Noname manuscript No.
(will be inserted by the editor)

Parsing Human Skeletons in an Operating Room

Vasileios Belagiannis[1,2] · **Xinchao Wang**[3] · **Horesh Beny Ben Shitrit**[3] · **Kiyoshi Hashimoto**[4] · **Ralf Stauder**[1] · **Yoshimitsu Aoki**[4] · **Michael Kranzfelder**[5] · **Armin Schneider**[5] · **Pascal Fua**[3] · **Slobodan Ilic**[1,6] · **Hubertus Feussner**[5] · **Nassir Navab**[1,7]

Received: date / Accepted: date

Abstract Multiple human pose estimation is an important yet challenging problem. In an Operating Room (OR) environment, the 3D body poses of surgeons and medical staff can provide important clues for surgical workflow analysis. For that purpose, we propose an algorithm for localising and recovering body poses of multiple human in an OR environment under a multi-camera setup. Our model builds on 3D Pictorial Structures (3DPS) and 2D body part localization across all camera views, using Convolutional Neural Networks (ConvNets). To evaluate our algorithm, we introduce a dataset captured in a real OR environment. Our dataset is unique, challenging and publicly available with annotated ground truths. Our proposed algorithm yields to promising pose estimation results on this dataset.

Keywords human pose estimation · part-based model · medical workflow analysis

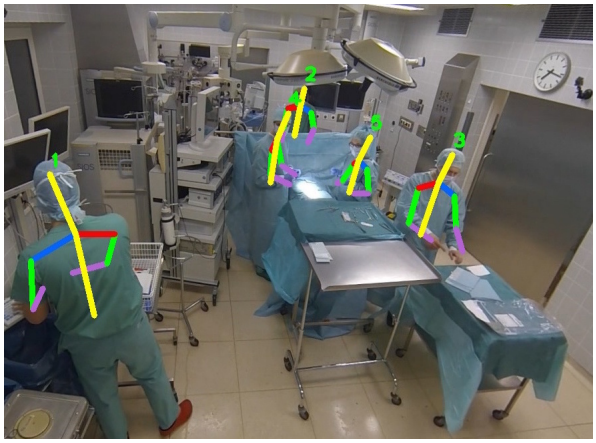


Fig. 1 Human pose estimation in the operating room: We introduce a unique dataset for multiple human 3D pose estimation from multiple views. Our results projected into a camera view.

1 Introduction

Recovering the body pose is a key task in many applications including surveillance, motion capture, activity recognition and human-machine interfaces. In this work, we address the problem of multiple human 3D pose estimation from multiple views in the scenario of an operating room. Our goal is to estimate the 3D body pose of the surgeons and medical staff.

We focus on the application of human body pose estimation in an operating room for the following reasons. Firstly, human pose estimation in the operating room is a crucial task and it may provide important clue for surgical workflow analysis. This claim is also supported by the fact that the body pose has been characterized as a very discriminative feature for action recognition, a related task to workflow modelling [26]. Secondly, estimating human pose in an operation room is challenging and the problem remains unsolved, because the environment is complex, dynamic and crowded; and people in the scene heavily occlude one another.

Surgical workflow models are built in order to derive and analyze statistical properties of a surgery for recovering the phase of the operation, staff training, data visualization, report generation and monitoring. Building a workflow model requires sufficient amount of data from different sources and sensors. For example, measurements are collected from in-

[1]Computer Aided Medical Procedures, Technische Universität München.
[2]VGG, University of Oxford.
[3]CVLAB, Ecole Polytechnique Fédérale de Lausanne (EPFL).
[4]Aoki Media Sensing Lab, Keio University.
[5]MITI, Klinikum rechts der Isar, Technische Universität München.
[6]Siemens AG.
[7]Johns Hopkins University.
Contact: belagian@in.tum.de.

struments, medical and monitoring devices [37]. A multi-view camera system that automatically estimates the 3D body pose of the surgeons and medical staff is another input modality to the framework of the surgical workflow modelling.

In this work, we propose an algorithm for estimating the 3D body pose of multiple individuals from a multi-view environment. Our approach is built on human tracking as well as 2D and 3D body pose estimation. Human tracking is performed with a multi-view tracker [12] that handles mutual occlusions between the target persons. After the localization and identification of the individuals using the human tracker, we propose to combine a 2D deep part detector with a 2D deep body regressor for generating a distribution of body part hypotheses for each localized individual in all views. We rely on Convolutional Neural Networks (ConvNets) to train the body part detectors and regressor. Finally, we parse the 3D body pose of each individual using the 3D Pictorial Structures (3DPS) model [7]. The 3DPS model is composed of unary, pairwise and ternary potential functions. The unary potentials incorporate the observations to our model, in our case we rely on the deep part detections across all views. The pairwise and ternary potentials model the human body prior. To ensure temporal consistency between the body poses over time, we have the temporal consistence potential function that incorporates with the human tracker output [10]. We demonstrate the performance of our approach on a new challenging dataset. To that end, we have set up a multi-view camera system inside a real operating room where we have recorded different simulated medical operations. Our multi-view dataset is unique and challenging. To our best knowledge, we are the very first ones to introduce such a dataset to the computer vision community. It is publicly available with annotated ground truths[1].

The contributions of our work are twofold. First, we introduce an operating room (OR) dataset which consists of 5 calibrated cameras with up to 5 individuals in the scene. The dataset is composed of 7000 frames per view with 2D and 3D human annotations in every tenth frame. Second, we propose an algorithm for multiple human pose estimation from multiple views. The algorithm combines 3D pictorial structures with deep learning. Finally, we demonstrate promising results in our new dataset.

2 Related work

In this section, we review the related work on human 3D pose estimation and focus on multiple human from multi-view approaches. We refer the reader to [36,43] for a general analysis of human motion analysis.

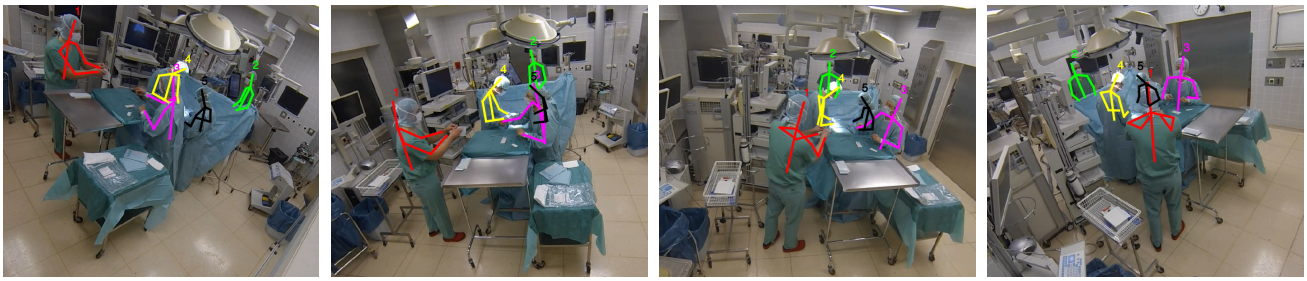
Defining the human body as a constellation of parts has been proved to be the effective way for 2D pose estimation [2,3,17,41]. The most notable part-based model are Pictorial Structures for 2D [4,18,20] or 3D body pose estimation [7,14]. The model has been successfully applied on multiple human pose estimation in 2D [3,17,41] and 3D [7] as well. However, existing human pose estimation datasets are limited to daily scene and sports, which are relatively simple because the scenes are not crowded and target individual can be easily distinguished in most of the frames. By contrast, we choose to apply our model in the operating room scenario which comprises significantly mutual occlusions between the target individuals and thus is much more challenging than the normal ones. In contrary to the concept of part-based models, the holistic models predict directly the body pose by learning a mapping between features and poses [1,6,23,25,29,44,47,59]. One very popular method to accomplish this task are random forests for human pose estimation from depth data [22,42]. However, the current depth sensors (i.e. Kinect) are not directly applicable to the operating room, due to the small working space which causes interference between the sensors.

Recently, deep learning approaches demonstrated promising results on many computer vision tasks, including human pose estimation. Convolutional Neural Networks (ConvNets), a popular deep learning algorithm, are the current state-of-the-art approach in human pose estimation [9,15, 32,38,48,49]. In this work, we also rely on ConvNets for training our body regressor and part detectors.

In 3D human pose estimation, there have been several approaches for multiple human pose estimation using monocular [5,31,60], stereo [21,40] or multi-view setup [27,33–35]. Moreover, the problem of 3D pose estimation has been often combined with tracking [31,60]. To improve the inferred 3D body in realistic environments, better appearance models have been introduced in [5], where the 3D pose is inferred by 2D pose lifting. In [21], a two-stage algorithm is applied on stereo input for human detection and pose recovery. Similar to our scenario, a multi-view system has been employed in [34,35] for multiple human body pose estimation. In [35], the proposed approach recovers the body poses of two individuals in a studio environment. In contrast, our model operates in an unconstrained environment, such as the operating room. Furthermore, our model is not bounded to a particular number of individuals. Closer to our approach are the frameworks from [33] and [27], although we rely neither on background subtraction [33] nor a massive number of input cameras [33].

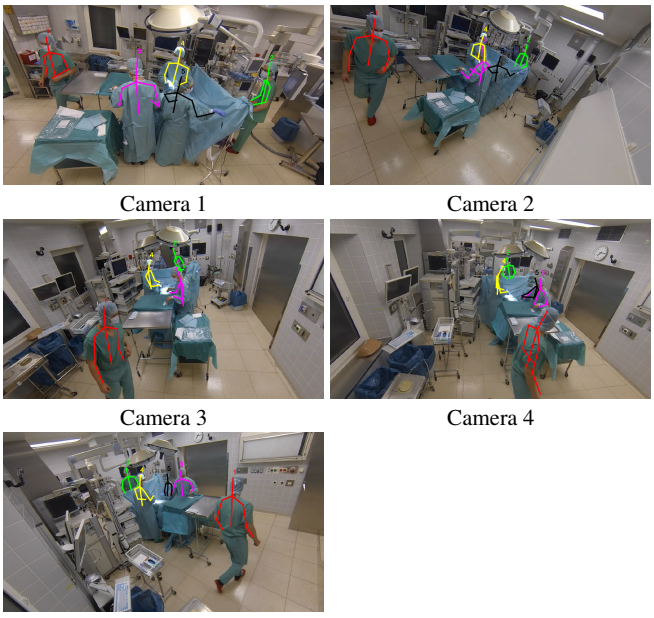
The most related work to our model is the 3D pictorial structures (3DPS) model. We follow similar formulation and apply the 3DPS model on human pose estimation in the operating room. Moreover, we propose to combine the 3DPS model with the 2D deep body regressor of [9] and a deep part detector for producing 2D body part proposals.

[1] http://campar.in.tum.de/Chair/MultiHumanOR



Camera 2 Camera 3 Camera 4 Camera 5

Fig. 2 Operating Room dataset: Our results on 3D pose estimation of multiple individuals projected in 4 out of 5 views of the Shelf dataset [7].



Camera 1 Camera 2

Camera 3 Camera 4

Camera 5

Fig. 3 Results on 2D Human Pose Estimation: Visual results of the 2D human pose estimation task are presented. The presented results are from the same time step across all camera views.

Our pose estimation framework takes people tracking results as input. In our implementation we have relied on the KSP tracker of [12], which outputs ground-plane trajectories. The KSP tracker has been shown to achieve the state-of-the-art tracking performance, and it has been recently extended to people re-identification [11,54], tracking interaction objects [55,56] and tracking cells [51] in biomedical imagery. However, our tracker may take input for any multi-object tracker but not limited to KSP.

The rest of the paper is organized as follows: In Section 3, we shortly present the 3DPS model and the deep body regressor and part detectors. The operating room dataset is presented in Section 4, followed by our experiments in Section 5 and our conclusion in Section 6.

3 Method

Our method adopts the 3D pictorial structures (3DPS) model. The 3DPS model is defined as a Conditional Random Field (CRF) that is composed of unary, pairwise and ternary potential functions. In the rest of this section, we first present the unary potential functions, which are formed by human body deep regressors for each camera view [9], combined with deep body part detectors. Next, we present the body prior model, as part of the pairwise and ternary potential functions. We follow the same formulation as in [7,8] to present the 3DPS model.

3.1 3D Pictorial Structures

Model In 3D pictorial structures (3DPS), a person is represented using an undirected graphical model. In our problem, we consider only the upper body due to the heavy occlusion of the lower body part (Fig. 3). We follow the same formulation with [8] and model the upper human body with n parts such that a 3D body configuration is given by $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Each body part $Y_i \in \Lambda_i$ is defined by the 3D position in the state space $\Lambda_i \subset \mathbb{R}^3$. Finally, the relation between the body parts is modelled using pairwise and ternary potential functions (Fig. 4). In the 3DPS model, the posterior for a body configuration $\mathbf{y} \in \mathbf{Y}$ is defined as:

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \mathbf{p}) = \frac{1}{Z(\mathbf{x}, \mathbf{w}, \mathbf{p})} \prod_{i}^{n} (\phi_i^{conf}(y_i, \mathbf{x}) \cdot \phi_i^{repr}(y_i, \mathbf{x}) \cdot \phi_i^{vis}(y_i, \mathbf{x}) \cdot \phi_i^{temp}(y_i, p_i))^{w_i} \prod_{(i,j) \in E_{tran}} \psi_{i,j}^{tran}(y_i, y_j)^{w_{ij}} \prod_{(i,j,k) \in E_{rot}} \psi_{i,j,k}^{rot}(y_i, y_j, y_k)^{w_{ijk}} \quad (1)$$

where the observation $\mathbf{x} \in \mathbf{X}$ corresponds to body part detections, $\mathbf{w} \in \mathbb{R}^D$ is the parameter vector and $\mathbf{p}$ a set of reference poses. The reference body poses $\mathbf{p}$ correspond to inferred poses from frames. Moreover, $Z(\mathbf{x})$ is the partition function and E_{tran} and E_{rot} are the graph edges that model the body constraints. The model consists of the detection

confidence $\phi_i^{conf}(y_i, \mathbf{x})$, reprojection error $\phi_i^{repr}(y_i, \mathbf{x})$, multi-view part visibility $\phi_i^{vis}(y_i, \mathbf{x})$ and the temporal consistence $\phi_i^{temp}(y_i, p_i)$ unary potential functions. The body prior is expressed by the pairwise and ternary potential functions, in terms of translation $\psi_{i,j}^{tran}(y_i, y_j)$ and rotation $\psi_{i,j,k}^{rot}(y_i, y_j, y_k)$ between body parts. Finally, the parameters w_i, w_{ij} and w_{ijk} of the model balance the influence of the 9 unary, 8 pairwise and 2 ternary potential functions ($D = 19$). In the following, we briefly present the potential functions. A detailed description of the 3DPS model can be found in [7,8].

3.2 State Space

The state space Λ_i comprises the locations that correspond to a candidate body part in the 3D space. Instead of discretizing the whole volume to generate our state space, we reduce the state space using body part detectors for each camera view. More specifically, we form the state space by conducting triangulation of all possible combinations of 2D body part detections of view pairs [7]. We assume that the cameras are calibrated and there is at least one pair of correct body part detections to fully recover a body part in 3D. The final global state space $\boldsymbol{\Lambda} = \{\Lambda_1, \Lambda_2, \dots \Lambda_n\}$ includes wrong hypotheses due to false positive detections. Since our method relies on tracking results, for each individual, we can significantly reduce the size of the state space based on the identity. Eventually, a separate state space is formulated for each individual based on the tracking input.

2D Human Model 3D Human Model (3DPS Model)

Fig. 4 Human model: On the left, our 2D human model is presented. It has 9 body joints which are regressed using a ConvNet. A confidence value is obtained for each regressed joint using a second ConvNet for classification. The symmetric joints count for a single class and thus we have in total 6 classes ($1-6, 2-5, 3-4, 7, 8, 9$). On the right, our 3D human model is presented. We model it using a CRF, where the blue edges correspond to pairwise potentials and the green ones correspond to ternary potentials. The pairwise potentials model the translation between the body parts, while the ternary model the rotation.

3.3 Body Part Detection

To sample 2D body parts for the state space generation, we rely on estimating a rough 2D body pose and then applying a body part detector around the estimated pose. These two steps are applied on each camera view and for each localized individual. Moreover, instead of relying on engineered features (e.g. HOG [16] or Haar-like [52]) to model the body parts' appearance in 2D, we learn the features using deep learning, which has demonstrated promising results for the task of human pose estimation [15,32,48,49]. The most well-established deep learning method is Convolutional Neural Networks (ConvNets) [30], which we employ in our model. The contributions of our ConvNets are twofold: Initially, a ConvNet is used to regress the 2D body pose given the image evidence. As we demonstrate in our experimental section, this rough 2D body pose estimation is usually accurate enough for our problem. Afterwards, a ConvNet body part detector is applied for each body part on the area that is constrained by the regressed body pose. Eventually, we use the detections for the generation of the state space as well as for the computation of the unary potential functions of the 3DPS model. In the following, we describe the 2D body pose regressor and body part detectors.

We use a ConvNet to regress a rough 2D body pose for each individual. The input to the regression network is a cropped image I_c from the camera view c that includes the localized individual. The output is the 2D body pose configuration given by a real-valued vector $\mathbf{y} = (y_1, y_2, \dots, y_N)$, with $y_i \in \mathbb{R}^2$. We also note the 2D body pose estimate with $\mathbf{y}$ for the case of the ConvNet, but it is different from Eq. (1). The architecture of the network is borrowed from [9], as well as, the robust loss function for training the ConvNet. We refer to [9] for further details on the network optimization and loss function. Finally, we use the 2D body pose regressor estimates as reference for defining the area from which we sample body part detections. In this way, we radically reduce the number of body part detections which we sample using another ConvNet.

The body part detector is formed by a ConvNet that classifies body parts based on our human model of Fig. 4. In particular, we train a ConvNet to classify among six different body parts, since we use the same class for the symmetric body parts. The structure of this ConvNet is presented in Fig. 5. The network is composed of four convolutional and two fully connected layers. In addition, we use dropout [45] to prevent over-fitting and regularize the network. The training of the network is performed using a soft-max loss at the end of the network and the backpropagation algorithm [30]. During prediction, we uniformly sample body part detections for a radius of 10 pixels around the regressed body part based on the output of the ConvNet regressor. In this manner, we generate body part detections for the state space generation and our unary potentials estimation.

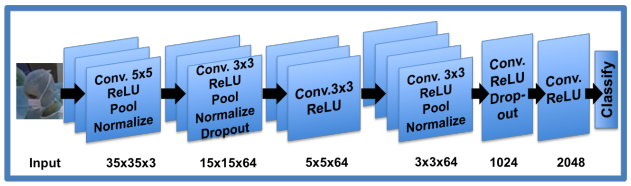


Fig. 5 Classification Network: The architecture of the ConvNet for the body part classification is relatively simple in comparison to the AlexNet [28]. However, our experiments demonstrate that the network capacity is sufficient for the body part detection in the operating room.

3.4 Potential Functions

The potential functions of the 3DPS model are particularly designed for multi-view setups. In the following, we present first the unary and then pairwise and ternary potential functions.

Unary potentials Given the generated state space, each 3D hypothesis has an average confidence that it defined by the pairs of the triangulated deep part detectors' confidence (i.e. classification ConvNet output). The detection confidence function $\phi_i^{conf}(y_i, \mathbf{x})$ is defined by the average confidence. The reprojection error $C(y_i; \mathbf{x})$ of every triangulated 3D hypothesis contributes to the reprojection error potential function that is computed as:

$$\phi_i^{repr}(y_i, \mathbf{x}) = \frac{1}{1 + \exp(C(y_i; \mathbf{x}))}. \quad (2)$$

The body part multi-view visibility potential $\phi_i^{vis}(y_i, \mathbf{x})$ accounts for the number of views in which a body part 3D hypothesis is detected. In order to compute the number of views, every body part 3D hypothesis is projected across all views and we search in a small area ($\sim$ 5 pixels radius) for body part detection. The accumulated number of visible views is normalized with respect to the total number of camera views. Eventually, the visibility potential is complementary to the reprojection error potential since it penalizes 3D hypotheses that occurred from ambiguous views or false positive detections. The last unary potential term is the temporal consistence function $\phi_i^{temp}(y_i, p_i)$, which acts as a regulariser between previously inferred body poses $\mathbf{p}$ and candidate body part 3D hypotheses. To prevent wrongly inferred body poses to influence the current hypotheses, the temporal consistence potential function has a threshold c (set to 10 cm) distance between the candidate hypotheses and the inferred body parts. Inferred poses that lie outside this radius do not contribute to the computation of the temporal consistence potential function. The temporal consistence potential function is given by:

$$\phi_i^{temp}(y_i, p_i) = \begin{cases} \frac{1}{1+\exp(d(y_i,p_i))} & \text{if } d(y_i, p_i) < c \\ \epsilon & \text{otherwise} \end{cases} \quad (3)$$

where $d(y_i, p_i)$ is the Euclidean distance between the 3D part hypothesis and previously inferred parts and ϵ a constant for numerical stability during inference.

Pairwise and ternary potentials The 3D body prior is encoded in pairwise and ternary potential functions. In detail, we model the kinematic body constraints in terms of translation and rotation between physical body parts using pairwise and ternary potentials respectively [8]. The translation potential corresponds to the translation of the part i to the local coordinate system of the part j and it is modeled with a multivariate Gaussian distribution as:

$$\psi_{i,j}^{tran}(y_i, y_j) = \mathcal{N}(y_{ij}^T \mid \mu_{ij}^T, \Sigma_{ij}^T), \quad (4)$$

where $y_{ij}^T = y_i - y_j$, μ_{ij}^T is the mean and Σ_{ij}^T is the covariance. The rotation potential function models a hinge joint (i.e. 1DoF) between two body limbs. In our problem this corresponds to the joint between the forearm and back arm. A unidimensional Gaussian distribution is used for the rotation and is given by:

$$\psi_{i,j,k}^{rot}(y_i, y_j, y_k) = \mathcal{N}(y_{ijk}^R \mid \mu_{ijk}^R, \sigma_{ijk}^R), \quad (5)$$

where $y_{i,j,k}^R = arccos(dot(y_i - y_j, y_k - y_j))$, μ_{ijk}^R is the mean and σ_{ijk}^R the variance. Finally, all type of potential functions are modelled using ground-truth information of multi-view annotated data.

Model parameters learning To learn the parameters $\mathbf{w}$ of the 3DPS model, we rely on regularised risk minimisation and use a Structured SVM (SSVM) solver [50]. The parameters $\mathbf{w}$ of the model balance the influence of the potential functions to the inference task. We follow the formulation of [8] and learn a weight for each potential function based on a set of training samples S with labels $y^s \in \{-1, 1\}$. A feature vector $\mathbf{\Phi}(\boldsymbol{\phi}^s, \boldsymbol{\psi}^s)$ with the concatenation of all potential functions is constructed for every training sample. We choose to minimize the $0-1$ loss function that is given by:

$$\min_{\mathbf{w}} \frac{1}{2} \parallel \mathbf{w} \parallel^2 + \frac{C}{S} \sum_{s=1}^{S} \xi^s$$
$$s.t. \quad \max(0, 1 - y^s \langle \mathbf{w}, \mathbf{\Phi}(\boldsymbol{\phi}^s, \boldsymbol{\psi}^s) \rangle) \leqslant \xi^s. \quad (6)$$

where ξ^s are the slack variables and C is a constant. The minimization of our energy is performed by the cutting plane algorithm [19].

3.5 3D Pose Inference

The last step for the 3D body pose recovery of different individuals is the inference. The hypotheses of each individual

lie on a separate state space and our goal is to seek for the hypotheses that maximize the posterior probability of Eq. (1) for each individual h, given as follows:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \mathbf{p}, \mathbf{h}) \tag{7}$$

where $\hat{\mathbf{y}}$ corresponds to the body pose of each individual h. To localization of each individual is performed by the human of tracker of [12]. Relying exclusively on tracking can result in drifts or mixed body parts [8], but in our problem the tracker did not fail.

The inference in the 3DPS model is performed using the max-product algorithm [13]. In addition, we profit from tracking and obtain the trajectory of each individual. To evaluate our approach, we propose the OR dataset that is captured in a real operating room.

4 OR Dataset

The operating room (OR) dataset is composed of five RGB cameras positioned in different locations of a real operating room. In Fig. 2, we show sample images from different camera views at the same time instant. The main goal of the dataset is to capture the human motion in different phases of a medical operation, in which there is active collaboration between the surgeons and staff. Note that we do not aim to recover the pose of the full body due to significant occlusions in the lower body. We aim to perform upper body 3D pose estimation of multiple individuals. As we have discussed in Section 1, the estimated body poses can contribute to the task of medical workflow modelling. Below, we provide details about the dataset formation.

Data acquisition We have mounted five GoPro® cameras on the walls of an operating room for capturing the OR dataset. The cameras are placed across the operating room wall not to interfere with the staff and also meet the sterilization requirements. Since, the GoPro® cameras do not offer an internal wired synchronization system, we have manually synchronized them after the recordings. The camera calibration has been done using the geometrical pattern of the floor [24,53]. To derive the ground-truth 2D body pose (Figure 4), we have manually annotated the image data for all camera views. Afterwards, we performed triangulation for generating the 3D body pose ground-truth. The accuracy of the annotation is around 50 millimetres (mm). In total, we performed two different recordings for creating training and testing datasets. Since the lighting is controlled in OR, the time difference between the recordings of the training and testing datasets does not have any effect on the recording environment. Finally, we performed the calibration, synchronization and annotations tasks for both recordings.

Scenario The dataset is composed of 5 individuals that interchange roles. The defined roles are two surgeons, an anaesthesiologist and two nurses. In the first recording all individuals are randomly associated to one role, while the role of each individual changes in the second recording. Hence, we create variations in the body motion for each role. In both recordings, the same medical operation is performed and thus there is a repetition in the performed actions.

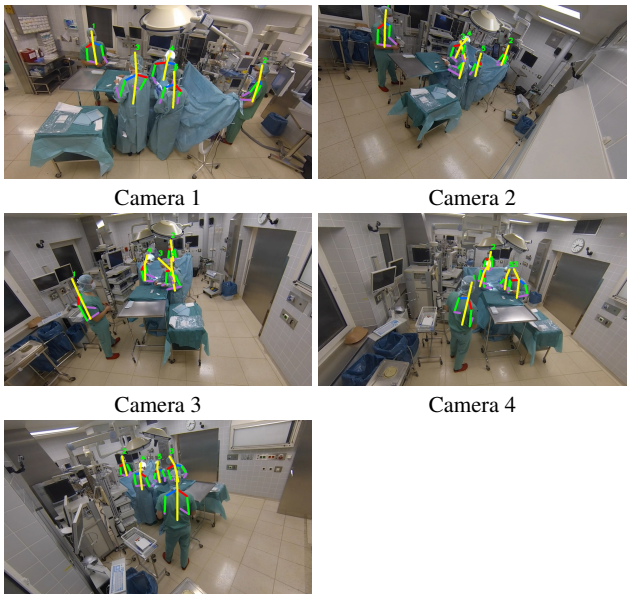


Fig. 6 Results on 3D Human Pose Estimation: Visual results of the 3D human pose estimation task are presented. The inferred 3D body poses are projected across all camera views.

Data partitioning The first recording comprises the training dataset, where a small subset of frames is used as validation dataset for hyper-parameters selection. The second recording forms the testing data. The training dataset includes 3000 images with up to 5 individuals for each camera view. Similarly, the testing dataset has up to 5 individuals in the scene, but it is composed of 4000 frames. In both cases, we provide annotation in every 10th frame. Note that the patient is a phantom. In the next section, we present the evaluation of the 3DPS model on the OR dataset for 2D and 3D human pose estimation.

5 Experiments

The task of multiple human pose estimation from multiple views has attracted notable interest recently [7,8]. In our work, we focus on the OR scenario, where the difficulty of the task increases due to the challenging environment. In this section, we evaluate 2D and 3D human pose estimation on the OR dataset. We create a comparison baseline for the OR dataset that can be used for future evaluations in this dataset.

Our model is composed of 9 body parts (Fig. 4) that model the upper body. To learn the potential functions and

parameters of the model, we use the training part of the OR dataset. At first, we train the body regression and body part detection ConvNets. The input RGB image to the networks has resolution 120×80, while the network parameters are similar to [9]. The learning rate and momentum are set to 0.01 and 0.9 respectively. The dropout is set to be 0.5 and the batch size is set to be 230 samples. Furthermore, we perform data augmentation in both classification and regression ConvNets. The initialisation of the ConvNets' parameters is done randomly using a Gaussian distribution with zero mean and standard deviation of 0.01. We learn the 3DPS model parameters and potential functions of the 3DPS model. The body prior is learned using the ground-truth data of the OR dataset.

The evaluation is divided into the following tasks: analysis of the state space, performance investigation with respect to the number of cameras, 2D human pose estimation and 3D human pose estimation. The analysis of the state space highlights how we profit from the tracking information and how we reduce the amount of computations in comparison to a global state space of all possible body part hypotheses. The examination of the performance w.r.t. to number of cameras provides an overview about the required number of cameras. In the 2D human pose estimation, we evaluate the general performance of the 2D body regression in conjunction with deep body part detector. Finally, the 3D human pose estimation evaluates the performance of our algorithm in each individual. For all evaluations, we rely on the *strict* PCP evaluation metric [39] for both 2D and 3D body pose estimation. In addition, we provide the error in millimetres (mm) for the 3D body pose results.

5.1 State Space Analysis

Given the identity of each individual obtained by our tracker, the number of body part hypotheses is significantly reduced since we do not triangulate body part detections of different individuals as in [7]. Consequently, the smaller state space accelerates the inference task that is performed in 1fps, given the body part detections. In Fig. 7, we present the number of 3D body part hypotheses versus the number of 2D joint detection samples for the OR dataset.

The number of recovered 3D hypotheses is the aggregation of the triangulation instances of all combinations of view pairs, given different number of body part detections. In the case of [7], the triangulation is performed between all individuals due to the unknown identity. As a result, their state space is much larger and the inference is computationally more expensive. It is true that our state space can result in missing body parts in case of occlusions and the missing body parts will be part of the state space of another individual. However, we did not experience this problem in practice.

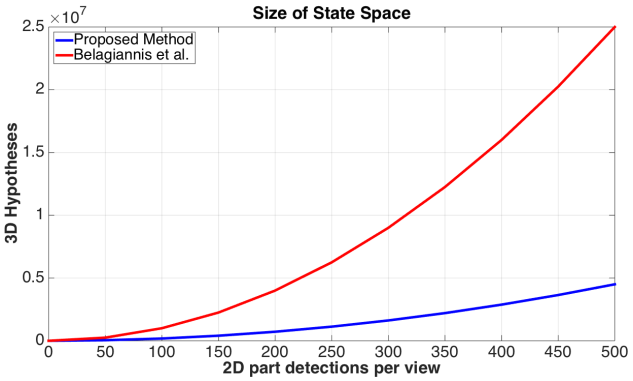


Fig. 7 State space: We show the number of recovered 3D candidates versus the number of 2D detection samples on both approaches. The number of 3D candidates is computed by summing up the triangulation instances of all combinations of all view pairs. The number of candidates is much lower than the one in [7].

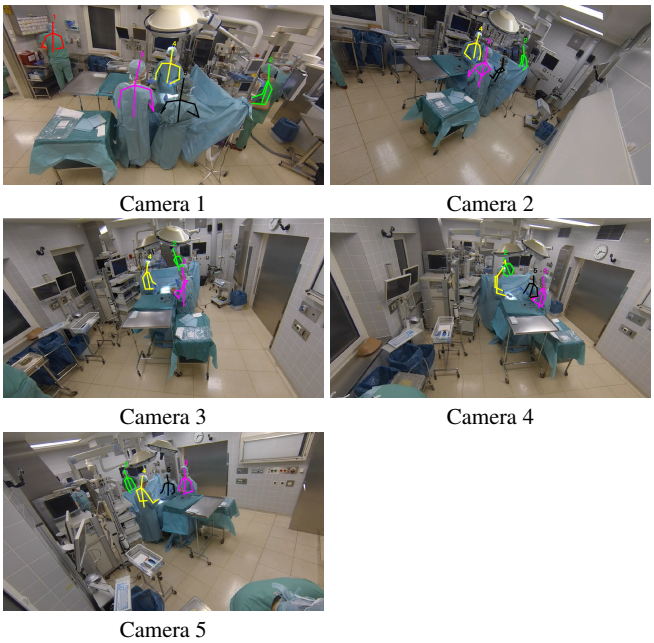


Fig. 8 More Results on 2D Human Pose Estimation: Visual results of the 2D human pose estimation task are presented. The presented results are from the same time step across all camera views.

5.2 Performance with Variable Camera Views

We examine the performance of our approach in 3D human pose estimation for different number cameras. The baseline for this experiment is defined by taking the minimum number of cameras that is two. Then gradually, we add more cameras to our framework and perform the evaluation using all combination of the available cameras. The average performance is reported in Fig. 9 for each individual separately. While adding more cameras improves performance, we observe that more than five cameras are not necessary

Table 1 2D Human Pose Evaluation: The evaluation on 2D human pose estimation is presented for each camera view. We have used the *strict* PCP performance metric. The last row summarizes the global PCP score.

	Camera 1	Camera 2	Camera 3	Camera 4	Camera 5
Head	97.90	91.67	98.15	93.85	97.66
Torso	81.76	92.89	91.46	92.08	91.66
Upper Arms	91.13	69.69	80.03	76.13	87.16
Lower Arms	50.59	48.73	46.56	42.79	51.55
Full Body	77.17	70.23	73.80	70.63	77.79
Total (global PCP)			73.70		

for our setup. Using four cameras, we already achieve good performance.

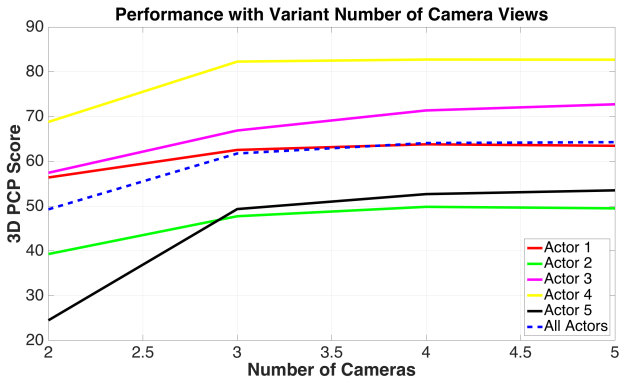


Fig. 9 Adaptive number of cameras: The 3D *strict* PCP score is presented for evaluating with different number of cameras. Each individual benefits differently by employing additional number of cameras, but in all cases adding more cameras brings additional performance.

5.3 2D Human Pose Evaluation

We evaluate the performance of the ConvNets to regress and localize the 2D body parts across each camera view. We found it important to train a different model for each camera view due to the high variance of the cameras' viewpoint. In this experiment, we focus on 2D body pose estimation of all individuals together in order to estimate a baseline for the 3D pose estimation. To this end, we estimate the PCP scores of all individuals jointly for each camera view to evaluate the regressor ConvNet. The results are summarized in Table 1. In Table 1, we observe similar performance of the body parts across the different camera views. The localization of the head and torso is quite precise for all cameras, while the lower arms are proven to be the most challenging body part to be correctly predicted. In general, the full body localization is similar for all camera views.

Moreover, we present the results of [58] in 2D human pose estimation, because it is a related approach. We have trained the model of [58] using the same data as with the ConvNets. The results are summarized in Table 2 and demonstrate the dominance of our approach. The HOG features [16] which form the base of [58] cannot capture effectively the operating room image data. Consequently, the Flexible Mixtures Parts (FMP) model [58] results in poor performance.

The error of the classification ConvNet for the body part detection is also similar for all camera views. In particular, we have 30.30% error for Camera 1 and 28.73% for Camera 5. Camera 2, 3 and 4 have slightly higher classification error that is 35.26%, 33.84% and 35.22%. We provide visual results of the 2D body pose prediction in Fig. 3 and 8.

5.4 3D Human Pose Evaluation

In this evaluation, we examine the 3D body pose results of each individual separately. We consider this evaluation as the most crucial for our approach and we summarize the results in Table 3. It is clear that the head and torso parts are the most easily inferred body parts for the individuals. On the other hand, the PCP score is low for the lower arms, as expected based on the 2D body pose results. The lower arms remain the most difficult part to infer, even with multiple camera views as input. The results on the upper arms are different between the individuals, with the Actor 4 having the best performance. In general, Actor 4 has the best results among the others, stemming from his ideal position that is well captured by Camera 1 and 5. Additionally, we provide the error in millimetres in Table 4.

Comparing the global PCP score between the 2D and 3D human pose estimation, we note that the 3D results are around 10% lower due to the higher dimensional output space. Inference in the 3D space is a more difficult and demanding task than in 2D space, but it does not result in significant lower performance. We provide visual results of the 3D human pose estimation in Fig. 1, 6 and 10.

In general, we consider our performance accurate enough for producing discriminative 3D body poses which will be valuable for the task of the medical workflow analysis. In future work, we plan to combine our approach with workflow estimation techniques [46,57].

6 Conclusion

We have introduced a unique dataset for human pose estimation that has been captured in a real operating room. The

Table 2 Comparisons in 2D Human Pose Estimation: We compare our results with the Flexible Mixture Parts (FMP) model [58] on full body 2D pose estimation (all individuals are included). The evaluation is presented for each camera view. We have used the *strict* PCP performance metric.

	Camera 1	Camera 2	Camera 3	Camera 4	Camera 5	Total (global PCP)
Our method	77.17	70.23	73.80	70.63	77.79	73.70
FMP [58]	19.40	17.23	17.78	16.98	19.69	18.03

Table 3 3D Human Pose Evaluation: The evaluation on 3D human pose estimation is presented for each individual. We have used the *strict* PCP performance metric. The last row summarizes the global PCP score.

	Actor 1	Actor 2	Actor 3	Actor 4	Actor 5
Head	84.24	73.62	94.43	97.75	89.39
Torso	84.51	84.17	89.87	98.25	82.49
Upper Arms	71.19	60.55	78.99	88.12	51.85
Lower Arms	33.97	8.79	47.09	61.75	22.81
Full Body	63.18	49.41	72.74	82.62	53.53
Total (global PCP)	64.29				

Table 4 3D Human Pose Evaluation (millimetres): The evaluation on 3D human pose estimation is presented for each individual. In this example, we present the error in millimetres (mm). The last row summarizes the total average error.

	Actor 1	Actor 2	Actor 3	Actor 4	Actor 5
Head	81.61	69.09	68.01	48.39	72.36
Torso	117.76	121.02	83.60	63.99	95.67
Upper Arms	106.50	114.48	80.82	72.83	98.18
Lower Arms	134.23	151.46	104.60	87.44	124.61
Full Body	113.47	120.33	87.07	72.15	102.27
Total (mm))	99.06				

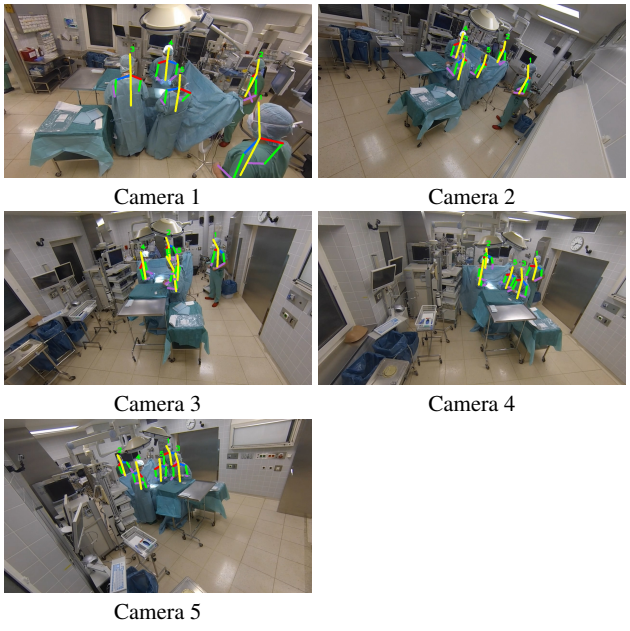


Fig. 10 More Results on 3D Human Pose Estimation: Visual results of the 3D human pose estimation task are presented. The inferred 3D body poses are projected across all camera views.

data has been acquired with a multi-view RGB camera system and simulates a medical operation using a phantom as patient. To perform the task of human pose estimation, we have presented our models for 2D and 3D inference applied on this dataset. In our evaluation, we have reported baseline score using our models and related approaches. Finally, our results demonstrate that our algorithms deliver discriminative body poses, which can be a valuable signal for surgical workflow analysis. In future work, we plan to use our body pose estimation models to support the task of predicting the phase of a medical operation.

Acknowledgements This work was supported in part by the Swiss National Science Foundation and by DFG - Deutsche Forschungsgemeinschaft under the project "Advanced Learning for Tracking and Detection in Medical Workflow Analysis". The authors would like to thank Iro Laina for helping with the data preparation.

References

1. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(1), 44–58 (2006)
2. Alahari, K., Seguin, G., Sivic, J., Laptev, I.: Pose estimation and segmentation of people in 3d movies. In: Computer Vision (ICCV), 2013 IEEE International Conference on, pp. 2112–2119. IEEE (2013)
3. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE (2008)
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 1014–1021. IEEE (2009)
5. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: Computer Vision and Pattern

Recognition (CVPR), 2010 IEEE Conference on, pp. 623–630. IEEE (2010)
6. Belagiannis, V., Amann, C., Navab, N., Ilic, S.: Holistic human pose estimation with regression forests. In: Articulated Motion and Deformable Objects, pp. 20–30. Springer (2014)
7. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures for multiple human pose estimation. In: Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on. IEEE (2014)
8. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures revisited: Multiple human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2015). DOI 10.1109/TPAMI.2015.2509986
9. Belagiannis, V., Rupprecht, C., Carneiro, G., Navab, N.: Robust optimization for deep regression. In: Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE (2015)
10. Belagiannis, V., Wang, X., Schiele, B., Fua, P., Ilic, S., Navab, N.: Multiple human pose estimation with temporally consistent 3D pictorial structures. In: Computer Vision–ECCV 2014, ChaLearn Looking at People Workshop. Springer (2014)
11. Ben Shitrit, H., Berclaz, J., Fleuret, F., Fua, P.: Multi-commodity network flow for tracking multiple people. Pattern Analysis and Machine Intelligence, IEEE Transactions on **36**(8), 1614–1627 (2014)
12. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(9), 1806–1819 (2011)
13. Bishop, C.M., et al.: Pattern recognition and machine learning, vol. 1. Springer New York (2006)
14. Burenius, M., Sullivan, J., Carlsson, S.: 3D pictorial structures for multiple view articulated pose estimation. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 3618–3625. IEEE (2013)
15. Chen, X., Yuille, A.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems (2014)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 886–893. IEEE (2005)
17. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: Computer Vision–ECCV 2010, pp. 228–242. Springer (2010)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1), 55–79 (2005)
19. Finley, T., Joachims, T.: Training structural svms when exact inference is intractable. In: Proceedings of the 25th international conference on Machine learning, pp. 304–311. ACM (2008)
20. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers **22**(1), 67–92 (1973)
21. Gammeter, S., Ess, A., Jäggli, T., Schindler, K., Leibe, B., Van Gool, L.: Articulated multi-body tracking under egomotion. In: Computer Vision–ECCV 2008, pp. 816–830. Springer (2008)
22. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 415–422. IEEE (2011)
23. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3d structure with a statistical image-based shape model. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 641–647. IEEE (2003)
24. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
25. Hofmann, M., Gavrila, D.M.: Multi-view 3d human pose estimation in complex environment. International journal of computer vision **96**(1), 103–124 (2012)
26. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Computer Vision (ICCV), 2013 IEEE International Conference on, pp. 3192–3199. IEEE (2013)
27. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3334–3342 (2015)
28. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
29. Lallemand, J., Pauly, O., Schwarz, L., Tan, D., Ilic, S.: Multitask forest for human pose estimation in depth images. In: 3DTV-Conference, 2013 International Conference on, pp. 271–278. IEEE (2013)
30. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
31. Lee, M.W., Nevatia, R.: Human pose tracking using multi-level structured models. In: Computer Vision–ECCV 2006, pp. 368–381. Springer (2006)
32. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision–ACCV 2014 (2014)
33. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(11), 2720–2735 (2013)
34. Luo, X., Berendsen, B., Tan, R.T., Veltkamp, R.C.: Human pose estimation for multiple persons based on volume reconstruction. In: Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 3591–3594. IEEE (2010)
35. Mitchelson, J.R., Hilton, A.: Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. In: BMVC, pp. 1–10 (2003)
36. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding **104**(2), 90–126 (2006)
37. Padoy, N., Blum, T., Feussner, H., Berger, M.O., Navab, N.: Online recognition of surgical activity for monitoring in the operating room. In: AAAI, pp. 1718–1724 (2008)
38. Pfister, T., Simonyan, K., Charles, J., Zisserman, A.: Deep convolutional neural networks for efficient pose estimation in gesture videos. In: Asian Conference on Computer Vision (2014)
39. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp. 588–595. IEEE (2013)
40. Plankers, R., Fua, P.: Articulated soft objects for multi-view shape and motion capture. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(10) (2003)
41. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2, pp. II–467. IEEE (2003)
42. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM **56**(1), 116–124 (2013)
43. Sigal, L., Black, M.J.: Guest editorial: state of the art in image- and video-based human pose and motion estimation. International Journal of Computer Vision **87**(1), 1–3 (2010)

44. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative density propagation for 3d human motion estimation. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 390–397. IEEE (2005)
45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
46. Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H., Navab, N.: Random forests for phase detection in surgical workflow analysis. In: Information Processing in Computer-Assisted Interventions, pp. 148–157. Springer (2014)
47. Taylor, G.W., Sigal, L., Fleet, D.J., Hinton, G.E.: Dynamical binary latent variable models for 3d human pose tracking. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 631–638. IEEE (2010)
48. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems (2014)
49. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on. IEEE (2014)
50. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. In: Journal of Machine Learning Research, pp. 1453–1484 (2005)
51. Turetken, E., Wang, X., Becker, C., Fua, P.: Detecting and tracking cells using network flow programming. arXiv preprint arXiv:1501.05499 (2015)
52. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 1, pp. I–511. IEEE (2001)
53. Wang, X.: Tracking interacting objects in image sequences. Ph.D. thesis, EPFL (2015)
54. Wang, X., Ablavsky, V., Ben Shitrit, H., Fua, P.: Take your eyes off the ball: Improving ball-tracking by focusing on team play. Computer Vision and Image Understanding **119**, 102–115 (2014)
55. Wang, X., Turetken, E., Fleuret, F., Fua, P.: Tracking interacting objects optimally using integer programming. In: ECCV, pp. 17–32 (2014)
56. Wang, X., Turetken, E., Fleuret, F., Fua, P.: Tracking interacting objects using intertwined flows. Pattern Analysis and Machine Intelligence, IEEE Transactions on (submitted)
57. Weede, O., Dittrich, F., Worn, H., Jensen, B., Knoll, A., Wilhelm, D., Kranzfelder, M., Schneider, A., Feussner, H.: Workflow analysis and surgical phase recognition in minimally invasive surgery. In: Robotics and Biomimetics (ROBIO), 2012 IEEE International Conference on, pp. 1080–1074. IEEE (2012)
58. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 1385–1392. IEEE (2011)
59. Yao, A., Gall, J., Gool, L.V., Urtasun, R.: Learning probabilistic non-linear latent variable models for tracking complex activities. In: Advances in Neural Information Processing Systems, pp. 1359–1367 (2011)
60. Zhao, T., Nevatia, R.: Tracking multiple humans in complex situations. Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(9), 1208–1221 (2004)