16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Robust 3D Human Pose Estimation Guided by Filtered Subsets of Body Keypoints

Alexandros Makris
FORTH
amakris@ics.forth.gr

Antonis Argyros
FORTH, University of Crete
argyros@ics.forth.gr

Abstract

We propose a novel hybrid human 3D body pose estimation method that uses RGBD input. The method relies on a deep neural network to get an initial 2D body pose. Using depth information from the sensor, a set of 2D landmarks on the body are transformed in 3D. Then, a multiple hypothesis tracker uses the obtained 2D and 3D body landmarks to estimate the 3D body pose. In order to safeguard from observation errors, each human pose hypothesis considered by the tracker is constructed using a gradient descent optimization scheme that is applied to a subset of the body landmarks. Landmark selection is driven by a set of geometric constraints and temporal continuity criteria. The resulting 3D poses are evaluated by an objective function that calculates densely the discrepancy between the 3D structure of the rendered 3D human body model and the actual depth observed by the sensor. The quantitative experiments show the advantages of the proposed method over a baseline that directly uses all landmark observations for the optimization, as well as over other recent 3D human pose estimation approaches.

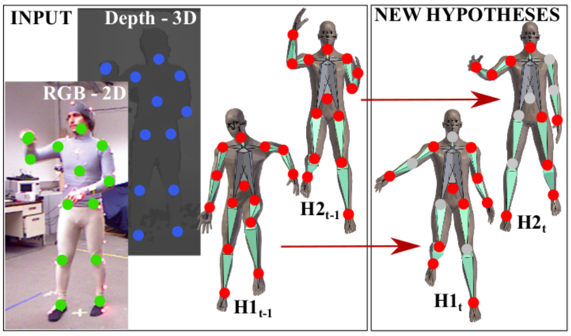

Figure 1. At each frame, the proposed method takes as input the previous pose hypotheses H_{t-1}, the 2D landmarks (green discs) extracted from the RGB image and the corresponding 3D landmarks (blue discs) calculated using depth. It then generates a set of hypotheses H_t for the current frame. For each hypothesis a different subset of the detected landmarks is used (red discs). The best hypothesis is selected by densely measuring its discrepancy from the observed depth.

1 Introduction

Vision-based human motion capture is an essential problem with many applications. Markerless unobtrusive methods have received a lot of attention from the computer vision community and considerable progress has already been achieved. However, accurate, fast and robust 3D human pose estimation in the wild is still an open problem.

1.1 Related Work

Human body pose estimation techniques may be classified into three broad classes, the bottom-up discriminative methods, the top-down generative methods and the hybrid ones. Generative methods can be very accurate, provide physically plausible solutions and do not require training. However, typically, they are computationally demanding, require initialization and can suffer from drift and track loss. Discriminative methods perform single frame pose estimation and do not require initialization. On the other hand, they rely on big collections of annotated training data and their solution is not always physically plausible. Hybrid methods integrate elements from both worlds in an effort to combine their merits.

Most recent human pose estimation methods rely on 2D keypoints extracted from RGB data [1, 2]. The accuracy of these methods is high, mainly due to the availability of large annotated datasets [3, 4]. By building on the 2D keypoints and relying on RGB information only, many recent approaches perform either 2D pose estimation [1] or 3D pose estimation [5–9]. To tackle the difficulties of lifting 2D keypoints to 3D, some methods directly regress 3D keypoints or volumetric representations [10]. Recent approaches proceed further to estimate both the pose and the shape of the human body [11–14]. In [15], they establish dense correspondences between images and the 3D human body model. The approaches that rely on RGB information only, either produce a scale normalized output or rely on prior assumptions to determine the models' scale. In both cases, their applicability in a number of domains (e.g., robotics) is limited.

To recover the full 3D human body pose in a real world coordinate frame, most approaches rely on RGBD sensors. The work in [16] that relies on random forest regression defined the baseline for these approaches. In [17,18] a generative approach is presented that relies on a single RGBD sensor and provides the full or partial body pose in real time. In [19] a deep learning approach using the depth map is presented. Instead of the depth map, the work in [20] uses a volumetric representation and a 3D CNN to obtain the hand or human pose in real-time. In [21] they embed local regions into a viewpoint invariant feature space to handle noise and occlusion. In [22] they propose a CNN approach that uses both color and depth information.

Recent body pose estimation methods use several strategies to take into account the geometric structure of the human body. Several methods describe the pose by a set of keypoints so that its structure is learnt implicitly during training [10, 20–22]. Other approaches extract the pose as a linear combination of prototype poses [19,23]. To enforce the accurate geometric structure several approaches employ 3D human body models. The pose parameters of these models are either inferred using bottom-up regression only [11,24] or estimated using a combination of bottom-up regression of body landmarks (e.g. 2D joint locations) and top down optimization [5,14,25]. In this work we employ a 3D model and we use the latter hybrid strategy to estimate its pose.

Several human pose estimation approaches focus specifically on handling occlusions. One direction is to treat visibility as a binary mask and exploit scene context to estimate it [26]. Other methods use templates for occluded versions of each body part [27] or introduce occlusion priors [28].

1.2 Our Approach

In this paper we propose a novel hybrid human body pose estimation method using RGBD input. The method relies on the *OpenPose* deep net architecture [1] to get an initial 2D pose. Using depth information from the sensor a set of 2D landmarks on the body are transformed in 3D. However, these estimations can be erroneous due to sensor errors, 2D joints miss-locations as well as because of (self) occlusions. Figure 2 shows a characteristic example of such errors. Therefore, to estimate the 3D human pose we employ a Multiple Hypothesis Tracker (MHT). Each generated hypothesis is defined by considering a subset of the available body landmarks. This way, it becomes possible to consider hypotheses that are not affected by the noisy measurements. The subsets are obtained by random weighted sampling. The weight of each landmark is calculated using a set of geometric constraints and temporal continuity criteria. Given a landmarks subset, we estimate the pose of a parametric 3D human body model by a gradient-based optimization scheme. Finally, the resulting 3D poses are evaluated by an objective function that densely calculates the discrepancy between the model and the observed depth.

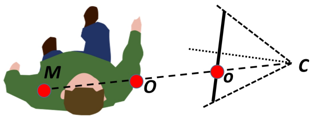

Figure 2. An RGBD camera at C observes sideways the left shoulder M of a human at image coordinates o. The perceived depth in the direction of o corresponds to O. Therefore, lifting a 2D keypoint o to 3D based on depth input will introduce outlying measurements.

Our contributions: Our main contributions are (a) a novel 3D human body tracking method based on 2D keypoint detections and RGBD data, (b) a multiple hypotheses tracking strategy to deal with occlusions, 3D sensor noise, and 2D detection errors and (c) geometric and temporal continuity rules to filter the keypoints that are fed to the optimizer for each hypothesis.

The method has been thoroughly tested quantitatively and qualitatively. The results demonstrate that the proposed method that filters keypoints using temporal and geometric constraints performs better than the baseline approach that uses all of them. Moreover, the proposed approach compares favorably to relevant state of the art approaches [16,17].

2 Method Description

2.1 Human Body Model

The employed human body model has 25 degrees of freedom (DOFs) and $K = 43$ parameters since we adopt the quaternion representation for 3D rotations. The global translation and rotation of the body is encoded by 7 parameters (3 for the 3D position and 4 for the quaternion). Each of the 9 joints are modeled using quaternions to represent the related $3D$ rotation. In this version of the model we do not enforce joint limits. We identify keypoints on the model skeleton that correspond to the locations of the joints that the 2D joint detector estimates.

2.2 3D Pose Estimation

For the purpose of 2D joint estimation we use the OpenPose method [1] which achieves state of the art results in difficult, real life datasets. This ability to generalize and produce good 2D keypoint detections from arbitrary images is key to the goals of this work.

Using the depth image and given the 2D estimations of the joints, we retrieve 3D landmark positions. Subsequently, we fit a 3D human body model to these 2D and 3D landmark positions. Under this general framework we implement two approaches:

- **LVM**: this is the baseline method where all the 2D keypoints are used to recover the 3D pose using the Levenberg-Marquardt optimization algorithm.
- **MHT**: this is the proposed optimization method based on a multiple hypothesis tracker that maintains and propagates a set of hypotheses for the 3D pose. Each hypothesis relies on a different subset of the observations (2D/3D landmark positions).

Optimization: Given a body pose $\mathbf{x}$ defined by the K parameters of the human body model and its forward kinematics function, we compute the 3D positions of the joint keypoints $M_i = (X_i, Y_i, Z_i)$, $i \in [1, I]$ in the camera coordinate frame, and their projections $m_i = (u_i, v_i)$, $i \in [1, I]$, $I = 18$, on the image plane.

Let $o_i = (u_i, v_i)$, $i \in [1, I]$, represent the detected 2D joints and f_i be a binary flag taking the value of 1 if the keypoint is detected and 0 otherwise. Let $O_i = (X_i, Y_i, Z_i)$, $i \in [1, I]$ be the 3D points associated with o_i. O_i are obtained for the detected keypoints for which there exists valid depth information. For these keypoints a binary flag F_i is set to 1; for the rest of the keypoints, F_i is set to 0. For a given pose, the total discrepancy $S(\mathbf{x}, O, o)$ between the observed and the model joints is given by:

$$S(\mathbf{x}, O, o) = \sum_{i=1}^{I} F_i \| M_i - O_i \| + (f_i - F_i) \| m_i - o_i \|. \quad (1)$$

The 3D pose x^* that is most compatible with the available observations can be estimated by minimizing the objective function of Eq.(1):

$$\mathbf{x}^* = \arg\min_{x} \{ S(\mathbf{x}, O, o) \}. \quad (2)$$

This is achieved using the Levenberg-Marquardt (LevMar) optimizer that minimizes this objective function after the automatic differentiation of the residuals. In our implementation, optimization has been performed by employing the Ceres Solver [29].

Multiple Hypothesis Tracker (MHT): MHT maintains a set of hypotheses N that are propagated through LevMar optimization using a subset of the observed keypoints. By relying on a subset of the keypoints, it is possible for the method to handle inaccurate 2D/3D estimations (see for example the situation in Fig. 2). The observation subset for each hypothesis is drawn by weighted random sampling. The 3D and 2D landmark selection weights W_i and w_i, respectively, are constructed using the following criteria:
Randomly: a ratio R, and r of selected 3D and 2D keypoints, respectively, are excluded from the optimization, so: $W_i^r = 1 - R$, $w_i^r = 1 - r$.
Model geometry: Detected 3D keypoints at step t imply body part locations. The geometry selection probability is calculated by measuring the discrepancy between the observed depth in these locations and the known model geometry: $W_i^g = \exp(-d_g/(2 * \sigma_g^2))$. $d_g \in [0, 1]$ is a normalized distance that takes into account two geometric constraints: the length between the observed keypoint i and its parent on the skeleton, and the 3D geometry of the line that connects these two keypoints. d_g is 0 when both length and line geometry comply with their respective 3D model values and increases when there is a discrepancy. For the 2D keypoints the selection probability is $w_i^g = l * W_i^g + (1 - l)$ with $l = 0.33$ so that on average for 1 out of 3 excluded 3D landmarks the corresponding 2D landmark is also excluded.
Temporal continuity: Detected 3D keypoints at step t are expected to be near their previous location: $W_i^h = \exp(-d_h/(2 * \sigma_h^2))$, where $d_h = min(||O_{i,t} - M_{i,t-1}^h||, d_{hmax})/d_{hmax}$ is the truncated normalized distance between the current detected landmark position and the corresponding model landmarks position at $t - 1$. σ_h is a standard deviation parameter. For the 2D keypoints: $w_i^h = l * W_i^h + (1 - l)$ with $l = 0.33$.
When multiple criteria are combined, the selection likelihood is given by: $W_i = min\{W_i^r, W_i^g, W_i^h\}$, $w_i = min\{w_i^r, w_i^g, w_i^h\}$. For each hypothesis the number of selected 3D and 2D landmarks are $S = \lfloor \sum_i W_i \rfloor$ and $s = \lfloor \sum_i w_i \rfloor$, respectively.

Hypothesis Evaluation: Each propagated hypothesis is evaluated using the depth observation likelihood. This measures the degree of matching between a rendered model pose and the depth observations as in [30]. The input of the method is an RGBD image and a model pose. A pre-processing step uses the estimated model position in the previous frame as reference and keeps only the observations that are within a predefined range around it. The observation consists of the resulting 2D depth and foreground maps $z = \{z_d, z_{fg}\}$. To calculate the likelihood for a hypothesis we perform rendering given the camera calibration. The result of rendering is a 2D depth map and the corresponding foreground map $\{r_d(\mathbf{x}), r_{fg}(\mathbf{x})\}$. Let P_i be the set of pixels that are labelled as foreground in both the observation and the model defined as $P_i = \{z_{fg} \wedge r_{fg}\}$ and P_u be the set of pixels that are labeled as foreground in either the model or the observation $P_u = \{z_{fg} \vee r_{fg}\}$. We denote as $\lambda = \mid P_i \mid / \mid P_u \mid$ the ratio of the number of elements of these two sets. The following function $\Delta(z, \mathbf{x})$ is then used to evaluate the discrepancy between a hypothesis $\mathbf{x}$ and the observation z:

$$\Delta(z, \mathbf{x}) = \lambda \frac{\sum_{p \in P_i} min(|z_{d,p} - r_{d,p}|, d_M)}{d_M \mid P_i \mid} + (1 - \lambda) \, . \quad (3)$$

This ranges from 0 for a perfect match to 1 for a mismatch. At each frame the hypothesis with the minimum $\Delta(z, \mathbf{x})$ is selected as the solution.

3 Experiments

We evaluated the following methods:

- **OpenNI** body pose estimation [16].
- **FHBT** Generative tracker [17]. This method reports results only for a subset of joints using a confidence metric.
- **LVM** Baseline method that performs Levenberg-Marquardt optimization over all detected joints.
- **MHT**$_r$ Proposed MHT tracker with randomly excluded keypoints.
- **MHT**$_g$ Proposed MHT tracker with randomly excluded keypoints plus geometric constraints.
- **MHT**$_h$ Proposed MHT tracker with randomly excluded keypoints plus geometric constraints plus temporal continuity (history) constraints.

Datasets: The proposed approach is evaluated both qualitatively and quantitatively. For the evaluation we used a subset of the Berkeley Multimodal Human Action Database (MHAD) [31]. This dataset features 12 human subjects of considerable variability with respect to age, size and body types. The subjects perform 11 different activities. As input we used the RGBD stream of one kinect sensor. The tracking results are compared against the ground truth obtained using a motion capture system.

Evaluation Metrics: To quantify the accuracy in body pose estimation we use two metrics. The first metric, D, is the average distance over a sequence of the estimated 3D points (joints) from their ground truth positions. The second metric, $C_j(d)$, is the joint success rate i.e. the percentage of joints in a sequence for which D is lower than d.

3.1 Quantitative Evaluation

Observation Subset Selection: The three strategies for selecting the observations subset that is used for each hypothesis have been evaluated and the results are presented in Table 1, Table 2 and Fig. 3. The best results are achieved with $\mathbf{MHT_h}$ (both geometric and temporal continuity constraints) considers 5 hypotheses and no further random exclusion of landmarks ($N = 5$, $r = 0$, $R = 0$). In this case, the mean distance error is 66 mm which is 26% lower than the baseline **LVM**. As illustrated in the plots of Fig. 3, $\mathbf{MHT_h}$ has superior success ratio for all threshold values compared to the other **MHT** variants and the **LVM** baseline.

In terms of average error, $\mathbf{MHT_h}$ only slightly outperforms **OpenNI** but its performance is much more stable as the error standard deviation is significantly lower. $\mathbf{MHT_g}$ ($N = 5$, $r = 0.05$, $R = 0.15$) which uses only geometric constraints for landmarks selection, also surpasses the baseline in all metrics. $\mathbf{MHT_r}$ ($N = 20$, $r = 0.05$, $R = 0.15$) only slightly surpasses the **LVM** baseline even with $N = 20$, showing that the use of geometric or temporal continuity information is key to the performance of **MHT** methods.

Table 1. $D(mm)$ for each exclusion ratios pair.

r	R	$\mathbf{MHT_r}$	$\mathbf{MHT_g}$	$\mathbf{MHT_h}$
0.0	0.0	N/A	85	**66**
0.0	0.05	88	82	68
0.0	0.10	87	83	70
0.0	0.15	86	82	75
0.05	0.10	87	83	70
0.05	0.15	**86**	**80**	74
0.05	0.30	88	85	83

Table 2. Mean error and standard deviation for the best performing variant of each evaluated method. ***FHBT*** reports results only for a subset of joints for wwhich the method is mostly confident, so performance is not directly comparable.

Method	Mean **D**(mm)	std **D**(mm)
LVM	89	38
$\mathbf{MHT_r}$	86	22
$\mathbf{MHT_g}$	80	29
$\mathbf{MHT_h}$	**66**	22
OpenNI	68	66
FHBT	58	41

Table 1 shows the mean distance error for different random exclusion ratios. The values that perform best for the $\mathbf{MHT_r}$ and $\mathbf{MHT_g}$ methods are $r = 0.05$ and $R = 0.15$ while $\mathbf{MHT_h}$ performs best when no random landmarks exclusion takes place since the geometric and temporal continuity criteria guarantee a good landmark subset selection.

The number of hypotheses influences both the tracking accuracy and the computational cost of the **MHT** methods. Since for the MHT_r method the keypoint exclusion is performed randomly, the error drops as N increases up to 20. Adding more hypotheses does not improve the performance of $\mathbf{MHT_r}$. $\mathbf{MHT_g}$ and $\mathbf{MHT_h}$ rely mostly on geometric and temporal continuity constraints to select keypoints, so their performance is superior to $\mathbf{MHT_r}$ even with very few hypotheses ($N = 5$) and does not change considerably for higher values of N.

Occlusions: We tested the behavior of all the methods in the presence of occlusions by randomly removing rectangular areas from the RGBD stream (see Fig. 4). The results for various occlusion levels are shown in Fig. 5. All methods are negatively influenced by the occlusions. However, the proposed approaches perform

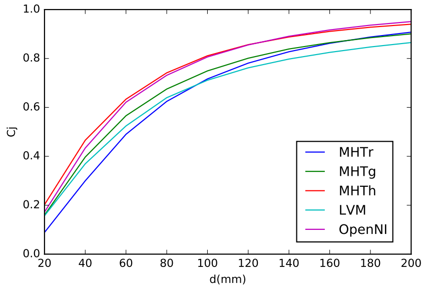

Figure 3. $C_j(d)$ for different threshold values.

Table 3. Average execution times / frame, fps.

Method	N	Exec. Times (ms)	fps
LVM	1	17	60
$\mathbf{MHT_r}$	20	262	4
$\mathbf{MHT_g}$	5	69	15
$\mathbf{MHT_h}$	5	69	15

better compared to the **LVM** baseline, which translates to higher accuracy on high occlusion levels.

Computational performance: The execution times for the methods are shown in Table 3. The current implementation is serial, so the computational cost increases linearly with the number of hypotheses. Since the computations for each hypothesis are independent, significant performance speedups are expected from their parallelization.

3.2 Qualitative Evaluation

The benefits of the proposed method are highlighted in the example frames of Fig. 6. The figure shows frames from the MHAD dataset. The baseline **LVM** method is compared to the proposed $\mathbf{MHT_h}$. The baseline method's detections are not accurate in many of the depicted frames mainly due to errors in the estimated joint $3D$ locations. The proposed $\mathbf{MHT_h}$ method is affected much less by such inaccuracies. More results are included in the supplementary material accompanying the paper: `https://youtu.be/jvLzGpnniWc`.

4 Conclusions

We presented a hybrid 3D body pose estimation method using RGBD input. The method uses a CNN to detect 2D human body landmarks and uplifts them to 3D by using the depth channel of the sensor input. To deal with noise, a robust multiple hypothesis tracker is used to evaluate and propagate pose hypotheses generated by subsets of the landmarks. The subsets are obtained by random weighted sampling, with the weight of each landmark being calculated using a set of geometric and temporal continuity criteria. The pose of a parametric 3D human body model is then estimated by a gradient based optimization scheme. The experiments show that our method significantly outperforms (in terms of accuracy) the baseline approach where all the keypoints are used in the optimization. The difference is even more significant when the sequence contains considerable occlusions. Future research directions include the incorporation of more elaborate criteria to select the relevant 2D and 3D landmarks and the simultaneous estimation of the body shape.

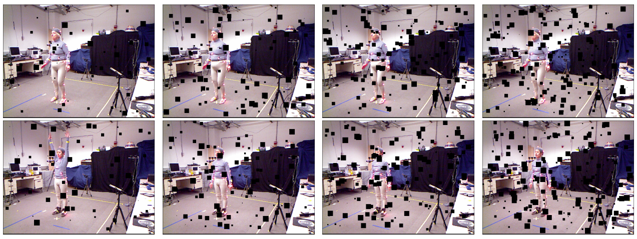

Figure 4. Example frames from the artificially occluded MHAD dataset. Each column corresponds to a different occlusion level with increasing occlusions from left to right.

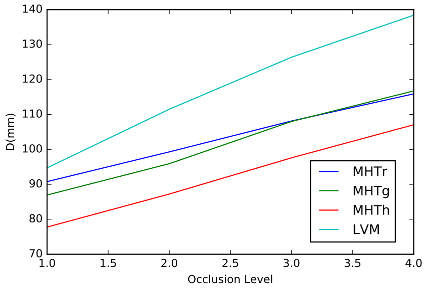

Figure 5. $D(mm)$ for different occlusion levels.

Acknowledgments: This work was supported by the H2020 EU project Co4Robots.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” in *CVPR*, 2017.

[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional Pose Machines,” *CVPR*, 2016.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” in *ECCV*, 2014.

[4] 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 Trans. on PAMI*, 2014.

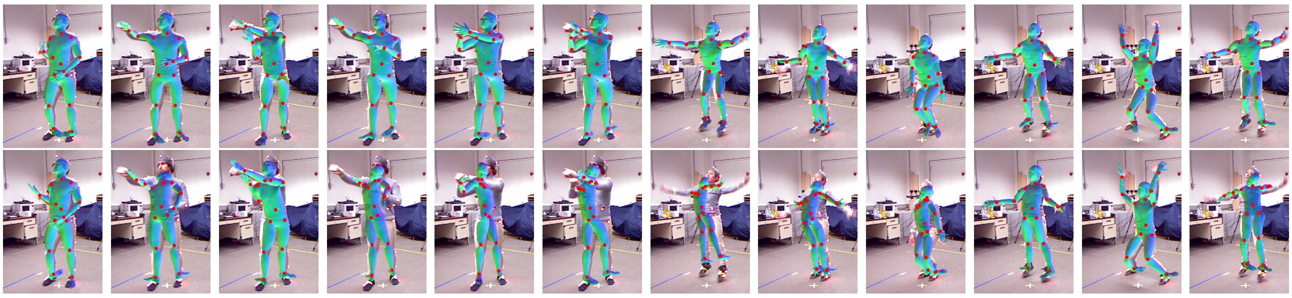

Figure 6. Results on the MHAD dataset. Proposed $\mathbf{MHT_h}$ method (top), baseline **LVM** method (bottom).

[5] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera," *TOG*, 2017.

[6] D. Tome, C. Russell, and L. Agapito, "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image," in *CVPR*, 2017.

[7] J. Martinez, M. J. Black, and J. Romero, "On Human Motion Prediction Using Recurrent Neural Networks," *CVPR*, 2017.

[8] D. Drover, R. MV, C.-H. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh, "Can 3D Pose be Learned from 2D Projections Alone?," tech. rep., 2018.

[9] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3D shape estimation from 2D landmarks: A convex relaxation approach," *CVPR*, 2015.

[10] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose," *CVPR*, 2017.

[11] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end Recovery of Human Shape and Pose," *CoRR*, 2017.

[12] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black, "Towards Accurate Marker-less Human Shape and Pose Estimation over Time," 2017.

[13] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele, "Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation," 2018.

[14] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *ECCV*, vol. 9909 LNCS, 2016.

[15] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," in *CVPR*, 2018.

[16] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, IEEE, 2011.

[17] D. Michel, A. Qammaz, and A. A. Argyros, "Markerless 3D Human Pose Estimation and Tracking based on RGBD Cameras," in *PETRA*, 2017.

[18] D. Michel and A. Argyros, "Apparatuses, methods and systems for recovering a 3-dimensional skeletal model of the human body," mar 2016.

[19] M. J. Marín-Jiménez, F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "3D human pose estimation from depth maps using a deep combination of poses," *Journal of Visual Communication and Image Representation*, aug 2018.

[20] G. Moon, J. Y. Chang, and K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map," in *CVPR*, 2017.

[21] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3D human pose estimation," in *ECCV*, vol. 9905 LNCS, 2016.

[22] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D Human Pose Estimation in RGBD Images for Robotic Task Learning," in *ICRA*, 2018.

[23] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis, "Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video," *CVPR*, 2016.

[24] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep Kinematic Pose Regression," in *ECCV Workshops*, sep 2016.

[25] A. Qammaz, D. Michel, and A. Argyros, "A Hybrid Method for 3D Pose Estimation of Personalized Human Body Models," in *WACV*, IEEE, mar 2018.

[26] T. Wang, X. He, and N. Barnes, "Learning Structured Hough Voting for Joint Object Detection and Occlusion Reasoning," in *CVPR*, IEEE, jun 2013.

[27] G. Ghiasi, Y. Yang, D. Ramanan, and C. C. Fowlkes, "Parsing Occluded People," in *CVPR*, 2014.

[28] U. Bonde, V. Badrinarayanan, and R. Cipolla, "Robust instance recognition in presence of occlusion and clutter," in *ECCV*, vol. 8690 LNCS, 2014.

[29] S. Agarwal, K. Mierle, and Others, "Ceres solver." `http://ceres-solver.org`.

[30] A. Makris, N. Kyriazis, and A. Argyros, "Hierarchical Particle Filtering for 3D Hand Tracking," in *CVPRW*, (Boston, Massachusetts), 2015.

[31] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive Multimodal Human Action Database," in *WACV*, IEEE, 2013.