Joint Camera Pose Estimation and 3D Human Pose Estimation in a Multi-Camera Setup

Jens Puwein[1], Luca Ballan[1], Remo Ziegler[2] and Marc Pollefeys[1]

[1]Department of Computer Science, ETH Zurich, Switzerland
[2]Vizrt

Abstract. In this paper we propose an approach to jointly perform camera pose estimation and human pose estimation from videos recorded by a set of cameras separated by wide baselines. Multi-camera pose estimation is very challenging in case of wide baselines or in general when patch-based feature correspondences are difficult to establish across images.
For this reason, we propose to exploit the motion of an articulated structure in the scene, such as a human, to relate these cameras. More precisely, we first run a part-based human pose estimation for each camera and each frame independently. Correctly detected joints are then used to compute an initial estimate of the epipolar geometry between pairs of cameras. In a combined optimization over all the recorded sequences, the multi-camera configuration and the 3D motion of the kinematic structure in the scene are inferred. The optimization accounts for time continuity, part-based detection scores, optical flow, and body part visibility.
Our approach was evaluated on 4 publicly available datasets, evaluating the accuracy of the camera poses and the human poses.

1 Introduction

Camera pose estimation is typically performed by establishing patch-based feature correspondences across images captured by the different cameras [1–4]. This task can be very challenging in case of cameras placed far apart from each other (wide baselines), or, in general, when no reliable correspondences can be found. This is the case, for instance, in Figure 1, where, due to the homogenous background and the wide baselines, it is prohibitive to establish patch-based correspondences. In such scenarios, different features need to be used, namely features incorporating a higher level representation of the scene.

In this paper, we propose to exploit the motion of an actor in the scene to establish correspondences between static intrinsically calibrated cameras. Subsequently, the extrinsic parameters of the cameras and the pose of the actor at each time instant are inferred jointly based on image measurements. The goal is to find the camera disposition and the motion of a kinematic structure inside the scene which best explain the measured optical flow and the probabilities of each joint to be in specific locations in the images.

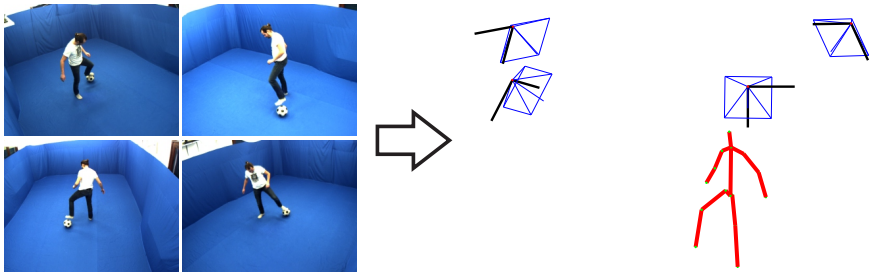

Fig. 1. Given the videos recorded by a set of fixed wide-baseline cameras, our approach recovers the extrinsic parameters of each camera in the scene together with the 3D pose of the moving person at each time instant.

Recent advances in human pose estimation allow for inference of the human pose even from a single image, without having to rely on any kind of foreground/background segmentation of the scene [5]. While these methods work very well for poses that are common in the training set, they often have shortcomings for others. Nevertheless, the results of these techniques can be leveraged to generate the high level correspondences necessary to provide an initial calibration of a static multi-camera setup. In this paper, we propose a method to identify the correctly detected joint positions in each frame of each camera. We then apply a standard structure-from-motion pipeline to these correspondences, taking pose ambiguities into account. Once an initial calibration and initial 3D joint positions are found, the 3D positions of the remaining joints are estimated using optical flow. The camera calibration and the 3D human poses are further optimized jointly leveraging the characteristic properties of multi-view videos, namely smooth 3D trajectories, consistency of 2D joint movements with respect to optical flow, and consistency with respect to discriminative scores of 2D joints. From the initialization to the final optimization, the method goes from single image 2D human pose estimation to the full joint estimation of 3D human poses and camera parameters in a multi-camera setup. Building only on a very general 2D human pose estimation approach, and starting with an extrinsically uncalibrated multi-camera sequence of a moving person, the camera calibration and the full 3D human poses at each time instant are computed.

2 Related Work

Human pose estimation has been tackled in various settings and with varying degrees of accuracy. At one end of the spectrum, there are the 2D human pose estimation approaches which aim at recovering the 2D position and orientation of the limbs or the positions of the joints of a human body from a single image [6, 5]. These approaches first compute the probabilities of each limb or joint being at a particular position, orientation, and scale in the image. Subsequently, a kinematic structure is fit on top of these observations to maximize a posterior probability.

When multiple images of the same scene and at the same time instance are available, some methods infer the full 3D pose of the articulated object, provided that the intrinsic and extrinsic parameters of each input image are known [7–9].

At the other end of the spectrum, when video content is available, time continuity is leveraged to resolve pose ambiguities generated by missing observations or occlusions in single cameras [10–12] or multiple cameras [13–17]. These methods rely on a known pose of the actor in the first frame of the sequence to carry out tracking for all the subsequent frames.

Pose estimation in uncalibrated multi-camera setups has also been explored in the past. However, methods dealing with such a problem typically rely on structure-from-motion techniques, which are first applied to the input videos in order to recover the camera locations and orientations at each time instance. A 3D human pose tracking approach is then employed to recover the motion of the actor in the scene [18].

Structure-from-motion is in fact the standard approach to infer the camera calibration parameters from images. It is typically based on establishing patch-based correspondences between images taken from different cameras. When this kind of correspondences cannot be established, like in case of wide-baseline cameras, higher level features, such as people and object trajectories, have been used. For instance, walking people can be treated as moving vertical poles of constant height, and their motion trajectories are used to calibrate the cameras [19–21]. The main restriction of these methods resides in the assumption that each camera is capturing upright, walking people. This is too restrictive in a more general setting. In contrast, several existing methods match people trajectories between multiple views and use this additional information for camera calibration [22–26].

In this work, we propose to do something similar, but instead of using only the position of a person, we exploit the location of each body part, generating a higher number of reliable correspondences and a larger spread in the images.

Sinha and Pollefeys [27] propose to calibrate a camera network using silhouette frontier points by sampling epipoles in each pair of images. However, this method requires accurate segmentations of the actor.

Izo and Grimson [28] propose to perform camera calibration by matching silhouettes of a persons walking cycles across views. At every frame, silhouettes are compared to example silhouettes that are coupled to camera parameters. The final sequence is obtained by combining per frame observations in a Hidden Markov Model (HMM). This method also requires accurate segmentations and it is moreover restricted to specific motions of the actor, such as a walk.

Recently, Ye et al. presented a 3D human pose and camera calibration tracking approach using three kinect sensors [29]. Manual initialization of the camera poses and the human pose is necessary. Subsequently, camera poses and human models are optimized jointly using an iterative procedure.

Our approach can deal with very wide baselines, it does not rely on segmentation, it does not depend on manual initialization, and it does not require the scene to have textured regions to establish correspondences between images. It finds an initial setup by estimating the 2D poses of the actor in each camera

independently and it tries to find a camera calibration and 3D human poses explaining the image observations.

3 Algorithm

The input to our method consists of a synchronized multi-view video sequence of a moving person. Cameras are assumed to be static, and their intrinsic parameters known a priori. Our goal is to estimate the full 3D pose of the person in the scene at each time instant together with the extrinsic parameters of each camera.

3.1 Initial Calibration

2D human pose estimation is first run on each camera and each frame independently. For this aim, we use the publicly available Matlab code for the Flexible Mixtures-of-parts (FMP) model [5]. FMP models humans as a tree-structured graphical model in which nodes correspond to joints and edges to body limbs. Unary terms model the appearance of each joint by means of HOG descriptors, and pairwise terms model the relative positioning of neighboring joints. Inference on this graphical model can be carried out very efficiently using dynamic programming and the generalized distance transform.

The resulting joint positions provide putative correspondences between cameras, which are then used for calibration. However, 2D human pose estimation usually does not differentiate between front and back facing people, or if it does, it does it very poorly. This is also the case for FMP. Hence, correspondences between symmetric body parts, like the arms and the legs, are ambiguous in the sense that it is not possible to differentiate between the left ones and right ones. To take this into account, both possibilities, front and back facing, have to be considered.

For each camera pair, the two-view geometry is estimated using RANSAC over the candidate joint correspondences [30]. During the sample selection, each view is chosen to be either front or back facing. When counting the inliers in each frame, the direction faced by the person which leads to the highest number of inliers is chosen.

Additionally, in order to avoid unstable configurations of minimal solutions when generating RANSAC hypotheses, correspondences are encouraged to be evenly distributed over the entire images. Therefore, when drawing the samples, sets of points originating from different joints and lying far apart temporally are assigned higher probabilities of being chosen. It is not necessary to consider all joints to establish correspondences. In fact, a wide spread of joint positions is obtained by using the head, the lower end of the spine, the wrists, and the ankles.

Cameras are added to a common world coordinate frame greedily, starting with a bundle adjustment of the camera pair with the most inlier correspondences [31]. Thereafter, in each step, the camera with the most inlier correspon-

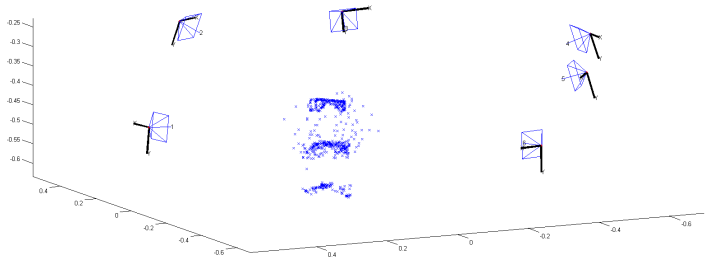

Fig. 2. Camera setup and joint positions used for the calibration of the INRA dancer dataset.

dences to the already included cameras is added, followed by a bundle adjustment. This process is repeated until all cameras are within a common world coordinate frame and refined using bundle adjustment.

The result is an initial estimate of the poses of all cameras in the setup. An example of a camera setup and the joint positions used for its calibration is shown in Figure 2.

3.2 Initial 3D Joint Positions

Body joints are triangulated using the initial camera calibration computed in the previous section. A triangulated joint position is considered valid if it can be triangulated by at least 3 cameras with a reprojection error below 5 pixels. In practice, triangulation is performed for each frame and each joint considering all the possible combinations of cameras that could verify that specific joint. The combination with the highest number of agreeing cameras is then kept.

For symmetric joints, like ankles and wrists, care has to be taken. For such pairs of joints, in a first step, the combination of cameras and front facing/back facing of the person that leads to the largest number of cameras verifying the joint is picked greedily. All remaining joint positions are used to potentially verify the second joint that is remaining. This leads to one 3D joint, two 3D joints, or no 3D joint of the same kind (e.g., left ankle and right ankle) per frame. Each joint might be either the left or right joint of the true 3D pose. In order to consistently label the 3D joints as left/right, an arbitrary left/right labeling is chosen for the first frame where both joints appear. This information is then propagated forward and backward through the whole sequence using optical flow.

Using the verified joints as anchors, the missing joint positions throughout the sequence can be inferred using optical flow.

3.3 Joint Optimization

The initial camera calibration and the initial 3D joint positions are refined in a combined optimization step, aiming at finding the correct camera configu-

ration and a consistent kinematic structure evolving over time, explaining the observations.

Let $\boldsymbol{\theta}_c$ denote the extrinsic parameters of camera c, and let $\mathbf{X}_i^t$ denote the 3D coordinates of joint i at time t. $\mathbf{X}_i^t$ and $\boldsymbol{\theta}_c$ are unknowns of the problem.

Since the goal is to find a single kinematic structure for the whole recorded sequence, the length of each body limb needs to be constant over time. To achieve this, an additional set of unknowns is introduced, namely $e_{(i,j)}$, indicating the length of the limb $(i,j) \in \mathcal{E}$ connecting joint i and joint j. Here, $\mathcal{E}$ represents the edges of the kinematic structure to estimate. To enforce constant limb lengths, the kinematic structure has to minimize the following error functional

$$E_{limb}(\mathbf{X}, \mathbf{e}) = \sum_t \sum_{(i,j)\in\mathcal{E}} \left(\|\mathbf{X}_i^t - \mathbf{X}_j^t\| - e_{(i,j)} \right)^2 . \tag{1}$$

To enforce time continuity, a constant velocity model for each joint in the structure is deployed by forcing the second derivative of $\mathbf{X}_i^t$ to be small. Formally, this is expressed by

$$E_{smooth}(\mathbf{X}) = \sum_{t,i} \|\ddot{\mathbf{X}}_i^t\|^2, \tag{2}$$

where $\ddot{\mathbf{X}}_i^t$ is approximated by central finite differences.

Concerning the image observations, both optical flow and part-based detection scores are used. It is assumed that the motion of the kinematic structure is coherent with the measured optical flow in each camera. Moreover the position of each joint in each frame should project to a 2D image position having a high detector score for the corresponding joint. To this aim, optical flow is computed for each video stream, and the detection scores are computed for each joint and each frame in each video.

Optical flow was computed using the OpenCV implementation of the algorithm introduced by Farnebäck [32, 33]. An example is shown in Figure 3. Let $OF_{c,t}(x,y)$ denote the optical flow measured in camera c at time t for a generic pixel (x,y), and let $\pi(\boldsymbol{\theta}_c, \mathbf{X})$ be the projection function mapping 3D points $\mathbf{X}$ to 2D image coordinates in camera c, specified by the calibration parameter $\boldsymbol{\theta}_c$. In order to force the motion of the kinematic structure to be consistent with the measured optical flow in each image, the following functional should be minimized

$$E_{OF}(\mathbf{X}, \boldsymbol{\theta}) = \sum_{c,t,i} \|OF_{c,t}(\pi(\boldsymbol{\theta}_c, \mathbf{X}_i^t)) - (\pi(\boldsymbol{\theta}_c, \mathbf{X}_i^{t+1}) - \pi(\boldsymbol{\theta}_c, \mathbf{X}_i^t))\|^2. \tag{3}$$

Let $Det_{c,t,i}(x,y)$ denote the detection probability for joint i measured in camera c at time t for a generic pixel (x,y). Probabilities $Det_{c,t,i}$ are computed using the sigmoid function and an SVM model trained on the 2D joint locations that were used to initialize the 3D points in Section 3.2 [34]. The feature vectors are constructed by concatenating HOG feature vectors [35] and color histogram feature vectors. HOG features are computed using cell size 8, block size 4 and a block overlap of 3. The color feature vectors are obtained by binning the HSV

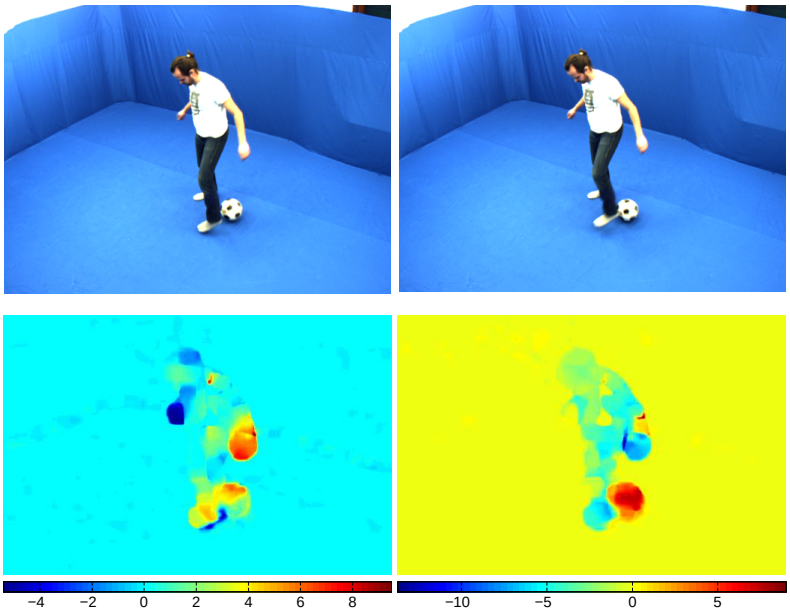

Fig. 3. Top row: two consecutive images from the Soccer Juggling sequence. Bottom row: optical flow in x-direction (left) and y-direction (right). The units of the color coding are given in pixels.

values of 25x25 patches independently into 8 bins per channel. After training, the SVM model is evaluated for all cameras and all images to obtain $Det_{c,t,i}$.

By applying the negative logarithm, probabilities $Det_{c,t,i}$ are transformed to negative log-probabilities. An example of the resulting detection scores for the left ankle is shown in Figure 4. The values obtained by subsequently taking the square root are denoted as $\overline{Det}_{c,t,i}(x,y)$. To enforce the joint positions to be consistent with the trained detector, the following functional needs to be minimized:

$$E_{Det}(\mathbf{X},\boldsymbol{\theta}) = \sum_{c,t,i} \overline{Det}_{c,t,i}(\pi(\boldsymbol{\theta}_c, \mathbf{X}_i^t))^2. \tag{4}$$

In order to guide the optimization and in order not to digress too much from the initial solution, reprojected joints should be close to the joint positions that were used for initialization, if available. More formally,

$$E_{Rep}(\mathbf{X},\boldsymbol{\theta}) = \sum_{c,t,i} \nu_{c,i}^t L_\delta(\|\pi(\boldsymbol{\theta}_c, \mathbf{X}_i^t) - \mathbf{x}_{c,i}^t\|). \tag{5}$$

should be minimized. $\nu_{c,i}^t$ is a binary variable indicating whether joint i in camera c at time t was consistent with multiple cameras and hence used for initialization. To account for outliers, the robust Huber cost function L_δ is used [3]. The threshold δ was set to 5 pixels.

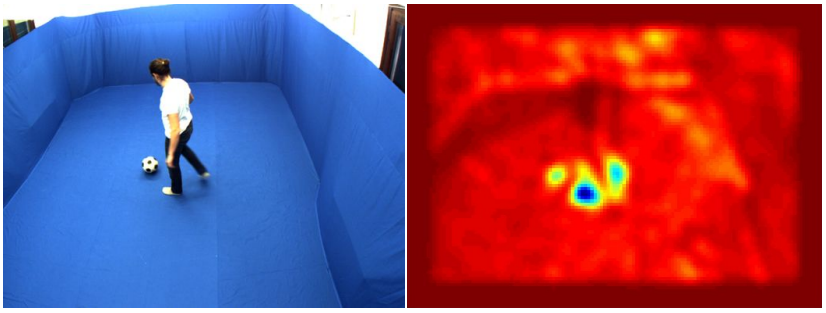

Fig. 4. Input image (left) and detection scores of the left ankle (right).

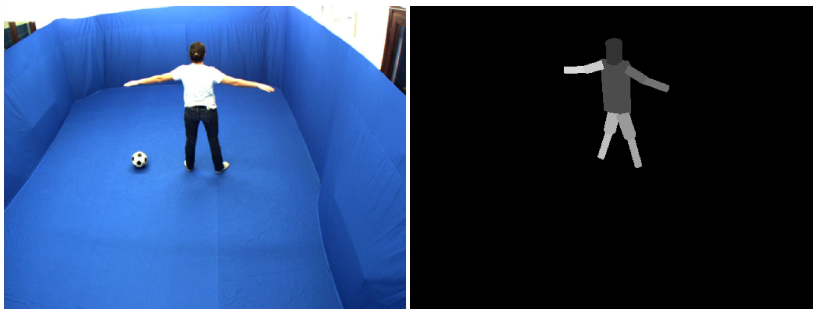

Fig. 5. The rendered 3D model of the human (right) and the corresponding image (left).

The final functional to minimize is a linear combination of the previously defined costs, i.e.,

$$\begin{aligned} E(\mathbf{X}, \mathbf{e}, \boldsymbol{\theta}) = \lambda_1 E_{limb}(\mathbf{X}, \mathbf{e}) + \lambda_2 E_{smooth}(\mathbf{X}) + \lambda_3 E_{OF}(\mathbf{X}, \boldsymbol{\theta}) \\ + \lambda_4 E_{Det}(\mathbf{X}, \boldsymbol{\theta}) + \lambda_5 E_{Rep}(\mathbf{X}, \boldsymbol{\theta}) \end{aligned} \tag{6}$$

where the λ_i are constants defined to balance the influence of each term. For the experiments, the values λ_i were chosen in a grid search on the first 100 frames of the Soccer Juggling sequence [14] and kept constant for all experiments. Since the 3D reconstruction is only given up to scale, the torso of the person is set to a fixed length to ensure that $E(\mathbf{X}, \mathbf{e}, \boldsymbol{\theta})$ is not affected by scale changes of the 3D structure.

The optimization is carried out using the Levenberg-Marquardt algorithm. Taking advantage of the sparse structure of the Jacobian of E makes the optimization much more efficient. To account for occlusions in both Equation 3 and Equation 4, a simple 3D model of a human is used, where every limb is modeled as a cylinder. By rendering the model in all images, it is easy to determine which joints are visible in which frames and in which cameras. Figure 5 shows an example of a rendered cylindrical model. Occluded joints are simply excluded from the sums in Eq. 3 and Eq. 4.

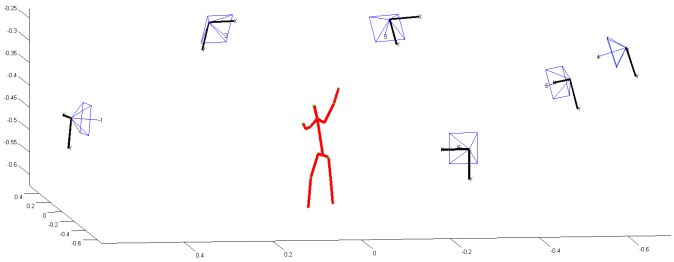

Fig. 6. Camera setup and 3D skeleton estimated for the INRA dancer dataset.

The optimization iterates between the Levenberg-Marquardt algorithm and recomputing the visibility term. Figure 6 shows the final camera setup and an example pose for the INRIA dataset.

4 Results

The presented approach was evaluated on 4 publicly available datasets, namely, the INRIA dancer dataset (201 frames, the first 6 cameras) [36], the HumanEva-II dataset (the first 500 frames, 4 cameras) [37], and the Soccer Juggling sequence (531 frames, 4 cameras) and the Sword Swing sequence (383 frames, 4 cameras), both from Ballan and Cortelazzo [14]. For all these datasets, a camera calibration is provided. The FMP model used in Section 3 was trained on the publicly available LEEDS sports dataset [38]. This model was then used for all experiments without any specific tuning. This shows that the presented method is general and applicable to a wide range of data. A few images from the LEEDS dataset are shown in Figure 7.

The geometric verification of joints using the camera parameters provides a valuable confidence measure for estimated joint positions. Even though the 2D human pose estimates in the individual cameras are noisy and often incorrect, the presented method corrects many errors by using only joint positions verified by the geometry of the camera setup to create an initial guess of 3D joint positions. The final optimization further optimizes joint positions by fixing edge lengths, enforcing smooth motions and consistency with image measurements. A comparison of a few poses obtained from 2D human pose estimation and poses obtained by projecting optimized 3D poses can be found in Figure 8.

Quantitative Evaluation The presented approach was evaluated quantitatively in terms of camera pose estimation error and human pose estimation error. Table 1 illustrates the resulting positional distances between estimated and groundtruth camera centers as well as the angular differences for the relative angles between all pairs of estimated cameras and all pairs of groundtruth cameras, respectively. The initial calibration obtained from Section 3 is compared with the calibration obtained from the final joint optimization of Section 3.

Fig. 7. Example training images of the LEEDS dataset.

Since the presented method returns a 3D reconstruction up to a similarity transformation (rotation, translation and scale), the result needs to be aligned with the groundtruth for comparison. This was done by computing the global similarity transformation minimizing the squared distances between the groundtruth and the estimated camera centers.

Concerning the error in the human pose estimation, both 3D and 2D errors were evaluated for the Soccer Juggling and the Sword Swing dataset. The very good results obtained by Ballan and Cortelazzo were inspected visually and used as groundtruth [14]. In both cases, left/right flips of limbs were ignored during the evaluation. The left arm was switched with the right arm, if the error decreased. The same holds for the legs.

To evaluate the 2D errors, 3D joint positions were projected into the images using the groundtruth camera calibration for the groundtruth 3D joint positions and the estimated camera calibration for the estimated 3D joint positions. To quantify the errors, the PCK measure introduced by Yang and Ramanan [5] was used. The PCK measure qualifies a detection as correct if the distance between the detected position and the groundtruth position is below $\alpha \max(w, h)$. w and h are the width and height of the axis-aligned bounding box containing all groundtruth joints in the respective image. Varying the PCK threshold α corresponds to varying the desired accuracy. PCK scores obtained by using the proposed approach are compared to the ones obtained by using the standard FMP approach [5]. The results for the Soccer Juggling dataset and the Sword Swing dataset are depicted in Figures 9 and 10, respectively. While head positions are estimated accurately in both methods, the errors of the remaining body parts are decreased significantly by the presented method.

The average errors in 3D joint positions for the Soccer Juggling dataset and the Sword Swing dataset are given in Tables 2 and 3, respectively. A plot illustrating the average 3D joint position errors per frame is shown in Figure 11. The aforementioned Tables 2 and 3 and Figure 11 also compare the 3D joint positions obtained after the initialization with the ones obtained after the final optimization, described in Sections 3.2 and 3.3. Especially for the Sword Swing dataset, the final optimization leads to a significant improvement of the 3D joint accuracy.

To evaluate the accuracy on the HumanEva-II dataset, the online evaluation tool has to be used [39]. For the S4 dataset, the walking cycle was evaluated (first 350 frames). The mean error over all joint positions after the final optimization

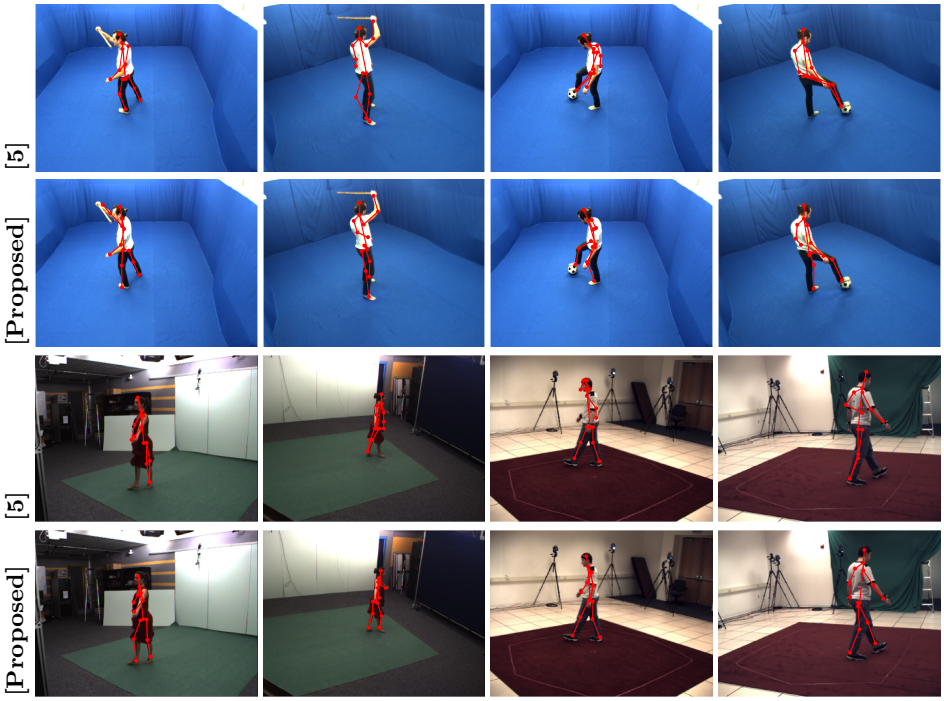

Fig. 8. Comparison with the baseline approach [5].

was 82mm. To the best of knowledge, the state-of-the-art result for the walking cycle was obtained by Gall et al. By tracking a full 3D model of the person an error of 28mm was achieved [40]. A plot showing the average 3D joint position errors per frame is given in Figure 12.

5 Conclusion

In this paper we presented a novel technique to calibrate a multi-camera setup by jointly estimating the extrinsic camera parameters and the 3D poses of a person in the scene, without relying on patch-based feature correspondences. 2D joint positions detected by 2D human pose estimation are used as higher level features to establish putative correspondences between cameras and to bootstrap the joint optimization of camera calibration and 3D poses. The final optimization takes advantage of the 3D articulated structure and temporal continuity and it enforces consistency with image measurements. The experimental evaluation on 4 publicly available datasets investigates the accuracy of the estimated camera poses and the 2D and 3D joint positions, showing the benefit of using the presented joint optimization.

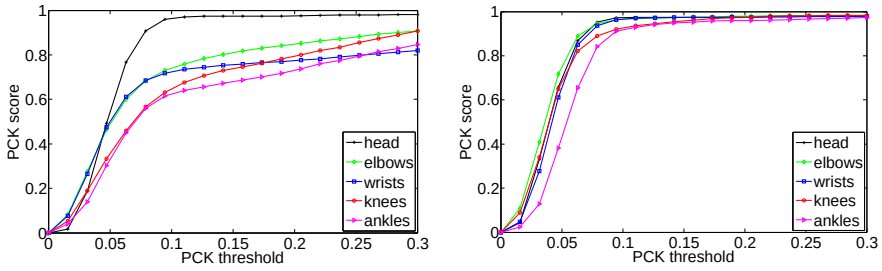

Fig. 9. PCK score obtained using the standard FMP model [5] (left), and the presented approach (right), on the Soccer Juggling dataset.

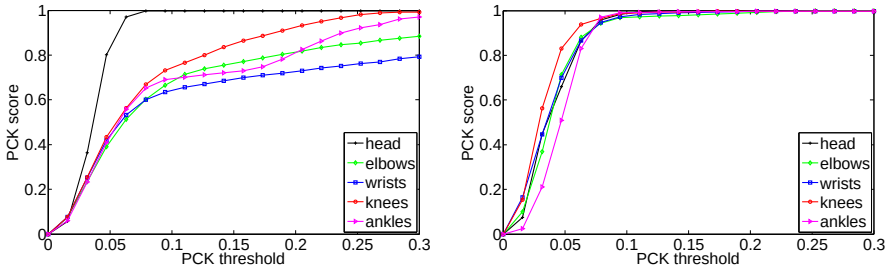

Fig. 10. PCK score obtained using the standard FMP model [5] (left), and the presented approach (right), on the Sword Swing dataset.

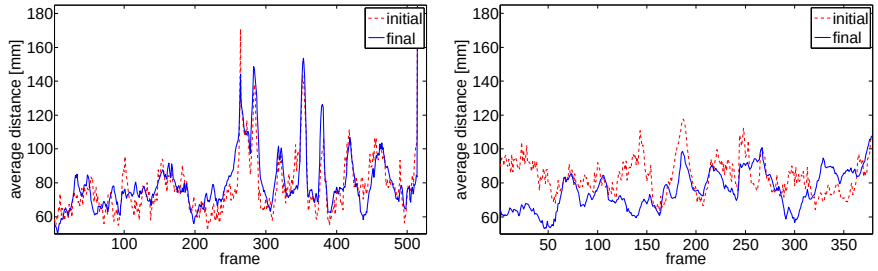

Fig. 11. Average per frame error of estimated 3D joint positions evaluated on the Soccer Juggling dataset (left), and on the Sword Swing dataset (right).

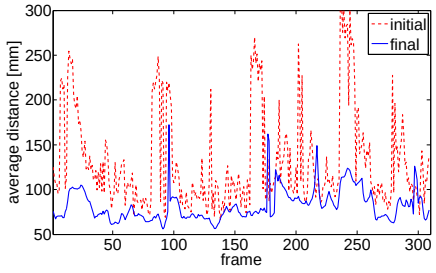

Fig. 12. Average per frame error of estimated 3D joint positions evaluated on the walking cycle of the HumanEva-II S4 dataset.

	Positional error [mm]		Angular error [deg]	
	initial	**final**	initial	**final**
Soccer Juggling	54 ± 11	**50 ± 13**	1.2 ± 0.7	**1.0 ± 0.5**
Sword Swing	71 ± 26	**58 ± 21**	0.9 ± 0.6	**1.0 ± 0.5**
INRIA	55 ± 13	**53 ± 13**	0.7 ± 0.5	**0.4 ± 0.3**
HumanEva-II	20 ± 7	**7 ± 2**	0.3 ± 0.2	**0.3 ± 0.3**

Table 1. Camera pose estimation error: average error ± standard deviation.

	Head	Elbows	Wrists	Knees	Ankles	**Total**
initial	68 ± 116	75 ± 119	87 ± 156	122 ± 129	115 ± 139	94 ± 127
final	**66 ± 115**	**79 ± 117**	**86 ± 154**	**123 ± 114**	**120 ± 144**	**96 ± 124**

Table 2. 3D joint position estimation error: average error ± standard deviation [mm], on the Soccer Juggling dataset.

	Head	Elbows	Wrists	Knees	Ankles	**Total**
initial	70 ± 19	93 ± 43	87 ± 51	69 ± 47	99 ± 43	84 ± 42
final	**34 ± 11**	**68 ± 38**	**64 ± 37**	**71 ± 34**	**94 ± 25**	**76 ± 41**

Table 3. 3D joint position estimation error: average error ± standard deviation [mm], on the Sword Swing dataset.

Acknowledgements. This project is supported by a grant of CTI Switzerland, the 4DVideo ERC Starting Grant Nr. 210806 and the SNF Recording Studio Grant.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding (CVIU) (2008)
3. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
4. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision. Springer (2004)
5. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2013)
6. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
7. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: Proceedings of the British Machine Vision Conference (BMVC). (2013)

8. Burenius, M., Sullivan, J., Carlsson, S.: 3d pictorial structures for multiple view articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
9. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures for multiple human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
10. Ramanan, D., Forsyth, D.A., Zisserman, A.: Strike a pose: Tracking people by finding stylized poses. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2005)
11. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (1998)
12. Salzmann, M., Urtasun, R.: Combining discriminative and generative methods for 3d deformable surface and articulated pose reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
13. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking loose-limbed people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2004)
14. Ballan, L., Cortelazzo, G.M.: Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In: International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT). (2008)
15. Liu, Y., Stoll, C., Gall, J., Seidel, H.P., Theobalt, C.: Markerless motion capture of interacting characters using multi-view image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
16. Ballan, L., Taneja, A., Gall, J., Gool, L.V., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: European Conference on Computer Vision (ECCV). (2012)
17. de La Gorce, M., Fleet, D., Paragios, N.: Model-based 3d hand pose estimation from monocular video. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). (2011)
18. Hasler, N., Rosenhahn, B., Thormhlen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
19. Krahnstoever, N., Mendonca, P.: Bayesian autocalibration for surveillance. In: IEEE International Conference on Computer Vision (ICCV). (2005)
20. Lv, F., Zhao, T., Nevatia, R.: Camera calibration from video of a walking human. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2006)
21. Chen, T., Del Bimbo, A., Pernici, F., Serra, G.: Accurate self-calibration of two cameras by observations of a moving person on a ground plane. In: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS). (2007)
22. Jaynes, C.: Multi-view calibration from planar motion for video surveillance. In: Second IEEE Workshop on Visual Surveillance (VS'99). (1999)
23. Stein, G.P.: Tracking from multiple view points: Self-calibration of space and time. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (1999)
24. Bose, B., Grimson, E.: Ground plane rectification by tracking moving objects. In: IEEE International Workshop on Visual Surveillance and PETS. (2004)
25. Meingast, M., Oh, S., Sastry, S.: Automatic camera network localization using object image tracks. In: IEEE International Conference on Computer Vision (ICCV). (2007)
26. Puwein, J., Ziegler, R., Ballan, L., Pollefeys, M.: PTZ camera network calibration from moving people in sports broadcasts. In: IEEE Workshop on Applications of Computer Vision (WACV). (2012)

27. Sinha, S., Pollefeys, M.: Camera network calibration and synchronization fromsilhouettes in archived video. Int. Journal of Computer Vision (2010)
28. Izo, T., Grimson, W.: Simultaneous pose estimation and camera calibration from multiple views. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). (2004)
29. Ye, G., Liu, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Performance capture of interacting characters with handheld kinects. In: European Conference on Computer Vision (ECCV). (2012)
30. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981)
31. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment - a modern synthesis. Vision algorithms: Theory and Practice (2000)
32. OpenCV. (http://opencv.org/) Accessed: 2014-08-19.
33. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Proceedings of the 13th Scandinavian Conference on Image Analysis. (2003)
34. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. (1999)
35. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2005)
36. Inria: Inria dancer, 4d repository. (http://4drepository.inrialpes.fr/public/datasets) Accessed: 2014-06-17.
37. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated humanmotion. Int. Journal of Computer Vision (2010)
38. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference (BMVC). (2010)
39. HumanEva. (http://vision.cs.brown.edu/humaneva/) Accessed: 2014-08-19.
40. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. Int. Journal of Computer Vision (2010)