ID-Match: A Hybrid Computer Vision and RFID System for Recognizing Individuals in Groups

Hanchuan Li[1,2], **Peijin Zhang** [1,3], **Samer Al Moubayed**[1], **Shwetak N. Patel**[2], **Alanson P. Sample**[1]

Disney Research[1]
Pittsburgh, USA
{samer, alanson.sample}
@disneyresearch.com

Computer Science & Engineering[2]
University of Washington
Seattle, USA
{hanchuan, shwetak}@cs.washington.edu

School of Computer Science[3]
Carnegie Mellon University
Pittsburgh, USA
xzhangpeijin@gmail.com

ABSTRACT

Technologies that allow autonomous robots and computer systems to quickly recognize and interact with individuals in a group setting has the potential to enable a wide range of personalized experiences. However, existing solutions fail to both identify and locate individuals with enough speed to enable seamless interactions in very dynamic environments that require fast, implicit, non-intrusive, and ubiquitous recognition of users.

In this work, we present a hybrid computer vision and RFID system that uses a novel reverse synthetic aperture technique to recover the relative motion paths of an RFID tags worn by people and correlate that to physical motion paths of individuals as measured with a 3D depth camera. Results show that our real-time system is capable of simultaneously recognizing and correctly assigning IDs to individuals within 4 seconds with 96.6% accuracy and groups of five people in 7 seconds with 95% accuracy. In order to test the effectiveness of this approach in realistic scenarios, groups of five participants play an interactive quiz game with an autonomous robot, resulting in an ID assignment accuracy of 93.3%.

Author Keywords

Human-Robot Interaction; Recognition; Synthetic Aperture; RFID; Computer Vision; Sensor Fusion

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Have you ever ran into someone and forgotten their name? Or even worse did not recognize them at all, only to find out you've met before! These panic filled moments are not only awkward and unpleasant for you, but can lead to hurt feelings and an aversion to future encounters by the other person. Now imagine an autonomous robot trying to have casual and meaningful encounters with people. If the robot calls someone by the wrong name or fails to recognize they have had a previous encounter, it can have disastrous effects when trying to build social relationships.

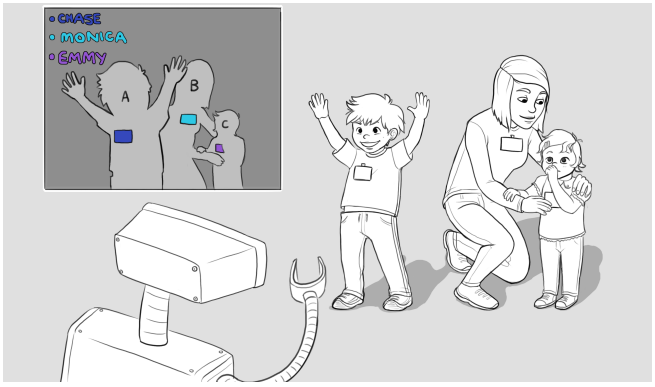

Figure 1. An illustration depicting and autonomous robot interacting with people wearing RFID tags. In order to provide a personalized experience the robot must quickly and precisely correlate the RFID and computer vision data to determine who and where the individuals are.

The ability for autonomous robots and computing systems to quickly recognize individuals is an important step to enabling natural encounters and personalized experiences that grow over multiple interactions. However, this becomes especially challenging in very dynamic environments that require fast, implicit, non-intrusive, and ubiquitous identification of the users.

Recognition systems based on computer vision (including stereo cameras, structured light, and LiDar) can robustly determine where people are in its field of view, but it still remains an open research challenge to identify who those individual people are. State-of-the-art face recognition systems have shown promising results when given large amounts still photos for training [26, 22]. However, face recognition requires a cumbersome user registration process consisting of photographing and manually annotating each participant for training purposes, which is not feasible when scaling to a large number of participants or for casual encounters. In addition, face recognition software is limited by constraints such as orientation, light condition and face resolution.

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858209

Alternatively, recognition approaches based on wireless signaling require active participation by the user, but significantly increases the reliability of identification. This typically requires the user to activate an app on their cell phone for each interaction or carry a wireless transponder [23, 15]. While these devices actively transmit data to identify who is within the vicinity of the robot or computing system, it is very difficult to determine precisely where the person is. This is primarily due to the nature of propagating electromagnetic waves, which makes it difficult to precisely locate a transmitter with an accuracy greater then 1-2 meter [17, 29, 19].

In this work, we propose a combination of light-based computer vision, and RF-based wireless communication to harness the best of both worlds (i.e. location and identification). An example scenario is depicted in Figure 1, where a family, wearing UHF RFID tags, approaches and interacts with a robotic character. To create a rich interactive experience the robot needs to be able to recognize people it has seen before, look at an individual in their eyes, and call him or her by their correct name. First the robot scans the family's long range UHF RFID tags and can infer that Chase, Monica, and Emmy are coarsely standing in front if it. At the same time, the robot's vision system analyzes the scene and determine that there are three people (i.e. skeletons) in the field of view. This raises what we are referring to as the "ID association problem". The challenge is to determining which name (Chase, Monica, and Emmy) belongs to which skeleton (A, B, or C), all while people are actively moving throughout the scene and with many additional RFID tags visible to the reader in the background.

To accomplish this, we have developed a method for correlating the time traces of the free-roaming synthetic apertures of the worn RFID tags (using a single RFID reader antenna), with the position traces of people (using a single depth camera). This new technique is further enhanced by using a support vector machine to correlate the changes in low-level RFID channel parameters such as (RSSI and Phase) as the tags are moved in space, to the motion of the individuals as seen by the depth camera. Finally, a probabilist voting system is implemented to assign ID to the people in the scene.

Our real-time system called *ID-Match* is capable of simultaneously recognizing and correctly assigning IDs to individuals in 4 seconds with 96.6% accuracy and people in groups of five in 7 seconds with 95% accuracy. Additionally, it is demonstrated that the system can operate in multi-path rich environments and distinguish between nearly identical motions between users without loss of accuracy. In order to test the effectiveness of *ID-Match* in realistic scenarios, groups of five participants play an interactive quiz game with an autonomous robot, and results show an ID assignment accuracy of 93.3%. Finally, the robustness of *ID-Match* is evaluated with a 7.5-hour test where 21 participants were autonomously recognized throughout a working day.

Contributions

We develop a novel hybrid computer vision and UHF RFID system capable of recognizing individuals walking in groups while wearing RFID tags. This real-time system is effective at operating in multi-path environments and under challenging usage scenarios.

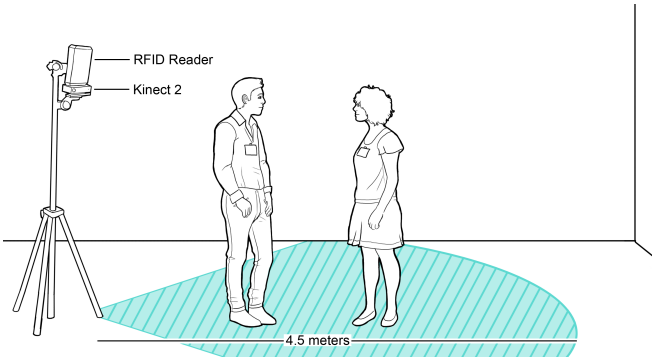

Figure 2. An illustration of a typical implementation of the *ID-Match* system consisting of a Kinect depth camera and UHF RFID reader.

- A reverse synthetic aperture radar technique for measuring the relative motion of UHF RFID tags for correlation to motion paths captured with a 3D depth camera
- Development of new features and a machine learning pipeline to correlate the tags low-level RFID channel parameters to body movements as measured with a 3D depth camera
- A real-time system capable of simultaneously recognizing five individuals and giving them a personalized robotic interaction

SYSTEM OVERVIEW

ID-Match is a real-time, hybrid computer vision and UHF RFID system that can simultaneously track, and individually identify multiple people wearing RFID tags in multipath rich, real world environments.

Figure 2 shows a typical implementation of the *ID-Match* system consisting of an Impinj Speedway Revolution UHF RFID reader [2] with a single antenna [1], along with a Kinect v2 depth camera [3]. Users wear low cost (7-15 cent each) UHF RFID tags in the form of clip-on name badges or lanyard. As people walk within view of the system the RFID tags are continuously read and the Kinect tracks the position of the unknown skeletons in 3D space. Since passive UHF RFID tags have a read range of up to 10 meters and a single reader can cover 50-150 square meters, the challenge is to determine which ID belongs to which skeleton.

ID-Match accomplishes this using two independent correlation pipelines, which are combined to provide a fast and accurate determination of the precise location and ID of individuals. The procedure is outlined below and is covered in greater detail in subsequent sections.

Reverse Synthetic Aperture Pipeline:

1. Use a reverse synthetic aperture technique to determine the relative radial path of each of the RFID tags to the RFID reader's antenna

2. Use the Kinect to track each person in 3D and determine his or her equivalent radial path in the coordinate frame of the RFID reader antenna
3. Compare and rank each of the tags synthetic aperture paths to the visual motion paths and place in a voting buffer

RF Motion SVM Pipeline:

1. Record the RSSI, phase, and channel number for each tag and extract seven RF motion features
2. Track the location of each skeleton using the Kinect and extract three motion features
3. Use SVM machine learning to classify similar RF motion features to physical motion features and place results in a voting buffer

Final Step:

1. Combined the two independent voting buffers and determine the final tag ID to skeleton correspondence

RELATED WORK

At the core of the *ID-Match* system, is a method for determining the fine grain change in distance between the reader and the tag using a reverse Synthetic Aperture Radar (SAR) approach. These RFID SAR paths are then compared to the position paths of the people as reported by a 3D depth camera. This core capability is augmented with a machine learning approach that builds upon [16], and classifies changes in low-level RFID communication channel parameters to the motion of a person in space.

Synthetic Aperture Radar (SAR) and Angle of Arrival (AoA) techniques are regularly used in wireless communication systems to locate active transmitters [29, 19, 14, 17, 7]. Many of these general approaches have been adapted to the UHF RFID space. Where RF channel parameters such as Received Signal Strength (RSSI) and RF Phase can be used to locate tags with accuracy on the order of several 10s-100s centimeters [20, 24, 13], for well-controlled environments such as anechoic chambers and well-structured portals.

In order to increase localization resolution of tags in uncontrolled, multipath environments several systems use SAR antennas on the RFID reader. However, these approaches face limitations when both the reader and the tag are moving. For example Miesen et al. [18], used an antenna on a linear actuator to create a synthetic aperture. However, this requires that the environment remains static while scanning occurs. Wang et al., [27] uses a spinning antenna to create a synthetic aperture but requires densely spaced marker tags place throughout the environment to disambiguate the mobile tag motion from reader antenna motion.

Alternatively, a modified AoA approach using multiple RFID readers, each with four antennas, demonstrates the ability to determine the trajectory of a single RFID tag in space [28]. Yang et al., [30] demonstrated the use of multiple RFID reader antennas that can locate multiple moving tags in harsh multipath environments. However, since they are not using a vision system that can track the objects path, that tags can only be located while traveling at a constant velocity along a known trajectory, i.e. on a conveyor belt.

A variety of mobile robotic systems have used RFID tags for navigation [11, 21] and to localize tagged objects [12, 8]. While these robots have computer visions systems, they are typically used for object manipulation instead of real-time object or human tracking. One notable exception is Grema et al. [25] which built an eight element phase array into a mobile robotic platform to determine the AoA of people wearing RFID tags. The robot employed a computer vision system to boost identification accuracy up to 83% for 4 people. In contrast, the *ID-Match* system can track five people using a single antenna with an accuracy of 93%.

Two other hybrid computer vision and RFID systems have been reported with the goal of identifying and locating tagged people. The closest prior work by Cafaro et al. [6], demonstrates the ability to determine which of two people are standing on the left and right of an interactive display. The second example by Goller et al. [10] shows an occupancy counting scenario where the authors are able to simultaneously determine the direction of travel for two people. Both of these systems use RSSI fingerprinting which requires extensive training and multiple reader antennas. As is the case with all fingerprinting based location schemes, changes in the RF environment will cause loss of accuracy and require the system to be retrained. Additionally, both systems are limited to tracking at most two people.

REVERSE SYNTHETIC APERTURE RFID

Synthetic Aperture Radar (SAR) techniques have been widely used to increase the imaging resolution of radar systems [7] and the localization accuracy of both radio transponders [17, 29] as well as UHF RFID tags [30, 27, 18]. This is accomplished by moving the objective antenna (i.e. reader antenna) along a known trajectory at a constant velocity, to synthesize a large array of antennas. Localization accuracy is increased since the objective antenna can take multiple samples of a stationary target from different angles.

In contrast, the goal of the *ID-Match* system is to locate and identify people wearing RFID tags as they naturally walk and play in their environment. This breaks traditional SAR approaches since both the objective antenna and the target object would be simultaneously moving at unknown velocities. This would normally result in blurry radar images and large localization errors.

In this work the SAR paradigm is reversed as shown in Figure 3, panel A. Here the RFID reader is placed in a fixed location and the motion of the RFID tag (when worn by a person moving) creates a synthetic aperture. Since the exact trajectory of the people wearing tags is not known (i.e. they are "free roaming") it is not possible to directly compute the exact location of the tags.

One of the key attributes of UHF RFID systems is that the tags do not actively generate and transmit radio waves. Instead the tags back-scatter (i.e. reflect) the RFID reader's carrier wave back to the reader in order to send data. This unique feature of the UHF RFID physical layer means that

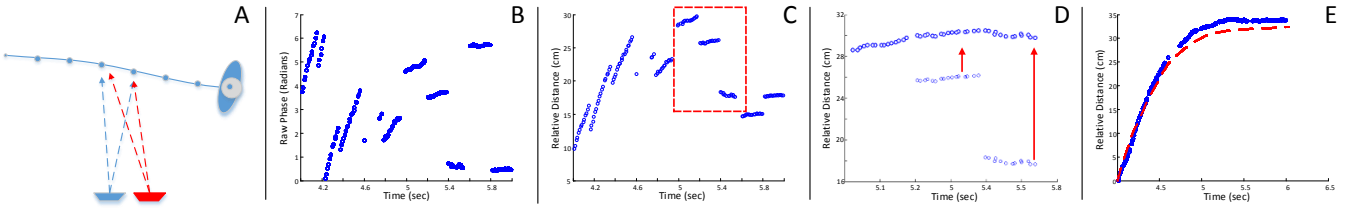

Figure 3. Overview of the signal processing algorithm used to associate RFID SAR data to Kinect motion data. Panel A shows a motion path of a person wearing and RFID tag while being continuously measured by the *ID-Match* system (top view). Panel B shows a two-second snapshot of the raw RFID phase data as a function of time. In panel C, the phase data has been converted to the radial distance from the antenna. In Panel D the phase discontinuities cause by frequency hopping are "stitched" back together. Finally, panel E shows the RFID SAR trajectory (blue circles) compared to the radial motion of the person as measured by the Kinect (dashed red line).

in a single read event, the RFID reader can precisely measure the phase angle between its transmitted signal and the reflected signal received from the tag. Figure 3, Panel B shows a two-second snapshot of the raw phase measurement as a function of time, for a person wearing an RFID tag walking towards the *ID-Match* system. The discontinuities in the reported phase are partially due to 2π radian wrapping as can be seen at 4.2 seconds. An additional source of error is due to an unknown phase offset introduced when the reader pseudo-randomly frequency hops from one carrier to another as required by government regulations [9]. This manifests in panel B as a grouping in the phase data. Where each group consists of a series of data points read consecutively at the same frequency.

Using equation 1, it is straightforward to calculate the relative radial distance from the tag to the reader, where ϕ is the phase angle reported by the RFID reader, f is the frequency of the RF carrier, and c is the speed of light.

$$Distance(relative) = \frac{\phi c}{4\pi f}; 0 < \phi < 2\pi \quad (1)$$

Figure 3, Panel C shows the phase data converted to distance. However, since the phase angle between the transmitted and reflected signal will rotate 2π radians for every λ wavelength, it is not possible to calculate the absolute distance, but instead rather the relative distance. Given the inertia of a person walking, and the high sampling rate of the reader, it is reasonable to assume that large discontinuities in distance are not due to human behavior, but rather an artifact of the RFID reader frequency hopping. Thus, given the channel number reported by the RFID reader it is possible to "stitch" the disconnected groups back together. This is shown in Figure 3 panel D, where the slope of the trailing points of one group are aligned with the slope of the leading points of the next group.

Working in conjunction with the RFID reader the *ID-Match* system uses a Kinect 3D depth camera to track the location of people with in its frame of view. In this work people wear RFID tags on their torso, as either name tags or lanyards. By computing the distance of the person's spine (as measured by the Kinect) to the RFID reader antenna, it is possible to correlate the physical path of the person, to the RFID tag's SAR path. It should be noted that due to phase wrapping, there is an equivalent distance wrapping, and the absolute distance of the tag to the reader cannot be determined. Thus for a given time window both the Kinect distance data and RFID SAR distance data are aligned at the beginning. The final result is shown in Figure 3, Panel E, which shows good agreement between the Kinect (red dashed line) and RFID reader (blue circles).

When multiple people are walking in front of the *ID-Match* system it will result in multiple RFID SAR paths and multiple Kinect motion paths. By taking the standard error between the sets of paths it is possible to determine which path belongs to which person. Details on the ranking algorithm are presented in the ID Association section.

PROBABILISTIC ID-BODY CORRELATION

In addition to the synthetic aperture approach described above it is possible to use low-level channel parameters along with machine learning techniques to determine the correlation between bodies in the view of the Kinect and tags within the view of the RFID reader. The key idea is that each time the RFID reader interrogates a tag, it reports channel parameters such as Received Signal Strength Indicator (RSSI), RF Phase, and Doppler shift, along with the channel number, representing a unique signature of the RF environment of that individual tag. By observing the changes in these channel parameters over time, it is possible to correlate the motion signature of a tag, with the motion of the person wearing it.

In order to give some intuition into how these RF parameters can be utilized Figure 4 shows a 30-second trace of RSSI and RF Phase trace of a single RFID tag. For the first 10 seconds the tag is held still, next the tag is waved at a slow speed for 10 seconds, and then at a faster rate for the remaining 10 seconds. The tag "still" state can be distinguished from motive states, by observing the rate of change of either the RSSI and/or RF Phase as a function of time, as can be seen in panel A.

It should be noted that the FCC regulations require RFID readers in the 915 MHz ISM band to pseudo-randomly change their transmit frequency in order to minimize interference with other devices. To satisfy this requirement, RFID readers "frequency hop" across 50 channels from 902 MHz to 928 MHz (in the USA) at a time interval of approximately 0.2 seconds. This results in the dramatic discontinuities in phase data as a function of time. To better reveal the un-

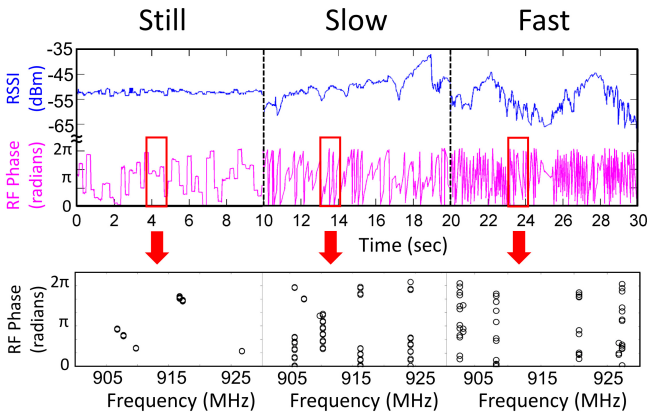

Figure 4. A plot of RSSI and RF Phase data for a RFID tag undergoing still, slow, and fast motion. In panel B, repotting phase vs channel frequency for a given time window reveals highly structured patterns that are positively correlated to tag motion.

derlying characteristics of phase hidden by frequency hopping, we take three 1-second slices of phase from the 3 different states and re-plot against channel frequency in Figure 4 (lower panel). For still states, phase is linearly correlated with the channel number resulting in lines with constant slope, that wrap between 0 and 2π. Motive states of the tag result in increased phase variation within each channel, which is positively correlated with the intensity of motion.

In prior work Li et al. [16] used RFID channel parameters including RSSI, RF Phase, and tag Read Rate to classify a tag as either being still, moving, cover by a hand, or swipe touch. Building upon this approach new features have been specifically designed for the task of correlating tag motion to the motion of people as measured by a depth camera.

RFID Phase Features

Note that all features are calculated using RF channel parameters within each segments. The first two RF phase features are based on the tags velocity using equation 2, which is calculated from consecutive tag reads on the same channel [20].

$$v_r = \frac{\lambda(\theta_1 - \theta_2)}{4\pi(t_1 - t_2)} \quad (2)$$

1. Radial Velocity: the average of v_r.
2. Absolute Radial velocity: the average of the absolute v_r.
3. Standard error of the simple linear regression of unwrapped phase versus channel frequencies. $stderror(linearfit(channel, unwrap(phase)))$
4. Average of phase change divided by frequency change when carrier signal frequency hops. $\sum_{i=1}^{k-1}((phase(i+1) - phase(i))/(frequency(i+1) - frequency(i)))/(k-1)$

RFID RSSI Features

Given the RSSI characteristics observed in Figure 4, change in distance will create variations in the RSSI signal which is employed in the following three features.

1. Average RSSI standard deviation within each channel
2. Average RSSI difference between consecutive samples within each channel
3. Absolute value of the RSSI

RFID Read Rate

The read rate of a given tag is primarily correlated to the amount of RF power it can receive. This typically means that RFID tags with a low read rate are on the edge of the read zone which is well outside the field of view of the Kinect. Thus, this feature is mainly employed to exclude tags that are unlikely to be a viable candidate for ID to body matching.

1. Read rate: number of packets received from each RFID tag per second

Kinect Motion Features

Both RSSI and phase are sensitive to movement in the radial direction to the reader. Thus, the position of the spines on the skeletons reported by the Kinect are transformed from the Kinect's Cartesian coordinate system to a polar coordinates centered at the reader. Once the skeleton tracking is started, the following three features are utilized.

1. Radial component of the skeleton's velocity relative to the RFID reader antenna
2. Azimuth (i.e. non-radial) component of the skeleton's velocity relative to the RFID reader antenna
3. Distance between the skeleton's spine and the RFID reader antenna

Data Segmentation & Machine Learning

A Support Vector Machine (SVM) classifier with the Radial Basis Function (RBF) kernel was trained by recruiting three participants to wear RFID tags while walking freely in the view of the depth camera and the reader (ground truth was taken manually). The data stream was segmented with a window size of 2/3 second which was advanced every 1/3 second resulting in a 50% overlap. This achieved a good balance between matching accuracy and latency. The classifier was only trained to two prediction classes, "Match" and "Not-a-Match". The parameters of the SVM classifier where optimized by maximizing the 10 fold cross validation results.

Since the system is trained for large movements that are generic to a majority of people, and since the machine learning features are differential and/or relative to the reader antenna, training does not have to be redone for new environments or people. In fact, training was done on one set of participants at one location, while all testing was done at several other locations with many other participants over the course of several weeks. Finally, In the following section, the output of the SVM classifier is combined with the results from the SAR method to provide final decisions on body to ID matches.

ID ASSOCIATION ALGORITHM

The true strength of the *ID-Match* system is that ID association is based on similarity ranking between a finite number

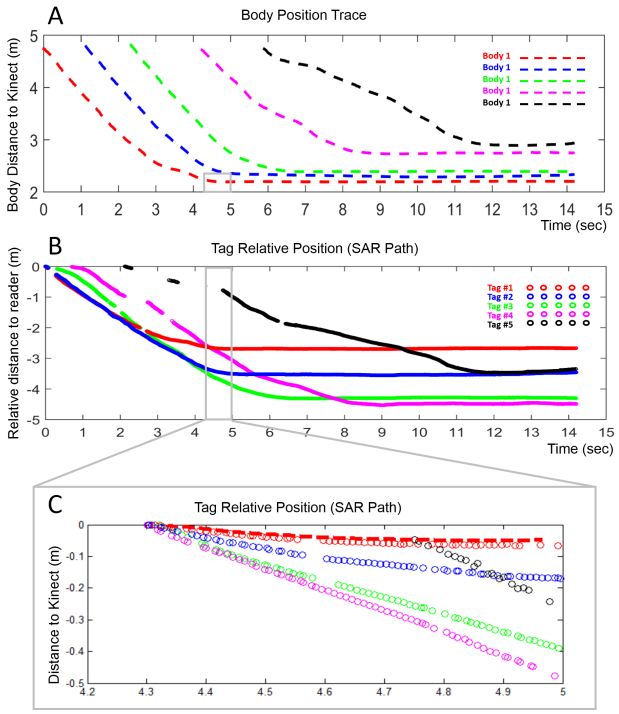

Figure 5. Panel A shows the distance in meters for five participants as tracked by the depth camera. Panel B shows the distance calculated using the SAR approach of the same five participants. Panel C shows a time slice of the motion path of 'Person 1' as measured by the depth camera (dashed red line) compared to all the SAR RFID tag motion paths over the same time period. Since 'Person 1' (dashed red line) closely matches 'Tag 1' (red circles) they have the lowest standard error and are considered a 'match'.

of possibilities, rather than the raw ability of the RFID reader and/or the Kinect to precisely locate tags and people. This is accomplished by measuring small variations in tag and body motion of people as they walk in order to differentiate between them. While these motion difference can be hard to visualize, they are statistically distinct. In order to illustrate how the ID association algorithm works we offer a "toy example" of five people walking in a single file line, towards the RFID reader and Kinect mounted on a tripod. While this does not represent a real usage scenario (which is presented in detail in subsequent sections) for the sake of an example, it simplifies the motion the people and ensures that there are no blocking events where one person is visually occluded by another.

Figure 5 panel A, shows the distance traces for five participants (i.e. Persons 1-5) walking toward the system as measured with the Kinect. Panel B shows the relative distance of the RFID tags (Tags 1-5) as worn by the five people over the same period of time. Since the RFID reader can detect the tags at a range of 7-8 meters they are visible to the system before the Kinects depth camera can track them (which occurs at 4.5 meters).

The *ID-Match* system is implemented in Matlab and is capable of assigning IDs to people in real time. This is accomplished by applying a 2/3 second sliding window (which is advanced every 1/3 seconds) to the incoming data stream from the Kinect camera and RFID reader. During each time segment the system analyzes the buffered data and applies the SAR and SVM techniques previously described.

In the case of the SAR approach, the standard error between each of the five SAR tag traces is computed across each of the five participants position traces. An example can be seen in Figure 5 panel C which shows an expanded view of a 2/3 second time slice from panels A and B. Here it can be seen that 'Tag 1' (red circles) is the best fit for 'Person 1' (dashed red line), which will result in the lowest standard error and highest confidence. Thus, every 1/3 seconds the system outputs new matching events and places them in the SAR results buffer for each person seen by the Kinect. It should be noted that SAR data is only considered valid for people in motion.

During the same time window, the 8 RF features for the five RFID tags and 3 Kinect features for the five bodies are passed to a trained SVM classifier which generates a probability estimation of each Tag and Person pair. For each individual body, if a given RFID tag has a probability with a margin of 30% or more when compared to all the other RFID tags, it is considered a matching event and the SVM results is placed in a result buffer.

Each of the five bodies is assigned their own final results buffer consisting of a FIFO of the last 10 prediction points from the SAR and SVM predictor. The arithmetic mode is taken as the final matching result, and the ID is assigned to the body. Since the prediction FIFO for each body seen by the system can be updated over time, misidentifications can be corrected for. Once sufficient confidence is obtained, the identity of the body (i.e. tag / body pair) is "locked in".

PERFORMANCE AND EVALUATION

In this section *ID-Match* is evaluated under a number of controlled scenarios so that the underlining behavior of the system can be better understood and quantified. In particular, we evaluate the ability of the system to: recognize individuals in groups of people, distinguish between multiple tags undergoing the same motion, and reacquire people after the identity match has been lost.

Recognition of Individuals In Groups

When multiple people, walking as a group, approach the *ID-Match* system. This creates a number of dynamic events where people in the foreground can visually occlude (and possibly RF occlude) people in the background. Additionally, since the RFID SAR motion paths are a relative measure of distance and not a trace of position over time, there is the potential that the paths may not be unique, causing confusion in the matching algorithm.

In this study, two groups of five participants wore RFID tags on their upper torso. They were asked to start outside of the Kinect's field of view (approximately 6-7 meters away from the system), and walk as a group up to the *ID-Match* system. They were instructed to stop in one of five general locations such that upon arrival in front of the *ID-Match* system they would be standing *five a breast* and could visually see the

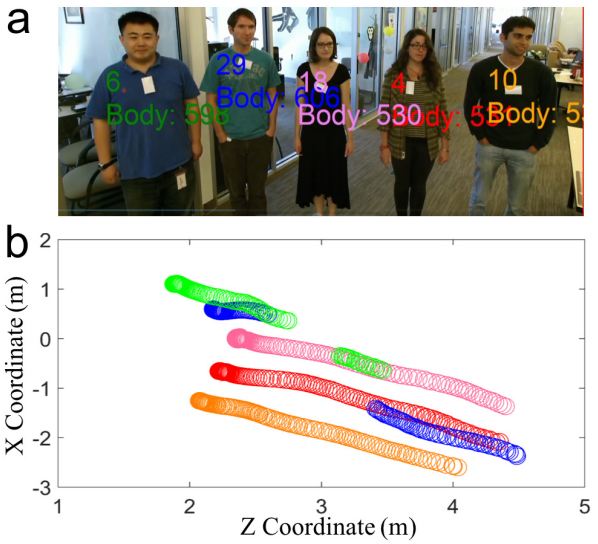

Figure 6. Panel A shows the RGB image captured by the Kinect as five people walk up to the *ID-Match* system. Overlaid on the image is the RFID tag ID that has been automatically associated with that person. Panel B shows a top view diagram of the paths of the participants as recorded by the Kinect.

People Wearing Tags	Assigned to →	RFID Tags Visible to the RFID Reader: ID 1	ID 2	ID 3	ID 4	ID 5	ID 6 (dummy)	ID 7 (dummy)	Error States: Pending (not matched)	Not Seen by Kinect
	Body 1	11	0	0	0	0	0	0	0	1
	Body 2	0	11	0	0	0	0	0	0	1
	Body 3	0	0	12	0	0	0	0	0	0
	Body 4	0	0	0	12	0	0	0	0	0
	Body 5	0	0	0	0	10	0	0	1	0

*Ground Truth: Body 1 = ID 1, Body 2 = ID 2, Body 3 = ID 3, Body 4 = ID 4, Body 5 = ID 5

Table 1. A confusion matrix showing the results for assigning IDs to five participants as they approach the *ID-Match* system.

system. Figure 6 panel a shows the final location of one set of participants of one trial as seen by the Kinect, and panel b shows the path they took as reported by the Kinect. For ground truth each person was assigned a "Body Number" (1-5) and the corresponding RFID ID (1-5) was also recorded. To emulate a more real world scenario and to ensure the problem space was not unduly constrained, two dummy tags were also introduced into the view of the reader (IDs 6 & 7). Each group of participants was asked to repeat the experiment 6 times for a total of 60 potential matching events. The two experiments are folded together into the truth table shown in table 1.

These results show and overall ID association accuracy of 95%. There were two trials where the Kinect was not able to see one of the participants as indicated in the two columns to the right in table 1. Excluding this data point results in an accuracy of 98.3% when both the Kinect and RFID reader can view all people. In one trial an *ID* was not assigned to a *body* leaving body 5 "pending", meaning the confidence threshold for a match has not been met. Observations of the trials suggest the errors are due to occlusions.

One of the important criteria of the proposed system is that it should be able to quickly identify and recognize individuals such that an autonomous robot can properly interact with them in a timely manor. However, for the *ID-Match* system calculating the exact acquisition time for groups of five people is a multi-dimensional problem consisting of (but not limited to); the walking rate of individuals, the total time all participants take to arrive at the destination zone, the fact the some people will enter the Kinect's depth field of view before others, the issue of people dynamically blocking the view of the Kinect, and the resulting need for target reacquisition and ID association.

Instead of quantitatively addressing each one of these variables individually the following plot qualitatively shows that, given all the above variables, *ID-Match* is indeed able to identify individuals in a timely manor. The red line in Figure 7 shows the accuracy of matching the correct *ID* to an individual as a function of time. Since there is variability in the time of arrival and walking speed, the data has been normalized to the point in time when each participant first walked into the view of the Kinect's depth camera. Results show that *ID-Match* is capable of correctly assigning IDs to individuals within 4 seconds with 96.6% accuracy.

The second blue line in the plot shows the accuracy of assigning the correct ID to all five participants in the group as a function of time. The reason this plot takes longer to converge is that the system must wait for all participants (regardless of how fast they walk) to enter the field of view. Results show that a group of five participants can be correctly matched within 7 seconds with 95% accuracy. It should be noted the line for the individual acquisition time has a higher accuracy than the group case because the two people not seen by the Kinect are excluded.

A detailed examination of the SAR and SVM voting buffers for the above data shows the predictive power of each technique individually. For the data shown in Figure 7, at t=4sec for each individual, the SVM classifier has an accuracy of 70.1% and the SAR approach has an accuracy of 89.4%, while the combination of the two results in a total of 96.6% accuracy.

Disambiguating Similar User Motion

One important question is how unique does the motions of the users and the paths they take, have to be for the ID Association algorithm to work properly. One concern is that groups of people walking in *lock step* may have nearly identical RFID SAR paths. Since the measurements RFID SAR of relative distance and not absolute distance it may not be possible to distinguish them from each other.

Given the complexity of precisely and repeatedly controlling the motion of five people, an alternative and arguably more stringent test has been implemented. Figure 8, panel A shows an image of five RFID tags (in the form of name tags) attached to a wooden rod. They are spaced evenly approximately 40 cm apart to mimic five people standing shoulder to shoulder. For this study ten participants were recruited and each person was asked to hold the wooden rod such that one of the tags was place firmly against their chest as if they

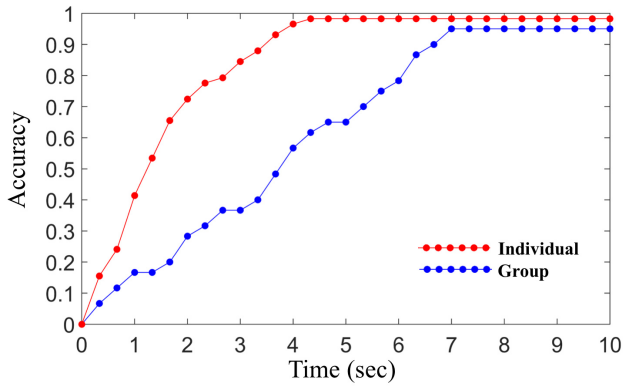

Figure 7. Prediction accuracy of the system overtime. The upper red line shows the accuracy of correctly identifying individuals. The lower blue line shows the accumulated accuracy of identifying all five members of the group starting the moment the first body enters the field of view of the depth camera.

were wearing the tag. Starting outside the field of view of the Kinect's depth camera the participants were ask to walk towards the *ID-Match* system. This procedure was repeated five times such that each participant held each tag on the wooden rod against his or her chest once, which resulted in a total of 50 runs. Again two dummy tags were placed in the environment to insure a more realistic experiment.

The results in Figure 9 show that *ID-Match* is capable of correctly matching the tag *ID* to the corresponding *body* position on the wooden rod with an accuracy of 96.7%. The probability of randomly guessing the correct answer for one instance is 1-out-of-7. Since each of the 60 trials was done by a single person (rather than a group) there were no occlusions, and the Kinects depth camera was able to track each person without errors. These results show that *ID-Match* is able to take advantage of the slight difference in radial direction of parallel movement to successfully assign the correct *ID* to the corresponding *body*.

Match Reacquisition

The Kinect depth camera has proven to be reasonably reliable at detecting the presence of a person, determining the position of their body and limbs (i.e. skeleton), and tracking the location of the person while in the field of view. However, all computer vision systems suffer from occlusions when the camera's view of the target person is blocked by an object or another person. In the case of the Kinect this means that when a person is occluded or goes out of view, their virtual skeleton (and corresponding skeleton ID) is discarded. Once the person reenters the field of view, they are given a new skeleton and skeleton ID. This means that each time the Kinect loses a person and they reenter the field of view the *ID-Match* system must reacquisition the person and perform ID association to determine who they are and where they are.

In order to investigate how effectively *ID-Match* reacquisitions tagged people once they have been lost by the Kinect's depth camera, the following experiment has been devised. Here three participants have been asked to walk in a

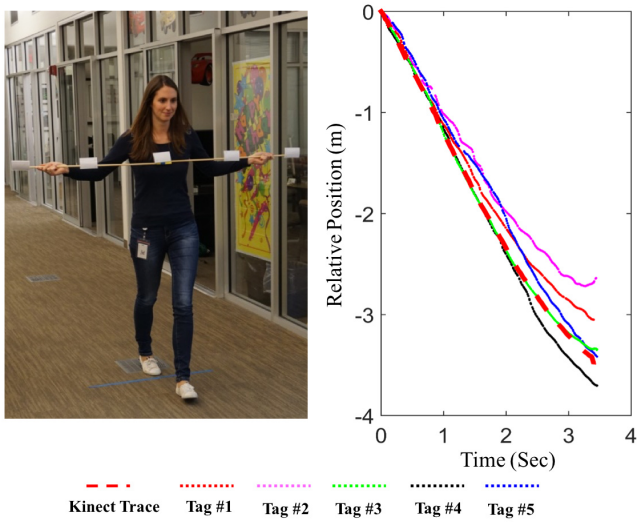

Figure 8. Image on the right of a person carrying a horizontal rod with five RFID tags mounted on it. The plot on the left shows the five RFID tag SAR trajectories along with the path of the person as measured by the Kinect (red dashed line). The results show the system is capable of assigning the correct ID while the person is walking even though all tags are undergoing nearly identical movements.

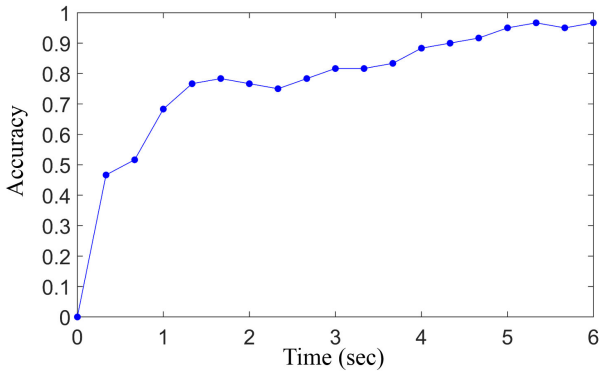

Figure 9. The accuracy of the system at matching the correct tag to the skeleton in the disambiguating similar motion evaluation. The blue line shows and average accuracy of 96.7% within 6 seconds, for the 60 trial.

circle marked on the floor with a diameter of 3 meters. As they walk around the perimeter of the circle the person in the foreground as seen by the Kinect blocks the people in the background. Thus, each person is potentially occluded twice per revolution. The three participants are asked to walk in a circle for 5 minutes resulting in 86 blocking and reacquisition events. Results for the percent accuracy for matching the correct ID to the correct person is shown in Figure 10. In this case the available time was only approximately five seconds (i.e. the time between a person becoming visible and then blocked again). This test shows that *ID-Match* performs quite well at reacquiring people once they have been lost by the Kinect, with a percent accuracy of 90.7% over the limited time interval of 5 seconds.

HRI AND OCCUPANCY TRACKING STUDIES

In an environment where the system is supposed to capture which users are passing a gate, users will enter, continue

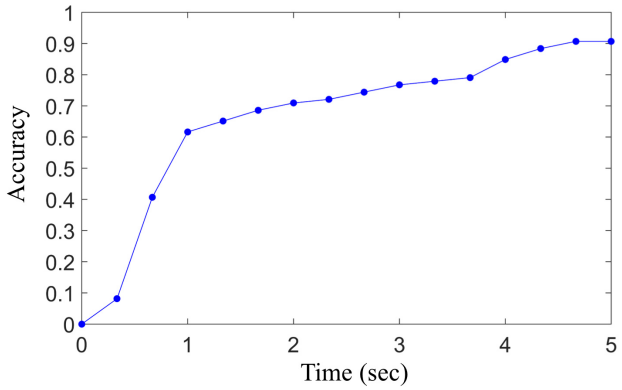

Figure 10. Accumulated accuracy overtime for re-acquisitioning a person after visual tracking is lost.

Figure 11. A snapshot of the Human-Robot Interaction setup showing 5 participants in front of the robot playing a quiz game.

moving, and leave the field of view of the system at different speeds and at different times, with no constraints on how long they would stay in the field of view.

In the following sections, we present two fundamentally different scenarios where we test the accuracy of *ID-Match* with less control on the behavior of the users.

Human Robot Interaction Study

In this application, the *ID-Match* system has been integrated into *Furhat* [4], an interactive anthropomorphic robotic head with an rear-projected 3D face. Computer animation delivers dynamic facial movements such as gaze (eye movement), facial expressions (happy, sad, confused, surprised etc.) and lip animation of basic phonemes which are synchronized to computer generated speech. Additionally *Furhat* is equipped with a pan-tilt neck which along with the animated eyes allows for convincing gaze simulation, and importantly for this work, allows the robot head to turn to address an individual person while looking at them in the eyes.

In this experiment *Furhat* has been programmed to autonomously host a multi-player quiz game [5]. In this interaction, users in groups freely approach the robot to compete in a gaming interaction scenario where the robot asks players different trivia questions and the users collect points when giving the correct answer. The interaction is constructed such that the robot needs to use the identity of the user and address the user by name (e.g. "Jack, here is the next question: what is..."). For the purposes of this game contestants were asked multiple choice questions and a commercially available speech recognition engine was used to capture users verbal answers, which were limited to "one", "two", "three", or "four".

The interaction was setup up in a large office room (4 x 6 square meters), with the robotic head placed on a pedestal in one of the corners, as shown in Figure 11. The RFID antenna and the depth camera was placed to maximize the field of view of the *ID-Match* systems and the coordinate frame of the Kinect was transformed to the robot's coordinate frame to account for the offset in location and pose between the two. Ten participants were recruited to interact with the robot in two groups of five users each. Seven of the users were male and three were female. All participants were graduate level students between 20 and 30 years old. Groups of five users carrying pre-registered RFID tags were instructed to enter the room and playing and quiz game with the robot. In order to allow the interaction to be natural, the participants were not told to stand in any predefined locations, order, or distance from the robot. However, the participants were asked to avoid standing directly in front of another person so as not to block the view of the other participants in the game. Each of the two groups of participants played the quiz game 6 times, giving a total number of 12 games, and 60 interactions (questions). Each of the 12 sessions lasted approximately 5 minutes.

Although the interaction was complex in terms of robot and task design, the *ID-Match* system task was simple: each time a new skeleton is tracked by the camera the system will attempt to match that skeleton to one of the RFID tags out of all the tags currently visible to the RFID reader. Figure 12 shows a plot of the average accumulated ID to human assignment accuracy of the system over time.

The "individual" (red line) shows the accuracy of the system at identifying each individual people approaching the robot in a group. It is important to mention here that at any moment in time, there was always two additional static tags in the environment that were not held by a person. The results show that the system is able to recognize each individual at 94.8% accuracy within 5 seconds starting from when the person enters the scene. This is consistent with the controlled studies presented earlier in the paper. This fast acquisition time gave the robot the ability to customize the interaction, and address the individual person by name, almost by the time they completed their approach and came to a stop by the robot. The"group" blue line in Figure 12 shows the accuracy of the system in identifying all individuals in the five-person group with time starting at the moment when the first person enters the scene. Here the system is capable of accurately identifying all individuals with 91.6% accuracy in 12 seconds. Although this group acquisition time is highly dependent on actions and speed of the participants, it is important to include this result as a benchmark for a group of people naturally approaches the robot for interactions.

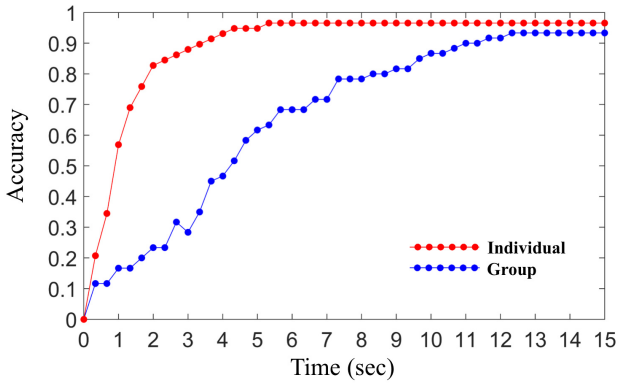

Figure 12. A plot showing the accuracy of the system over time. The top red line shows the accuracy of detecting each person in the FoV of the camera at any point in time, plotted against the time since they appeared. Lower blue line shows the accuracy of the system detecting the whole group (detecting the 5 different users) over time, starting from the moment the first user enters the FoV of the camera.

Occupancy Monitoring Scenario

The second uncontrolled user study evaluated *ID-Match* 's ability to passively monitor the flow of people in an office as they pass through a "virtual gates" and checkpoints. The systems task is to quickly recognize the identity of the passersby while they are visible to the camera and determine in which direction they are moving over a long period of time.

Physical Space

The study was conducted in an office setting as shown in Figure 13 consisting of a central hallway with offices one side and workstations in an open floor plan on the other side. The *ID-Match* system was configured to log data for a full working day (i.e. continuously for 7.5 hours). Out of 40 employees on site during that day, 16 were assigned RFID tags and registered in the system, the other employees had no other RFID tags. In the same area where the system was setup, two additional static RFID tags were placed and were visible to the reader at all times. Additionally, 3 of the 16 employees with RFID tags had workstations within the read range of the RFID antennas, being visible to the reader at almost all times, but not visible to the camera. This causes a greater level of uncertainty when attempting to match the IDs to people. The Kinect camera was set up pointing towards the hallway with a 30-degree side view as demonstrated in 13, panel A. In order to increase system reliability a second RFID antenna was used to insure tags were properly read as participants either walked towards and away from the Kinect.

As mentioned earlier, 40 participants took part in the experiment, 16 of them were carrying a registered RFID tag. The environment and the users were not controlled during the day. The participants were not informed about the purpose of our study in order to maintain their usual behavior. The only requirement placed on the participants was to wear the tags around their chest position, at the time they entered the office till the time they left for the day.

During the day-long experiment, 302 instances of people passed through the hallway where the experiment was set up.

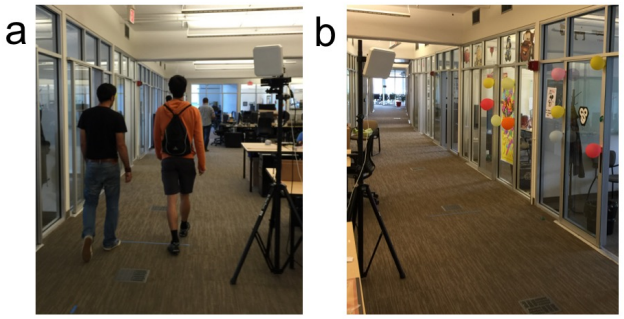

Figure 13. Snapshots of the physical setup of the occupancy study.

129 of these were people who carried an RFID tag, and 173 did not carry a tag. The system's task was to recognize the identity of any visible skeleton, whether it belongs to one of the registered users, or whether that skeleton did not have a registered tag in the system. Out of the 129 instances of people wearing tags, the *ID-Match* system was able to correctly recognize 122 users (at an accuracy of 94.5%). For people not wearing tags, the system incorrectly assigned 5 identifications, yielding a 2.9% false positive rate, while the rest was correctly identified as not wearing tags.

By using the 3D skeleton tracking of the Kinect, the system was able to recognize the direction of movement of the correctly identified skeletons at a 100% accuracy. Although this study is set up in a highly uncontrolled environment, the main objective of the study is to show the ability to function equally for registered users and unregistered users, giving bigger context for the application space where the system can be deployed.

CONCLUSION

This work presents a novel hybrid computer vision and RFID system that is capable of seamlessly matching the identity of an individual (as stored on an RFID tag) to the 3D image of that person as captured by a depth camera. This real-time system is capable of determining the identity of individuals within 4 seconds at an accuracy of 96.6% and groups of five people in 7 seconds with 95% accuracy.

Users studies show that the *ID-Match* system is indeed capable of robustly identifying people with enough speed and accuracy to enable a humanoid autonomous robot to naturally interact with up to five people simultaneously. The system has also been shown to be effective at passively monitoring both tag and un-tag participants without requiring active participation for identification and tracking.

In order to demonstrate that this approach is scalable to other usage scenarios several controlled lab experiments are presented that cover many of the edge conditions and worst case scenarios. Including participant re-acquisition after visual tracking is lost, and the systems ability to distinguish between nearly identical tag motion. Ultimately *ID-Match* is a novel sensor fusion technique providing a valuable method for enabling automated computing systems to quickly and accurately recognize multiple people simultaneously.

REFERENCES

1. 2015. Impinj Far Field Antenna Datasheets. (2015). https://support.impinj.com/hc/en-us/article_attachments/200774708/IPJ-A1000-A1001-USA_Antenna_Spec.pdf
2. 2015. Impinj Speedway Reader Product Brief Datasheet. (2015). https://support.impinj.com/hc/en-us/article_attachments/202942578/Impinj_SpeedwayReaders_ProductBrief_080715.pdf
3. 2015. Kinect for Windows. (sep 2015). https://dev.windows.com/en-us/kinect
4. Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems*. Springer, 114–130. DOI: http://dx.doi.org/10.1007/978-3-642-34584-5_9
5. Samer Al Moubayed and Jill Lehman. 2015. Regulating Turn-Taking in Multi-child Spoken Interaction. In *Intelligent Virtual Agents*. Springer, 363–374. DOI: http://dx.doi.org/10.1007/978-3-319-21996-7_40
6. Francesco Cafaro, Alessandro Panella, Leilah Lyons, Jessica Roberts, and Josh Radinsky. 2013. I See You There!: Developing Identity-preserving Embodied Interaction for Museum Exhibits. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1911–1920. DOI: http://dx.doi.org/10.1145/2470654.2466252
7. John C Curlander and Robert N McDonough. 1991. *Synthetic aperture radar*. John Wiley & Sons.
8. T. Deyle, H. Nguyen, M. Reynolds, and C.C. Kemp. 2009. RF vision: RFID receive signal strength indicator (RSSI) images for sensor fusion and mobile manipulation. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. 5553–5560. DOI: http://dx.doi.org/10.1109/IROS.2009.5354047
9. Federal Communication Commission (FCC). 2011. Title 47: Telecommunication, Part 15 Radio Frequency Devices. www.fcc.gov. (January, 31 2011). http://www.fcc.gov/
10. M. Goller, C. Feichtenhofer, and A. Pinz. 2014. Fusing RFID and computer vision for probabilistic tag localization. In *RFID (IEEE RFID), 2014 IEEE International Conference on*. 89–96. DOI: http://dx.doi.org/10.1109/RFID.2014.6810717
11. W. Gueaieb and Md.S. Miah. 2008. An Intelligent Mobile Robot Navigation Technique Using RFID Technology. *Instrumentation and Measurement, IEEE Transactions on* 57, 9 (Sept 2008), 1908–1917. DOI: http://dx.doi.org/10.1109/TIM.2008.919902
12. D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose. 2004. Mapping and localization with RFID technology. In *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, Vol. 1. 1015 – 1020 Vol.1. DOI: http://dx.doi.org/10.1109/ROBOT.2004.1307283
13. Cory Hekimian-Williams, Brandon Grant, Xiuwen Liu, Zhenghao Zhang, and Piyush Kumar. Accurate localization of RFID tags using phase difference. In *RFID, 2010 IEEE International Conference on*. http://dx.doi.org/10.1109/RFID.2010.5467268
14. Jeffrey Hightower, Roy Want, and Gaetano Borriello. 2000. SpotON: An indoor 3D location sensing technology based on RF signal strength. *UW CSE 00-02-02, University of Washington, Department of Computer Science and Engineering, Seattle, WA* 1 (2000).
15. Sherry Hsi and Holly Fait. 2005. RFID Enhances Visitors' Museum Experience at the Exploratorium. *Commun. ACM* 48, 9 (Sept. 2005), 60–65. DOI: http://dx.doi.org/10.1145/1081992.1082021
16. Hanchuan Li, Can Ye, and Alanson P. Sample. 2015. IDSense: A Human Object Interaction Detection System Based on Passive UHF RFID. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2555–2564. DOI: http://dx.doi.org/10.1145/2702123.2702178
17. Hui Liu, H. Darabi, P. Banerjee, and Jing Liu. 2007. Survey of Wireless Indoor Positioning Techniques and Systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 37, 6 (nov. 2007), 1067 –1080. DOI: http://dx.doi.org/10.1109/TSMCC.2007.905750
18. R. Miesen, F. Kirsch, and M. Vossiek. 2013. UHF RFID Localization Based on Synthetic Apertures. *Automation Science and Engineering, IEEE Transactions on* 10, 3 (July 2013), 807–815. DOI: http://dx.doi.org/10.1109/TASE.2012.2224656
19. L.M. Ni, Yunhao Liu, Yiu Cho Lau, and A.P. Patil. 2003. LANDMARC: indoor location sensing using active RFID. In *Pervasive Computing and Communications, 2003. (PerCom 2003). Proceedings of the First IEEE International Conference on*. 407–415. DOI: http://dx.doi.org/10.1109/PERCOM.2003.1192765
20. P.V. Nikitin, R. Martinez, S. Ramamurthy, H. Leland, G. Spiess, and K.V.S. Rao. 2010. Phase based spatial identification of UHF RFID tags. In *RFID, 2010 IEEE International Conference on*. 102–109. DOI: http://dx.doi.org/10.1109/RFID.2010.5467253
21. Sunhong Park and S. Hashimoto. 2009. Autonomous Mobile Robot Navigation Using Passive RFID in Indoor Environment. *Industrial Electronics, IEEE Transactions on* 56, 7 (July 2009), 2366–2373. DOI: http://dx.doi.org/10.1109/TIE.2009.2013690

22. Patrick J. Grother; George W. Quinn; P J. Phillips;. 2010. *Report on the Evaluation of 2D Still-Image Face Recognition Algorithms*. Technical Report. National Institute of Standards and Technology. http://www.nist.gov/manuscript-publication-search.cfm?pub_id=905968

23. Mahsan Rofouei, Andrew Wilson, A.J. Brush, and Stewart Tansley. 2012. Your Phone or Mine?: Fusing Body, Touch and Device Sensing for Multi-user Device-display Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1915–1918. DOI: http://dx.doi.org/10.1145/2207676.2208332

24. S. Sarkka, V. Viikari, M. Huusko, and K. Jaakkola. 2011. Phase-Based UHF RFID Tracking with Non-Linear Kalman Filtering and Smoothing. *Sensors Journal, IEEE* PP, 99 (2011), 1. DOI: http://dx.doi.org/10.1109/JSEN.2011.2164062

25. V. Cadenat T. Germa, F. Lerasle N. Ouadah. 2010. Vision and RFID data fusion for tracking people in crowds by a mobile robot. *Computer Vision and Image Understanding* 114, Issue 6 (jun 2010), Pages 641651. DOI: http://dx.doi.org/10.1016/j.cviu.2010.01.008

26. Y. Taigman, Ming Yang, M. Ranzato, and L. Wolf. 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. 1701–1708. DOI: http://dx.doi.org/10.1109/CVPR.2014.220

27. Jue Wang and Dina Katabi. 2013. Dude, Where's My Card?: RFID Positioning That Works with Multipath and Non-line of Sight. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM (SIGCOMM '13)*. ACM, New York, NY, USA, 51–62. DOI:http://dx.doi.org/10.1145/2486001.2486029

28. Jue Wang, Deepak Vasisht, and Dina Katabi. 2014. RF-IDraw: Virtual Touch Screen in the Air Using RF Signals. *SIGCOMM Comput. Commun. Rev.* 44, 4 (Aug. 2014), 235–246. DOI: http://dx.doi.org/10.1145/2740070.2626330

29. Jie Xiong and Kyle Jamieson. 2013. ArrayTrack: A Fine-Grained Indoor Location System.. In *NSDI*. 71–84. http://dl.acm.org/citation.cfm?id=2482635

30. Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time Tracking of Mobile RFID Tags to High Precision Using COTS Devices. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom '14)*. ACM, New York, NY, USA, 237–248. DOI: http://dx.doi.org/10.1145/2639108.2639111