Object Disambiguation for Augmented Reality Applications

Wei-Chen Chiu[1]
walon@mpi-inf.mpg.de
Gregory S. Johnson[2]
gregory.s.johnson@intel.com
Daniel Mcculley[2]
daniel.b.mcculley@intel.com
Oliver Grau[2]
oliver.grau@intel.com
Mario Fritz[1]
mfritz@mpi-inf.mpg.de

[1] Max Planck Institute for Informatics
Saarbrücken, Germany
[2] Intel Corporation

Abstract

The broad deployment of wearable camera technology in the foreseeable future offers new opportunities for augmented reality applications ranging from consumer (e.g. games) to professional (e.g. assistance). In order to span this wide scope of use cases, a markerless object detection and disambiguation technology is needed that is robust and can be easily adapted to new scenarios. Further, standardized benchmarking data and performance metrics are needed to establish the relative success rates of different detection and disambiguation methods designed for augmented reality applications.

Here, we propose a novel object recognition system that fuses state-of-the-art 2D detection with 3D context. We focus on assisting a maintenance worker by providing an augmented reality overlay that identifies and disambiguates potentially repetitive machine parts. In addition, we provide an annotated dataset that can be used to quantify the success rate of a variety of 2D and 3D systems for object detection and disambiguation. Finally, we evaluate several performance metrics for object disambiguation relative to the baseline success rate of a human.

Method

We seek a monocular system that operates markerless and exploits state-of-the-art object detectors in order to disambiguates objects as parts of a machine. Figure 1 shows an overview of our system.

We use the sparse 3D information generated by the SLAM system[1] in order to reproject the 2D object detections[2] to 3D. As all preceding frames are connected by SLAM track, we accumulate the reprojected 2D object detections over time. In addition to 3D detection clouds, we also require a 3D machine layout that specifies the relative locations of each object. Such description are often provided by the machine specifications, but it doesn't have to be metric or a complete model in our method, which provides easy deployment and adaptation to new scenarios.

In order to match the 3D layout with N objects g_n to the observed detections d, we define an energy function that is taking into account the object appearance ($E_{appearance}$), deformation of the layout ($E_{deformation}$), scale (E_{scale}), viewpoint ($E_{viewpoint}$) as well as amount of matched objects (optional part in the deformation energy). The energy on scale and viewpoint capture an expectation of typical viewpoints the machine is viewed in. We seek the best match by finding an assignment of detections $d_1,\ldots,d_N$ as well as a projection matrix M so that the following objective:

$$\underset{d_1,d_2,\ldots,d_N,M}{\arg\min} \; E_{deformation} + E_{appearance} + E_{scale} + E_{viewpoint} \tag{1}$$

where

$$\begin{aligned} E_{deformation} &= \frac{\sum_{n=1}^{N}\delta_n}{N}\sum_{n=1}^{N}\delta_n \cdot log(\left\|\bar{M}(P_{g_n}) - P_{d_n}\right\|) \\ E_{appearance} &= -\sum_{n=1}^{N}\delta_n \cdot A_{d_n} \end{aligned} \tag{2}$$

P_{g_n} and P_{d_n} are the 3D coordinate of g_n and d_n, while A_{d_n} is the detection score of the match d_n. δ_n is for handling the non-matched machine parts, where $\delta_n = 1$ if $\left\|\bar{M}(P_{g_n}) - P_{d_n}\right\|$ smaller than a threshold and $\delta_n = 0$ otherwise. In both E_{scale} and $E_{viewpoint}$, the scale factor s and three view points included in the 3D transformation $M(\cdot)$ are hard-constrained according to their distributions learnt from the training videos.

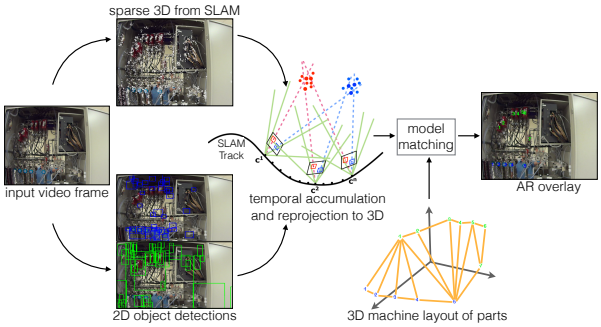

Figure 1: Overview of our system for object disambiguation.

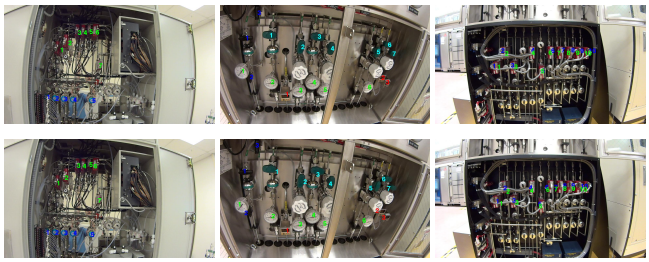

Figure 2: Example results. First row are for the groundtruh of each machine. Second row are the corresponding results from our method.

In order to minimize the objective, we follow a RANSAC pipeline by randomly selecting candidate alignments between the detections and the machine layout which results in an initial geometric transformation.

Experiments

In order to evaluate our approach, we propose the first benchmark for an object disambiguation task in maintenance work that is composed of an annotated dataset. Furthermore, instead of using traditional Pascal metric, we are interested in a metric that captures the object disambiguation performance of a human if provided with the produced overlay. Therefore we propose a set of candidate metrics and then evaluate which one is closest to actual human judgement on the task. Our proposed metric gives a more realistic estimate of the system performance than a traditional Pascal object detection metric that consistently underestimates the system performance. Figure 2 shows example results of our system in comparison to the groundtruth annotations.

References

[1] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.

[2] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost van de Weijer, Andrew D Bagdanov, Maria Vanrell, and Antonio M Lopez. Color attributes for object detection. In *CVPR*, 2012.