PANOPTICON: Person recognition and tracking through occlusion using extended Kalman filter

Boiar Qin and Ningchuan Wan

***Abstract*—Building upon some prior research in the areas of face and body recognition, we created PANOPTICON, a fusion of the Kinect camera and Erratic robot that recognizes and tracks humans and predicts their location, even when the targets' faces or bodies are not detected in camera view. Using face and body recognition algorithms with an extended Kalman filter, we demonstrate that a reasonable approach to the problem of tracking through occlusion can be implemented.**

I. Introduction

Robots and cameras are becoming more prevalent and increasingly applicable in a variety of day-to-day situations. With the introduction of Microsoft's Kinect, which allows a game player to use his own body as a controller, it is a short step to further enhancing casual gaming or realizing the possibility of robot security.

Currently gaming consoles like the Wii and Xbox involve manually signing in in order to play a game or access a player's profile. By adding in a facial recognition component, it would be possible for the console to recognize a player and sign them in automatically. Even more seriously, during a time where surveillance of airports and public spaces is critical to national safety, there is a need for a persistent and economic means of security that supercedes human fallibility. Robotic security guards could be manufactured with a fraction of the time it would take to raise and train a human guard, allowing security forces to be more flexible with their resources.

However, even with facial recognition, it is not likely that a robotic intelligence would be any better than a human one, given that robots are not innately able to predict movement and location. Humans, whom at a very young age are able to learn that objects and people do not simply disappear into thin air, are superior to robots in this sense. In the PANOPTICON project, we have taken on the challenge of "teaching" a robot to keep track of persons, even when their faces are unable to be seen, or when they have briefly disappeared behind large objects.

Using a Kinect mounted on a Videre Erratic mobile robot, we scan a room using realtime 3D data from the Kinect and use body and facial detection algorithms to recognize that people are in view of the camera. Using an Extended Kalman Filter, we are also able to track people through occlusion situations with a success rate of 92%. Additionally, by using the facial recognition algorithm in occlusion tracking gives a 51% success rate.

Boiar Qin and Ningchuan Wan are with the Department of Computer Science, Cornell University. {bq27, nw79}@cornell.edu

II. Related Work

Prior research has been done on all of the main parts of our project, however, previous results have yet to be combined in a satisfactory way.

Face detection has been researched for over a decade and is quite sophisticated now. It is present in many web applications and can be done over frontal and profile views. Current face detection methods are quite successful, working roughly 95% of the time [1].

Body detection has been done using skin-color detection and methods like heuristic comparison with generic body models [2]. However since the Kinect can gather 3-D data using infrared sensors, detailed depth maps of a scene can be taken and compared against the depth map of a human body.

Using Open NI and NITE, Kinect-based person followers have been written. [3]. These person followers must detect the body, get a “skeleton” and follow it, keeping a certain distance between the robot and human.

Occlusion tracking has been done quite effectively by Girondel et al. [4] for multimedia/video game purposes with a regular Kalman filter. However, person-detection is done by comparison of skin colors, so this method has room to be updated.

Additionally, most papers regarding occlusion tracking use a stationary camera. PANOPTICON builds upon previous research done on face and body detection, and for occlusion tracking, involves a camera mounted on a mobile robot.

III. Approach

Approach consists of body detection, face recognition, and the extended Kalman filter.

Open NI is used to detect bodies within view of the Kinect. Once a body has been detected, Open CV is used to detect faces within the body. The face is compared to faces stored in a database. If there is at least 0.9 confidence in the identity of a face, then the face is labeled as such—otherwise it is called “Unknown.” New faces and identities are added to the database by commanding the robot to 'train [name]' on a person's face within view. 100 images are associated with each person. Only face data is used to recognize a person, since information gained from clothing or height is not very long-lived or robust (people can change clothes, or stand too close to the camera).

12 features are taken from the body and face detection algorithms and used with the Kalman filter: The coordinates of the body and face are analyzed and the most extreme coordinates are taken to form a rectangular bounding-box. This gives 8 features (two points to define a rectangle, with each point defined by an x- and y-

coordinate). The distance from camera to person is another feature, and the velocity of the body centroid (v_x, v_y, v_z) are collected with respect to milliseconds. $\underline{x}^T = \{x_{body\text{-}left}, x_{body\text{-}right}, y_{body\text{-}top}, y_{body\text{-}bottom}, x_{face\text{-}left}, x_{face\text{-}right}, y_{face\text{-}top}, y_{face\text{-}bottom}, z, v_x, v_y, v_z\}$.

The control vector has only two elements—forward distance and radians turned by the robot. $\underline{u}^T = \{r, \theta\}$.

The 12 features are used differently, depending on whether the face and body are actually seen. If both face and body can be seen, then the normal input and control vectors are used in the filter. If the body is present, but face is occluded, the face position is estimated based on current body position and previous face location. If the body is occluded, then no face is seen either, so the previous measurements are used to estimate body and face position.

The evolution matrix A used was:

$$\begin{matrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & T & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & T & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & T & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & T & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & T & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & T & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & T \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{matrix}$$

where T is the time elapsed between updates. KFilter software [5] was used to implement the extended Kalman filter.

The robot can be given the command 'target [name]', whereupon, it will scan the surrounding area to determine if the target is there. Upon finding the target's face (with at least 0.9 confidence), the robot will drive follow the target, even if the target moves away from its original position.

In order to quantitatively determine the effectiveness of this approach, video data of many situations was collected and manually analyzed, frame by frame.

IV. Results

538 frames were analyzed, for a total of 829 instances of categorizable situations. Each situation involved two properties—the presence of a person(present and facing the camera as in Figure 1; present with face occluded as in Figure 2; and present with body occluded as in Figures 3 and 4) and the correctness of identification (correctly labeled with name, incorrectly labeled with name, labeled as unknown, not labeled). A control situation (no person present) was deemed correct or incorrect, depending on whether a person was labeled in the frame.

	Empty (127)	Facing (453)	Face Occ (157)	Body Occ (92)
Correct	48.8%	79.7%	45.2%	62.0%
Incorrect	51.2%	6.6%	35.7%	19.6%
Unknown		4.4%	17.2%	1.1%
None		9.3%	1.9%	17.4%

Table 1. Results from frame analysis. Numbers in parentheses represent total instances of situation

V. Examples

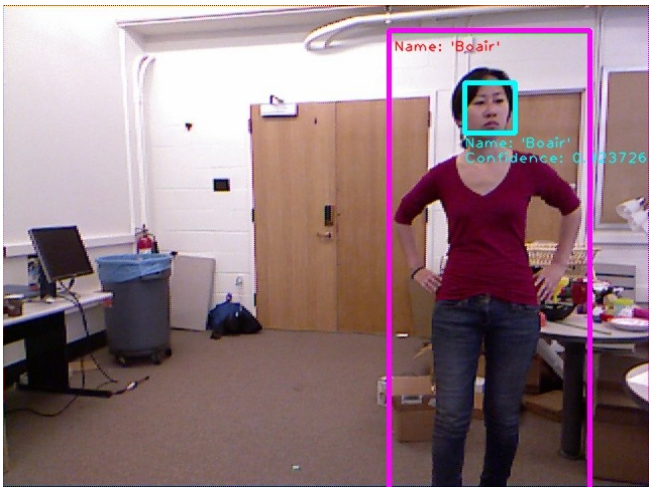

Figure 1. Correctly labeled, known person, front-facing

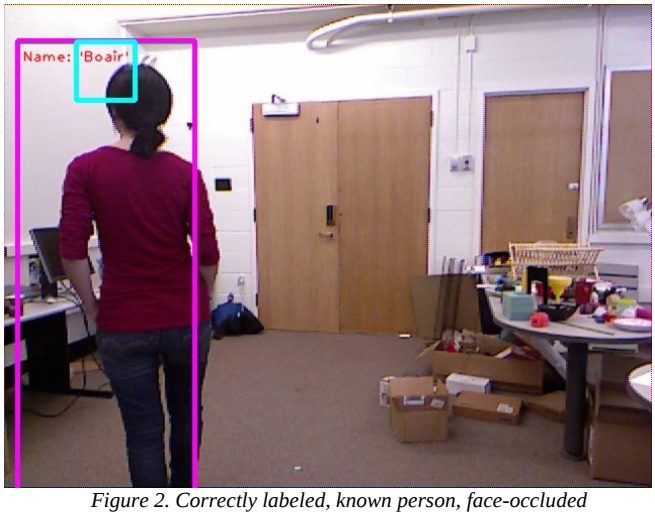

Figure 2. Correctly labeled, known person, face-occluded

Figures 3, 4. Correctly labeled, known persons, body-occluded

Figures 5, 6. Correctly labeled, unknown persons, some occlusion

VI. Analysis and Discussion of Results

Overall, the approach taken did moderately well at finding occluded persons. People facing the camera were labeled correctly about 80% of the time, while people with occluded faces were labeled correctly 45.2% of the time, and people with occluded bodies were labeled correctly 62.0% of the time. The success of this method is slightly diminished by the control situations, where there was about a 50-50 chance that it would be labeled correctly or incorrectly. Since the labeling of a person depends very highly on the labeling of a person's face, this way of quantifying results is thus highly dependent on the face detection and recognition algorithms. Quite often, a pole or the back of a chair is mistaken for a body, or a wastebin and a stack of boxes in the background is mistaken for a human face.

The use of the extended Kalman filter seems to have given a boost to identification of body-occluded persons over face-occluded persons. Although a higher percentage of people are identified in the face-occlusion situation, they are often classified incorrectly. This may have occurred because there are a higher number of situations where face-occlusion occurs. Additionally, in the case where occlusion occurs by turning the face, during the turning, the person's face may be mistaken for another person's, which results in the rest of the frames being labeled as the second, incorrect person. This problem may be a result of the face recognition software itself, or of the usage of only 100 reference images per person, most of which are taken against a normal, but somewhat cluttered background.

VII. Conclusions

Based on the results seen, it is not likely that any kind of robotic guard will be feasible in the near future. However, most of the weaknesses in the approach seem to lie in the reliance on Open CV's face recognition software. If face recognition isn't successful, than even if body detection has been done right, it will not be labeled correctly. However, if a more robust implementation of face recognition is available, it can easily be inserted into PANOPTICON's code to be used with body detection and occlusion tracking.

In order to boost the success of face recognition, it would be best to train on images taken against a plain background so that none of the items in that background are mistakenly classified as faces. More than 100 reference images would also be useful, as well as a variety of different lighting situations.

Body detection is slightly more of a black box. At times detection can be fickle, dropping and creating a “new” user even when the same person has been standing in front of the camera unoccluded. It would be worth the time to “consolidate” previously seen persons with supposed “new users” in order to cut down on the possibility of one person having multiple bounding boxes. Care must also be taken to ensure that this does not combine the bounding boxes of two people crossing behind or in front of each other.

Although there is plenty of room for improvement and a long road before it, this project represents the inception and execution of a high-level idea which uses machine learning to achieve a practical means. Overall, we feel that given a semester's time, PANOPTICON has achieved a reasonable level of success.

Acknowledgments

The authors would like to thank Ashutosh Saxena for excellent instruction regarding concepts in artificial intelligence and machine learning, and for allowing the use of lab equipment and robots; Akram Helou and Marcus Lim for aiding with ROS installation; Norris Xu for assistance with robot movement and interfacing; and especially Colin Ponce for taking the time to hold weekly status meetings and for helping develop the focus of PANOPTICON.

References

[1] H. Schneiderman, T. Kanade. “A Statistical Method for 3D Object Detection Applied to Faces and Cars”. Carnegie Mellon University, 2000.

[2] A. Micilotta, E. Ong, R. Bowden. “”Real-time Upper Body Detection and 3D Pose Estimation in Monoscopic Images”. University of Surrey, 2006.

[3] G. Gallagher. Kinect-based Person Follower. MIT, 2010.

[4] V. Girondel, A. Caplier, and L. Bonnaud, "Real time tracking of multiple persons by Kalman filtering and face pursuit for multimedia applications". Universite Pierre Mendes France, 2004.

[5] V. Zalzal. KFilter: A C++ Extended Kalman Filter Library.