Object Disambiguation for Augmented Reality Applications

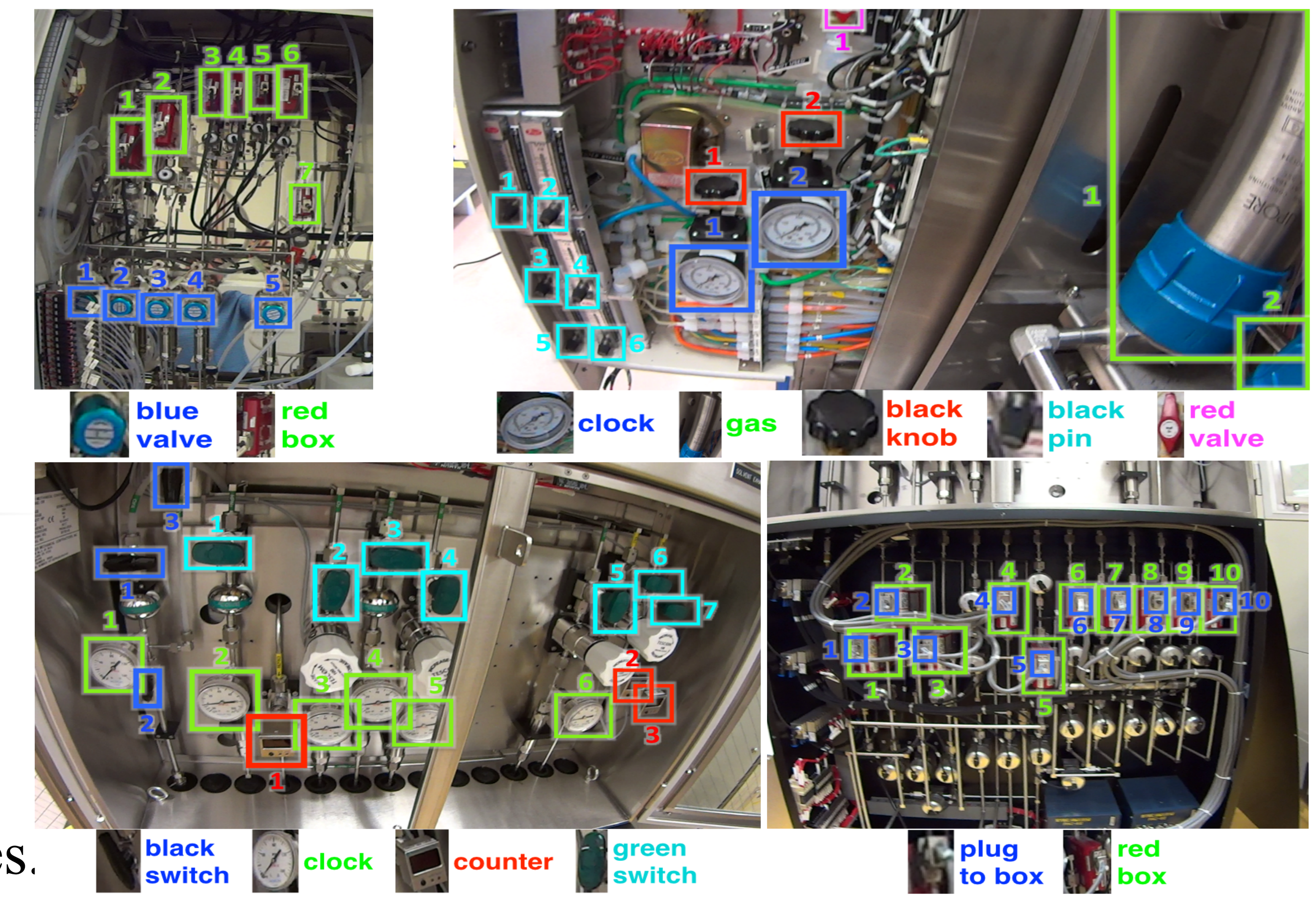
Wei-Chen Chiu[1], Gregory Johnson[2], Dan McCulley[2], Oliver Grau[2], Mario Fritz[1]
Max Planck Institute for Informatics[1], Intel Corporation[2]

Goals

- Robust monocular object recognition and identification system that leverages 3D contextual information.
- Augmented Reality application for guided maintenance to disambiguate potentially repetitive machine parts.

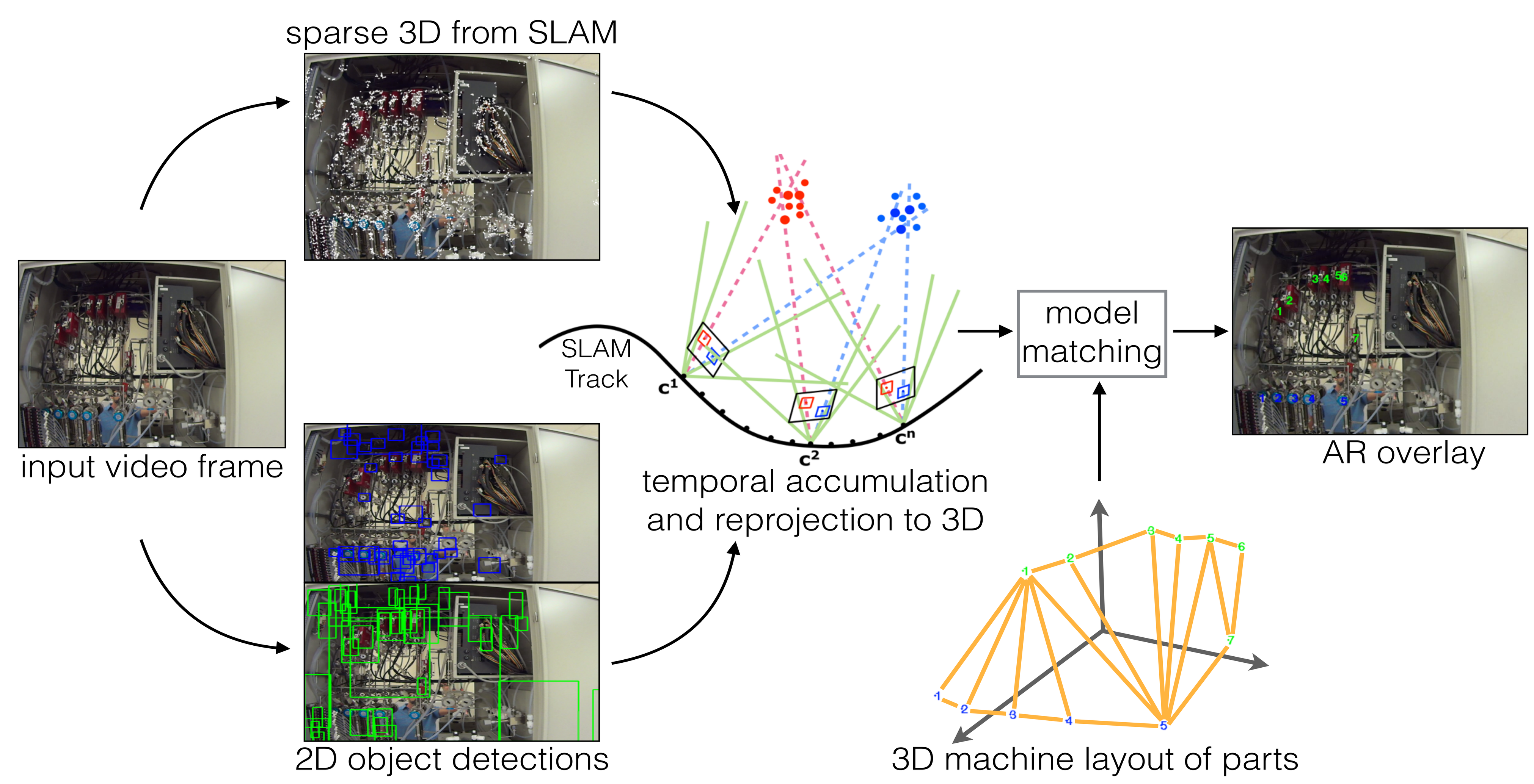
Benchmark

- We propose the first benchmark for an object disambiguation that is composed of an annotated dataset.
- Composed of 14 videos with different viewing scenarios on 4 machines with 13 partially shared components. In total 249 frames with 6244 parts are annotated by bounding boxes and unique identities.



Approach

We seek a monocular system that operates markerless and exploits state-of-the art object detectors in order to disambiguate objects as parts of a machine. For disambiguating we fuse the object detector output with a SLAM system that allows us resolve ambiguities by reasoning over spatial context.



2D Object Detection

We evaluate on different 2D detectors, including linemod2D, cascade detectors with Haar, HoG or LBP features, and extended deformable part based model (DPM) with LAB color features.

	LINE-MOD	Haar cascade	HoG cascade	LBP cascade	color-DPM
avg. precision	10.81%	8.37%	13.38 %	8.90 %	36.73 %

Object Disambiguation

- Based on the SLAM and the 2D object detector, we reproject the detections back to 3D and temporally accumulate them into point clouds.
- We acquire the prior knowledge of the 3D machine layout that specifies the relative locations of each part.
- We apply the RANSAC to iteratively estimate the geometric transformation M between 3D layout with N objects g_n and the observed detections d w.r.t deformation of the layout, object appearance, expectation of viewpoints and scales, as well as amount of matched objects.

$$\underset{d_1,d_2,\ldots,d_N,M}{\arg\min} \quad E_{deformation} + E_{appearance} + E_{scale} + E_{viewpoint}$$

where

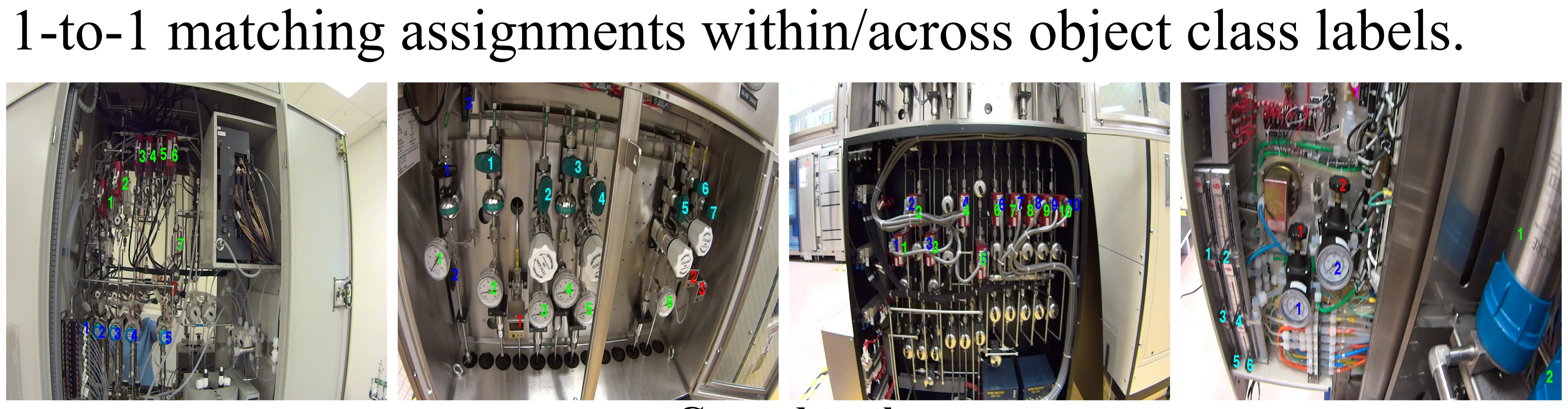
$$E_{deformation} = \frac{\sum_{n=1}^{N} \delta_n}{N} \sum_{n=1}^{N} \delta_n \cdot log(\|\bar{M}(P_{g_n}) - P_{d_n}\|)$$

$$E_{appearance} = -\sum_{n=1}^{N} \delta_n \cdot A_{d_n}$$

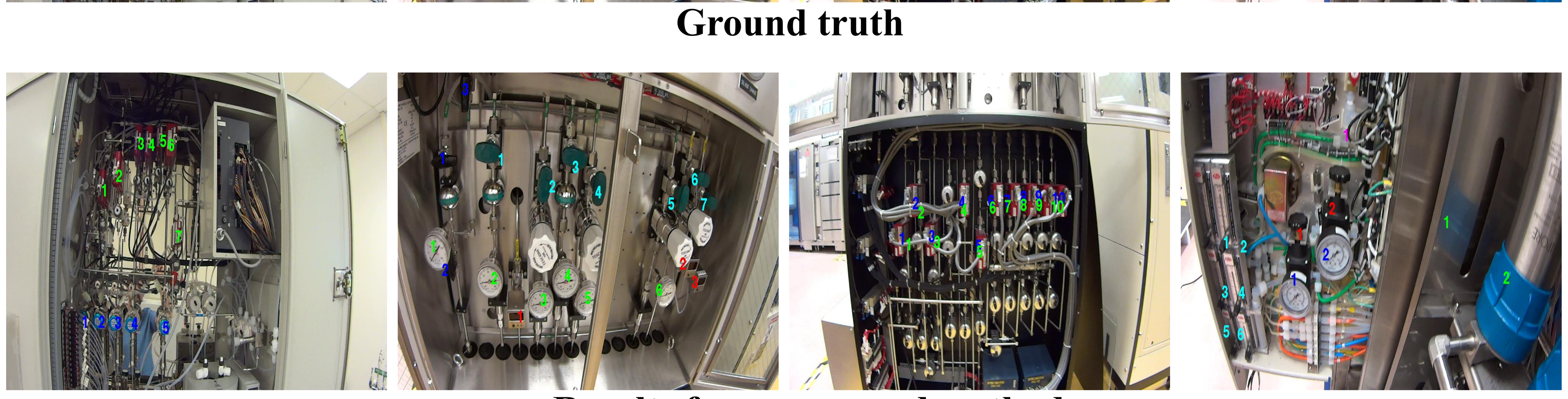
$\delta_n = 1$ if $\|\bar{M}(P_{g_n}) - P_{d_n}\|$ smaller than a threshold ε, and $\delta_n = 0$ otherwise

Experimental Results

For seeking a metric which can capture the object disambiguation performance of a human if provided with the produced overlay. We investigate different metrics: Pascal, nearest neighbor and 1-to-1 matching assignments within/across object class labels.



Ground truth



Results from proposed method

	machine 1	machine 2	machine 3	machine 4	average
Human Judge.	74.12%	100.00%	99.68%	70.57%	86.09%
Pascal	60.92%	98.68%	95.60%	25.10%	70.08%
NN (within)	57.05%	94.76%	88.06%	72.88%	78.19 %
NN (across)	56.07%	91.97%	65.20%	56.84%	67.52 %
1-to-1 (within)	77.55%	99.18%	99.68%	79.25%	88.92%
1-to-1 (across)	74.63%	96.92%	93.10%	72.45%	84.28 %

	machine 1	machine 2	machine 3	machine 4	average
full model	74.63%	96.92%	93.10%	72.45%	84.28 %
no appearance	67.29%	93.32%	64.05%	51.06%	68.93%
no deformation	83.89%	95.05%	61.44%	40.30%	70.17%
no scale	67.29%	98.53%	53.94%	43.57%	65.84%
no viewpoint	38.01%	88.89%	43.04%	10.21%	45.04%
no scale & no viewpoint	38.01%	88.89%	43.04%	10.21%	45.04%
no non-matched	74.61%	74.16%	64.10%	55.65%	67.13%

Object Disambiguation DataSet (ObDiDas) is available at
http://datasets.d2.mpi-inf.mpg.de/object-disambiguation/