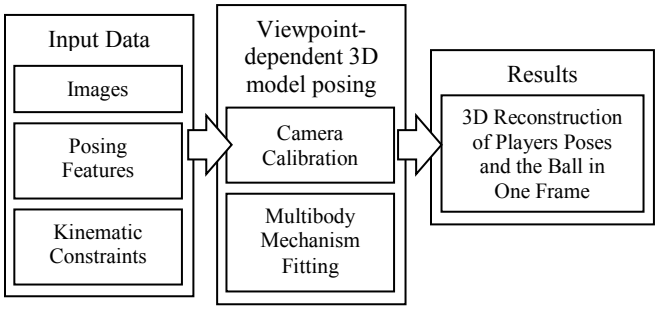

Fig. 2. Viewpoint-dependent 3D model posing general procedure.

The key-frames are processed in two steps: (1) camera calibration and (2) multibody mechanism fitting. Both processes are assisted by constrained IK (Inverse Kinematics).

3.1. Constrained Inverse Kinematics

Figure 3 shows the 3D kinematic model and the posing features that control its configurations through IK. These posing features correspond to the positions of pelvis, head, hips, knees, ankles, shoulders, elbows and wrists. For a specific set of posing feature values different body poses can be obtained, depending on the adopted IK approach.

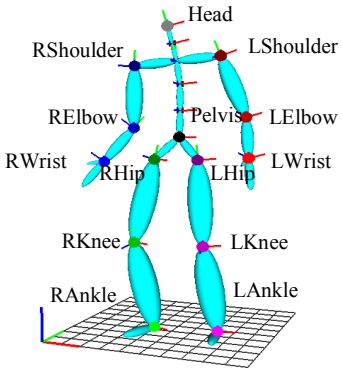

Fig. 3. The kinematic structure of the human body and its posing features.

In our case, taking into account that we want to solve the problem of 3D human body posing on 2D images, with ambiguities derived from the perspective projection, it is especially helpful to constrain the poses to those expected in the context of sports. For the IK adjustment, we consider five kinematic chains: (1) the trunk, which contains the pelvis, hips and head posing features, along with the pelvis and spine body segments, (2) the left lower limb, which contains the left hip, knee and ankle posing features, along with the left leg and foot segments, (3) the left upper limb, which contains the left shoulder, elbow and wrist posing features, along with the left clavicle, arm and hand body segments, (4) the right lower limb, and (5) right upper limb.

The kinematic chains are adjusted sequentially [8]. Biomechanical joint constraints are included in order to reduce the mobility of joints to those that make sense with human body joint motion ranges (see Figure 4 for an example).

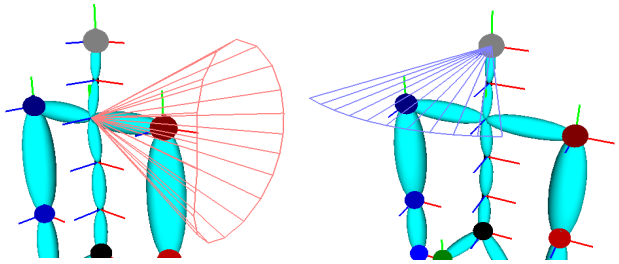

Fig. 4. On the left, the swing-circumduction limits of the left sternoclavicular joint, and on the right, the twist limits of the head.

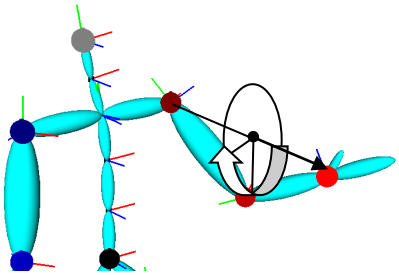

Fig. 5. The swivel angle of the left arm.

The posing features update the human pose in specific ways: (1) the pelvis position controls the motion of the whole body and the rest of posing features as a rigid body, (2) the head position controls the spine, while maintaining the rest of posing features static, (3) the shoulders control the motions of clavicles and arms, maintaining the arm relative poses static, (4) the ankle and wrist positions control their corresponding arms and legs, giving more preference to the correct matching of the limb end-effectors than to the matching of their intermediate ones (elbows and knees), and (5) the intermediate posing features control the swivel angles of upper and lower limbs (Figure 5). The latter is especially required in our context, as the position of hands and feet is of important relevance since they interact with objects in the scene such as sticks, balls and the floor. The floor model is used to correct the posing features in order to avoid their penetration in it. After the corrections, the IK approach makes the kinematic model adapt accordingly.

3.2. Camera calibration

A calibration of the watched scene is required to apply the proposed methods for the posing of players. It can be obtained by computing the intrinsic parameters of the camera (the focal length and the principal point), and the extrinsic parameters (the rotation and translation) with respect to a selected coordinate frame. For this, we follow the approach from [9], but without considering distortion, as this usually is not present in the cameras used for recording sports games and simplifies the calibration procedure.

The intrinsic and extrinsic parameters are calculated in a single-step process. The user introduces manually 4 points in the image that correspond to a rectangle in the floor plane of the scene, plus its longitudinal and transversal sizes in metric units. This information is sufficient to compute the homography between the image plane and the floor plane using the DLT (Direct Linear Transform) algorithm [10]. Once the homography has been computed and calibrated, it is possible to extract the rotation and translation from the resulting matrix.

Finally, a refinement step is applied to optimize simultaneously the reprojection error over the set of camera parameters. The Levenberg-Marquardt non-linear optimization method [11] is used for this.

3.3. Multibody Mechanism Semi-Automatic Fitting

Once the camera calibration is obtained for a given key-frame, we manually set the 2D relevant points of each player. We then take the body dimensions as parameters to be estimated through Levenberg-Marquardt algorithm [11]. The error measurement at each iteration is calculated from the distances of the body joint projections with respect to their corresponding manually marked 2D key-points. For that, at each iteration of the optimization process:

- We set the kinematical structure in the neutral configuration (standing pose).
- We then assume the body trunk to be a rigid body and fit only its 3D model to its corresponding points (pelvis, head and shoulders) through EPnP procedure [12]. From this step we apply the estimated position and orientation of the trunk to the pelvis joint of the kinematical structure.
- Afterwards, we proceed to fit the four body limbs. For that, we infer the depth of the rest of 2D key-points, taking as reference those derived from the trunk posing. We use these depth values and constrained IK, in order to control the posing parameters of the limbs.
- Finally, we calculate the projections of the 3D key points and measure the error with respect to their corresponding manually marked 2D key-points. If the error is beyond the considered maximum value for convergence and the number of iterations is not beyond a considered limit, we continue iterating.

Once this procedure has finished, the user can refine the results through a classical 3D posing scheme, by varying the positions of the 3D key-points, obtaining poses estimated through constrained IK, and also through Forward-Kinematics, if necessary. A free-viewpoint camera can also be used in this step as a complement to refine the depth variations with respect to the video-camera viewpoint.

4. EXPERIMENTAL RESULTS

Figure 6 shows some reconstruction examples obtained with our approach from TV footages of traditional European sports, such as Hurling and Gaelic Football (from Ireland) and Handball and Jai Alai (from the Basque Country region, part located in Spain and part in France).

It can be observed how the obtained overlapping between the 3D model projections and their corresponding image regions have a visually acceptable quality. In a posterior stage, one could analyze the body joint motions, including points of view different from those of the video-camera.

During the manual setting of the 2D key-points, we have observed that the correct placement of the floor's reference rectangle is determinant for a good quality of the multibody

mechanism fitting result. The better we place its four 2D corner points and the rectangular sizes with respect to the real world reference, the less interactions will be required to refine the body poses obtained automatically with the optimization algorithm explained in subsection 3.3. The main reason for this is that floor placement discrepancies with respect to real world result mainly in depth discrepancies with respect to the video-camera viewpoint. Depending on the differences between the player locations with respect to the floor, the floor penetration avoidance procedure can lead to different posing results. In order to decrease this discrepancy, we recommend to check that the height metrics observable in the camera calibration step have reasonable dimensions with respect to the expected player heights (see Figure 7).

Both, the camera calibration and the semi-automatic multibody mechanism fitting procedures, assisted by constrained IK, obtain plausible initial poses with respect to the video-camera viewpoint, thus requiring less interaction for a further refinement, when compared to the direct manual posing in 3D from the beginning. This initialization is especially helpful for users that are not used to control the virtual camera and human body poses, with respect to the video-camera viewpoint, directly in 3D.

Fig. 6. Examples of obtained results on TV footages of Hurling, Gaelic Football, Handball and Jai Alai, the first two sports originally coming from Ireland and the next two from the Basque Country region (Spain/France).

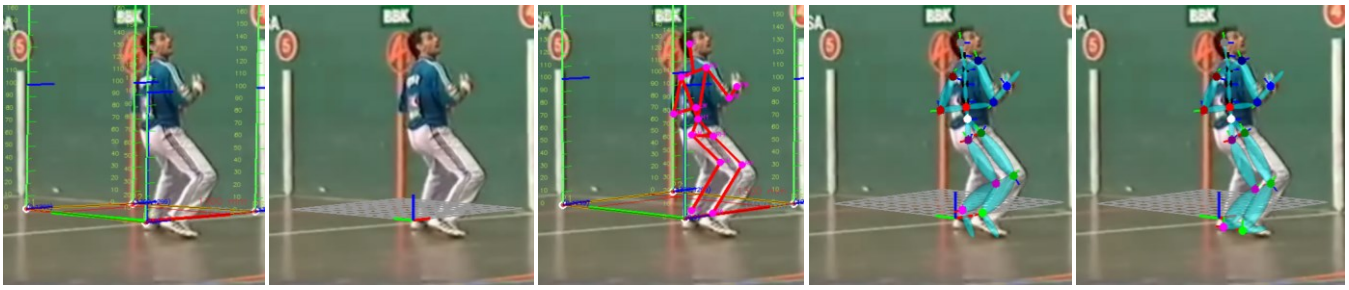

Fig. 7. From left to right, (1) the 2D manual setting the four floor corner points with the resulting perpendicular lines derived from the calibration, (2) the resulting 3D floor model fitted on the image, (3) the 2D manual setting of the player's key-points, (4) the resulting 3D model fitted on the image and (5) the refinement of the feet-floor contact.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a semi-automatic method for 3D reconstruction of sports players from TV footages, which can include camera motion, several players, considerable partial occlusions, motion blur and image noise, without camera calibration information available. Our method can estimate the camera calibration parameters, the positions and poses of the players and their body part dimensions, requiring less manual intervention from the user, when compared to other alternatives. Experimental results show that this technique obtains reconstructions that can help to analyze techniques of past players and the evolution of sports through time for Intangible Cultural Heritage preservation and promotion.

In case of multi-camera recordings, different points of view can help to constrain the motions to be captured. More constraints can also be added from the semantic point of view, relating the achievable poses with those expected with the observed specific action. In the future, we plan to study the accuracy differences of our approach in the monocular, the multi-camera and the semantically-constrained cases with respect to other motion capture systems.

6. ACKNOWLEDGEMENTS

We would like to thank Javier Barandiaran and Marcos Nieto from Vicomtech-IK4 for their explanations and technical support in camera calibration procedures.

REFERENCES

[1] M.L. Stefano, P. Davis, and G. Corsane, Eds., *Safeguarding Intangible Cultural Heritage: Touching the Intangible*, Boydell & Brewer, 2012.

[2] T.B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, *Visual Analysis of Humans: Looking at People*, Springer, 2011.

[3] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D Pose Estimation and Tracking by Detection", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 623-630, 2010.

[4] P. Agarwal, S. Kumar, J. Ryde, J.J. Corso, and V.N. Krovi, "An Optimization Based Framework for Human Pose Estimation in Monocular Videos," *Proc. International Symposium on Visual Computing*, Rethymnon, Crete, Greece, 2012, Part I, LNCS 7431, pp. 575-586.

[5] X. Wei, and J. Chai, "VideoMocap: Modeling Physically Realistic Human Motion from Monocular Video Sequences," *ACM Transactions on Graphics (Proc. SIGGRAPH 2010)*, vol. 29, no. 4, 2010.

[6] M. Fastovets, J.-Y. Guillemaut and A. Hilton, "Athlete Pose Estimation from Monocular TV Sports Footage", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1048-1054, 2013.

[7] K. Shoemake, "Animating Rotation with Quaternion Curves," *Newsletter ACM SIGGRAPH Computer Graphics (Proc. SIGGRAPH 1985)*, vol. 19, no. 3, 1985.

[8] L. Unzueta, M. Peinado, R. Boulic, and Á. Suescun. "Full-Body Performance Animation with Sequential Inverse Kinematics," *Graphical Models*, vol. 70, pp. 87-104, 2008.

[9] M. Nieto, J.D. Ortega, A. Cortes, and S. Gaines, "Perspective Multiscale Detection and Tracking of Persons," in *Proc. International Conference on MultiMedia Modeling (MMM 2014)*, Dublin, Ireland, 2014, Part II, LNCS 8326, pp. 92-103.

[10] R. Hartley, and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2003.

[11] P.R. Gill, W. Murray; and M.H. Wright, "The Levenberg-Marquardt Method," *Practical Optimization*, Emerald Group Publishing Limited, pp. 136-137, 1982.

[12] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An Accurate O(n) Solution to the PnP Problem," *International Journal Of Computer Vision*, vol. 81, pp. 155-166, 2009.