Universidad
Rey Juan Carlos

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

DL Deep Learning. 2, 3

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 [7]. 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 [5]. 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 [10], has favored an increasing interest in research within this field. The great variety of tasks that can be addressed through human pose estimation has lead to the development of many different approaches. Some solutions are designed considering 2D or 3D single images [2], while others take video streams into consideration [3, 12]. Regarding the output provided by the design, there are also different possibilities ranging from 2D and 3D joints locations [9] 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 [4, 11]. 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 [6], 3D estimations are needed in order to achieve a complete visual understanding of the scene [10]. 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)[8], 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;

Esta parte, queda pendiente. Hacerla cuando tengamos la estructura completa The rest of this paper is structured as follows. In Section **??**, a review of related work and state-of-the-art methods is presented. In Section **??**, 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 **??**. Finally, results are discussed and summarized in Section **??**.

Chapter 2

Related Work

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

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.