2019 International Conference on 3D Vision (3DV)

Motion Capture from Pan-Tilt Cameras with Unknown Orientation

Roman Bachmann[1] Jörg Spörri[2] Pascal Fua[1] Helge Rhodin[1,3]

[1]EPFL, Lausanne, Switzerland
[2]Balgrist University Hospital, University of Zurich, Zurich, Switzerland
[3]The University of British Columbia, Vancouver, Canada

firstname.lastname@epfl.ch

Abstract

In sports, such as alpine skiing, coaches would like to know the speed and various biomechanical variables of their athletes and competitors. Existing methods use either body-worn sensors, which are cumbersome to setup, or manual image annotation, which is time consuming. We propose a method for estimating an athlete's global 3D position and articulated pose using multiple cameras. By contrast to classical markerless motion capture solutions, we allow cameras to rotate freely so that large capture volumes can be covered. In a first step, tight crops around the skier are predicted and fed to a 2D pose estimator network. The 3D pose is then reconstructed using a bundle adjustment method. Key to our solution is the rotation estimation of Pan-Tilt cameras in a joint optimization with the athlete pose and conditioning on relative background motion computed with feature tracking. Furthermore, we created a new alpine skiing dataset and annotated it with 2D pose labels, to overcome shortcomings of existing ones. Our method estimates accurate global 3D poses from images only and provides coaches with an automatic and fast tool for measuring and improving an athlete's performance.

1. Introduction

In many sports, like alpine skiing, coaches would like to know performance metrics such as Center of Mass, speed and various biomechanical variables at every point in time, giving them accurate feedback about potential increases or losses in speed and precision. This can be used to enhance the athlete's performance by comparing athletes and finding optimal motion trajectories. Existing methods, like optical barriers in skiing, only offer average speeds within segments, while other methods using Inertial Measurement Unit and/or GPS sensors [10, 11, 12, 14, 27] are cumbersome to wear. Using motion capture suits is also not feasible in high-speed settings with large capture volumes.

Recent deep-learning-based monocular human pose estimation methods are able to reconstruct articulated 3D pose from moving cameras [22, 23, 28, 41, 30, 26, 35, 29, 43, 40, 38, 42], however, only relative to the camera pose and without accurate scale and depth information [13]. Such relative poses contain no information on the athletes global position and speed, the unquestionably most important metric for racing sports. We therefore have the goal of estimating an athlete's global 3D pose at every point in time using just video frames from multiple cameras arranged around the track. One way to get those poses is to manually annotate every frame. This manual annotation is however very tedious and time-consuming, so instead we chose to train a pose estimation network to predict 2D joint locations without the athletes needing to wear markers. Normally, pose estimation algorithms are only trained on human pose databases which don't feature motions of particular sports. We focus on alpine skiing, where no suitable dataset exists. Existing alpine skiing datasets [37, 10, 34] are very limited in the number of athletes and locations that they feature, making methods trained on them not generalize well. To remedy those problems we created a new alpine skiing dataset, containing 1982 manually annotated frames from various recordings and in diverse weather conditions.

To go from videos to global articulated 3D poses, we propose the following multi-stage approach. Because athletes are often very small in the captured images, we first train a network to predict a tight bounding box around them. Those crops are then given to the pose estimation network that was mainly trained on the new skiing dataset. The 2D detections from all cameras are then combined in a bundle adjustment approach to reconstruct the global 3D pose. Key to our solution was the expression of the athletes's motion in terms of a discrete cosine basis, which enforces smoothness constraints explicitly, and tracking features on the static background as an additional cue for constraining the camera motion. This strategy is closely related to panorama stitching and structure from motion and more general than

2475-7888/19/$31.00 ©2019 IEEE
DOI 10.1109/3DV.2019.00042

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

related methods that utilize known line markings on sports fields [7, 8, 32]. We tried off-the-shelf structure from motion methods, however, these fail on skiing footage due to the large zoom, large distance and view angle between cameras, and lack of discriminative patterns on the ski slope. Moreover, optimizing camera motion freely as in [39] lead to underconstrained systems of equations and diverging behaviour. We evaluate the performance of taking fully calibrated cameras and our method for estimating the camera rotations with varying number of cameras. Our results are significantly more accurate across various biomechanical variables when compared to using the monocular reconstruction of [34].

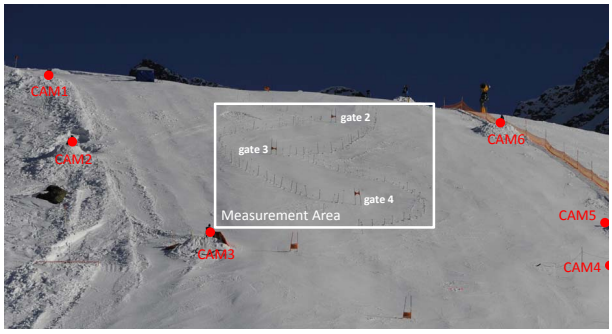


Figure 1: Setup overview of the multi-view skiing dataset.

2. Related work

As this paper builds upon work in the field of 3D human pose estimation, we outline in the following, the most important advances in this area and also explain our need for a new task-specific skiing dataset.

Global 3D human pose estimation. Using at least two cameras from different perspectives, it is possible to obtain a global 3D pose estimate and potential ambiguities in scale can be resolved [24, 25]. It is now common to estimate 2D pose with deep neural networks [6, 18] and infer skeleton pose with model-based optimization [33, 17]. More recently, Pavlakos *et al*. [29] propose to extend pictorial structure models by taking CNN generated 2D heatmaps and resolving the 3D structure in a quantized grid by maximizing a likelihood term explaining the 2D detections. This line of work requires known camera pose and intrinsic parameters.

Puwein *et al*. [31] jointly estimate a 3D human pose and the position and orientation of several fixed wide-baseline cameras using a bundle adjustment method that minimizes an energy function comprising reprojection errors, a smoothness term and optical flow consistency between the motion of the estimated kinematic structure and the videos. Similarly, Elhayek *et al*. [9] estimate both pose and camera locations simultaneously, with the difference that some cameras are fixed, while a small subset can freely move. They minimize an energy function containing a negative likelihood term describing the similarity of the model parameters to the measured data, as well as smoothness terms for both the human pose and the cameras.

More difficult is the reconstruction from moving cameras with totally unknown orientation. Several papers [7, 8, 32] leverage common line markings of sports fields as known reference points for pan-tilt-zoom (PTZ) camera calibration. Those methods can leverage the geometric constraints that games like football are played on a two-dimensional surface with a limited spatial extent, but don't generalize to sports with unconstrained environments, such as ski racing.

Using multiple hand-held and unsynchronized cameras, Hasler *et al*. [16] first construct a global model of the environment, then synchronize the cameras using sound, and finally use a silhouette-based approach to find a 3D mesh of the subject. In our case, reconstruction on very zoomed in, fast moving motion would not yield enough overlap to employ the structure from motion approach described.

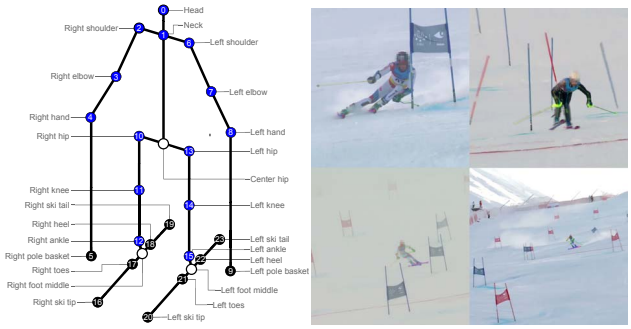
Most closely related to ours is the approach of Takahashi *et al*. [39], using multiple unsynchronized and uncalibrated cameras. They use a bundle adjustment method that leverages the limb lengths as priors on the human body and takes into account that the 2D pose estimations contain some amount of error. However, using their approach to freely optimize camera rotations failed in our large capture volume with very fast moving athletes.

Ski datasets. While there exists extensive datasets for human poses in various settings like the MPII Human Pose [5] dataset or the Human3.6M [19] dataset, they feature only very few, if any, skiing images of amateurs and lack annotation of the skis and poles. Professional athletes in a racing scenario are even more rare, which would make accurate inference impossible in those cases. To train or refine a 2D detector, we have therefore decided to create a new alpine skiing dataset featuring semi-professional athletes for which videos are publicly available.

For the purpose of evaluating 3D pose estimation methods and comparing to related work, we used a manually annotated multi-view (*MV-Ski*) pan-tilt-zoom alpine skiing dataset [37, 10, 34]. It features 6 professional athletes on a Giant Slalom slope with three turns, filmed by six cameras that are arranged in a circle around the center of the track as shown in Figure 1. 2D joint locations were manually annotated. Calibration points around the track served to calculate the camera parameters, specifically the intrinsic and extrinsic camera matrices. From this, global ground truth 3D poses were triangulated.

While the *MV-Ski* dataset is well suited for developing semi-supervised models [34], the fact that it only features 6 athletes in similar suits performing the run on the same slope with the same camera angles makes methods trained on it unable to generalize to different skiing settings.

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.



(a) Skier joints (b) Example images

Figure 2: Figure a shows the 24 annotated joints, with the subset of body joints marked in blue. The white circles are non-annotated helper joints. Figure b shows four annotated images of the new dataset featuring various conditions.

3. New single-view alpine skiing dataset

To facilitate generalization of marker-less ski motion capture to new environments, we create a large single view (*SV-Ski*) dataset for alpine skiing. We downloaded 16 alpine skiing videos that were posted on Youtube under the Creative Commons license, featuring mainly semi-professional ski racers from many different perspectives in various weather conditions. Those videos were split into 147 training and 11 validation sequences of various lengths, from which frames were sampled in fixed intervals ranging from 0.3 to 10 seconds, depending on the discipline. In total, 1982 images were sampled and annotated with 24 2D key points, as depicted in Figure 2a, of which 1830 were used as training and 152 as validation images. The dataset comprises at least 32 unique athletes in 5 unique locations and various conditions (see Figure 2b) and is made available online [1] for further research.

Calibration pole augmentation. As this newly created Alpine dataset does not feature any calibration poles like the *MV-Ski* dataset, evaluating a 2D pose estimation algorithm that was only trained on this will produce significant outliers, particularly for the ski poles. One way to improve robustness on the *MV-Ski* dataset is to augment SV-training images with randomly superimposed cutouts of various calibration poles, see Figure 3. At training time, we uniformly sample $\mathcal{U}(0, 20)$ randomly selected poles and place them uniformly over the image. The poles are scaled by $\mathcal{U}(0.5, 2.5)$ and rotated $\mathcal{U}(-15, 15)$ degrees. We compare this method to adding one *MV-Ski* sequence to the training of OpenPose.

[1] Single view alpine skiing dataset: https://cvlab.epfl.ch/ski-2dpose-dataset/

Figure 3: Left: Example image showcasing calibration poles. Right: Augmented alpine skiing image.

4. Method

Our goal is to take as input a set of synchronized video streams of the same athlete filmed from different angles and estimate the global and articulated 3D pose. We target an easy-to-setup solution: The cameras are assumed to be intrinsically calibrated, have known relative position, but unknown orientation. The advantage is that this setup allows us to use consumer cameras without specialized hardware for angular readout — handheld recording is conceivable. The position requirement might sound restrictive, however, when using up to three cameras a simple distance measure between pairs of cameras is sufficient to determine their relative position. Furthermore, the intrinsics for fixed-focal length cameras only have to be calibrated once. Still, professional PTZ-cameras that have been calibrated for different zoom levels can be used as well.

We develop our approach for estimating the gloabal articulated 3D pose in two steps. In the first, we assume the rotation matrices as known, while in the second we jointly optimize for 3D pose and camera rotations. To go from images to 3D pose, we propose a multi-staged approach as shown in Figure 4, where 2D pose detections are generated from cropped images around the athlete and then 3D poses are optimized to best fit all localized 2D joints. Generating 2D estimations first allows us to analyze potential detection weaknesses when using a new dataset, before developing a method for 3D joint detection.

First, we train and run an object detection network [21] on each video stream to generate a tight square bounding box around the primary athlete, which effectively excludes persons in the background. Outliers are filtered out and bounding box detections are temporally smoothed. The 2D pose estimation network [6] is subsequently trained and run on the square crop, generating joint heatmaps, from which 2D joint key points are extracted. The 2D detections from all cameras are then incorporated into a bundle adjustment method which reconstructs the underlying 3D pose of the skier. This optimization includes our core contribution on conditioning camera motion on tracked features.

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

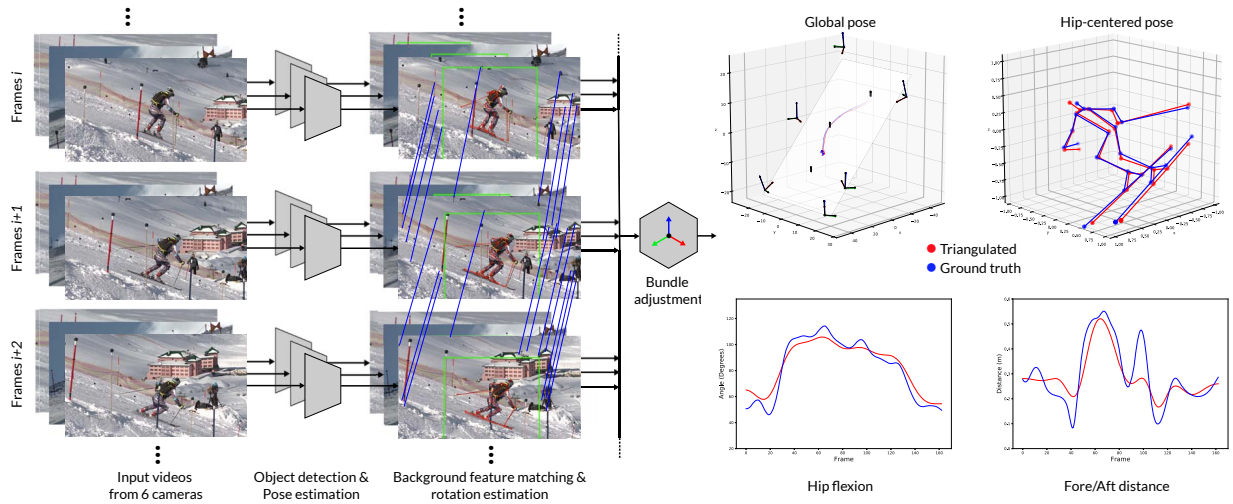


Figure 4: Method overview – Images from up to 6 different cameras are preprocessed to find 2D joint locations by first cropping a bounding box around the athlete and then running a pose estimation network on it. 2D poses from all cameras and all frames are used in a bundle adjustment approach to triangulate the athlete's global 3D pose. Additionally, camera rotations can be estimated by tracking background features over consecutive frames. Coaches can extract useful information about biomechanical variables like the hip flexion or fore/aft distance along the athlete's trajectory from the 3D pose.

4.1. Athlete detection

Sports videos are often captured from a distance and contain sequences of very different zoom levels, resulting in the athlete size in the image ranging from frame-filling to only making up a very small portion of the frame. While pose estimation networks like OpenPose run images through their network at multiple scales to account for this fact, athletes can be so small in images that detection fails completely. Even when the skier fills a good portion of the frame and the input scale is right for the pose estimation network, other people on the slopes or other high-contrast objects in the background can lead to wrong detections when only interested in the pose of the main subject. We therefore first detect a tight bounding box around the athlete, resulting in the pose estimation network always receiving examples of the same scale.

In general, object detection is the process of localizing occurrences of certain classes in images and drawing a tight bounding box around them. In recent years, several Deep Learning approaches made great advances in terms of accuracy and detection speed. Liu *et al*. propose the Single Shot MultiBox Detector (SSD) [21] generating scores for the presence of an object in predefined bins and then adjusting the bins to better match the object shape. Detections from multiple feature maps and different resolutions are then combined to allow for detection of various sizes in one single network stage. For this task we chose the SSD [21] network for its good performance and low computational overhead during both training and test time.

As the SSD network is a single-image detector and does not incorporate temporal information from the fact that we are dealing with videos, detections between frames can suffer from jitter and outliers if other people are present in the scene. To remove strong outliers, we replace all detections whose center deviates beyond the bounding box of the last correctly detected frame by the latter. Jitter and varying sizes of the bounding boxes are dealt with by applying temporal Gaussian smoothing to the center location of the crops and their respective side lengths with parameters $\sigma_{center} = 10$ and $\sigma_{side} = 50$ frames. Finally, side lengths are scaled by a factor of 1.5 for the newly created *SV-Ski* and 2 for the *MV-Ski* dataset and all crops get resized to 500 x 500 pixels. This was done because the SSD network tended to produce too small bounding boxes on the unseen *MV-Ski* dataset. Training specifications for the SSD network are described in the supplementary material.

4.2. 2D pose estimation

Given an input image $\boldsymbol{I} \in \mathbb{R}^{w \times h \times 3}$ of width w and height h, the task of 2D pose estimation is to compute x and y coordinates for every joint $j \in \{1, ..., N_J\}$. The OpenPose network by Cao *et al*. [6] returns for every joint j a confidence/heat map $\boldsymbol{C}_j \in \mathbb{R}^{w \times h}$ and for each limb/bone $l \in \{1, ..., N_L\}$ a Part Affinity Field (PAF) $\boldsymbol{B}_l \in \mathbb{R}^{w \times h \times 2}$, with every point in $\boldsymbol{B}_l$ encoding a vector describing limb orientations. Using the PAF as an indicator for which joints in the heatmaps belong together, the poses of multiple people can be efficiently differentiated. We finetune this model for the extended set of ski pose joints, including ski tip and tail, and pole positions. Because in our problem we are focusing solely on the pose estimation of a single athlete on

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

the slope and not multiple people, we don't rely on some of the multi-person detection advantages that PAF's bring to the table. Indeed, for this task we take the maximum location $\boldsymbol{p}_j^* \in \mathbb{R}^2$ as $\boldsymbol{p}_j^* = \arg\max_{w,h} \boldsymbol{C}_j$ of each confidence map for each joint j. Training specifications for the OpenPose network are described in the supplementary material.

Refining SSD bounding boxes using OpenPose. Because of the Gaussian smoothing of detected SSD bounding boxes, drifts in the crop with respect to the athlete's center may still be present. Athletes often extend their arms and poles outwards for balance in difficult terrain, which may cause the thin ski poles to exit the bounding box when drifts in the smoothed crops are present. To remedy this, we run OpenPose on the generated crops and take the median of the computed joint positions as the new center for each frame. Then, we apply a weaker Gaussian smoothing pass to the center locations and side lengths with parameters $\sigma_{center} = 5$ and $\sigma_{side} = 5$, and re-run OpenPose.

4.3. 3D pose estimation

The last step in our approach is estimating the 3D poses of the skier using a bundle adjustment optimization method with the detected 2D joint locations. We take the OpenPose output $\boldsymbol{p}_j^{f,c} \in \mathbb{R}^2$ from all cameras $c \in \{1, ..., N_C\}$, over all frames $f \in \{1, ..., N_F\}$, for each joint $j \in \{1, ..., N_J\}$ and reconstruct the respective underlying 3D joint positions $\boldsymbol{P}_j^f \in \mathbb{R}^3$ in global space. Let us denote the complete 3D pose at time f as $\boldsymbol{P}^f \in \mathbb{R}^{N_J \times 3}$.

Cosine basis parametrization. One way to go about the bundle adjustment would be to directly optimize for the points $\boldsymbol{P}$ in 3D space. We would jointly try to optimize $f \times N_J \times 3$ unrestricted parameters, meaning poses in neighboring frames are not guaranteed to transition in a smooth motion. This is usually countered by penalizing accelerations measured by finite differences.

Instead of letting our 3D points be completely free and add a smoothness term after the fact, we chose to describe them using a parametrization that is inherently smooth over time [4]. We model the motion using the inverse discrete cosine transform (IDCT), meaning that it is the result of a sum of N_Π cosine waves, scaled by coefficients $\boldsymbol{\Pi}_{j,d} \in \mathbb{R}^{N_\Pi}$ for each joint $j \in \{1, ..., N_J\}$ and its dimension $d \in \{x, y, z\}$. The 3D pose can then be reconstructed using

$$\boldsymbol{P}_{j,d}^f = \frac{\boldsymbol{\Pi}_{j,d}^0}{2} \sum_{n=1}^{N_\Pi - 1} \boldsymbol{\Pi}_{j,d}^n \cos\left[\frac{\pi n}{N_\Pi}\left(f + \frac{1}{2}\right)\right] , \quad (1)$$

for $f \in \{0, ..., N_F - 1\}$.

The lower we set N_Π, the smoother the motion will be, as we are only using low-frequency cosine waves, but we might not be able to reproduce actual fast changes in movement. On the other hand, if we increase the number of coefficients, we could approximate more complex motions but risk picking up high-frequency noise. With known rotation matrices, we set N_Π to 25, while with uncalibrated cameras we set it to 11. These values were chosen empirically for a good balance between smoothness and accuracy.

4.4. Using known camera rotations

In this section, we assume that the parameters for all frames f and all cameras c are known. Specifically, this means we know the intrinsic matrix $\boldsymbol{K}^{f,c} \in \mathbb{R}^{3\times3}$, the matrix describing world to camera rotation $\boldsymbol{R}^{f,c} \in \mathbb{R}^{3\times3}$ and camera location $\boldsymbol{t}^{f,c} \in \mathbb{R}^3$. Using the extrinsics $[\boldsymbol{R}^{f,c} \mid \boldsymbol{t}^{f,c}]$, the transformation of a world coordinate point $\boldsymbol{P}_j^{f,w}$ to camera c's coordinate frame is given by

$$\boldsymbol{P}_j^{f,c} = \boldsymbol{R}^{f,c}\boldsymbol{P}_j^{f,w} + \boldsymbol{t}^{f,c} . \quad (2)$$

The projection $\hat{\boldsymbol{p}}_j^{f,c} \in \mathbb{R}^3$ (in homogeneous coordinates) of point $\boldsymbol{P}_j^{f,c}$ onto camera c's image plane is then given by

$$\hat{\boldsymbol{p}}_j^{f,c} = \boldsymbol{K}^{f,c}\boldsymbol{P}_j^{f,w} . \quad (3)$$

The homogeneous point $\hat{\boldsymbol{p}}_j^{f,c}$ can then be transformed to the Euclidean point $\tilde{\boldsymbol{p}}_j^{f,c} \in \mathbb{R}^2$ by dividing by the last coordinate. Finally, denote the complete projection from world coordinates to an image plane as

$$\pi_c(\boldsymbol{P}_j^{f,w}) = \tilde{\boldsymbol{p}}_j^{f,c} . \quad (4)$$

3D reconstruction is done using a bundle adjustment approach, where we optimize an energy function

$$\arg\min_{\boldsymbol{\Pi}} E(\boldsymbol{\Pi}, \boldsymbol{K}, \boldsymbol{R}, \boldsymbol{t}) , \quad (5)$$

that includes a reprojection error, as well as priors on the human body defined as

$$E(\boldsymbol{\Pi}, \boldsymbol{K}, \boldsymbol{R}, \boldsymbol{t}) = \lambda_{rep}E_{rep} + \lambda_{limbs}E_{limbs} . \quad (6)$$

Reprojection term. The 3D joint location estimations are iteratively updated by gradient descent such that when projected to each camera plane, they are as close as possible to the 2D joint locations. If we had perfectly consistent 2D localizations, a simple least-squares bundle adjustment process with decent initializations would yield very good results. In our case, 2D detections sometimes contain high per-joint pixel errors, and we use a robust norm that also incorporates the detection confidence, similar to Takahashi *et al.* [39]. We write the reprojection energy term as

$$E_{rep}(\boldsymbol{\Pi}, \boldsymbol{K}, \boldsymbol{R}, \boldsymbol{t}) = \frac{1}{N_F N_C N_J} \sum_{f=1}^{N_F} \sum_{c=1}^{N_C} \sum_{j=1}^{N_J} g(\pi_c(\boldsymbol{P}_j^{f,w}), \boldsymbol{p}_j^{f,c}), \quad (7)$$

with $\boldsymbol{P}_j^w = \mathrm{IDCT}(\boldsymbol{\Pi}_j)$ in DCT encoding. The distance

$$g(x, y) = (n(0) - n(e_{rep}(x, y)))e_{rep}(x, y) \quad (8)$$

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

re-weights the scaled reprojection errors

$$e_{rep}(x, y) = ||(x - y)\boldsymbol{C}(y)||_2 \ , \tag{9}$$

where $n(x)$ denotes the normal distribution's probability density function $N(0, \sigma^2)$ and $\boldsymbol{C}(y)$ the heatmap probability value at point y. Using this norm with $\sigma^2 = 100$, outlier points have negligible influence on the energy function.

Human prior term. We would like all limbs $(i, j) \in Limbs$ to consistently have the same lengths $\ell(i, j)$ over time. To this end, we minimize the difference between the estimated and the known limb lengths,

$$E_{limbs}(\boldsymbol{\Pi}) = \frac{1}{N_F} \sum_{f=1...N_F} \sum_{i,j} \left(\|\boldsymbol{P}_i^{f,w} - \boldsymbol{P}_j^{f,w}\|_2 - \ell(i, j) \right)^2, \tag{10}$$

with $\boldsymbol{P}_j^w = \text{IDCT}(\boldsymbol{\Pi}_j)$ in DCT encoding. The limb lengths $\ell(i, j)$ were taken from the ground truth data, but can also be measured manually on the athletes.

Optimization and parameters. When optimizing for absolute 3D positions, all points were initialized in the center between all cameras, with an additional random spread of $\mathcal{U}(-10, 10)$ meters. We used the L-BFGS [20], a quasi-Newton optimization algorithm, with step length 0.05, running it for 100 outer iterations, with at most 20 inner iterations per optimization step. The energy terms were scaled by $\lambda_{rep} = 80$ and $\lambda_{limbs} = 1$.

4.5. Estimating camera rotations

In the same way we optimized the 3D pose positions, it is possible to freely optimize other parameters such as the camera's rotation [39]. We again use the IDCT as in Equation 1 to compactly describe the Euler angles of the camera rotations using $N_\Gamma = 11$ coefficients $\boldsymbol{\Gamma}$. The objective then becomes

$$\underset{\boldsymbol{\Pi}, \boldsymbol{\Gamma}}{\arg\min}\, E(\boldsymbol{\Pi}, \boldsymbol{K}, \boldsymbol{\Gamma}, \boldsymbol{t}) \ . \tag{11}$$

Initialization. Like with the 3D pose, the camera angles can be parametrized by a low-dimensional cosine basis and iteratively updated to the correct ones by gradient descent. A problem with this approach is however initialization. If the cameras face in randomly initialized directions, it is unlikely that the optimization objective can converge to a desirable solution. We instead propose a bootstrapping step, where in every gradient descent iteration only the 3D poses are optimized, while the cameras are adjusted to always point to the center of the estimated poses.

More specifically, in the beginning we initialize all 3D pose positions around the center of all cameras with a random spread of $\mathcal{U}(-1, 1)$ meters to be largely independent of any specific sport. For every camera we then compute the look-at rotation matrix, with the target being the mean location of the 3D pose and the camera's up direction being the global z-axis. A problem with the look-at matrix is, that the person is originally not necessarily in the middle of the image. To solve this, we first compute the horizontal and vertical relative position of the skier in the 2D image. From the camera intrinsics, we know the Field of View (FoV) and can then pan and tilt the rotation matrix in the opposite direction of the calculated horizontal and vertical FoV shift.

When optimizing with this method for 25 outer iterations, we get a very rough estimate for the real 3D pose positions and camera rotation matrices, which serves as an initialization for joint optimization of all parameters.

Homography camera rotation differences. From this initial estimate, we could potentially optimize for 3D poses and camera rotations jointly, but preliminary tests (Table 2) have shown that doing this with only the 2D pose estimates as information for the optimization does not yield accurate results. What we propose instead is to use the background information and the fact that we are dealing with PTZ-cameras to our advantage. Since PTZ-cameras are fixed in space and can therefore only rotate, detected 2D points, say of consecutive frames f and $f + 1$ in camera c, are related [15] by a homography matrix $\boldsymbol{H}^{f,c}$ satisfying

$$\boldsymbol{H}^{f,c} = \boldsymbol{K}^{f+1,c} \Delta \boldsymbol{R}^{f,c} (\boldsymbol{K}^{f,c})^{-1} \ , \tag{12}$$

with $\Delta \boldsymbol{R}^{f,c}$ being the relative rotation between the images and $\boldsymbol{K}^{f,c}$ and $\boldsymbol{K}^{f+1,c}$ the intrinsics of the first and second frame, respectively.

Because we assume the intrinsics as known, to find the rotation we only need to compute the homography. For this, we collect features in consecutive images using an ORB detector [36, 3] and match them by minimizing the Hamming norm [2]. We only consider those points lying outside the bounding box predicted by the SSD network, to exclusively track features in the static background. The corresponding points are then used to find the homography matrix [1], excluding outliers with RANSAC. Finally we compute the rotation between any consecutive frames f and $f + 1$

$$\Delta \boldsymbol{R}^{f,c} = (\boldsymbol{K}^{f+1,c})^{-1} \boldsymbol{H}^{f,c} \boldsymbol{K}^{f,c} \ . \tag{13}$$

For cameras with a fixed position, $\Delta \boldsymbol{R}^{f,c}$ is a pure rotation matrix. In practice, slight camera movements and noisy inputs to the homography computation will not produce perfect rotation matrices, but we found that extracted Euler angles are not far off their ground truth, as shown for one example sequence in Figure 5. Notice that since the camera has a large focal length, pitch and yaw changes in the x and y axes are much more significant than rolling in the z axis, and are also estimated more accurately. Outliers in the rotation differences are removed by a median filter with size 7 and smoothed using a Gaussian with standard deviation 3.

Bundle adjustment with rotation differences. In our energy function, we add a term

$$E_{rot}(\boldsymbol{\Gamma}) = \frac{1}{N_F N_C} \sum_{f=1}^{N_F - 1} \sum_{c=1}^{N_C} \|\Delta \boldsymbol{R}^{f,c} - \left(\boldsymbol{R}^{f+1,c} \left(\boldsymbol{R}^{f,c} \right)^\intercal \right)\|_2, \tag{14}$$

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

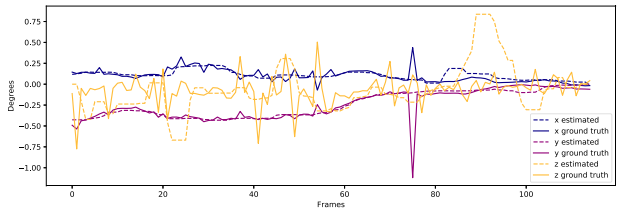


Figure 5: Estimated change in rotation $\Delta \boldsymbol{R}$ for one camera.

with $\boldsymbol{R}^c = \mathrm{IDCT}(\boldsymbol{\Gamma}^c)$ in DCT encoding, minimizing the norm between the measured $\Delta \boldsymbol{R}^{f,c}$ and estimated consecutive camera rotations $\boldsymbol{R}^{f+1,c} \left(\boldsymbol{R}^{f,c}\right)^\intercal$. By minimizing this term, we enforce that the estimated relative camera motion matches the one measured by optical flow, which enables the bundle adjustment to find the absolute rotation more easily. Note that this formulation is robust to measurement noise and the estimated homographies not being pure rotations, since the Euler-angle representation ensures that $\boldsymbol{R}$ is a proper rotation matrix. We further optimize for $\boldsymbol{\Pi}$ and $\boldsymbol{\Gamma}$ over 1500 outer iterations. The energy function was weighted with $\lambda_{rep} = 500$, $\lambda_{limbs} = 1$ and $\lambda_{rot} = 10000$.

5. Results

In this section, we report the performance of our motion capture algorithm on the task of ski performance analysis. First, we quantify the accuracy gain brought about by the new *SV-Ski* dataset, optical-flow guided rotation estimation, and other model choices. Second, we compare against existing monocular methods for 3D human pose estimation on the public *MV-Ski* dataset.

Metrics. We report the widely-used Percentage of Correct Key points (PCK) metric, the fraction of predicted joint positions that is within one head-neck distance to the ground truth, as well as the Mean Per Joint Position Errors (MPJPE), the mean euclidean distance to the ground truth. For 2D keypoints, MPJPE is measured in normalized image coordinates ranging from 0 to 1, for 3D in meters (m).

In addition we, analyze the mean absolute error (MAE) of skiing-specific metrics that are widely used in performance analysis [37, 10], such as the center of mass (CoM), knee angle, and lean angle, where the latter is measured in the plane orthogonal to the skiing direction. These are defined formally in the supplementary material. For the MPJPE and CoM, we analyze the *Global* error computed in world coordinates, the *Centered* error measured relative to the hip, and in *Normalized* coordinates where the scale of the prediction is adjusted to the ground truth in the least squares sense before error computation. Monocular methods can only estimate the latter, as scale and depth is ambiguous without knowing the athletes height. All metrics are computed over the available test sequences and the mean and standard deviation (std) across all frames is reported.

(a) SV-Ski training

Test dataset	PCK	MPJPE ± std
SV-Ski all	92.69	0.0195 ± 0.0638
SV-Ski body	94.88	0.0132 ± 0.0416
MV-Ski all	51.37	0.1064 ± 0.1297
MV-Ski body	58.26	0.0835 ± 0.1092

(b) SV-Ski Augmented training

Test dataset	PCK	MPJPE ± std
SV-Ski all	95.77	0.0807 ± 0.0488
SV-Ski body	97.91	0.0625 ± 0.0346
MV-Ski all	65.51	0.0137 ± 0.1236
MV-Ski body	73.01	0.0092 ± 0.1074

(c) SV-Ski Aug. + weight init. MPII

Test dataset	PCK	MPJPE ± std
SV-Ski all	**96.76**	**0.0119 ± 0.1268**
SV-Ski body	**98.36**	**0.0081 ± 0.1077**
MV-Ski all	70.10	0.0755 ± 0.0429
MV-Ski body	76.83	0.0577 ± 0.0268

(d) SV-Ski Aug. + init. MPII + MV-Ski

Test dataset	PCK	MPJPE ± std
SV-Ski all	96.51	0.0119 ± 0.0431
SV-Ski body	97.81	0.0087 ± 0.0296
MV-Ski all	**78.11**	**0.0627 ± 0.1275**
MV-Ski body	**83.12**	**0.0507 ± 0.1133**

Table 1: **2D pose estimation results** on *SV-Ski* and *MV-Ski* with four different dataset configurations used for training.

Test sets. *MV-Ski-test* contains two runs of a skier not contained in the training set, totalling to 1674 frames. *SV-Ski-test* comprises 152 images that are strictly excluded from training.

5.1. 2D pose estimation

We trained OpenPose using four different dataset configurations. First we only trained it on the newly created SV-Ski dataset, which we then augmented with calibration poles. We then initialized the network using pretrained weights from the MPII Human Pose dataset, and finally added one *MV-Ski* sequence using four camera angles.

In Table 1 we analyze different training and test splits, using *all* keypoints and also just the *body* joints, the 14 joints 0-4, 6-8, and 10-15 shown in Figure 2a. In the case of the *SV-Ski-test* set we have information about joint visibility and invisible joints were not counted in the PCK results. The data augmentation and taking one *MV-Ski* sequence for training brought the biggest gain in accuracy. The accuracy on the *SV-Ski-test* set improved overall, but not as much because it was already very high. See the supplementary material for the same comparisons, displayed graphically.

5.2. 3D pose estimation

In Table 2 we highlight the best results obtained from both the calibrated and uncalibrated cases. For those, the bundle adjustment used all 6 cameras and 2D pose estimates by the best performing dataset configuration that includes one additional *MV-Ski* sequence. For the local metrics, like the centered MPJPE and the biomechanical variables, the two methods both yield comparable, high accuracies, with the uncalibrated method only slightly worse. Still, the latter is mostly well within the standard deviations of the calibrated methods. On the global metrics, like the MPJPE, CoM distances and speed, we see a larger discrepancy between the two methods, meaning that the poses found when estimating the camera rotations are locally accurate, but globally contain small deviations of the whole pose.

As baselines, we also report all metrics on bundle adjustment approaches, where we don't use the cosine basis parametrization (Ours-A and Ours-B). In the uncalibrated

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

Metric	Ours-calibrated	Ours-uncalibrated	Ours-A (calibrated)	Ours-B (uncalibrated)	C	D [34]
Global MPJPE [m]	**0.092 ± 0.091**	0.701 ± 0.219	0.096 ± 0.120	7.590 ± 4.946	n/a	n/a
Global Body MPJPE [m]	0.060 ± 0.046	0.688 ± 0.201	**0.056 ± 0.033**	7.490 ± 4.792	n/a	n/a
Centered MPJPE [m]	**0.077 ± 0.087**	0.090 ± 0.085	0.087 ± 0.122	0.459 ± 0.960	n/a	n/a
Centered Body MPJPE [m]	**0.045 ± 0.030**	0.071 ± 0.053	0.050 ± 0.034	0.355 ± 0.383	n/a	n/a
Normalized MPJPE[m]	0.075 ± 0.083	0.087 ± 0.082	0.087 ± 0.117	0.232 ± 0.238	**0.07**	n/a
Normalized Body MPJPE[m]	**0.039 ± 0.025**	0.051 ± 0.035	0.042 ± 0.029	0.132 ± 0.103	n/a	0.081
Global CoM Error [m]	0.05 ± 0.04	0.78 ± 0.24	**0.05 ± 0.02**	8.93 ± 5.40	n/a	n/a
Global speed MAE [m/s]	**0.74 ± 1.76**	1.87. ± 2.64	0.45 ± 1.08	31.35 ±22.45	n/a	n/a
Knee flexion MAE [deg]	3.96 ± 3.16	4.83 ± 3.18	4.45 ± 3.98	14.21 ±11.84	**2.3 ± 6.1**	7.39
Hip flexion MAE [deg]	3.92 ± 2.82	4.74 ± 3.45	4.26 ± 3.08	15.30 ±11.38	**2.6 ± 5.3**	5.74
Lean angle MAE [deg]	3.91 ± 2.60	3.68 ± 2.56	4.48 ± 4.61	8.09 ± 7.01	**3.3 ± 3.3**	n/a
Fore/aft angle MAE [deg]	6.77 ± 5.20	**5.75 ± 5.67**	8.72 ± 8.42	12.26 ±11.72	n/a	n/a
Fore/aft distance MAE [m]	0.07 ± 0.05	0.07 ± 0.06	0.09 ± 0.08	0.30 ± 0.38	**0.03 ± 0.05**	n/a

Table 2: Comparison of our results in calibrated and uncalibrated cases to methods proposed by Oštrek *et al*. (unpublished data) (C: *Monocular 3D pose estimation*) and Rhodin *et al*. [34] (D: *Semi-supervised*). As a baseline, we also provide results when not using a cosine basis and directly optimizing for 3D pose coordinates (Ours-A and Ours-B). In the uncalibrated case (Ours-B), we show the performance when optimizing camera rotations without enforcing Equation 14.

case (Ours-B) we also directly optimize for the rotation matrices without getting rotation measures from the homography approach. As can be seen in Table 2, our new uncalibrated approach outperforms the baseline by an order of magnitude in all metrics, with only slightly lower gains in the calibrated case. Note that the bundle adjustment in Ours-B easily gets stuck in a local degenerate minimum. While this approach worked for previous work [39], the high speeds, large capture volume and zoomed in cameras in the *MV-Ski* dataset impede convergence to better solutions. Using background cues and the cosine basis provided a strong guide to avoid those local minima. The other main benefits in that case are the improved smoothness of the motion and speedup of the bundle adjustment. Using the cosine basis parametrization, we usually need only about half the number of iterations to reach a similar performance.

An analysis showcasing the performance of both calibrated and uncalibrated cases, across dataset configurations and number of cameras used, can be found in the supplementary material. We see the largest gains in accuracy when going from two to three cameras. With more than that, reconstruction quality still improves, but with diminishing returns. We also show how improved 2D detection quality directly translates into higher 3D triangulation accuracy.

Comparison to existing methods. In Table 2 we compare our best results with the methods proposed by Oštrek *et al*. (unpublished data) and Rhodin *et al*. [34]. Oštrek *et al*.'s method computes all biomechanical variables indirectly from images via monocular 3D pose estimation trained on the *MV-Ski* dataset. Rhodin *et al*. estimate a monocular 3D pose using a semi-supervised method by constraining the model to predict the same pose in all views and needing only few labelled images. Both our best calibrated and uncalibrated methods perform only slightly worse than Oštrek *et al*. (unpublished data), while yielding lower standard deviations and using only a single sequence from the *MV-Ski* dataset. Both our methods outperform the semi-supervised method by Rhodin *et al*. [34].

Velocity estimation. In contrast to existing monocular approaches [34], our method allows to estimate the athletes instantaneous velocity as the change in CoM position between two frames, even if camera rotations are unknown. Using calibrated cameras, we get mean absolute errors of 0.74 ± 1.76 m/s, while when estimating rotations, it rises to 1.87 ± 2.64 m/s. This is still relatively low, given the high speed of the professional athletes, that ranges between $15 - 20$ m/s in the test sequences and the large capture volume of more than $30 - 50$ m distance between cameras.

6. Conclusion

We developed a practical method for reconstructing the global articulated 3D pose from bare videos taken from multiple rotating cameras. Our key contribution is joint optimization of 3D human pose and camera rotation by incorporating additional constraints from tracked features and resulting homographies. Our empirical evaluation shows that training 2D keypoint detection on the large *SV-Ski* dataset and subsequent multi-view 3D reconstruction is as accurate or better as training 3D pose estimation directly on the available small-scale multi-view datasets [34], while promising improved generalization capability to new scenes. The improvement brought about by our contributions are quantified in terms of widely used reconstruction metrics as well as biomechanical variables that are common for performance analysis of professional athletes. By contrast to monocular solutions, we are able to provide accurate global measurements without the need for cumbersome camera rotation calibration, which makes this method directly applicable for ski coaches.

Acknowledgement. This work was supported in part by the Swiss National Science Foundation.

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

References

[1] Camera Calibration and 3D Reconstruction. https://docs.opencv.org/3.0-beta/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html. [Online; last accessed on Jun 10, 2019]. 6

[2] Feature Matching. https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_feature2d/py_matcher/py_matcher.html. [Online; last accessed on Jun 10, 2019]. 6

[3] ORB (Oriented FAST and Rotated BRIEF). https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_feature2d/py_orb/py_orb.html. [Online; last accessed on Jun 10, 2019]. 6

[4] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear spatiotemporal basis models. *ACM Transactions on Graphics (TOG)*, 31(2):17, 2012. 5

[5] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[6] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CVPR*, 2017. 2, 3, 4

[7] J. Chen and J. J. Little. Sports Camera Calibration via Synthetic Data. *CoRR*, abs/1810.10658, 2018. 2

[8] J. Chen, F. Zhu, and J. J. Little. A Two-point Method for PTZ Camera Calibration in Sports. *WACV*, 2018. 2

[9] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt. Outdoor Human Motion Capture by Simultaneous Optimization of Pose and Camera Parameters. *Computer Graphics Forum*, 34, 12 2014. 2

[10] B. Fasel, J. Spörri, M. Gilgien, G. Boffi, J. Chardonnens, E. Müller, and K. Aminian. Three-Dimensional Body and Centre of Mass Kinematics in Alpine Ski Racing Using Differential GNSS and Inertial Sensors. *Remote Sensing*, 8, 09 2016. 1, 2, 7

[11] M. Gilgien, J. Spörri, J. Chardonnes, J. Kröll, P. Limpach, and E. Müller. Determination of the centre of mass kinematics in alpine skiing using differential global navigation satellite systems. *Journal of Sports Sciences*, 33(9):960-9, 2015. 1

[12] M. Gilgien, J. Spörri, P. Limpach, A. Geiger, and E. Müller. The effect of different Global Navigation Satellite System methods on positioning accuracy in elite alpine skiing. *Sensors (Basel)*, 14(10):18433-53, 2014. 1

[13] S. Günel, H. Rhodin, and P. Fua. What face and body shapes can tell about height. *arXiv preprint arXiv:1805.10355*, 2018. 1

[14] Y. Gwangjae, J. J. Young, K. Jinhyeok, H. K. Jin, Y. K. Hye, K. Kitae, and B. P. Siddhartha. Potential of IMU Sensors in Performance Analysis of Professional Alpine Skiers. *Sensors (Basel)*, 16(4):463, 2016. 1

[15] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. 6

[16] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. pages 224 – 231, 07 2009. 2

[17] C. Huang, F. Gao, J. Pan, Z. Yang, W. Qiu, P. Chen, X. Yang, S. Shen, and K. Cheng. Act: An autonomous drone cinematography system for action scenes. 2018. 2

[18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. 2016. 2

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 2

[20] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45:503–528, 1989. 6

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: Single Shot MultiBox Detector. *ECCV*, 2016. 3, 4

[22] J. Martinez, R. Hossain, J. Romero, and J. Little. A Simple Yet Effective Baseline for 3D Human Pose Estimation. 2017. 1

[23] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. 2017. 1

[24] T. Moeslund and E. Granum. A Survey of Computer Vision-Based Human Motion Capture. 81(3), March 2001. 2

[25] T. B. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. 104(2):90–126, 2006. 2

[26] F. Moreno-noguer. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. 2017. 1

[27] B. Nemec, T. Petrič, J. Babič, and M. Supej. Estimation of alpine skier posture using machine learning techniques. *Sensors (Basel)*, 14(10):18898-914, 2014. 1

[28] G. Pavlakos, X. Zhou, K. Derpanis, G. Konstantinos, and K. Daniilidis. Coarse-To-Fine Volumetric Prediction for Single-Image 3D Human Pose. 2017. 1

[29] G. Pavlakos, X. Zhou, K. D. G. Konstantinos, and D. Kostas. Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations. 2017. 1, 2

[30] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep Multitask Architecture for Integrated 2D and 3D Human Sensing. 2017. 1

[31] J. Puwein, L. Ballan, R. Ziegler, and M. Pollefeys. Joint Camera Pose Estimation and 3D Human Pose Estimation in a Multi-camera Setup. In *ACCV*, 2014. 2

[32] J. Puwein, R. Ziegler, L. Ballan, and M. Pollefeys. PTZ Camera Network Calibration from Moving People in Sports Broadcasts. pages 25–32, 01 2012. 2

[33] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues. 2016. 2

[34] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua. Learning Monocular

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.

3D Human Pose Estimation from Multi-view Images. *CVPR*, 2018. 1, 2, 8

[35] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-Net: Localization-Classification-Regression for Human Pose. 2017. 1

[36] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society. 6

[37] J. Spörri. *Reasearch Dedicated to Sports Injury Prevention - the 'Sequence of Prevention' on the Example of Alpine Ski Racing*. Habilitation with Venia Docendi in Biomechanics, 2016. 1, 2, 7

[38] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional Human Pose Regression. 2017. 1

[39] K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata. Human Pose As Calibration Pattern; 3D Human Pose Estimation With Multiple Unsynchronized and Uncalibrated Cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2, 5, 6, 8

[40] B. Tekin, P. Marquez-neila, M. Salzmann, and P. Fua. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. 2017. 1

[41] D. Tome, C. Russell, and L. Agapito. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In *arXiv preprint, arXiv:1701.00295*, 2017. 1

[42] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes - the Importance of Multiple Scene Constraints. June 2018. 1

[43] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. We. Weakly-Supervised Transfer for 3D Human Pose Estimation in the Wild. 2017. 1

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on April 26,2023 at 20:48:46 UTC from IEEE Xplore. Restrictions apply.