Acquisition and evaluation of depth data from humans, in robotized industrial environments

W Nieto¹, M Arias-Correa¹, and C Madrigal-González¹
¹ Laboratorio de Óptica y Visión Artificial, Instituto Tecnológico Metropolitano, Medellín, Colombia
E-mail: mauricioarias@itm.edu.co

Abstract. Industry 4.0 places great importance on collaborative robotics in industrial production and, therefore, on the safe interaction between humans and robots. In previous decades, robots that were part of flexible manufacturing systems remained isolated from human operators through physical enclosures, but current human-robot collaboration activities require a simultaneous concurrent workspace for both actors. Collaborative robotics security systems have used (among others) artificial vision systems based on red, green and blue additive color model images, as proven solutions in detecting humans within the work area of robots. In recent work, the combination of red, green, and blue additive color model imaging with depth imaging systems has demonstrated low sensitivity to lighting changes and a high positional match between distance data and color pixels in both types of acquired images. However, a detailed review resulted in the absence of databases of industrial robotic environments in this type of images. In this paper, an extensive database is delivered in the format already explained that contains people in robotic industrial settings. Furthermore, evaluating the database using machine learning techniques enables computer vision researchers to gain a better understanding of human detection using a Kinect sensor and convolutional neural networks.

1. Introduction
Collaborative robotics is a trend of Industry 4.0, which allows human-robot interaction in a common workspace. This kind of interaction increases the probabilities of human suffering accidents [1]. Early detection of humans in the space of the robot's task through vision systems improves their security. Developing a security system for robotized industrialized environments based on red-green-blue and depth (RGB-D) images (combination of an image in color channels red, green and blue, with its corresponding depth information for each pixel), trained with neural networks, requires a large image dataset [2]. The first step in this process is the search for datasets consisting of RGB-D images acquired in industrial robotized environments.

In [2–4], authors present datasets acquired with Kinect, consisting of images of household and office objects, utensils, furniture, and grooming implements. There are also RGB-D datasets obtained in interior spaces, consisting of images of bedrooms, bathrooms, study rooms and kitchens, which are very useful in navigation tasks of mobile robots supported by artificial vision [5–7], but are different from datasets used with the simultaneous localization and mapping (SLAM) technique [8,9]. In order to detect and recognize the activities of humans, datasets with RGB-D images have been created, in which men and women assume different positions with their hands [10]. In [11], a pair of Kinect cameras were used in opposite directions to each other, to acquire the images of a group of people who performed recreational activities such as raising their hands, clapping, throwing objects and moving.
Gestures, as well as body posture, are of particular interest in [12]. In this work, you can find RGB-D images related to finger movements, lip movements, and other facial expressions, which generate 20 categories of gestures, stored as depth images, skeleton images, red-green-blue (RGB) images, and even audios. A database of RGB-D images of the human body in different activities of daily life is presented in [13], which is particularly interesting due to the different types of occlusions of the human figure, therefore, the difficulty to detecting people using this database is greater than with the other databases cited in this work.

Table 1 shows the review of the RGB-D databases available on the web. Some of them contain images of household objects, offices, follow-ups of people, people in daily activities, among others. Based on this review, it was established that there are no open-access datasets, containing RGB-D images acquired in robotized industrial environments.

**Table 1.** RGB-D image databases, available on the web

| Nº  | Dataset name                                      | URL                                      | Type                      | Date  | Images | Video | Features                                      |
|-----|--------------------------------------------------|------------------------------------------|---------------------------|-------|--------|-------|-----------------------------------------------|
| 1   | RGBD object dataset                              | http://www.cs.washington.edu/rgbd-dataset. | Household objects         | 2014  | X      |       | 3D objects reconstruction                     |
| 2   | Berkeley 3-D object (B3DO) dataset               | http://rll.berkeley.edu/bigbird/         | Offices, rooms and desks  | 2013  | X      |       | 3D objects reconstruction                     |
| 3   | Object Disappearance for object discovery        | http://wiki.ros.org/Papers/IROS2012_Martheson_Marthi_Parr | Videos in several rooms  | 2012  | X      |       | Object recognition for service robot trajectory |
| 4   | Kinflu-scenes (ply colored mesh files dataset)   | https://cs.stanford.edu/people/karpathy/discovery/ | Desktop and domestic objects | 2013  | X      |       | Objects detection in 3D scenes                |
| 5   | Cornell-RGBD-dataset                             | http://pr.cs.cornell.edu/sceneunderstanding/ | Office scenes (tagged)    | 2013  | X      |       | Path recognition for personal robots.         |
| 6   | NYU Depth V1                                     | https://cs.nyu.edu/~silberman/datasets/nyu_depth_v1.html | House with messy interior | 2012  | X      |       | Interpretation of surfaces in domestic environment |
| 7   | SceneNN                                          | http://www.scenenn.net.                  | Domestic environments     | 2016  | X      |       | Object position interpretation                |
| 8   | RGB-D SLAM dataset and Benchmark                 | https://vision.in.tum.de/data/datasets/rgbd-dataset | Path images with a Kinect on service robot | 2011  | X      |       | Reference point for SLAM visual systems       |
| 9   | Microsoft 7-scenes dataset                       | https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/?from=hp%3A%2F%2Fsearch.microsoft.com%2Fen-us%2Fprojects%2F7-scenes%2F | Trajectory tracking       | 2013  | X      |       | For evaluation of dense tracking, mapping or relocation techniques |
| 10  | Cornell activity datasets, CAD-60 & CAD-120      | http://pr.cs.cornell.edu/humanactivities/data.php | Daily activities of people | 2012  | X      |       | Human activities recognition                  |
2. **Dataset creation process**

The Microsoft Kinect is an off the shelf sensor chosen to acquire RGB-D images in a real robotized industrial environment because of its availability and very low cost (compared to similar sensors). The creation process of the dataset will be described in this section.

2.1. **Red-green-blue and depth images acquisition device**

Kinect sensor has an acquisition system [14] described in Figure 1(a), consisted of an infrared (IR) projector, a color (RGB) camera and the infrared camera. In Figure 1(b) the external appearance of a Kinect sensor is shown.

The Kinect’s depth sensor uses an infrared laser that passes through a diffraction grating, which generates a pattern of projected points (IR projection). The relative geometry between the IR projector and the IR camera, as well as the pattern of projected IR points, are known, therefore, the correspondence between projected and acquired points can be determined, which, in turn, allows 3D reconstruction [15]. Kinect simultaneously captures images in color (RGB) and depth, at a maximum rate of 30 frames per second (fps). The integration of RGB-D images generates a point cloud that contains approximately 300

| Dataset Name                                      | URL                                                                 | Description                                                                 | Year | Image Type | Presence of Occlusion |
|---------------------------------------------------|---------------------------------------------------------------------|------------------------------------------------------------------------------|------|-------------|------------------------|
| Berkeley multimodal human action database         | http://teleimmersion.citris.uc.org/berkeley_mha                    | Physical activity of people with sound                                       | 2013 | X           | Human pose analysis and movement information |
| ChaLearn looking at people                        | http://gesture.chalearn.org/2013-multi-modal-challenge/data-2013-challenge | Gestures of groups of people.                                                | 2013 | X           | Gestures recognition   |
| Dataset of a human performing daily life activities in a scene with occlusions | https://team.inria.fr/larsen/software/datasets/                      | Daily life activities                                                       | 2015 | X           | Tracking with and without occlusion        |
| 3D mask attack dataset                            | https://www.idiap.ch/dataset/3dmad                                  | Faces                                                                       | 2013 | X X         | It focuses in eyes’ position                |
| Eurecom Kinect face dataset                       | http://rgbd.eurecom.fr/                                             | Faces                                                                       | 2012 | X           | Some parts of the face with and without occlusion |
| Datasets involving humans: Head and face          | http://www.fanelli.li/#db                                           | Human head and face                                                         | 2013 | X           | Head position                      |
| Top view person re-identification (TVPR) dataset  | http://vrai.dii.unipv.m.it/re-id-dataset                            | Human body                                                                  | 2016 | X           | Body’s upper view                 |
| Manipulation action (MANIAC) dataset              | http://www.dpi.physik.uni-goettingen.de/cns/index.php?page=maniacae-dataset | Hands                                                                       | 2014 | X           | Objects manipulation               |
| ChaLearn gesture challenge dataset                | http://gesture.chalearn.org/data                                    | Human body                                                                  | 2012 | X           | GeBody gestures                   |
| UR Fall detection dataset                         | http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html                   | Human body                                                                  | 2014 | X X         | People falling.                   |
000 points per frame [16–18]. The V-FOV (vertical field of view) and the H-FOV (horizontal field of view), will be calculated as part of this process.

![Image of Kinect sensor](image)

**Figure 1.** Acquisition system of a Microsoft's Kinect sensor, (a) Parts in the Kinect’s acquisition system, (b) Kinect’s external appearance with case.

2.2. Images acquisition process

To acquire images of a robotized industrial process, Kinect’s vertical field of view (V-FOV) and horizontal field of view (H-FOV) were calculated. These values can be seen in Figure 2, with Kinect positioned at 1.50 meters in height (0 degrees inclination), acquisition range between 0.8 m and 3.5 m and lighting supplied by 4 fluorescent tubes T8 (1200 lm), the maximum V-FOV is 0.365 m and the maximum H-FOV is 0.448 m. The process is a dishware’s pick and place operation and comprises a manipulator robot, a conveyor belt, and an operator.

![Image of V-FOV and H-FOV](image)

**Figure 2.** Calculated V-FOV and H-FOV for a Kinect in a pick and place industrial process.

The RGB-D images acquired by the Kinect have 640 x 480 pixels and have been classified according to the number of people present in the image as explained in Table 2.

| Class      | Images acquired |
|------------|-----------------|
| 1 person   | RGB 562         |
|            | Depth 562       |
| 2 people   | RGB 1174        |
|            | Depth 1174      |
| 3-4 people | RGB 3548        |
|            | Depth 3548      |
| 0 people   | RGB 5260        |
|            | Depth 5260      |
In Figure 3, are shown examples of both RGB-D images acquired together depending on the number of present people and also its occlusion conditions. In Figure 3(a) one person, Figure 3(b) one person with occlusion, in Figure 3(c) two people, Figure 3(d) two people with occlusion, in Figure 3(e) three people and Figure 3(f) three people with occlusion.

![Figure 3](image)

**Figure 3.** RGB-D images acquired depending on the people present and occlusion (a) one person (b) one person with occlusion (c) two people (d) two people with occlusion and (e) three people and (f) three people with occlusion.

### 3. Dataset evaluation

The dataset created was named RGBD-DHaRIo (RGBD by RGB-D format, and DHaRIo by dataset for humans in robotized industrial environments), and its usefulness was evaluated through the training of a PointNet architecture, whose entries consist of point clouds obtained from depth images. Prior to that, it was necessary to segment the relevant elements in the RGB images. The stages of the evaluation process are presented below.

#### 3.1. Detection of relevant elements in red-green-blue and depth images

Detection of relevant elements is a concept that refers to the performance of both the stage of detection of relevant objects in an image, and the stage of precise segmentation of the regions of those images [19–21]. In [22], the authors present a model that allows to detect outgoing objects in RGB images with high accuracy and high computational efficiency, based on a model trained through convolutional neural networks (CNNs). As can be seen in Figure 4, that model of [22] can use a RGB image from the RGBD-DHaRIo dataset (Figure 4(a)) as an input and obtain its relevant contours (Figure 4(b)). Subsequently, an algorithm will delimit the contours of the relevant elements by means of bounding rectangles (Figure 4(b)) whose coordinates expressed in the format: (x, y -coordinates of the upper left corner, width of rectangle, height of rectangle), will be stored in a tags file.

The tag file generated from the relevant elements of the RGB images is read by another algorithm that maps the coordinates to the corresponding depth images (RGB-D that were acquired simultaneously) as can be seen in Figure 4(c).
3.2. People detection

In [23], an approach called PointNet for object classification is proposed as a type of CNN for object classification that works directly with point clouds without prior treatment. In our work, PointNet receives the point clouds corresponding to segmented regions of depth images in dataset and generates the number of people present in the image. PointNet has been trained with 19597-point clouds and evaluation were carried out with 4899. Training with 250 epochs, displayed a 96% of accuracy and the loss function was 0.096735.

The dataset’s evaluation procedure is described in Figure (5). RGB-D images acquired with Kinect sensor are the input to the block named “Detection Process”. Contours and coordinates of the relevant elements in the RGB image are obtained in the first stage of processing, then boundary boxes are mapped to Depth images and coordinates define regions to obtain point clouds. Third stage takes the point clouds as inputs for the PointNet model (previously trained). The exit of PointNet, is a classification which will tell us if in the image there is, one person, there are two people or three people working with the robot in the robotized environment.

4. Conclusions

A dataset consisting of RGB-D images of people in robotized industrial environments, called RGBD DHaRlo, was presented in this work. The dataset consisting of 10544 RGB images and 10544 depth images, as well as 10520 images of industrial environments without people, will be available in Mendeley databases for free. There are no datasets like this, available for the community of computer vision, so this is a first contribution from our research.

In the other hand, the procedure for detecting people in industrial robotized environments using RGB-D images as inputs and a PointNet model as a classification method, was successful. There are plenty of similar procedures and methodologies, but none explodes the advantages of adding depth information and becoming this way more accurate in people detection. Nevertheless, for practical purposes, PointNet should be trained with a greater number of images and for a greater number of epochs. This procedure is another contribution to the community. Finally, it is important state that the quality of point clouds obtained will depend on the accuracy of the sensor used. Kinect is a cheap and useful sensor but there are other available sensors with better overall performance.
References

[1] Villani V, Pini F, Leali F and Secchi C 2018 Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications Mechatronics 55 248

[2] Aldoma A, Tombari F, Di Stefano L, Vincze M 2012 A global hypotheses verification method for 3D object recognition European Conference on Computer Vision (ECCV) Lecture Notes in Computer Science vol 7574, eds Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (Berlin: Springer)

[3] Janoch A, Karayev S, Jia Y, Barron J T, Fritz M, Saenko K and Darrell T 2011 A category-level 3-D object dataset: Putting the Kinect to work IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (Barcelona: IEEE) pp 1168-1174

[4] Lai K, Bo L and Fox D 2014 Unsupervised feature learning for 3D scene labeling IEEE International Conference on Robotics and Automation ICRA (Hong Kong: City University of Hong Kong) pp 3050-3057

[5] Anand A, Koppula H S, Joachims T and Saxena A 2011 Contextually guided semantic labeling and search for three-dimensional point clouds The International Journal of Robotics Research 30(1) 19

[6] Couprie C, Farabet C, Najman L and LeCun Y 2013 Indoor semantic segmentation using depth information ArXiv abs/1301.3572

[7] Karpathy A, Miller S and Li F F 2013 Object discovery in 3D scenes via shape analysis IEEE International Conference on Robotics and Automation (Germany: Karlsruhe Institute of Technology) pp 2088-2095

[8] Engelhard N, Colas F, Cremers D, Siegwart R, Burgard W, Engelhard N and Colas F 2011 Towards a benchmark for RGB-D SLAM evaluation Proceedings of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conference (Los Angeles: Cornell University)

[9] Glocker B, Izadi S, Shotton J and Criminisi A 2013 Real-time RGB-D camera relocalization IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (Australia: IEEE) pp 173–179

[10] Sung J, Ponce C, Selman B and Saxena A 2012 Unstructured human activity detection from RGBD images IEEE International Conference on Robotics and Automation (United States: IEEE) pp 842-849

[11] Ofl C, Chaudhry R and Kurillo G 2013 Berkeley multimodal human action Database (MHAD) IEEE Workshop on Applications of Computer Vision WACV (United States: IEEE) pp 53–60

[12] Escalera S, González J, Baró X, Reyes M, Lopes O, Guyon I, Athitsos V and Escalante H 2013 Multimodal gesture recognition challenge 2013: dataset and results 15th ACM on International Conference on Multimodal Interaction (Australia: Association for Computing Machinery) pp 445-452

[13] Diby A and Charpilety F 2015 Pose estimation for a partially observable human body from RGB-D cameras IEEE/RSJ International Conference on Intelligent Robots and Systems IROS (Germany: IEEE) pp 4915-4922

[14] Cai Z, Han J, Liu L and Shao L 2017 RGB-D datasets using Microsoft Kinect or similar sensors: A survey Multimedia Tools and Applications 76 4313

[15] Zeng W and Zhang Z 2012 Multimedia at work Microsoft Kinect sensor and its effect IEEE Multimedia 19(2) 4

[16] Khoshelham K and Elberink S O 2012 Accuracy and resolution of kinect depth data for indoor mapping applications Sensors 12(2) 1437

[17] Kim C, Yun S, Jung S-W and Won C S 2015 Color and depth image correspondence for Kinect v2 Advanced Multimedia and Ubiquitous Engineering (Berlin: Springer) pp 111-116

[18] Lai K, Bo L, Ren X and Fox D 2013 RGB-D object recognition: Features, algorithms, and a large scale benchmark Consumer Depth Cameras for Computer Vision (London: Springer) pp 167-192

[19] Yin Li, Xiaodi Hou C K 2014 The secrets of salient object segmentation supplementary materials IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (United States: IEEE) pp 280-287

[20] Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X and Shum H Y 2011 Learning to detect a salient object IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2) 353

[21] Borji A 2015 What is a salient object? A dataset and a baseline model for salient object detection IEEE Transactions on Image Processing 24(2) 742

[22] Hou Q, Cheng M M, Hu X, Borji A, Tu Z and Torr P H S 2017 Deeply supervised salient object detection with short connections IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu: IEEE) pp 3203-3212

[23] Qi C R, Su H, Mo K and Guibas L J 2017 PointNet: Deep learning on point sets for 3D classification and segmentation IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu: IEEE) pp 652-660