Capturing the sporting heroes of our past by extracting 3D movements from legacy video content

Jon Goenetxea[1], Luis Unzueta[1], Maria Teresa Linaza[1], Mikel Rodriguez[1],
Noel O'Connor[2], Kieran Moran[3]

[1] Fundación Centro de Tecnologías de Interacción Visual y Comunicaciones – Vicomtech-IK4, Spain
[2] Insight Centre for Data Analytics, Dublin City University, Ireland
[3] Applied Sports Performance Research Group, School of Health and Human Performance, Dublin City University, Ireland
{jgoenetxea, lunzueta, mtlinaza, mrodriguez}@vicomtech.org
{noel.oconnor, kieran.moran}@dcu.ie

Abstract Sports are a key part of cultural identity, and it is necessary to preserve them as important intangible Cultural Heritage, especially the human motion techniques specific to individual sports. In this paper we present a method for extracting 3D athlete motion from video broadcast sources, providing an important tool for preserving the heritage represented by these movements. Broadcast videos include camera motion, multiple player interaction, occlusions and noise, presenting significant challenges to solve the reconstruction. The approach requires initial definition of some key-frames and setting of 2D key-points in those frames manually. Thereafter an automatic process estimates the poses and positions of the players in the key-frames, and in the frames between key-frames, taking into account collisions with the environment and human kinematic constraints. Initial results are extremely promising and this data could be used to analyze the sport's evolution over time, or even to generate animations for interactive applications.

Keywords: Motion capture, human body posing, intangible cultural heritage, video legacy.

1 Introduction

While the specific role of sport in society has been debated for many decades, it is widely accepted that it is an important part of human and social development [1]. It can contribute to social cohesion, tolerance and integration and is an effective channel for physical and socio-economic development [2]. As a universal language, sport can be a powerful medium for social and economic change: it can be utilized to bridge cultural gaps; solve conflicts and educate people in ways that very few other activities can.

The ability of sport to shape society is still evident today and in our recent history where it has been used openly and actively by many nations to preserve unique social and cultural identities. Traditional Sports and Games (TSG) represent social values that have taken many years to reach equilibrium in their environment. The language, the land and local customs have modelled them into the forms that we know today. Such TSG can form the backbone of a community and many elements of traditional culture (e.g. language, cuisine, dress, music, dance, the arts), so they have to be promoted to foster community spirit, bring people together and generate a sense of pride in a society's cultural roots.

Since the beginning of the modern industrial society, many TSG have been transformed into very codified and regulated sports, often becoming professional spectacles, or have been lost following the domination of these codified sports (e.g. soccer, tennis, volleyball). Thousands of TSG worldwide were taken out of their place or lost, along with the rich Heritage they represent. Therefore, initiatives are required to both engage individuals in their TSG, thereby increasing levels of participation, and record accurately the techniques used in the recent past to help preserving knowledge of these sports and games.

Ideally, sports actions should be captured in 3D in order to best represent their complexity, although this usually involves specialised cameras and environmental set ups (see Section 3). However, significant amounts of 2D digital content related to sports are available, mainly from broadcast sources. These video legacy archives store many examples of the skills of past and current games. This paper describes an approach for 3D human body pose reconstruction from TSG video legacy recorded with non-calibrated monocular cameras in order to recover the main skills of representative TSG players (National Heroes). This reconstruction generates an animated 3D skeleton, representing the movements of the player and poses in a 3D environment. Once the skills of the players are extracted to an independent 3D environment, such 3D data can be used to generate digital virtual content to be shown in virtual museums, virtual immersive systems or videogames. This method is not limited to video sequences but it is also applicable to single images.

The paper is organized as follows. Section 2 gives an introduction to the related work in this matter. Section 3 explains the method proposed to extract the skills of a player from a video sequence. In section 4, the algorithm used to extract the 3D human pose from each image is explained. Section 5 shows the experimental results for monocular TV sports footage. Finally, in section 6 we discuss the obtained results and the future work.

2 Related Work

Motion capture techniques available in the literature can be classified in two different groups: marker based motion capture and markerless motion capture. The former approaches are based on the tracking of certain markers located over the tracked subject. The tracking data is captured using various infrared cameras, located around the tracked subject. In most of the cases, the used cameras have to be calibrated and synchronized. Although such approaches are very accurate, they are not applicable for broadcast videos. In this scenario, the tracked subject does not have any kind of marker attached during the recording of the video.

Secondly, most of the markerless approaches are based on synchronized, multi-view image sources. For instance, [1] describes a multi-view approach for markerless full body tracking. Even though the accuracy and good results of this solution, it also requires good quality images and complex initialization steps.

In the case of broadcast video analysis, the input data is always a monocular video and the recorded players are wearing neither specific markers nor clothes with special markers for motion analysis. The extraction of full body movements from a monocular video is a specially challenging issue or/and task [4], [5] and [6], as depth data cannot be directly measured from a monocular image. Also, the proportions of the human body vary largely from one individual to another, so those proportions could not be measured accurately. Some approaches like [7] rely on statistical body models in order to estimate the depth for the different poses. Other approaches like [8] and [9] are constrained to a 2D movement tracking. Even if these approaches have a good motion tracking results, they cannot extract 3D motion from the video sequence.

The approach proposed in [10] does not need prior learning or to previously set up key-points. However, it is constrained by the need for static cameras (e.g. no panning or zooming) and good image quality to make a background extraction. The image quality in broadcast videos is not always good, and the camera is not static in most of them, so it is especially challenging scenario.

In our recent work [11] we presented 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 that work, the conversion from 2D to 3D is processed in each key-frame separately, leveraging the estimations from key-frame to key-frame. This paper overcomes such approach by including all the considered key-frames in the same processing loop simultaneously and combining multiple cues for the estimation of the in-betweens.

3 Video-based Motion Capture

The proposed motion capture process has two main phases: (1) manual selection and setup of key-frames, and (2) automatic calculation and reconstruction of movements between key-frames. First, some key-frames are manually selected from the video sequence (for example, the first and the last frames of the sequence). Depending on the length of the video sequence and the complexity of the movements, more key-frames can be selected in between. For each key-frame, the floor definition and the location of the player must be set up.

A reference element is needed to define the floor. This element has to be something in the scene with known sizes. Field marks and billboards are good examples of reference elements. Such element is used to describe a rectangle over the field, defining four reference points over this element. Four markers are located over the reference points (Fig. 1.a). To reconstruct the 3D scene, the distances between the reference points are needed (real world distances).

The location of the player is defined by placing a set of 2D markers over the main body joints (head, shoulders, elbows, wrists, pelvis, hips, knees and ankles). A 2D human structure is used to define the location of the body joints in the image (Fig. 1.c). The markers are used as the reference value in the reconstruction process (Fig. 1.d).

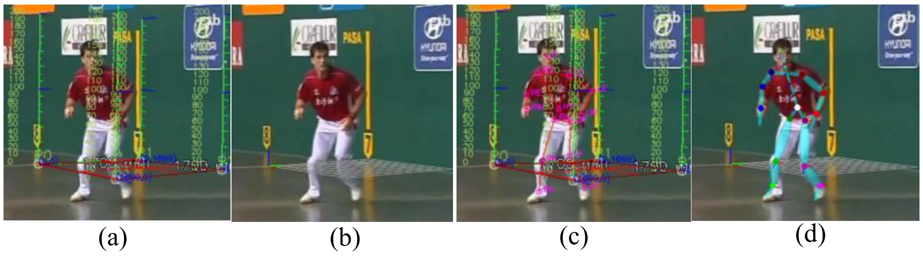

Fig. 1 (a) Definition of the 2D markers of the floor; (b) Resulting 3D reconstruction of the floor; (c), Definition of the main joint locations of the player using the 2D joint marker structure; (d) Resulting floor and player 3D reconstructions.

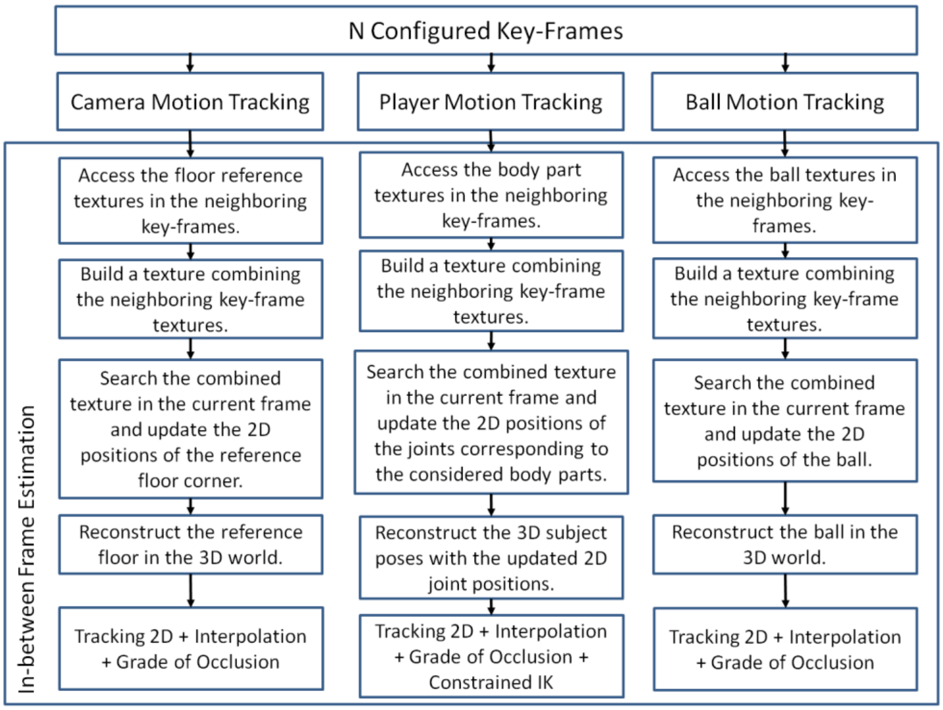
N Configured Key-Frames
Camera Motion Tracking
Player Motion Tracking
Ball Motion Tracking
In-between Frame Estimation
Access the floor reference textures in the neighboring key-frames.
Build a texture combining the neighboring key-frame textures.
Search the combined texture in the current frame and update the 2D positions of the reference floor corner.
Reconstruct the reference floor in the 3D world.
Tracking 2D + Interpolation + Grade of Occlusion
Access the body part textures in the neighboring key-frames.
Build a texture combining the neighboring key-frame textures.
Search the combined texture in the current frame and update the 2D positions of the joints corresponding to the considered body parts.
Reconstruct the 3D subject poses with the updated 2D joint positions.
Tracking 2D + Interpolation + Grade of Occlusion + Constrained IK
Access the ball textures in the neighboring key-frames.
Build a texture combining the neighboring key-frame textures.
Search the combined texture in the current frame and update the 2D positions of the ball.
Reconstruct the ball in the 3D world.
Tracking 2D + Interpolation + Grade of Occlusion

Fig. 2 Diagram showing the element reconstruction process. The reconstruction of the player is segmented in six body parts: left arm (left upper arm and forearm), right arm (right upper arm and forearm), left leg (left thigh and shank), right leg (right thigh and shank), trunk (the graphical models from pelvis to head, ignoring the graphic of the head) and head.

Secondly, the 2D positions of the scene elements (camera, ball and player) are computed in the key-frame in-betweens. The motion of the camera, the pose of the player and the trajectory of the ball are estimated for each frame using those 2D positions as input data for the 3D reconstruction. To estimate the 2D position of the scene elements in the key-frame in-betweens, we propose using a combination of different cues (the texture-based tracking of visible body parts and objects, and the 2D projections of the interpolation of their 3D positions and orientations from key-frame to key-frame), which are weighted in accordance to the grade of occlusion. In the case of the player reconstructions we also include constrained Inverse Kinematics (IK) in this process (Fig. 2).

4 Reconstruction of the 3D Poses

The 3D reconstruction algorithm uses a 3D kinematical structure (Fig. 3) to extract the 3D pose of the player in each frame. The pose is computed fitting the joint projections of the kinematical structure in the camera plane with the configured 2D locations of the joints. Fig. 4 shows the process followed to compute the 3D human body posing. The manually selected key-frames including floor configuration and the location of the player are used as input data.

First, the parameters of the camera are computed using the floor definition from the manually configured key-frames. Such parameters are needed to re-project the kinematical structure into the camera plane. The camera intrinsic (focal length and principal point) and extrinsic (rotation and translation) parameters are computed using the homography between the image and the plane as shown in [12]. The homography is generated using a Direct Linear Transformation (DLT) as proposed in [13]. The lens distortion is not taken into account, as the distortion of the lens can be considered negligible in broadcast videos.

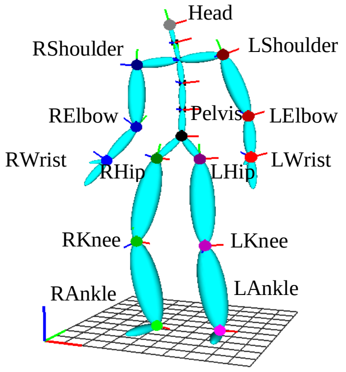

Fig. 3 3D representation of the kinematical structure with the posing features.

Secondly, the multi-body mechanism fitting process fits the 3D human kinematical structure with the 2D human joint definitions, defining the location of the body joints in the 3D space. The posing features of the model (circles) are used to change the pose of the structure. Such features correspond to the main body joints as head, shoulders, elbows, wrists, pelvis, hips, knees and ankles. The orientations of the body joints are constrained to avoid non-feasible poses.

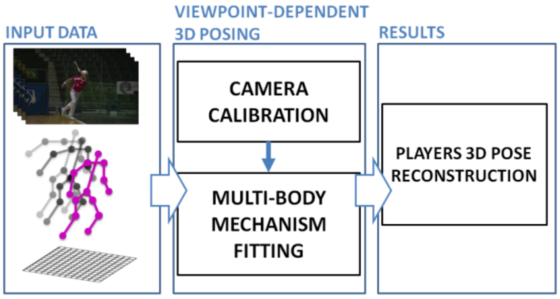
INPUT DATA
VIEWPOINT-DEPENDENT 3D POSING
RESULTS
CAMERA CALIBRATION
MULTI-BODY MECHANISM FITTING
PLAYERS 3D POSE RECONSTRUCTION

Fig. 4 The reconstruction process schema.

Pose changes are computed using the IK procedure proposed in [14]. In this approach, five kinematic chains are defined, all of them including the body segments linked with the involved posing features. The chains are shown in Fig. 5: (1) pelvis, hips and head posing features affecting also spine segments; (2) left hip, knee and ankle posing features; (3) left shoulder, elbow and wrist posing features; (4) right hip, knee and ankle posing features; and (5) right shoulder, elbow and wrist posing features.

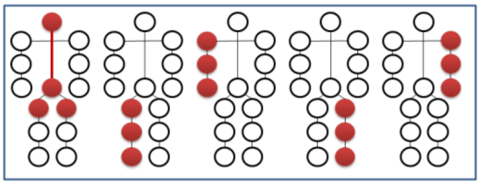

Fig. 5 Kinematical chains defined for the kinematical structure. From left to right: 1) pelvis, hips and head defining the trunk, 2) left hip, knee and ankle defining left leg, 3) left shoulder, elbow and wrist defining left arm, 4) right hip, knee and ankle defining right leg, 5) right shoulder, elbow and wrist defining right arm.

Each of the posing features moves the body segments of the kinematic chain in a different way. The movement of the joints is constrained to limit their mobility to a range according to the corresponding movement of the human body.

✓ The whole body is moved as a rigid body moving the posing feature of the pelvis.

- ✓ The spine can be controlled using the posing feature of the head.
- ✓ The upper limbs are controlled as a rigid body using the shoulder posing feature. This control rotates the limb using the clavicle segment as the rotation reference.
- ✓ The upper and lower limbs can be controlled as articulated limbs using the posing features of the wrist or ankle.
- ✓ The intermediate posing features (elbows and knees) control the swivel angles of the involved limb.

Moreover, the floor configuration avoids the penetration of any element under the plane of the floor. If this happens, the IK approach can correct the position of the involved element chain.

The sizes of the body segments of the player are unknown, so the fitting process must estimate such sizes using N key-frames and their configuration. This process is done iteratively using the Levenberg-Marquardt (LM) algorithm [15] (Fig. 6).

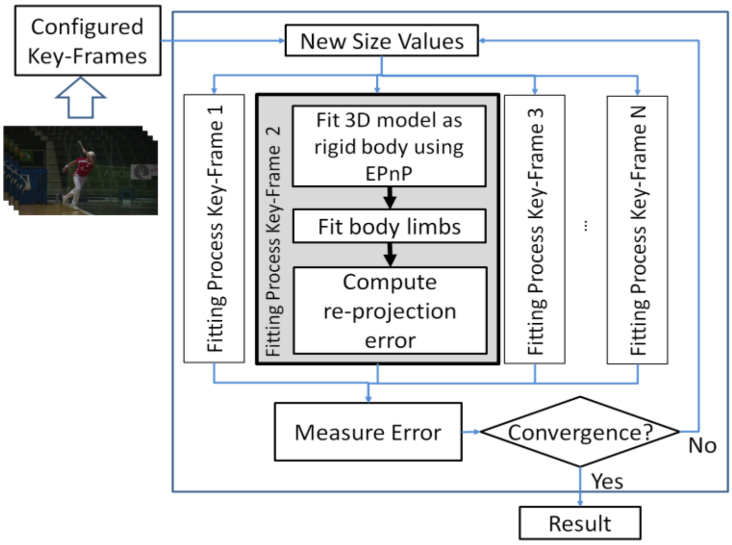
Configured Key-Frames
New Size Values
Fitting Process Key-Frame 1
Fitting Process Key-Frame 2
Fit 3D model as rigid body using EPnP
Fit body limbs
Compute re-projection error
Fitting Process Key-Frame 3
...
Fitting Process Key-Frame N
Measure Error
Convergence?
No
Yes
Result

Fig. 6 Levenberg-Marquardt implementation taking the configured key-frames as input data, and giving the estimated body measures for the tracked player.

For each iteration, LM suggests a set of size values for the body segments to resize the 3D kinematical structure. Then, the 3D structure is fitted following a three step procedure:

- The pelvis, head and shoulders of the model are fitted with the configured 2D markers for the pelvis, head and shoulders on the basis of the Efficient Perspective-n–Point (EPnP) algorithm [16]. As a result, the whole kinematical model is moved as a rigid object.
- The limbs are fitted to their corresponding 2D markers. As the depth value is not measurable, the end-point posing features (wrists and ankles) are moved

with a constant depth. The kinematical constrains relocates the rest of the posing features.

- The re-projection error is measured re-projecting the kinematical structure into the camera plane and comparing the new projected points with the previously configured 2D markers.

This fitting process is made for all the key-frames, and the re-projection error of all of them is combined to compute a global measurement error. The iteration finishes when the global error is less than a predetermined threshold, or the number of iterations is bigger than a predetermined number.

To refine the result of the automatic fitting procedure, the model pose could be further relocated moving the posing features manually. This interaction could be made using forward or inverse kinematics. This relocation takes into account kinematical constraints and floor plane collisions.

5 Experimental Results

Fig. 7 shows some reconstruction examples using the proposed approach. As it can be seen, the re-projection of the reconstruction over the original image has a visually acceptable quality.

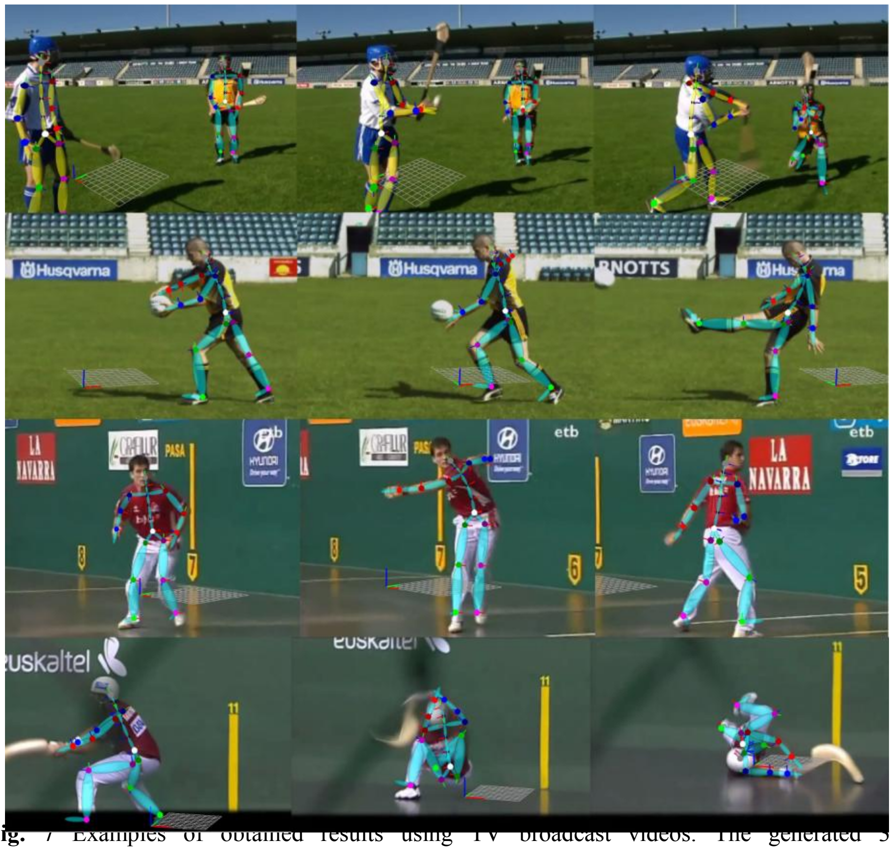

Fig. 7 Examples of obtained results using TV broadcast videos. The generated 3D reconstruction is overlapping the frame image as a colored skeleton.

In the first video sequence, two Hurling players are recorded and tracked during a *frontal block*. In this movement, both players are doing completely different movements (one attacking and the other blocking). It must be mentioned that the player on the left has the feet out of the image, making it more difficult to detect his legs. As shown in the first image sequence, the movement of both players is reconstructed. Each of the players has a different color skeleton.

In the second sequence, a Gaelic Football player is recorded during a *punt kick*. Even though most parts of the body are clearly shown, his right arm is occluded during the movement. However, the results show that the right arm of the player is tracked even with the occluded movements.

The third sequence displays a Basque Pelota player during a *right-handed carom with spin*. This video has wide camera motion, bad image quality and also the body parts of the player are occluded during the spin. The player rotates to the right first, and then to the left during the skill. As shown, the orientation of the reconstructed skeleton matches the orientation of the player in spite of the occlusions during the spin and the movement of the camera.

Finally, the last sequence shows a Jai-Alai player during a *backhand rebound* while falling. This time, the image is blurred and the majority of the body is occluded because of the point of view. Although this last video sequence is the most challenging due to the complex movement, the results show how a very complex and highly occluded movement can be reconstructed.

During the experimentation, it was observed that the configuration of the field plane had a significant effect on the body pose reconstruction. The more accurate the field definition, the better the reconstruction, and less refinement changes needed.

The semi-automatic multi-body mechanism fitting, based on constrained IK and the camera calibration procedure, has demonstrated that plausible initial poses can be reconstructed from the camera view. Using this initial pose, the refinement step needs less interaction than a full manual 3D pose configuration done from the beginning.

6 Discussion and Future Work

This paper describes a method to extract human body motion from video broadcast sources related to TSGs. This method provides good results with less interaction than other alternatives.

Broadcast videos have been recorded with un-calibrated cameras, and may include camera motion, multiple player interaction, occlusions and image noise. The proposed approach estimates the configuration parameters of the camera and body part dimensions of the player by an iterative process. The experimental results demonstrates that the 3D reconstruction extracted using this technique could be used to analyze the evolution of movements or techniques over the time, or even to generate animations for interactive applications like video games.

In multi-view recordings, the inclusion of additional views can improve the reconstruction including depth measures. The movement reconstruction can be improved adding specific constraints to the model according to the player's expected movements. In the future, we plan to study the multi-camera and the semantically-constrained cases with respect to other motion capture systems. Also, the next phase includes a validation step to determine the accuracy of the 3D reconstruction by comparing it to a 'gold standard' (e.g. VICON motion analysis system). However, the current results visually indicate the capacity of the system to accurately track human movement.

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

References

1. Schnitzer, M., Stephenson Jr., M., Zanotti, L., Stivachtis, Y.: Theorizing the role of sport for development and peace building, In: Sport in Society 16:5, pp. 595--610, (2013)
2. Tonts, M.:Competitive sport and social capital in rural Australia, In: Journal of Rural Studies Volume 21, Issue 2, pp. 137–-149 (2005)

3. Quah C.K., Ko, M., Ong, A., Seah, H.S., Gagalowicz, A.: Video-Based Motion Capture for Measuring Human Movement, In: Digital Sport for Performance Enhancement and Competitive Evolution: Intelligent Gaming Technologies. IGI Global, Hershey (2009)
4. Hen, Y.W., Paramesran, R.: Single Camera 3D Human Pose Estimation: A Review of Current Techniques, In: International Conference for Technical Postgraduates. (2009)
5. Sminchisescu, C.: 3D Human Motion Analysis in Monocular Video Techniques and Challenges, In: IEEE International Conference on Video and Signal Based Surveillance, (2006)
6. Brubaker, M.A., Sigal, F., Fleet, D.J.: Video-Based People Tracking. Handbook of Ambient Intelligence and Smart Environments (Part II), pp. 57--87 (2010)
7. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating Human Shape and Pose from a Single Image, In: IEEE 12th International Conference on Computer Vision (2009)
8. Fastoverts, M., Guillemaut J.Y., Hilton, A.: Athlete Pose Estimation from Monocular TV Sports Footage. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1048--1054 (2013)
9. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking People by Learning Their Appearance. Pattern Analysis and Machine Intelligence 29, pp. 65--81 (2007)
10. Agarwal, P., Kumar, S., Ryde, J., Corso, J.J., Krovi, V.N.: An Optimization Based Framework for Human Pose Estimation in Monocular Videos. In: Bebis, G., Boyle, R., Parvin, B., Advances in Visual Computing. LNCS vol. 7431, pp. 575—586
11. Unzueta, L., Goenetxea, J., Rodriguez, M., Linaza, M.: Viewpoint-Dependent 3D Human Body Posing for Sports Legacy Recovery from Images and Video. In: European Signal and Processing Conference, accepted for publication (2014).
12. Nieto, M., Ortega, J.D., Cortes, A., and Gaines, S.: Perspective Multiscale Detection and Tracking of Persons, In: International Conference on MultiMedia Modeling, Dublin, Ireland, 2014, Part II, LNCS 8326, pp. 92-103
13. Hartley, R., and Zisserman, A.: Multiple View Geometry in Computer Vision, In: Cambridge University Press, (2003)
14. Unzueta, L., Peinado, M., Boulic, R., and Suescun, A.: Full-Body Performance Animation with Sequential Inverse Kinematics, In: Graphical Models, vol. 70, pp. 87--104, (2008)
15. Gill, P.R., Murray, W., and Wright, M.H.: The Levenberg-Marquardt Method, In: Practical Optimization, Emerald Group Publishing Limited, pp. 136-137, (1982)
16. Lepetit, V., Moreno-Noguer, F., and Fua, P.: EPnP: An Accurate O(n) Solution to the PnP Problem, In: International Journal Of Computer Vision, vol. 81, pp. 155--166, (2009)