A simple yet smart head module for mobile manipulators*

Martin Aguilar
Faculty of Mechanical Engineering and Production Science
Escuela Superior Politécnica del Litoral, ESPOL
Campus Gustavo Galindo, Guayaquil, Ecuador
marsagui@espol.edu.ec

Diego Ronquillo
Faculty of Mechanical Engineering and Production Science
Escuela Superior Politécnica del Litoral, ESPOL
Campus Gustavo Galindo, Guayaquil, Ecuador
diromano@espol.edu.ec

Jan Rosell
Inst. of Industrial and Control Eng.
Universitat Politècnica de Catalunya
Barcelona, Spain
jan.rosell@upc.edu

Leopold Palomo-Avellaneda
Inst. of Industrial and Control Eng.
Universitat Politècnica de Catalunya
Barcelona, Spain
leopold.palomo@upc.edu

Raúl Suárez
Inst. of Industrial and Control Eng.
Universitat Politècnica de Catalunya
Barcelona, Spain
raul.suarez@upc.edu

***Abstract*—Mobile manipulators working in semi-structured human environments need smart vision capabilities to perceive the world and interact with human operators. With this purpose in mind, this paper presents the development of a robotic head module composed of an OAK-D camera mounted on a pan-and-tilt WidowX XM430 structure. The OAK-D camera provides high-resolution images, including stereo vision and depth sensing, and advanced capabilities based on embedded artificial intelligence functions. These features combined with the high-range motion of the WidowX XM430, allow the head module to have advanced visual tracking capabilities. The implementation has been done using ROS (Robot Operating System), which allows the head module to be easily integrated to any mobile manipulator.**

***Index Terms*—Mechatronic Systems, Perception and Sensing, Motion Control Systems, Autonomous Robots.**

I. Introduction

Numerous sectors have been transformed by the usage of robots, which has resulted in considerable gains in production, efficiency and safety. Robots must have the capacity to sense and comprehend their environment in order to be useful in the applications for which they are designed. This is specially true for mobile manipulators conceived to work in semi-structured human environments. One of the most popular methods to give robots perception capabilities is using cameras, which may allow the robot to recognize and track things of interest, e.g. robots used in production must be able to find and pick up specific products [1], whilst those employed in surveillance or human-robot collaborative tasks must be able to recognize and track people [2].

In this paper a robotic head for a mobile manipulator is proposed, which is composed of an OAK-D camera mounted on a WidowX XM430 pan-and-tilt structure. This way, a 360-degree view is given to the camera, greatly enhancing the robot perception capabilities. The open-source robotics middleware ROS (Robotic Operating System [3]) has been used to operate the pan-and-tilt and the OAK-D camera, allowing a simple and smooth management of the robot head motions and camera view. Additionally, ROS gives the ability to efficiently handle and examine the camera data, eases the use of image analysis modules like those for the detection and tracking of people, and facilitates the integration with other modules of the robotic system. The head module has been mounted on the MADAR dual-arm mobile manipulator [4], as shown in Fig. 1.

This work was partially supported by the Spanish Government through the project PID2020-114819GB-I00

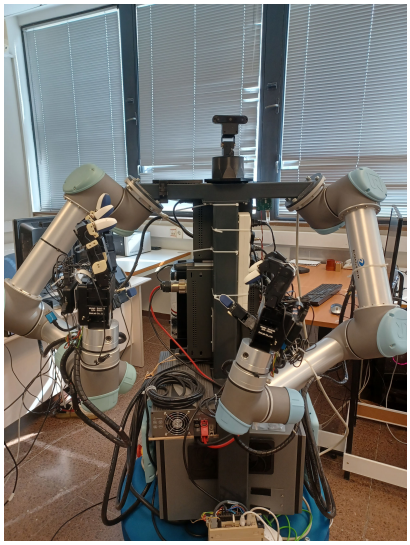

Fig. 1: The head module mounted on the MADAR robot.

II. The pan and tilt camera system

This section describes the hardware set-up, the software requirements and the steps taken to develop the pan-and-tilt with the OAK-D camera head for advanced visual tracking: pan-and-tilt control; tracker functionalities implemented as ROS services; pan-and-tilt and OAK-D physical integration.

(a) WidowX Xm430 (b) OAK-D camera

Fig. 2: Hardware components.

A. Hardware set-up

The selected hardware components are (Fig. 2):

a) **WidowX XM430 Turret by Interbotix**[1]: It is a motorized pan-and-tilt structure with two Dynamixel XM430-W350 servomotors with the following features: maximum torque 3.5 Nm, maximum speed 45 rpm, position resolution 0.088 degrees. These features provide a high performance allowing to move precisely and provide exact feedback, being useful in applications that need high accuracy and stability.

b) **OAK-D AI camera**[2]: It is a powerful camera based on Intel Myriad X processor, a highly effective system-on-chip developed for computer vision and artificial intelligence (AI) applications. The stereo vision in the OAK-D camera enables it to record 3D depth information. It also has a 12-megapixel color camera for high-resolution images, and a variety of sensors, including an accelerometer, gyroscope, and magnetometer, which offer further information to support the interpretation of the visual data. The OAK-D camera is a flexible device for a variety of computer vision and AI applications, including gesture and facial recognition, object identification and tracking [5], [6]. Many computer vision applications require accurate depth perception and comprehensive visual information, which stereo vision and high-resolution images provided by this camera make possible.

B. Software requirements

The following functionalities have been identified as being of interest to be provided by the robot head software:

- Face detection and tracking
- Body detection and tracking
- ROS frame tracking
- User interface with real-time camera visualization
- Operation in simulation and real environments

C. Pan-and-tilt motion control

When developing any system involving motor movements, it is important to consider the ranges of motion and torque limits of the motors. In this work a motion control algorithm has been implemented to carefully take into account these physical limitations of the motors while designing the tracking services. When attempting to track an object or a person with the camera mounted on the pan-and-tilt, the physical limitations of the motors may preclude the success of the task because the motors may not be able to move fast enough or may not be able to reach the commanded angle. Taking in consideration the ranges of motion of the motors, the motion control algorithm smoothly modifies the motion commands maintaining them always within limits.

D. Frame tracker

A crucial part of the developed tracking system is the capability to follow a reference frame. In this work, a frame tracker service has been implemented to track the origin of a reference frame given as a request, so as to maintain it at the center of the image. When the system receives a request to monitor a frame, the frame tracker service first uses the ROS *TF2*[3] transforms to ascertain the relative pose $(x, y, z, \alpha, \beta, \gamma)$ of the target frame with respect to the reference frame. Then, it calculates the distance d between the current position of the camera and the target position and, finally, the precise values of the pan and tilt angles are computed, e.g. for the pan angle Θ:

$$\Theta_{\text{target}} = \Theta_{\text{current}} + \operatorname{atan}\left(\frac{y}{d}\right),$$

and an analogous computation is done for the tilt angle. Motor control signals are then used to move the camera. To ensure that the item is correctly tracked, the movement stops when it is inside a predefined tolerance range.

A coordinate tracker service has also been implemented as a variant of the frame tracker service, where the service request consists of three float numbers that correspond to the X, Y and Z coordinates. The frame tracker function is called by the coordinate tracker after it creates a frame in RVIZ with the given coordinates. This allows to monitor objects based on their spatial coordinates, enabling an easier tracking in some applications.

E. Face and body detector and tracker functionalities

Some computer vision methods, including deep neural networks and Haar cascades, are able to detect the traits that are distinctive to human faces and, therefore, have been used for the identification and monitoring of human faces in digital pictures or video streams, i.e. for face detection and face tracking. In this work, face detection was developed using *DepthAI* framework[4], the embedded software of the OAK-D camera composed of a set of tools and APIs for interacting with the depth and visual data provided by the camera, and the *OpenVINO* toolkit[5], that allows pre-trained deep learning models to be optimized and installed for face detection onto the OAK-D camera.

The DepthAI pipeline is composed of a group of nodes and linkages between them, used to process the data captured by

[1] https://www.trossenrobotics.com/widowx-x-series-robot-turret.aspx#documentation

[2] https://docs.luxonis.com/projects/hardware/en/latest/pages/BW1098OAK.html

[3] http://wiki.ros.org/tf2

[4] https://github.com/luxonis/depthai-python

[5] https://github.com/openvinotoolkit/openvino

the camera, i.e. each node performs a specific function, and the links connect the nodes to form a data flow. Flexibility in the processing of visual and depth data from the camera can be obtained by adding or removing nodes, changing their order, or modifying their parameters to suit the specific computer vision applications. For the case of face detection, the implemented method captures images from a color camera, preprocesses the image using an *ImageManip* node, and runs a *MobileNet*-based face detection neural network on the image. The results are then displayed on the image frame in real-time. After that, it uses a ROS publisher to send the image frame with information of the center of the bounding box if the face is detected.

Using the center of the bounding box, the face tracker function is performed. It receives the position of the face and calculates the pan and tilt angles required to move the camera in order to show the face at the center of the image. First, the service calculates the angle per pixel, K_{pixel}, for the horizontal and vertical axes according to the pixel size, focal length, field of view of the camera, and the resolution of the camera sensor, as follows:

$$K_{\text{pixel}} = \text{atan}\left(\frac{\text{pixel\_size}}{2 \times \text{focal\_length}}\right) \times \left(\frac{\text{FOV}}{180}\right) \times \text{resolution}.$$

Then, the pan and tilt angles required to move the camera in order to show the face at the center of the image are computed using K_{pixel} and the position (in pixels) of the face (X_p, Y_p) in the image, e.g. for the pan angle:

$$\Theta_{\text{target}} = \Theta_{\text{actual}} \mp K_{\text{pixel}} \times Y_p,$$

and an analogous computation is done for the tilt angle. Finally, the pan and tilt angles are published and used to move the pan-and-tilt either in simulation or in real.

A Python script was used to analyze camera-captured photos and identify people using the *Open Model Zoo* (OMZ) model to find bodies in the surroundings. OMZ is a part of the OpenVINO toolkit that consists in a repository of pre-trained models for various computer vision tasks, such as object detection, classification or segmentation. The pre-trained object detection model from the OMZ can be loaded onto the OAK-D camera and executed using the DepthAI framework, taking advantage of the hardware acceleration of the camera to achieve efficient and accelerated object detection at the edge.

A color camera node, an image-editing node, and a mobile net detection network node are included in the Depth AI pipeline that is built by the script. The color camera node starts the pipeline by taking pictures, which are then preprocessed by the image manipulation node to make them fit the specifications of the input needed by the detection network node. After the photos have been processed, if a body is found within one, the bounding box around it is indicated, and its central coordinates are released. The image is then published, and the identified bodies are centered in the image view using the same service that was used for the face tracker.

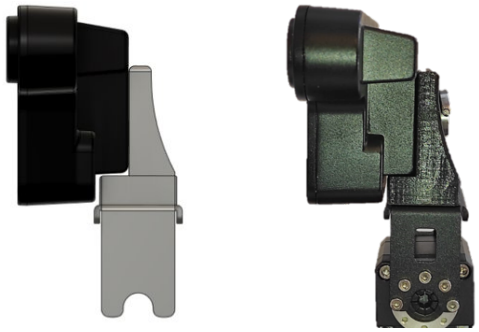

Fig. 3: Designed piece for pan-and-tilt and OAK-D integration.

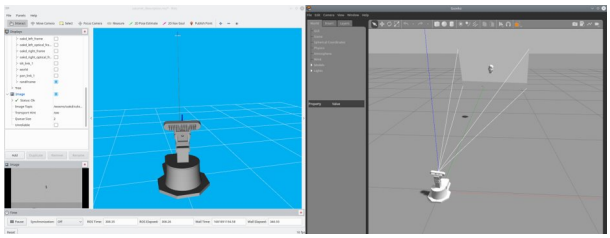

Fig. 4: Simulated environment performing frame tracking.

F. Physical integration of the pan-and-tilt and the camera

It was required to place the OAK-D camera in a vertical configuration (i.e. having the camera lens at a 90-degree angle to the contact surface of the pan and tilt mechanism) to increase its field of view. Consequently, a specific piece was specially designed for the integration of both equipments (Fig 3). This new part creates a safe and dependable connection between the pan-and-tilt module and the OAK-D camera, bridging the space between the two, and guaranteeing that the camera is positioned properly while maintaining stability when in use. The system tracking features are improved by this integration since the complete range of motion may be used to offer a broader picture of the surroundings. The design process involved 3D modeling and simulation of each component using FreeCAD[6] to ensure accurate measurements and fit. The mechanical components were fabricated using 3D printing to achieve the desired shape and accuracy.

III. Results

The following images showcase the various capabilities of detecting and tracking of the pan-and-tilt mechanism with OAK-D camera. Results are shown using the coordinate tracker in simulation (Fig. 4), the face tracker (Fig. 5 left) and body tracker (Fig. 5 right) in the real scenario. Also, the developed user interface is shown in Fig. 6 with the extra option of border detection, which has also been included with the functionalities described in the previous sections. The code of all the developed modules is available at: https://gitioc.upc.edu/labs/ros_widowx_xm430.git .

[6]www.freecad.org

Fig. 5: Face and body detection and tracking. Videos: https://youtu.be/1CSUpKkCwjQ) (left), https://youtu.be/e4wjRndYuCs (right).

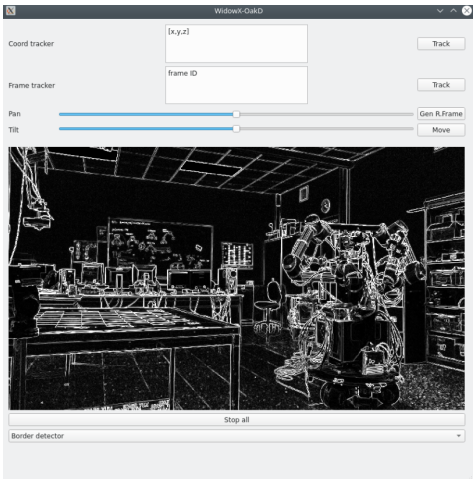

Fig. 6: User interface.

IV. Conclusions and Future Works

The contribution of this paper is the open-source development of a robot head module with advanced AI vision capabilities with a ROS interface. The inclusion of the OAK-D camera in the robot head makes the module a potent tool with leading computer vision and machine learning capabilities. This camera can capture images of the highest quality and, at the same time, process data in real time. Therefore, it can detect and recognize objects and people, making the robot head a versatile perception module for different applications. Counting with the pan-and-tilt mechanism the head module has a wide range of motion and flexibility. This mechanism allows the robot head to move the image view both horizontally and vertically, giving it the ability to view its whole surroundings from various perspectives. Hence, it can acquire more information and draw conclusions more accurately based on its observations, improving its perception and decision-making ability. Moreover, thanks to the integration of the camera and the pan-and-tilt mechanism, the robot head can track and follow moving objects. This function is very helpful when the robot must continually watch over a certain region, object, color or person. Overall, the creation of a robot head equipped with an OAK-D camera installed on a pan-and-tilt mechanism has a great potential, creating new opportunities for the effective development of more advanced and dynamic robotic applications.

Although object tracking using a pan–and-tilt camera setup is quite a well-known task in robotics, the subject continues to be a hot research topic by, e.g., proposing advanced predicting tracking capabilities for grasping purposes [7], or by using advanced AI camera features to track objects in complex backgrounds [8]. In this line, this work seeks to contribute in making perception smart for robotic manipulation tasks in human-robot collaborative environments. For this, we have installed the head module on top of the MADAR robot, a dual-arm mobile manipulator equipped with anthropomorphic hands and conceived to be a robot co-worker. With this set-up, we are currently working in the coordination of the head module with the arm and hand motions in order to continuously have a good view of the manipulation task being performed. Also, the coordination with the mobile base will be developed in order to improve the robot navigation capabilities, including the following of objects or people.

References

[1] J. Chen, H.-W. Huang, P. Rupp, A. Sinha, C. Ehmke, and G. Traverso, "Closed-loop region of interest enabling high spatial and temporal resolutions in object detection and tracking via wireless camera," *IEEE Access*, vol. 9, pp. 87 340–87 350, 2021.

[2] K. Jeyatharan, H. E. M. H. B. Ekanayake, and K. D. Sandaruwan, "Behavior based tracking for human following robot," in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, 2021, pp. 371–376.

[3] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, vol. 3, 2009, p. 5.

[4] R. Suárez, L. Palomo-Avellaneda, J. Martínez, D. Clos, and N. García, "Manipulador móvil, bibrazo y diestro con nuevas ruedas omnidireccional," *Revista Iberoamericana de Automática e Informática industrial*, vol. 17, no. 1, pp. 10–21, 2020.

[5] D. Perazzo et al., "OAK-D as a platform for human movement analysis: A case study," in *SVR'21: Symposium on Virtual and Augmented Reality*.

[6] T. Wakayama, G. A. Garcia Ricardez, L. El Hafi, and J. Takamatsu, "6D-pose estimation for manipulation in retail robotics using the inference-embedded OAK-D camera," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 1046–1051.

[7] R. Nebeluk, K. Zarzycki, D. Seredynski, P. Chaber, M. Figat, P. D. Domanski, and C. Zielinski, "Predictive tracking of an object by a pan–tilt camera of a robot," *Nonlinear Dynamics volume*, vol. 111, p. 8383–8395, 2023.

[8] M. Jiang, K. Shimasaki, S. Hu, T. Senoo, and I. Ishii, "A 500-fps pan-tilt tracking system with deep-learning-based object detection," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 691–698, 2021.