MÁSTER UNIVERSITARIO
EN VISIÓN ARTIFICIAL

TRABAJO FIN DE MÁSTER

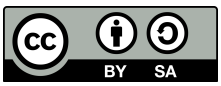
3D Human Pose Estimation for Assistive Robotics

Autor: David Pascual Hernández

Tutor: José María Cañas Plaza

Cotutor: Inmaculada Mora Jiménez

Curso académico 2019/2020



Agradecimientos

Resumen

Summary

Contents

List of Figures

List of Tables

Acronyms

CNN Convolutional Neural Network. 2, 5–8

DL Deep Learning. 2, 3, 5–8

ML Machine Learning. 1

Chapter 1

Introduction

Azul: lista de cosas que faltan por contar; Rojo: a revisar y/o falta;

In recent years, the increase in the computational capacity to collect and store data is producing a revolution in all areas of society (*e.g.* health, sports, education, finance, marketing, security or transport). Robotics is naturally involved in this revolution due to its enormous range of applications, from space to medicine, always with the aim of improving and making our lives more comfortable.

As part of this revolution, Machine Learning (ML) is also undergoing an unprecedented development, enabling many tasks to be performed in real-time, such as fraud detection, speech and face recognition or autonomous driving, just to mention a few ones [33]. In this sense, even though all areas in robotics are influenced by developments in ML, computer vision is perhaps the field which has experienced the most vertiginous growth. Furthermore, the growing interest in assistive robotics has enabled advances in some particular subfields of computer vision, such as human pose estimation, human detection and gesture recognition [30]. In particular, human pose estimation is a very handy capability for assistive robots, such as personal or home robots, as it can serve as an input for solving higher level tasks, for instance, fall detection. Broadly speaking, a precise identification of human pose provides a better understanding of the scene, which is a major requirement for any human-robot interface.

Motion-Capture systems are a well known solution for human pose estimation [**Vicon**]. They typically provide a very precise 3D estimation but require some markers in the person to be tracked. There are many technologies for human pose estimation developed for Ambient Assisted Living []. They use cameras or other sensors around the scenario to build their estimations. Typically the cameras are close to the ceiling and cover the area

to monitor, having a high point of view [**OpenPose**]. The larger the area to monitor, the greater the number of involved cameras. In order to avoid the high costs of managing several cameras, to avoid the need of any installation and to make the system more portable to many homes it is advisable to endow the home robot with the human pose estimation algorithm using only its sensors or cameras. Here the point of view of the cameras is not so high and there may be temporary occlusions, but the robot itself may move around the scenario for clearer views if needed.

The spectacular development of computer vision has been possible due to the push from Deep Learning (DL) techniques, a field of ML enabling to create complex models trained with huge amounts of examples associated to the task to be solved. Thanks to the inclusion of DL based techniques, specially Convolutional Neural Networks (CNNs), the performance of human pose estimation methods has significantly increased in the last years. This factor, in conjunction with the need for more reliable ways of understanding real scenes in applications such as video surveillance, human activity monitoring and human-computer interaction [51], has favored an increasing interest in research within this field. The great variety of tasks that can be addressed through human pose estimation has led to the development of many different approaches. Some solutions are designed considering 2D or 3D single images [13], while others take video streams into consideration [22, 63]. Regarding the output provided by the design, there are also different possibilities ranging from 2D and 3D joints locations [45] to dense estimations [1].

While the irruption of DL has had an undeniable beneficial impact in the development of the field of human pose estimation, the needed amount of real-world data for designing models achieving a good performance is still an issue, as capture and annotation processes are costly and very time consuming. This is specially true for 3D images, where motion capture systems are usually needed for collecting reliable annotations [25, 53]. As a consequence, there is a strong gap in the number of publicly available large-scale datasets in favor of the two-dimensional ones. While 2D estimations might be enough for solving some particular tasks like action recognition [32], 3D estimations are needed in order to achieve a complete visual understanding of the scene [51]. Fortunately, the emergence of robust RGBD sensors at affordable prices allows the integration of real 3D information into human pose estimation pipelines.

In this work, we propose and implement an end-to-end pipeline for augmenting 2D human pose estimations into 3D estimations in real time. This perceptive algorithm

is intended to be run on-board in many assistive or home robots: it works in real time in an off-the-shelf computer, it works with regular RGBD cameras, which are common in robotics, and when those cameras are placed at typical robot height, not close to the ceiling. We have validated it experimentally and performed a wide and detailed comparison with some of the most relevant human pose estimation algorithms, using Berkeley's Multimodal Human Action Database (BMHAD)[42], which is a publicly available and well-known international dataset.

Iterar contribuciones en función de dónde apuntemos??

Our main contributions are summarized as follows:

- We evaluate the performance and accuracy of multiple state-of-the-art methods for 2D human pose estimation;
- We propose and evaluate a straight-forward pipeline for augmenting 2D estimations into 3D estimations in real-time with the inclusion of an RGBD sensor;
- We evaluate and compare our 3D human pose estimation method with a state-of-the-art DL algorithm;
- ~~We build a web-based 3D visualization tool for human activity monitoring;~~
- We demonstrate our proposed pipeline performance with a practical use case for assistive robotics;

The rest of this report is structured as follows. In Section 2, a review of related work and state-of-the-art methods is presented. In Section 3, each step in the proposed pipeline is described in detail. The experiments carried out to evaluate the performance of this pipeline are presented in Section 4. Finally, results are discussed and summarized in Section 5.

Chapter 2

Related Work

From a computer vision perspective, the human body can be considered as a collection of rigid parts connected between them by a number of joints [22]. Following this approach, the goal of human pose estimation methods is to determine the two-dimensional or three-dimensional location of these joints and parts from an image or a sequence of images. Classical approaches usually address the modeling and estimation of these articulated poses building upon the seminal works of [19, 17] in pictorial structures and part-based models. Such is the case of Andriluka *et al.* [3], who combines strong part detectors with pictorial structures, or Yang and Ramanan [61] when proposing a mixture of non-oriented pictorial structures. Other works have further explored these methods [50, 47, 29, 58], even trying to address multi-person scenarios [14] and introducing spatio-temporal cues in video sequences [11, 16, 18, 65].

More holistic approaches started to gain popularity with the inclusion of DL techniques. In [57], Toshev and Szegedy explore the application of CNNs to estimate the body joints locations considering a cascade of Deep Neural Networks refining a coarse initial estimation. In [56], Tompson *et al.* proposed a joint training of a hybrid architecture composed of a CNN (as the part detector) with a part-based spatial model inspired on Markov Random Fields, significantly outperforming previous methods. Since these works, most research in human pose estimation has shifted towards DL-based solutions [15, 6, 31, 8, 12, 10, 60, 4, 28, 39, 24, 59]. Newell *et al.* [39] propose a novel CNN architecture for this particular task forcing *bottom-up* and *top-down* processing with intermediate supervision. In [59], Wei *et al.* inherit the *pose machines* architecture proposed by Ramakrishna *et al.* [49], but introducing CNNs as feature detectors. Their sequential architecture, based on multiple stages which take as input belief maps from the previous stages, enables learning

long-range dependencies among parts. A more general approach for spatial localization tasks is proposed by Gkioxari *et al.* [21]. Because of their intuitiveness and performance, we have chosen [39], [59] and [21] methods as 2D human pose estimators for our work. We will go deeper in these methods in Section 3. One of the most successful works published in the last years is the one proposed by Chen *et al.* [10]. In order to deal with joint occlusions and overlapping of human body parts, they propose the usage of structure-aware CNNs and Generative Adversarial Networks to train a pose generator only yielding plausible poses, implicitly learning priors about the human body structure. This approach is out of the scope of this paper because, to the best of our knowledge, there is no open source code available.

The aforementioned works are focused on 2D pose estimation from individuals on single images. However, several authors have also tried to apply DL-based techniques by taking advantage of the temporal information provided by sequences of images [27, 46, 54, 20]. Such is the case of Pfister *et al.* [46], who introduce in their pipeline dense optical flow estimations to warp estimated per-joint heatmaps in consecutive frames. In that way, the trained CNN is forced to learn the temporal relationships between sequences of poses. Another approach to human pose estimation in 2D video is proposed by Girdhar *et al.* [20]. In their work, they propose a 3D extension of the Mask Region-based CNN architecture to couple spatio-temporal information. DL-based methods have also been employed in multi-person scenarios [7, 26, 43, 23] and dense pose estimations, which map image pixels of the human body to its corresponding 3D surface [1].

Regarding the 3D estimation methods, there are different approaches depending on the source of information. Estimating 3D locations from their 2D projections is a highly ill-posed problem, as very different 3D poses can generate very similar 2D projections. However, the *a priori* knowledge about the human body and plausible poses, have allowed researchers to tackle this problem by means of hybrid approaches composed of two-stages: 2D estimation and 3D reconstruction from 2D projections. For instance, Andriluka *et al.* [2] propose a complete framework for multi-person pose detection in videos using tracking-by-detection and 3D exemplars. Regarding the 3D reconstruction from 2D projections, Ramakrishna *et al.* [48] propose an optimization proxy which jointly estimates 3D coordinates from 2D locations of anatomical landmarks and camera parameters while enforcing anthropometric regularity. Most recent methods introduce CNNs for 3D estimation. Chen and Ramanan [9] make use of the Convolutional Pose

Machiness (CPMs) proposed by Wei *et al.* [59] for 2D estimation and then match the resulting pose with a library of 3D exemplars. In [5], Bogo *et al.* estimate 2D poses with the DeepCut model proposed by Pischulin *et al.* [47] and then fit a statistical body shape model to the resulting 2D joints. However, DL-based methods have not only been used for the 2D pose estimations, but they have also proved to be a very straightforward and effective solution for inferring 3D from 2D poses. Thus, Zhou *et al.* [65] propose the inclusion of a *depth regression module* taking the 2D heatmaps generated by a CNN as input, providing a 3D pose estimation as the output. A simple but effective approach is proposed by Martinez *et al.* [36] by training a deep neural network to estimate 3D body joint locations from the corresponding 2D positions. For an adequate performance of the trained model, they consider the inverse transform of the camera to preprocess the 3D ground-truth before training, thus making the 2D to 3D problem similar across different cameras. Tome *et al.* [55] propose a more sophisticated solution which not only predicts 3D poses but also uses them to improve their previous 2D estimation, blurring the separation between the two previously mentioned stages. Several authors [40, 34, 44, 41, 37] have chosen direct inference instead of tackling the problem in these stages. Such is the case of the work proposed by Mehta *et al.* [37]. In their work, not only they infer 3D poses but also add temporal filtering and fit a kinematic skeleton model in order to take advantage of temporal correlation between frames.

The inclusion of depth measurements allows for higher accuracy and better understanding of the scene being analyzed. In [64], Youding *et al.* use video stream from a time-of-flight sensor for detecting and tracking the position of anamotical landmarks using a probabilistic inferencing algorithm. Schwarz *et al.* [52] try to solve the same task using geodesic distances and optical flow. A different approach is presented in [62], where Ye *et al.* estimate body pose by matching and refining pre-captured exemplars. Recently, DL-based solutions have been proposed to address pose estimation from depth maps. Marín-Jimenez *et al.* [35] train a CNN estimating 3D body poses as a linear combination of prototype poses. Moon *et al.* [38] introduce the novelty of designing their model as a 3D CNN, which estimates the likelihood per voxel for each joint in the pose. They test their model performance not only for addressing human pose estimation, but also hand pose. In [66], Zimmermann *et al.* use a 2D keypoint detector to estimate the 2D pose, which is next used together with the depth map as input to train a CNN yielding predictions of real world 3D coordinates. These 3D predictions are used for robotic tasks

learning.

Chapter 3

Proposed method

Chapter 4

Experiments

Chapter 5

Conclusions

Bibliography

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2018, pp. 7297–7306.

[2] M Andriluka, S Roth, and B Schiele. "Monocular 3D pose estimation and tracking by detection". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. ieeexplore.ieee.org, June 2010, pp. 623–630.

[3] M Andriluka, S Roth, and B Schiele. "Pictorial structures revisited: People detection and articulated pose estimation". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. June 2009, pp. 1014–1021.

[4] V Belagiannis and A Zisserman. "Recurrent Human Pose Estimation". In: *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. ieeexplore.ieee.org, May 2017, pp. 468–475.

[5] Federica Bogo et al. "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image". In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 561–578.

[6] Adrian Bulat and Georgios Tzimiropoulos. "Human Pose Estimation via Convolutional Part Heatmap Regression". In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 717–732.

[7] Zhe Cao et al. *Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields*. 2017.

[8] Joao Carreira et al. "Human pose estimation with iterative error feedback". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. openaccess.thecvf.com, 2016, pp. 4733–4742.

[9] Ching-Hang Chen and Deva Ramanan. "3d human pose estimation= 2d pose estimation+ matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 7035–7043.

[10] Yu Chen et al. “Adversarial posenet: A structure-aware convolutional network for human pose estimation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. openaccess.thecvf.com, 2017, pp. 1212–1221.

[11] Anoop Cherian et al. “Mixing body-part sequences for human pose estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2353–2360.

[12] Xiao Chu et al. “Multi-context attention for human pose estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 1831–1840.

[13] Qi Dang et al. “Deep learning based 2d human pose estimation: A survey”. In: *Tsinghua Science and Technology* 24.6 (2019), pp. 663–676.

[14] Marcin Eichner and Vittorio Ferrari. “We Are Family: Joint Pose Estimation of Multiple Persons”. In: *Computer Vision – ECCV 2010*. Springer Berlin Heidelberg, 2010, pp. 228–242.

[15] Xiaochuan Fan et al. “Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. openaccess.thecvf.com, 2015, pp. 1347–1355.

[16] A Fathi and G Mori. “Human Pose Estimation using Motion Exemplars”. In: *2007 IEEE 11th International Conference on Computer Vision*. ieeexplore.ieee.org, Oct. 2007, pp. 1–8.

[17] Pedro F Felzenszwalb et al. “Object detection with discriminatively trained part-based models”. en. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.9 (Sept. 2010), pp. 1627–1645.

[18] V Ferrari, M Marin-Jimenez, and A Zisserman. “Progressive search space reduction for human pose estimation”. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. ieeexplore.ieee.org, June 2008, pp. 1–8.

[19] Martin A Fischler and Robert A Elschlager. “The representation and matching of pictorial structures”. In: *IEEE Trans. Comput.* 1 (1973), pp. 67–92.

[20] Rohit Girdhar et al. “Detect-and-track: Efficient pose estimation in videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2018, pp. 350–359.

[21] Georgia Gkioxari, Alexander Toshev, and Navdeep Jaitly. “Chained Predictions Using Convolutional Neural Networks”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 728–743.

[22] Wenjuan Gong et al. “Human Pose Estimation from Monocular Images: A Comprehensive Survey”. en. In: *Sensors* 16.12 (Nov. 2016).

[23] Eldar Insafutdinov et al. “Arttrack: Articulated multi-person tracking in the wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 6457–6465.

[24] Eldar Insafutdinov et al. “DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 34–50.

[25] Catalin Ionescu et al. “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[26] Umar Iqbal, Anton Milan, and Juergen Gall. “Posetrack: Joint multi-person pose estimation and tracking”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 2011–2020.

[27] Arjun Jain et al. “MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation”. In: *Computer Vision – ACCV 2014*. Springer International Publishing, 2015, pp. 302–315.

[28] Lipeng Ke et al. “Multi-scale structure-aware network for human pose estimation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. openaccess.thecvf.com, 2018, pp. 713–728.

[29] Martin Kiefel and Peter Vincent Gehler. “Human Pose Estimation with Fields of Parts”. In: *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 331–346.

[30] Marco Leo et al. “Deep learning for assistive computer vision”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 0–0.

[31] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. “Human Pose Estimation Using Deep Consensus Voting”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 246–260.

[32] Mengyuan Liu and Junsong Yuan. “Recognizing human actions as the evolution of pose estimation maps”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1159–1168.

[33] Weibo Liu et al. “A survey of deep neural network architectures and their applications”. In: *Neurocomputing* 234 (2017), pp. 11–26.

[34] Diogo C Luvizon, David Picard, and Hedi Tabia. “2d/3d pose estimation and action recognition using multitask deep learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2018, pp. 5137–5146.

[35] Manuel J Marın-Jiménez et al. “3D human pose estimation from depth maps using a deep combination of poses”. In: *J. Vis. Commun. Image Represent.* 55 (Aug. 2018), pp. 627–639.

[36] Julieta Martinez et al. “A simple yet effective baseline for 3d human pose estimation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. openaccess.thecvf.com, 2017, pp. 2640–2649.

[37] Dushyant Mehta et al. “VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera”. In: *ACM Trans. Graph.* 36.4 (July 2017), 44:1–44:14.

[38] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. “V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2018, pp. 5079–5088.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked Hourglass Networks for Human Pose Estimation”. In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 483–499.

[40] A Nibali et al. “3D Human Pose Estimation With 2D Marginal Heatmaps”. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. ieeexplore.ieee.org, Jan. 2019, pp. 1477–1485.

[41] B X Nie, P Wei, and S Zhu. “Monocular 3D Human Pose Estimation by Predicting Depth on Joints”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. ieeexplore.ieee.org, Oct. 2017, pp. 3467–3475.

[42] Ferda Ofli et al. “Berkeley mhad: A comprehensive multimodal human action database”. In: *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. IEEE. 2013, pp. 53–60.

[43] George Papandreou et al. “Towards accurate multi-person pose estimation in the wild”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4903–4911.

[44] Georgios Pavlakos et al. “Coarse-to-fine volumetric prediction for single-image 3D human pose”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 7025–7034.

[45] Xavier Perez-Sala et al. “A survey on model based approaches for 2D and 3D visual human pose recovery”. In: *Sensors* 14.3 (2014), pp. 4189–4210.

[46] Tomas Pfister, James Charles, and Andrew Zisserman. “Flowing convnets for human pose estimation in videos”. In: *Proceedings of the IEEE International Conference on Computer Vision*. openaccess.thecvf.com, 2015, pp. 1913–1921.

[47] L Pishchulin et al. “Poselet Conditioned Pictorial Structures”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. June 2013, pp. 588–595.

[48] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. “Reconstructing 3D Human Pose from 2D Image Landmarks”. In: *Computer Vision – ECCV 2012*. Springer Berlin Heidelberg, 2012, pp. 573–586.

[49] Varun Ramakrishna et al. “Pose Machines: Articulated Pose Estimation via Inference Machines”. In: *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 33–47.

[50] B Sapp, C Jordan, and B Taskar. “Adaptive pose priors for pictorial structures”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. June 2010, pp. 422–429.

[51] Nikolaos Sarafianos et al. “3d human pose estimation: A review of the literature and analysis of covariates”. In: *Computer Vision and Image Understanding* 152 (2016), pp. 1–20.

[52] L A Schwarz et al. “Estimating human 3D pose from Time-of-Flight images based on geodesic distances and optical flow”. In: *Face and Gesture 2011*. ieeexplore.ieee.org, Mar. 2011, pp. 700–706.

[53] Leonid Sigal, Alexandru O Balan, and Michael J Black. “Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion”. In: *International journal of computer vision* 87.1-2 (2010), p. 4.

[54] Jie Song et al. “Thin-slicing network: A deep structured model for pose estimation in videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 2017, pp. 4220–4229.

[55] Denis Tome, Chris Russell, and Lourdes Agapito. “Lifting from the deep: Convolutional 3d pose estimation from a single image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2500–2509.

[56] Jonathan J Tompson et al. “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z Ghahramani et al. Curran Associates, Inc., 2014, pp. 1799–1807.

[57] Alexander Toshev and Christian Szegedy. “Deeppose: Human pose estimation via deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. openaccess.thecvf.com, 2014, pp. 1653–1660.

[58] Fang Wang and Yi Li. “Beyond physical connections: Tree models in human pose estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 596–603.

[59] Shih-En Wei et al. “Convolutional pose machines”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. cv-foundation.org, 2016, pp. 4724–4732.

[60] Wei Yang et al. “Learning feature pyramids for human pose estimation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. openaccess.thecvf.com, 2017, pp. 1281–1290.

[61] Y Yang and D Ramanan. “Articulated pose estimation with flexible mixtures-of-parts”. In: *CVPR 2011*. June 2011, pp. 1385–1392.

[62] M Ye et al. “Accurate 3D pose estimation from a single depth image”. In: *2011 International Conference on Computer Vision*. ieeexplore.ieee.org, Nov. 2011, pp. 731–738.

[63] Mao Ye et al. “A survey on human motion analysis from depth data”. In: *Time-of-flight and depth imaging. sensors, algorithms, and applications*. Springer, 2013, pp. 149–187.

[64] Youding Zhu, B Dariush, and Kikuo Fujimura. “Controlled human pose estimation from depth image streams”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. ieeexplore.ieee.org, June 2008, pp. 1–8.

[65] Dong Zhang and Mubarak Shah. “Human pose estimation in videos”. In: *Proceedings of the IEEE International Conference on Computer Vision*. cv-foundation.org, 2015, pp. 2012–2020.

[66] C Zimmermann et al. “3D Human Pose Estimation in RGBD Images for Robotic Task Learning”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. ieeexplore.ieee.org, May 2018, pp. 1986–1992.