Real-Time Human Head Imitation for Humanoid Robots

Dario Cazzato
Interdisciplinary Centre for Security,
Reliability and Trust,
Université du Luxembourg
29, Avenue J.F Kennedy,
L-1855 Luxembourg
dario.cazzato@uni.lu

Claudio Cimarelli
Interdisciplinary Centre for Security,
Reliability and Trust,
Université du Luxembourg
29, Avenue J.F Kennedy,
L-1855 Luxembourg
claudio.cimarelli@uni.lu

Jose Luis Sanchez-Lopez
Interdisciplinary Centre for Security,
Reliability and Trust,
Université du Luxembourg
29, Avenue J.F Kennedy,
L-1855 Luxembourg
miguel.olivaresmendez@uni.lu

Miguel A. Olivares-Mendez
Interdisciplinary Centre for Security,
Reliability and Trust,
Université du Luxembourg
29, Avenue J.F Kennedy,
L-1855 Luxembourg
joseluis.sanchezlopez@uni.lu

Holger Voos
Interdisciplinary Centre for Security,
Reliability and Trust,
Université du Luxembourg
29, Avenue J.F Kennedy,
L-1855 Luxembourg
holger.voos@uni.lu

ABSTRACT

The ability of the robots to imitate human movements has been an active research study since the dawn of the robotics. Obtaining a realistic imitation is essential in terms of perceived quality in human-robot interaction, but it is still a challenge due to the lack of effective mapping between human movements and the degrees of freedom of robotics systems. If high-level programming interfaces, software and simulation tools simplified robot programming, there is still a strong gap between robot control and natural user interfaces. In this paper, a system to reproduce on a robot the head movements of a user in the field of view of a consumer camera is presented. The system recognizes the presence of a user and its head pose in real-time by using a deep neural network, in order to extract head position angles and to command the robot head movements consequently, obtaining a realistic imitation. At the same time, the system represents a natural user interface to control the Aldebaran NAO and Pepper humanoid robots with the head movements, with applications in human-robot interaction.

CCS Concepts

• Human-centered Computing → Human computer interaction (HCI) • Computer systems organization → Embedded and cyber-physical systems → Robotics

Keywords

Human-robot interaction; natural user interface; head pose estimation; behavior generation.

1. INTRODUCTION

The possibility to have robots imitating human behaviour is a key topic and an active research field. Applications range from affective robotics to gaming, from autism spectrum disorder to realistic character modelling in simulation scenarios, from socially assistive robotics to entertainment [24, 25, 26]. A robot able to imitate emotions and/or basic movements also presents advantages in affective human-robot interaction (HRI), not only in terms of acceptance level but also since the robot can become more expressive, eliciting responses and actively modifying the user's emotional state [2]. Moreover, a precise human imitation can automatically represent a natural user interface (NUI) to control the robot, a very active topic in the state of the art [1, 23]. In light of this, it is not surprising that many works focused on making human-robot interactions more natural and the robot more socially and contextually aware [4].

Marker-based capture systems are typically employed to observe human motion because of their reliability. They work by attaching to the human operator reflective patches that are precisely tracked over time, usually by a multi-camera system. At this aim, the work of [1] presents a robot that imitates a human dancer whose movements have been extracted from a motion capture system. A method to reproduce realistic motions by mapping their three-dimensional appearance from a human performer to the android has been proposed in [7], again by employing a motion capture system for the perception. Other works that use the same technology to animate an Aldebaran NAO robot can be found in [5, 6]. If they can provide a very precise and reliable solution, such systems are very costly. Moreover, many systems limit their usage to indoor setups or require a tedious calibration procedure thus, in many specific application contexts, a computer vision based system could be more desirable [8].

Stereo-vision system for the ARMAR-IIIb robot has been employed in [11]. A real-time human imitation system based on non-invasive image processing techniques has been proposed in [3], but authors use input coming from RGBD images. A Microsoft Kinect has been used as optical motion capture sensor for arm control in [22,27]. Head pose angles have been estimated using the Kinect for a teleoperation scenario for the Furhat robot head in [28], while a learning scenarios for people with autism spectrum disorder has been proposed in [29]. Similarly, a Kinect has been employed to extract user facial expression and 3D head pose in order to reproduce both of them on a robotic head (Muecas) [9], to estimate hand shape and orientation for object grasping (with two additional force sensors), or to replicate a full body control (on the DARwIn-OP robot) [10]. Apart than robotics applications, this RGBD sensor has been massively used for face analysis and human-machine interaction studies [30, 31, 32, 33], but a constraint in terms of hardware could be removed if a simple RGB sensor is employed.

This work represents an attempt to fill the gap between the wide literature on robot control for human imitation and recent advances in pose detection with deep neural networks operating by processing only RGB images [12]. In particular, a real-time human head imitation system for the Aldebaran NAO and Pepper robots is proposed. The system processes images coming from a consumer webcam in order to extract the user 3D head pose. Yaw and pitch angles are the input for an imitation module that can directly move the head of the robots. The system has been implemented in both simulation and real scenario, showing reliable and real-time performance. At the same time, the system represents a NUI to control the two robots with the head movements, opening to several HRI scenarios. The rest of the manuscript is organized as follows: in Section 2, the system is described. In Section 3 experiments obtained in both simulation and real environment are shown and discussed, while Section 4 has the conclusion.

2. SYSTEM DESCRIPTION

In Fig. 1, a block diagram of the proposed system is shown. In particular, our system is composed of two modules opportunely coupled: the first one is the head pose estimation system that processes images coming from a consumer webcam and output the 3D user head pose. The second component is the imitation module that is responsible to transfer the command movements to a physical or simulated robot. In order to simplify the communication, this block directly dialogues with the NAOqi. Next subsections will describe the components in details.

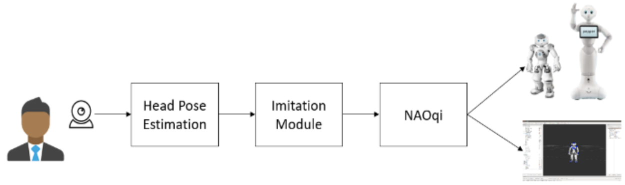

Figure 1. A block diagram of the proposed system.

Head Pose Estimation

Head pose estimation is the problem of estimating the three degrees of freedom of a human head, referred in the literature as yaw, pitch and roll [18] (see Fig. 2). In the proposed system, the presence of the face in each input image is detected by using a pre-trained deep learning module with reduced ResNet-10 SSD, a deep residual network [13].

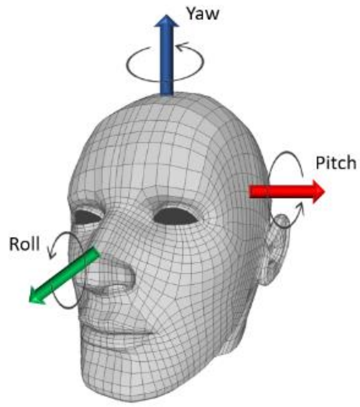

Figure 2. Head pose angles representation.

Introducing this preliminary step gave a double advantage in terms of less false detection and misdetections at the same time. The image is cropped in correspondence of the face position and the patch is processed by OpenFace [15], an open source tool designed for a complete facial behavior analysis that provides not only facial landmark detection and head pose estimation, but also facial action unit recognition and gaze estimation. OpenFace works by computing, first of all, the 2D position of 68 facial landmarks that are detected and tracked by using Conditional Local Neural Fields (CLNF) [17], a probabilistic model that can learn non-linear and spatial relationships between the input pixels and the probability of a landmark being aligned, furtherly optimised with a Non-uniform Regularised Landmark Mean-Shift technique. Refer to [16] for more details. The 3D position of the head with regards to the camera reference system is estimated by employing the iterative Perspective-n-Point algorithm based on Levenberg-Marquardt optimization [14]. In particular, knowing the 2D-3D correspondences and the camera intrinsic calibration matrix K, defined as:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

for each correspondence between image plane points (subscript IP) and 3D points (subscript 3D) we have:

$$s \begin{bmatrix} x_{IP} \\ y_{IP} \\ 1 \end{bmatrix} = K \begin{bmatrix} r_{11} r_{21} r_{31} t_1 \\ r_{21} r_{22} r_{23} t_2 \\ r_{31} r_{23} r_{33} t_3 \end{bmatrix} \begin{bmatrix} x_{3D} \\ y_{3D} \\ z_{3D} \\ 1 \end{bmatrix}$$

Knowing different correspondences led to the overdetermined system whose least square solution represents the 6-DOF pose under consideration.

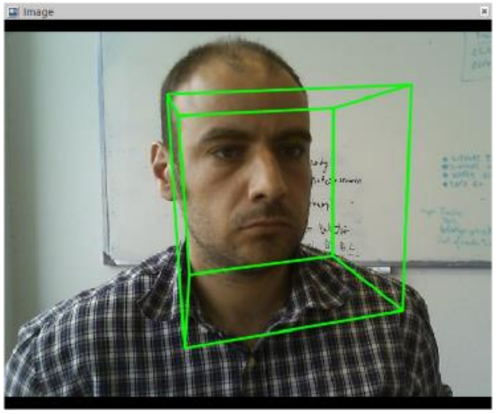

Figure 3. An output of the Head Pose Estimation module.

An output of the head pose estimation on a user image is shown in Fig. 3, where a cube orientated with yaw, pitch and roll angles as the user head pose has been drawn.

Imitation Module and NAOqi

This component is responsible for converting the user head pose angles in a command that properly moves the robot head. It has been implemented as a ROS node [19] to ease the communication with the NAOqi, a distributed object framework that encapsulates robot functionalities, giving a programming interface to communicate with different sensors and actuators. In particular, a proxy to *ALMotion* module that provides methods that facilitate moving the robot has been employed. For each published ROS message, interpolation of yaw and pitch head joints are moved to a target angles. In fact, the imitation module directly interfaces with the actuator, i.e. the physical NAO, a Pepper robot, or their version in a simulation environment. About the latter, the complete models (URDF) for the robots and the NAOqi have been used to represent a full and realistic body control in the Rviz environment. Finally, in both simulation and real scenarios, a smoothing on the robot head movements can be obtained by reducing maximum motors speed.

Aldebaran-Robotics NAO and Pepper

Aldebaran-Robotics NAO is a humanoid robot with 5 DOF joints [20], while Pepper has 20 DOF joints. In both cases, the head is able to rotate on both yaw and pitch axes. NAO has multiple sensors and controllers, in particular, head and jaw cameras, chest sonar sensors, movement motors on neck, hands and feet, color LEDs on the eyes and the tactile sensors on the head and feet [21]. In the case of the Pepper, also a tablet on the chest, 3D depth sensors behind the eyes and six laser sensors on the legs (that end on a moving platform) are present.

3. EXPERIMENTAL SETUP AND RESULTS

First of all, as an additional motivation for adopting the chosen head pose estimator, an analysis of errors in the yaw and pitch angles of various state of the art methods on the publicly available Biwi Kinect [33] and BU [36] datasets is reported in Tab. 1. The first dataset contains RGBD data, while the second one only RGB. For RGB data, the methods proposed in [34] and [35] have also been reported. It can be observed that OpenFace [15] can even outperforms methods based on RGBD data. Note that only yaw and pitch angles are compared, since they represent the angle of interest for controlling the robot head.

Table 1. Comparison of head pose estimation errors (in degrees and for yaw and pitch angles).

Method	Biwi Kinect (Yaw/Pitch)	BU (Yaw/Pitch)
Fanelli *et al.* [33]	9.2/8.5	-/-
Saragih *et al.* [34]	8.2/8/2	3.0/3.5
Asthana *et al.* [35]	13.9/14.7	3.8/4.6
Baltrušaitis *et al.* [15]	**7.9./5.6**	**2.8/3.3**

For the experimental assessment, two qualitative scenarios have been prepared. In the first one, a Rviz simulation environment with a NAO robot has been designed; afterward, a Pepper has been used to test the system in real scenarios. Seven different human users were asked to sit in front of a consumer webcam with a distance between 50-70 cm from the sensor. Users were different in appearance and in terms of hairstyle, beard, eyeglasses, etc., without any given constraint. Fig. 4 reports the employed experimental setup, where is visible one of the seven participants sitting down in front of a PC with a webcam on the top of the screen and a Pepper robot few meters ahead. During this experimental session, a free robot imitation scenario has been created placing the robot sit in front of the user while facing at the same direction. An evaluation of the interaction quality has been asked to the users, in terms of realism of the robot simulation. The feedback shows that the overall results are very encouraging, and provide realistic and precise head movements from the robots.

Two examples of interaction during the experiments with both simulated and real robot are shown in Fig. 5 and Fig. 6: at the left, the output of the head pose estimation, while at the right, the simulated or the real robot imitating the human head pose. A video with a summary of the experiments is available at https://youtu.be/HJnpwOnZcJA.

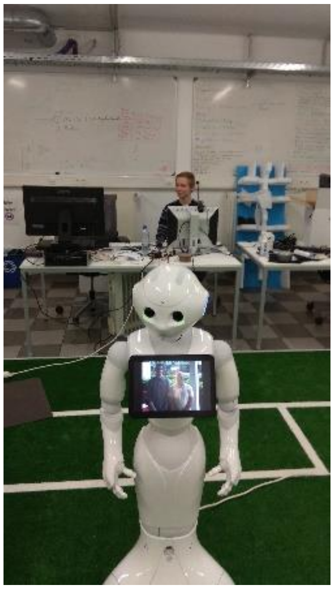

Figure 4. The employed experimental setup.

As it can be observed, the interaction is in general very fluid and the robot can imitate the user head position in real-time. Only in some case, when the yaw angle approaches ±90° degrees or the head is completely facing up or down, the estimation is less precise so that the robot cannot precisely follow the user.

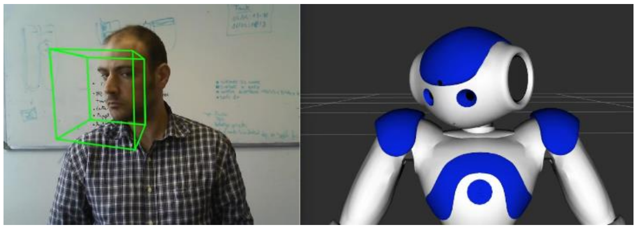

Figure 5. An output of the experiments: result of the head pose estimation (left), and the simulated NAO robot imitating the human head pose (right).

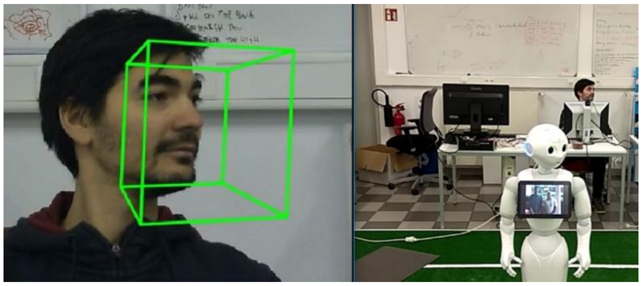

Figure 6. An output of the experiments: result of the head pose estimation (left), and the Pepper robot imitating the human head pose (right).

About implementation details, a ROS [19] node has been created to ease the communication between the head pose estimation algorithm and the robot imitation system respectively. These nodes have been programmed in Python programming language. Head pose estimation has been executed on a PC with Intel Xeon CPU E3-1505M v6 3.00GHz processor, with 32GB of RAM and NVIDIA Quadro M1200 GPU. In the case of simulation, all the system has been executed on the same machine. For the experiments with the real robot, the head pose estimation and the imitation modules have been executed in the aforementioned machine. Thus, the imitation module directly communicates with the NAOqi in order to move the physical robot.

The head pose estimation module represents our potential bottleneck, but it can produce an estimate at more than 33 fps with input images of resolution of 640X480. In the light of this, during the simulation scenario, the delay between an estimated head pose and the robot movement was negligible. Operating in the same LAN, also working with a physical robot led to robot movements without any perceivable delay. Instead, robots engines have been slowed down of 70% in order to provide a smooth interaction.

4. CONCLUSION

In this work, a system to reproduce the head movements of a user in the field of view of a consumer camera has been presented. The proposed work tried to fill the gap between methods for human imitation from robots and state of the art deep neural networks. The latter is used to estimate the user head pose in real-time, and the proposed system can directly transmit the head pose angles to lead the robot head movements. Obtained results show that the system represents a potential natural user interface to control the NAO and Pepper robots with the head movements, as well as a human head imitation system for the two humanoids robot. Future work will investigate the possibility of employing the obtained results to realize an assistive application, thanks to the possibility to remote control the robot and accessing to his cameras. Moreover, integration of a human skeleton tracker from RGB images in order to realize a full humanoid body control interface and a complete imitation system.

5. REFERENCES

[1] Pollard, N. S., Hodgins, J. K., Riley, M. J., & Atkeson, C. G. (2002). Adapting human motion for the control of a humanoid robot. In Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292) (Vol. 2, pp. 1390-1397). IEEE.

[2] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), 39-58.

[3] Ou, Y., Hu, J., Wang, Z., Fu, Y., Wu, X., & Li, X. (2015). A real-time human imitation system using kinect. International Journal of Social Robotics, 7(5), 587-600.

[4] Han, J., Campbell, N., Jokinen, K., & Wilcock, G. (2012, December). Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 679-683). IEEE.

[5] Koenemann, J., & Bennewitz, M. (2012, March). Whole-body imitation of human motions with a nao humanoid. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 425-426). ACM.

[6] Thobbi, A., & Sheng, W. (2010, December). Imitation learning of arm gestures in presence of missing data for humanoid robots. In 2010 10th IEEE-RAS International Conference on Humanoid Robots (pp. 92-97). IEEE.

[7] Matsui, D., Minato, T., MacDorman, K. F., & Ishiguro, H. (2018). Generating natural motion in an android by mapping human motion (pp. 57-73). Springer Singapore.

[8] Neunert, M., Bloesch, M., & Buchli, J. (2016, July). An open source, fiducial based, visual-inertial motion capture system. In 2016 19th International Conference on Information Fusion (FUSION) (pp. 1523-1530). IEEE.

[9] Cid, F., Moreno, J., Bustos, P., & Núnez, P. (2014). Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. Sensors, 14(5), 7711-7737.

[10] Lee, J. H. (2012, December). Full-body imitation of human motions with kinect and heterogeneous kinematic structure of humanoid robot. In 2012 IEEE/SICE International Symposium on System Integration (SII) (pp. 93-98). IEEE.

[11] Do, M., Azad, P., Asfour, T., & Dillmann, R. (2008, December). Imitation of human motion on a humanoid robot using non-linear optimization. In Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots (pp. 545-552). IEEE.

[12] Liu, Z., Zhu, J., Bu, J., & Chen, C. (2015). A survey of human pose estimation: the body parts parsing based methods. Journal of Visual Communication and Image Representation, 32, 10-19.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[14] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6), 381-395.

[15] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-10). IEEE.

[16] Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6.

[17] Baltrusaitis, T., Robinson, P., & Morency, L. P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 354-361).

[18] Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. IEEE transactions on pattern analysis and machine intelligence, 31(4), 607-626.

[19] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., ... & Ng, A. Y. (2009, May). ROS: an open-source Robot Operating System. In ICRA workshop on open source software (Vol. 3, No. 3.2, p. 5).

[20] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., ... & Maisonnier, B. (2009, May). Mechatronic design of NAO humanoid. In 2009 IEEE International Conference on Robotics and Automation (pp. 769-774). IEEE.

[21] Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., ... & Maisonnier, B. (2008). The nao humanoid: a combination of performance and affordability. CoRR abs/0807.3223.

[22] Rodriguez, I., Astigarraga, A., Jauregi, E., Ruiz, T., & Lazkano, E. (2014, November). Humanizing NAO robot teleoperation using ROS. In 2014 IEEE-RAS International Conference on Humanoid Robots (pp. 179-186). IEEE.

[23] Reddivari, H., Yang, C., Ju, Z., Liang, P., Li, Z., & Xu, B. (2014, September). Teleoperation control of Baxter robot using body motion tracking. In 2014 International conference on multisensor fusion and information integration for intelligent systems (MFI) (pp. 1-6). IEEE.

[24] Boucenna, S., Anzalone, S., Tilmont, E., Cohen, D., & Chetouani, M. (2014). Learning of social signatures through imitation game between a robot and a human partner. IEEE Transactions on Autonomous Mental Development, 6(3), 213-225.

[25] Taheri, A. R., Alemi, M., Meghdari, A., Pouretemad, H. R., & Holderread, S. L. (2015). Clinical application of humanoid robots in playing imitation games for autistic children in Iran. Procedia-Social and Behavioral Sciences, 176, 898-906.

[26] Guneysu, A., & Arnrich, B. (2017, August). Socially assistive child-robot interaction in physical exercise coaching. In 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 670-675). IEEE.

[27] Shahverdi, P., & Masouleh, M. T. (2016, October). A simple and fast geometric kinematic solution for imitation of human arms by a NAO humanoid robot. In 2016 4th International Conference on Robotics and Mechatronics (ICROM) (pp. 572-577). IEEE.

[28] Agarwal, P., Al Moubayed, S., Alspach, A., Kim, J., Carter, E. J., Lehman, J. F., & Yamane, K. (2016, August). Imitating human movement with teleoperated robotic head. In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 630-637). IEEE.

[29] Zheng, Z., Das, S., Young, E. M., Swanson, A., Warren, Z., & Sarkar, N. (2014, May). Autonomous robot-mediated imitation learning for children with autism. In 2014 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2707-2712). IEEE.

[30] Cazzato, D., Leo, M., & Distante, C. (2014). An investigation on the feasibility of uncalibrated and unconstrained gaze tracking for human assistive applications by using head pose estimation. Sensors, 14(5), 8363-8379.

[31] Kondori, F. A., Yousefi, S., Li, H., Sonning, S., & Sonning, S. (2011, November). 3D head pose estimation using the Kinect. In 2011 International Conference on Wireless Communications and Signal Processing (WCSP) (pp. 1-4). IEEE.

[32] Cazzato, D., Leo, M., Distante, C., Crifaci, G., Bernava, G., Ruta, L., ... & Castro, S. (2018). An Ecological Visual Exploration Tool to Support the Analysis of Visual Processing Pathways in Children with Autism Spectrum Disorders. Journal of Imaging, 4(1), 9.

[33] Fanelli, G., Weise, T., Gall, J., & Van Gool, L. (2011, August). Real time head pose estimation from consumer depth cameras. In Joint Pattern Recognition Symposium (pp. 101-110). Springer, Berlin, Heidelberg.

[34] Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision, 91(2), 200-215.

[35] Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1859-1866).

[36] La Casica, M., Sclaroff, S., & Athitsos, V. (2011). Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D model. Boston University Computer Science Department.

Columns on Last Page Should Be Made As Close As Possible to Equal Length

Authors' background

Your Name	Title*	Research Field	Personal website
Dario Cazzato	**Research Associate**	**Computer Vision & Robotics**	https://wwwfr.uni.lu/snt/
Claudio Cimarelli	**PhD candidate**	**Computer Vision & Robotics**	https://wwwfr.uni.lu/snt/
Jose Luis Sanchez-Lopez	**Research Associate**	**Automation and Robotics**	https://wwwfr.uni.lu/snt/
Miguel A. Olivares-Mendez	**Research Scientist**	**Automation and Robotics**	https://wwwfr.uni.lu/snt/
Holger Voos	**Full Professor**	**Automation and Robotics**	https://wwwfr.uni.lu/snt/

***This form helps us to understand your paper better, the form itself will not be published.**

***Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor**