Multimodal Person Tracking and Attention Classification

Marek P. Michalowski
Carnegie Mellon University
Pittsburgh, PA 15213

Reid Simmons
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

The problems of human detection, tracking, and attention recognition can be solved more effectively by integrating multiple sensory modalities, such as vision and range data. We present a system that uses a laser range scanner and a single camera to detect and track people, and to classify their attention relative to a socially interactive robot.

Categories and Subject Descriptors: I.2.9 [Artificial Intelligence]: Robotics–*Operator interfaces, Sensors*; I.4.8 [Image processing and computer vision]: Scene analysis–*Sensor fusion*

General Terms: Algorithms, Measurement

Keywords: Human-robot interaction, social robotics

1. INTRODUCTION

In the case of social, face-to-face, human-robot interaction, the perception of humans is important because it facilitates a bi-directional flow of information: the robot must *understand* what is being conveyed by a human's attentional behavior, as well as *direct* its own verbal and non-verbal behavior to the human in an appropriate manner. For example, in the case of a robot whose purpose is to serve as a receptionist, there are a number of capabilities that depend on a good model of human attention: polite *greeting*; appropriate *verbalizing*; attentive *listening*; determining how to be of *service*; and *evaluating* effectiveness.

Various sensory modalities exhibit different strengths and weaknesses in terms of detail, resolution, and accuracy. Fortunately, the signal is often correlated between sensors, and the noise is often uncorrelated. It is possible for multiple sensors to be used to strengthen each others' complementary estimates while reducing the effects of noise or errors in any single modality. Our work aims to integrate spatio-temporal tracking of humans by a laser scanner with detected and tracked faces from a vision system.

2. SYSTEM DESIGN

The system is implemented on the Roboceptionist at Carnegie Mellon University [2]. Our system architecture was motivated by three design criteria: the robot should correctly attend to people (by turning its monitor and face in the direction of a person's face) as they approach and stand around it, the robot should recognize when people arrive and depart so as to maintain context for ongoing interactions, and the robot should verbalize (e.g. greet visitors) in an appropriate manner. The first requirement calls for location information for people in the environment, the second calls for tracking people's continuity of identity, and the third calls for inference about a more abstract property of the participants – that is, their level of engagement or their direction of attention.

HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA.
ACM 1-59593-294-1/06/0003.

2.1 Sensors

The robot receives data from a laser range scanner, a PTZ camera, and a keyboard interface used for interaction. The robot is situated behind a desk in a booth, the laser is mounted in the desk behind a slit at human knee-level, and the camera is installed on the front and top of the robot near human chest-height. The laser scanner performs a 180° radial sweep, giving the lengths of scan vectors radiating from the sensor. Adjacent range values that deviate from a learned background model are clustered and tracked using a Kalman filter (see [5]). The vision software returns a list of tracked faces with coarse pose information (whether the face is frontal or not). Frontal faces are detected using the method in [6] and used to initialize (and re-initialize) a mean shift color tracker described in [1]. These operators take different amounts of time to process a frame, and the system is multithreaded so that slow operators (e.g. face detection) run infrequently and faster operators that benefit from smooth motion (e.g. color tracking) run more often.

2.2 Sensor fusion

Goodridge [3] discusses levels of representation in sensor fusion. Our work uses both feature-level and symbol-level data fusion in order to detect people, track their location, maintain continuity of identity, and classify them according to their level of engagement with a social robot.

On the **feature level**, the goal of the system is to obtain a model of nearby people in terms of **spatial location and pose** so that the robot can visibly and appropriately direct its attention to people around its desk. That is, the robot should be able to determine where people are standing on the floor plane relative to the robot, where their heads are located in space, and the direction of head orientation. In order for output from the sensors to be correlated, the location features of detected objects must be transformed into a common reference frame. 2D polar coordinates from the laser and 2D image-space coordinates from the camera are transformed into 3D Cartesian coordinates, translated, and rotated to compensate for the sensors' positions and orientations relative to the robot. We use a nearest-neighbor

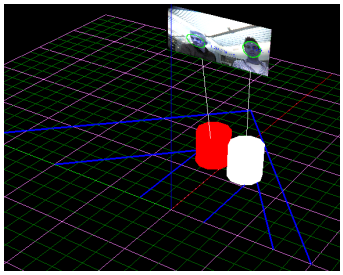

Figure 1: Visualization of sensor fusion between laser and camera in tracking people.

matching between the locations of people returned by the laser scanner and the projections onto the floor of the locations of faces returned by the camera (Fig. 1).

On the **symbol level**, the system serves two purposes:

To track identity – Frequently, a tracker will mistakenly determine that an object has disappeared and that a new one has been detected. From the perspective of social interaction, such an error results in a loss of continuity and context in an ongoing interaction. This system uses tracking information from one modality, if available, to attempt to compensate for discontinuities in the other.

To classify people according to their attention – For our purposes, we have designed a categorical model of attention: *present* (far from the robot), *attending* (idling closer to the robot), *engaged* (next to the booth), and *interacting* (actively participating in an exchange with the robot). It should be noted that this model is neither motivated by, nor reflective of, the psychology of human attention. Rather, it is a generative model designed for the purposes of specifying different behaviors of the robot towards people at different levels of engagement (see [2]).

Each sensory modality is able to contribute independent estimates about these attention levels with respect to detected people. For example, the classification can be done by laser scanner alone, based on physical location, or by camera alone, based on head pose estimation. However, after sensor data is fused on the feature level, the classification of a person's attention must be a single label that somehow integrates these separate estimates. This symbolic fusion is done in a hierarchical manner inspired by work in the cognitive study of the perception and mediation of social attentional cues [4]. In our system, the robot uses spatial location as an initial estimate of attention. If a frontal face has been associated with that location, the classification is adjusted to a higher level of engagement.

3. EVALUATION

Prior to the development of this system, the robot's attentive, tracking, and verbalizing behaviors depended on laser data alone. We observed failures in terms of each of the three design criteria described above: the robot had direction but not height information necessary to attend to a person's face, tracking discontinuities resulted in the robot treating an existing interactor as a new arrival to the booth, and the robot talked excessively to people who approached the booth but were turned away. The system described in this paper is successful in attenuating these three problems.

The continuous tracking of a person's identity was improved by fusing data from multiple sensors rather than using a single one. In a week of operation, 6486 matches were made between laser objects and camera faces, of which 4% were broken by a failure in laser tracking; the rest were broken by failures in face tracking (as the camera's field of view is smaller than the laser scanner's operating range). 40% of laser tracking failures were corrected by pairing the remaining face with a new laser object, and 48% of face tracking failures were corrected by pairing the remaining laser object with a new face. It is most important that we corrected failures that would have resulted in loss of context for current interactors. Of the 242 failures in laser tracking, 183 had involved a person interacting with the robot. The sensor fusion system achieved a 34% reduction in the number of interactions that would have been confusingly cut short by the single-modality tracking failures.

The sensor fusion system's classification of attention is more accurate than a classification based on location alone. Typing on the keyboard may be considered "ground truth" that a person is Interacting. If fused sensor data can more accurately predict when someone is about to interact with the robot, then verbal greetings will be more appropriate and less annoying (for people who approach the booth but do not intend to interact). In two weeks of operation, using the laser scanner alone, only 19% of 5315 people who were classified as Interacting (and were therefore greeted by the robot) ended up typing to it. With this system, on the other hand, 25% of people were correctly predicted to begin typing to the robot. The sensor fusion system decreased by 69% the number of inappropriately greeted people from the use of the laser scanner alone.

4. CONCLUSION

We have described a system that fuses laser and camera data on two levels: on a feature level, it detects and tracks the location of people, and on a symbol level, it classifies people according to their engagement with a social robot. The system has successfully attenuated a number of problems we had observed with our robot: it is now able to attend to people based on height in addition to location; it loses track of interactors less frequently, reducing confusion when context is lost; and it is less annoying to people who approach the booth but do not intend to interact. In future work, we will explore how a robot can use the ability to evaluate human attention in order to measure the effects of its behaviors on attracting and directing this attention.

5. REFERENCES

[1] G. R. Bradski. Computer vision face tracking as a component of a perceptual user interface. In *Applications of Computer Vision*, 1998.

[2] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang. Designing robots for long-term social interaction. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS '05)*, August 2005.

[3] S. G. Goodridge. *Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction*. PhD thesis, North Carolina State University, 1997.

[4] S. R. Langton, R. J. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 2(2), Februrary 2000.

[5] R. Simmons et al. GRACE: An autonomous robot for the AAAI Robot Challenge. *AAAI Magazine*, 24(2):51–72, 2003.

[6] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, 2001.