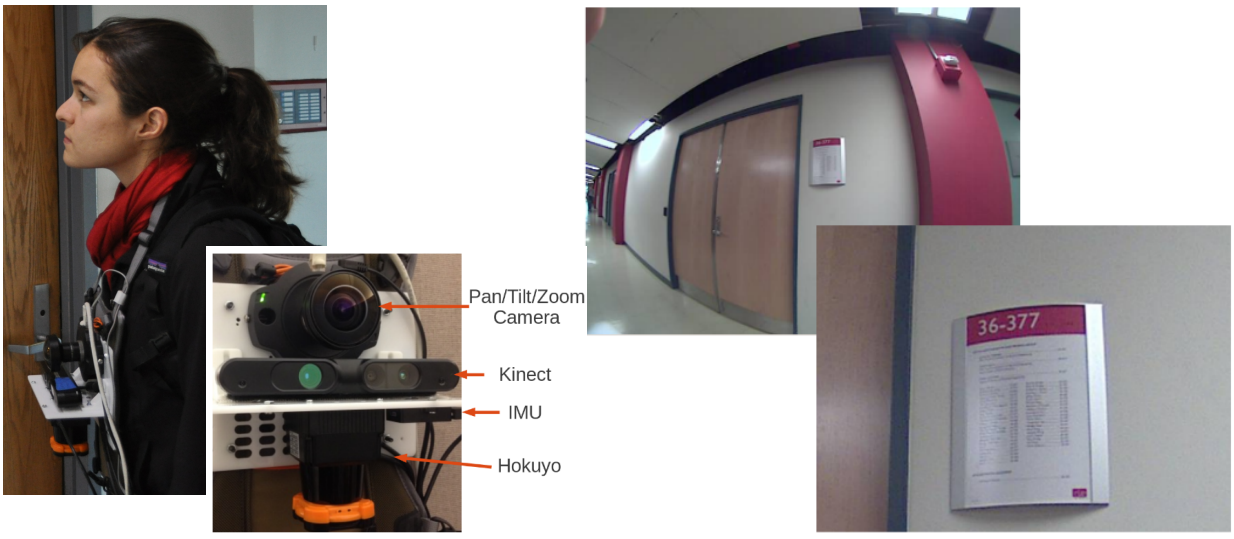
Environmental Text Spotting for the Blind using a Body-worn CPS

Hsueh-Cheng Wang, Rahul Namdev, Chelsea Finn, and Seth Teller
Robotics, Vision, and Sensor Networks Group,
Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT

Background.

Motivations. Information about environmental text is useful in many task domains. Examples of outdoor text include house numbers and traffic and informational signage; indoor text arises in building directories, aisle guidance signs, office numbers, and nameplates. However, such information is inaccessible to the blind or visually impaired (BVI). Different from traditional CPS (such as medical implants), we propose an non-invasive body-worn CPS as a substitute for the eyes to allow interactions among sensors, algorithms, and BVI users. A key design goal is to develop fast text detection methods for wide-field imagery, and accurate text decoding to support real-time decision-making, such as generation of navigation cues.

Challenges. Unlike scanned documents with high-resolution, scene text often occurs in small portion of entire field of view, which lack of enough pixels for resolution-demanding text decoding. Text observed by a moving camera will generally be subject to motion blur or limited depth of field (i.e. lack of focus). Blurry and/or low-contrast images as well as real-time constraint make it challenging to our goal. Neither increasing sensor resolution/ CPU bandwidth, nor deblur/superresolution algorithms alone, are likely to solve the problems; instead, improved methods combining both physical entitles and computational elements are required. Similar to classical CPS challenges, a real-time system with high-bandwidth and low latency networks that allows message passing and data marshalling among computation and physical processes is needed.



(a) (b)

Figure 1. A body-worn cyber-physical system. (a) A wearable sensor suite. (b) Wide-angle and foveated views.

Proposed Research.

Body-worn CPS. To achieve such goals, we take advantage of established frameworks and tools in robotic and machine perception, such as LCM (Lightweight Communications and Marshalling) and ROS (Robot Operation System). Equipped on-board sensors such as laser range scanners (LIDARs) or Kinect as shown in Figure 1, simultaneous localization and mapping (SLAM) can estimate accurate egomotion with respect to 3D scene surface of the surroundings. With the feedback loops, the proposed CPS 1) provides spatial prior for where text usually occurs, such as on vertical planar surface about shoulder height, 2) estimates where text decoding is likely to work well, e.g. on fairly close, fairly fronto-parallel scene surfaces observed by a sensor that was not moving too quickly during observation, and 3) incorporates temporal/spatial information to either integrate multiple observation to increase accuracy, or suppress previously-inspected regions to significantly reduce duplicative computation.

Human-CPS Interaction. Inspired by human foveal/parafoveal processing, we use pan/tilt/zoom (PTZ) cameras to first localize text in wide field of view, and then zoom in to foveate suspected regions due to the fact that the resolution requirement for text detection is different from decoding. Given the relative distance and viewing angle between the observer and a text region from SLAM, a PTZ camera can adjust its focus to remove out-of-focus blur and physically perform an inverse movement to observer, therefore reducing motion blur while tracing a text region. Using an electronic braille, blind users can not only access where text likely occurs in current field of view, but also control the PTZ cameras to foveate the region of their interests.

Potential Impact. Our work can lead to practical applications in many domains. In health care for the blind and visually impaired, a semi-autonomous “eyeball” can provide safety and navigation cues in unconstrained environments, such as signs and directions indoors or outdoors, labels of items in a supermarket, name tags of people in a conference room, and many other situations. In the field of intelligent transportation systems, text spotting can potentially improve visual media geo-localization, which determine where an image/video was taken. In augmented reality, fast and accurate detection of foreign-language street signs, store signage, menus etc. would greatly improve travelers' abilities to explore their surroundings.

Agenda. Possible milestones for the next two years include:

- developing a more robust and fast text detector against common false positive textures, such as window frames or brick walls in May, 2014,
- implementing an anti-motion blur devices in August, 2014,
- incorporating the electronic braille as an interface between our system and a blind user in Dec, 2014,
- carrying out evaluations for reliability in April, 2015,
- improving cosmetic design to be generally accepted in August, 2015, and
- exploring other practical applications by the end of 2015.