Demo: AEYE—Visual Large-scale Industrial Interaction Processing

Christoph Heindl
Gernot Stübl
Thomas Pönitz
Andreas Pichler
christoph.heindl@profactor.at
Visual Computing
PROFACTOR GmbH
Steyr, Austria

Josef Scharinger
josef.scharinger@jku.at
Institute of Computational Perception
Johannes Kepler University
Linz, Austria

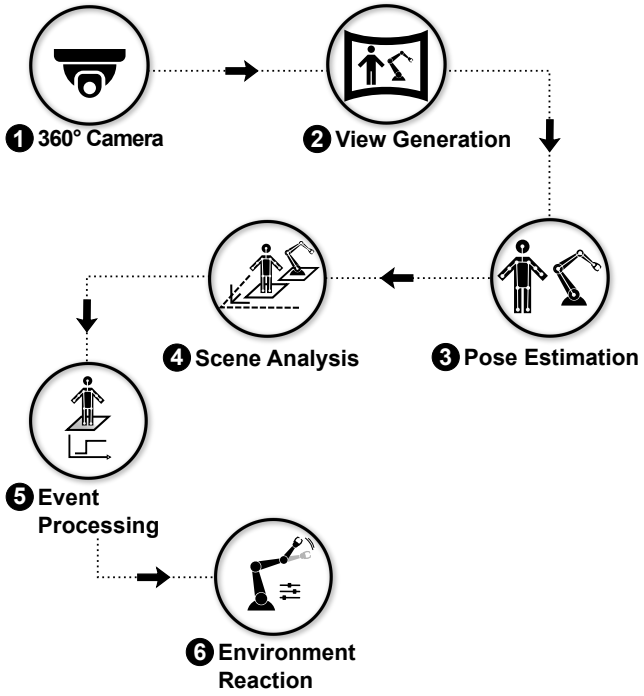

Figure 1: System Overview. Our approach considers panoramic images as input (1); these are synthesized to form one or more rectilinear views (2). Deeply learned neural networks predict human and robot pose keypoints (3). Detected keypoints are lifted to 3D metric space using homographies of planes (4). A number of virtual regions raise location-aware events (5) based on geometric relations to surrounding entities. (6) These events in turn lead to application dependent environmental reactions.

INTRODUCTION

The cost pressure in industrial production has led to an increasing number of isolated machine networks, running fully automated. The advent of collaborative robotics has broken up closed architectures through interaction with humans. Thus, we need to rethink how these networks are automated. Conventional sensor technology is insufficient for the following reasons: (a) sensors exhibit low cognitive abilities; (b) human activities are not captured; (c) sensor adaptation to environmental changes require major rewiring and reprogramming efforts.

Recent image-based approaches to sensing apply active depth cameras to capture human poses and measure their position relative to other objects in the environment [3]. Such approaches are, however, inherently limited to the available depth range (usually a few meters). In addition, the light emitted by these cameras is easily scattered by shiny surfaces (typical in industrial environments) rendering all measurements unusable.

Figure 1 depicts our approach. We propose the replacement of many conventional hardware sensors by a single non-intrusive vision system that coordinates human/machine events from a bird's eye perspective. In contrast to common sensors, our system builds on recent results from the deep-learning field; this enables it to answer complex image-related queries robustly. An expandable set of virtual regions triggers location-aware events based on the geometric relationship to surrounding objects. We suggest working exclusively with color input streams, rather than relying on range-limited depth input. Many of the system's responses are lifted into metric 3D space by fusing detection results with known scene configurations.

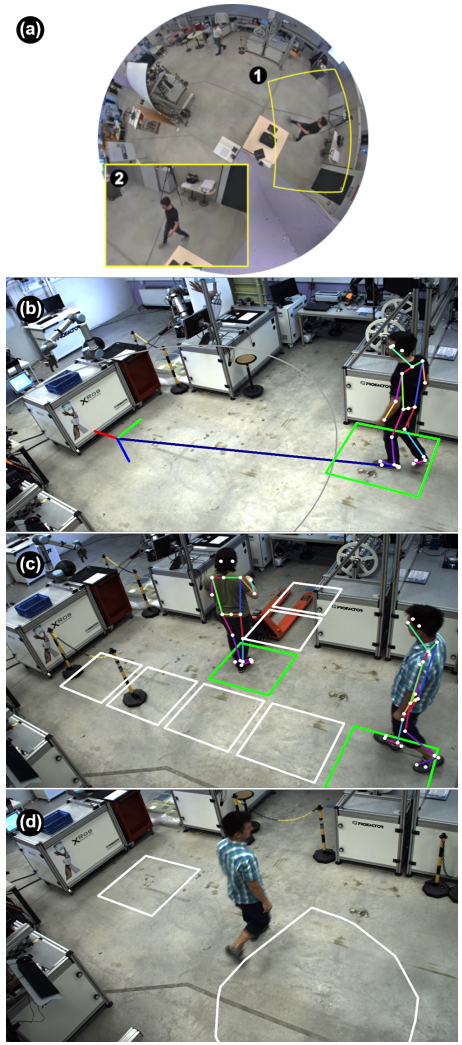

Figure 2: (a) Synthetic view generation from panoramic images. (b-d) Use cases (UCs); (b) UC1 Entering the (green) region enables the robot; worker's proximity to the robot controls its movement velocities. (c) UC2 Regions are only sensitive to humans, but are not influenced by other objects. (d) UC3 Free-form region definition by humans.

METHOD

Our system first creates one or more synthetic rectilinear views from panoramic color images (see Figure 2a). Next, we perform human and robot pose estimation [1, 2] on each view (see superimposed skeletons of Figure 2b-d). These detections are then transformed into metric space by mapping pixel locations to world coordinates using plane homographies. In particular, a homography between ground and camera image plane allows us to convert pixel to metric ground coordinates. Since such a mapping is valid only for keypoints semantically close to the ground (such as feet positions), we additionally incorporate a statistical human body model that adds additional keypoints to be mapped. We use these extra points to predict body orientation and to stabilize localization in the presence of occlusions. Once world coordinates are established, our system scans for events that arise from the interaction of humans with a predefined set of virtual regions placed in floor coordinates (see Figure 2). Each such event might then trigger one or more environmental reactions, depending on the application.

DEMONSTRATION

We demonstrate the interaction potential of the AEYE system in three Use Cases (UCs):

UC1 focuses on human-robot interaction (see Figure 2b). A human entering the rectangular region enables the robot to start. The speed of the robot is automatically adjusted depending on the proximity of the worker, enabling new collaboration patterns.

UC2 demonstrates several people interacting with multiple regions. The focus is on system robustness in the presence of occlusion and the system's ability to react to people only—avoiding false triggers caused by other objects (see Figure 2c).

UC3 shows free-form region definition using human movements and action detection as input (see Figure 2d). Note that action detection is planned, but not part of the system. Currently, start and stop gestures are triggered by a wireless presenter in this use case.

A demonstration video is available online at **https://youtu.be/mcUcA-kWMQI**. The demonstration at the conference takes less space. The physical robot is substituted by a robot simulation on a screen. All use cases will be presented and will allow direct visitor participation.

REFERENCES

[1] Christoph Heindl, Thomas Ponitz, Andreas Pichler, and Josef Scharinger. 2018. Large Area 3D Human Pose Detection Via Stereo Reconstruction in Panoramic Cameras. In *Proceedings of the OAGM Workshop 2018*.

[2] Christoph Heindl, Sebastian Zambal, Thomas Ponitz, Andreas Pichler, and Josef Scharinger. 2019. 3D Robot Pose Estimation from 2D Images. In *International Conference on Digital Image & Signal Processing*.

[3] Thomas Kosch, Yomna Abdelrahman, Markus Funk, and Albrecht Schmidt. 2017. One size does not fit all: challenges of providing interactive worker assistance in industrial settings. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1006–1011.