MVIP: A Dataset for Industrial Part Recognition

Paul Koch[a,c], Marian Schlueter[a], Clemens Briese[a], Vivek Chavan[a], and Joerg Krueger[b]

a) Fraunhofer IPK, Pascalstraße 8-9, 10587 Berlin, Germany

b) TU Berlin, Straße des 17 Juni 135, 10623 Berlin, Germany

c) corresponding author: paul.koch@ipk.fraunhofer.de

Abstract

We present MVIP, a novel dataset for multi-modal and multi-view application oriented industrial part recognition. Here we combine a calibrated RGBD multi-view dataset with additional object context such as physical properties, natural language, and super-classes. Our main goal with MVIP is to study and push transferability of various state-of-the-art methods within related downstream tasks towards an efficient deployment of industrial classifiers. Additionally, we intent to push with MVIP research regarding several modality fusion topics, (automated) synthetic data generation, and complex data sampling methods – combined in a single application oriented benchmark.

1. Introduction

With MVIP we contribute a novel multi-view (MV) and multi-modal (MM) dataset for application oriented industrial part recognition. For MVIP's data acquisition, we designed a digitisation and recognition station (see Fig. 1) as one could expect to be found on site at an industrial facility. The station is equipped with a vast set of task relevant sensors (ten calibrated RGBD cameras and a scale) in order to investigate; A) which data (sensors) benefits the industrial part recognition, B) how to design and efficiently train a robust MV and MM model for industrial part recognition. In addition to the color, depth, and weight modalities, other modalities such as package-size (width, height, length), natural language tags (descriptions), and super-classes (general class spanning a common subset of classes, e.g. tool, car-component, etc.) are available in the dataset. At industrial applications, such meta data is often present in analog or digital catalogs in order to guide workers to identify industrial parts. This additional meta data allows further research regarding modality-fusion, training or sampling methods, and hybrid search engines within the top K classification results. The calibrated MV-dataset allows 3D reconstruction of scenery and objects (see Fig. 3), which enables research regarding (automated) synthetic data generation and 3D based object recognition.

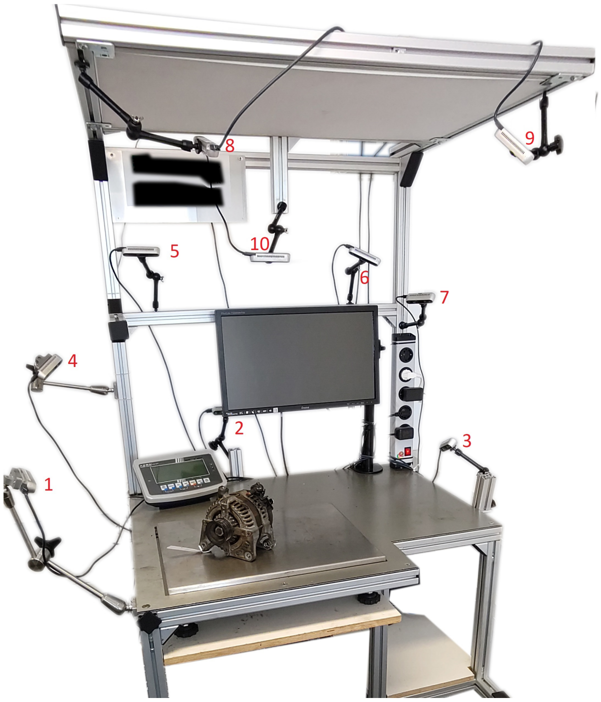

Figure 1. Digitisation station used for MVIP

2. The MVIP Dataset

MVIP is captured on a digitisation station as illustrated in Fig. 1. Ten RGBD cameras are mounted on the table construction, which are all facing a common point on the integrated scale. An ArUco-Board is surrounding the scale (see Fig. 3), thereby a 6D-Pose can be determined for each camera at any given time, given the camera's intrinsic parameters. Thus, the cameras are calibrated to each other within each captured image set (see. Fig. 2). Due to the arrangement of camera perspectives and calibration, one set of images covers most of the object surface and allows 3D construction of the objects (Fig. 3). The front of the ta-

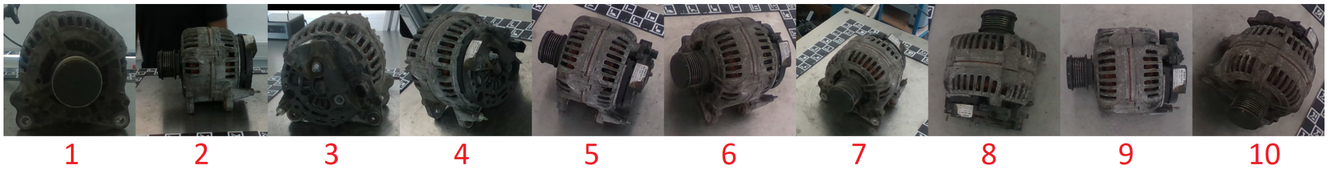

Figure 2. An indexed ROI cropped MVIP image set featuring ten simultaneously captured Views.

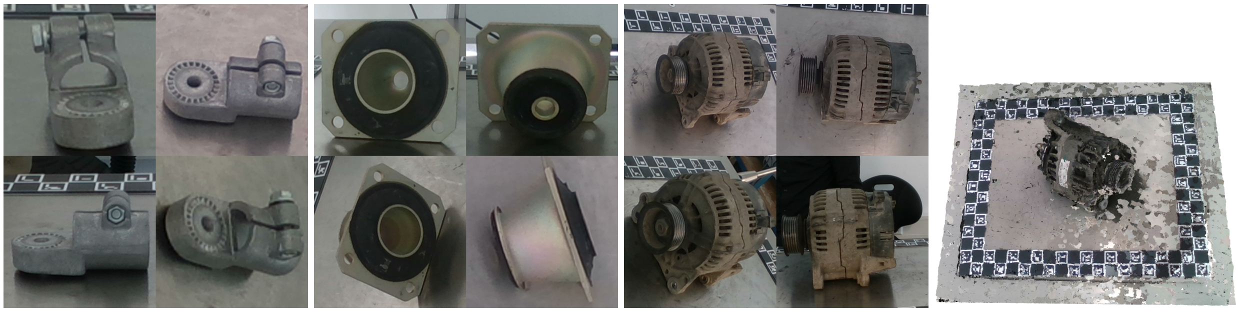

Figure 3. A subset of three objects featured in MVIP. On the right a 3D-Reconstruction generated via a single image set from MVIP

ble is not equipped with cameras to leave room for a worker.

The industrial components featured in MVIP are set to be at least the approximate size of a fist (in a subjective assessment of the worker) and maximum $350\ mm \times 450\ mm \times 300\ mm$ with weight of $< 15\ kg$. Each object featured in the dataset is rotated during digitisation 12 times (approximately 30° steps), given a subjective ”natural laying” position (See Fig. 1 for illustration). Since some objects have multiple ”natural laying” positions, the 12 rotations are repeated according to the subjective assessment of the worker. For test (5) and validation (5) purposes, ten additional image sets are captured, where the worker randomly moves the object on the scale according to a natural laying position. In addition to image data, the dataset features meta data for all objects; weight, package-size (length, width, height), object class, super-classes (general class spanning a common subset of classes), natural language (NL) tags and generated view-wise object segmentation masks (thus, also ROI Bounding Boxes). Moreover, the calibrated MV RGBD design of the dataset enables anyone to employ our toolbox methods for 3D-Object-Point-Cloud and 3D-Scene-Reconstruction, 6D-Object-Pose Estimation, and Synthetic-Data generation. In total MVIP features 308 classes of industrial components, which are grouped into 18 super-classes. From eight categories (shapes, colors, materials, textures, conditions, size, weight, and density) MVIP uses 77 NL-Tags to describe the classes. MVIP has a total of approximately 71.276 image sets. Each set has the images: RGB, depth, HHA, and segmentation mask. Additionally each set is associated to a specific background set (empty scene with equal scene condition). The images are captured at a resolution of 1280×720 pixels with RealSense D435 and D415 Cameras.

Toolbox, Data & Website

A toolbox for dataset handeling and a download link to MVIP can be found at Github: https://github.com/KochPJ/MVIP-Toolbox

The project website can be found here: https://www.ipk.fraunhofer.de/de/zusammenarbeit/referenzen/eiba.html

Acknowledgements

We would like to thank everyone at the Fraunhofer IPK involved in the MVIP project. A special thanks to Maulik Jagtap for his contributions to the dataset.

This work is funded by the German Federal Ministry of Education and Research (BMBF) with the EIBA project 033R226 in the ReziProK programm over the FONA platform for sustainable research.

SPONSORED BY THE

Federal Ministry
of Education
and Research

EIBA