I See *You* There! Developing Identity-Preserving Embodied Interaction for Museum Exhibits

Francesco Cafaro[1], Alessandro Panella[1], Leilah Lyons[1,2], Jessica Roberts[2], Josh Radinsky[2]
[1]Computer Science [2]Learning Sciences
University of Illinois at Chicago
{fcafar2, apanel2, llyons, jrober31, joshuar} @uic.edu

ABSTRACT

Museums are increasingly embracing technologies that provide highly-individualized and highly-interactive experiences to visitors. With embodied interaction experiences, increased localization accuracy supports greater nuance in interaction design, but there is usually a tradeoff between fast, accurate tracking and the ability to preserve the identity of users. Customization of experience relies on the ability to detect the identity of visitors, however. We present a method that combines fine-grained indoor tracking with robust preservation of the unique identities of multiple users. Our model merges input from an RFID reader with input from a commercial camera-based tracking system. We developed a probabilistic Bayesian model to infer at run-time the correct identification of the subjects in the camera's field of view. This method, tested in a lab and at a local museum, requires minimal modification to the exhibition space, while addressing several identity-preservation problems for which many indoor tracking systems do not have robust solutions.

Author Keywords

Localization; Identification; Tracking; Ambient Displays; RFID; Cameras; Embodied Interaction; Museum Exhibits

ACM Classification Keywords

I.4.8 [Image Processing and Computer Vision]: Scene Analysis – Sensor Fusion, Tracking; I.4.9 [Image Processing and Computer Vision]: Applications; K.3.1 [Computers and Education]: Computer Uses in Education – Collaborative learning

General Terms

Algorithms, Design, Experimentation, Human Factors

INTRODUCTION

Over the past two decades, there has been increased interest in how exhibits might promote interactive learning and sociability – in other words, supporting not just how visitors interact with the exhibits, but also how they interact with one another while in the presence of exhibits [19]. When exhibits are computer-based, then, designers must attend to how to scale both the input and the output of the exhibit to support multiple simultaneous visitors. Shared output is most commonly achieved by using very large displays. Supporting simultaneous, multi-user *input* is more of an open challenge, however – many approaches have been tried, from using mobile devices as opportunistic user interfaces [13] to using multi-touch tables [9] to using tangible user interfaces [2].

Embodied interaction [12] is yet another input approach that has been gaining popularity in museums. Various sensing technologies can be used to track the movements of human bodies as a means of providing input to the exhibit. Embodied interaction is appealing because it does not require visitors to use devices (e.g., trackballs, mobiles, tables, light pens, etc.) to provide input to the system, and because it offers an engaging kinesthetic experience.

The richness of the *learning* experience offered by embodied interaction exhibits, however, is directly affected by how much control visitors have over the interaction, and the transparency of that control. In other words, engaging meaningfully with the exhibit is dependent on whether visitors can link the effects seen in the shared exhibit to their own individual actions [1]. Supporting such interactive control via embodied interaction requires a system that accurately tracks visitor movements in space (so the exhibit can provide more nuanced feedback, thus promoting deep engagement and learning). Because these exhibits are multi-user, pro-learning embodied interaction also requires preserving the identity of each user, in order to accurately attribute actions to individual visitors (so that visitors can understand who is producing the observed feedback).

Indoor tracking of individual users has remained something of a challenge, however, despite the proliferation of sensing technologies applied to the problem. There is usually a tradeoff between fast, accurate tracking and the ability to preserve the unique identity of users. There can also be pragmatic tradeoffs: some systems require extensive modifications to the indoor space (by installing multiple beacons, cameras, or antennae grids), require extensive calibration time, are prohibitively expensive (especially for small museums or cultural institutions), or require users to carry expensive or cumbersome equipment (e.g. beacons, Wi-Fi devices) which may be perceived as invasive and which museums do not wish to restock in the event of theft or loss (these tradeoffs are described in more depth below).

CHI 2013, April 27–May 2, 2013, Paris, France.
Copyright © 2013 ACM 978-1-4503-1899-0/13/04...$15.00.

To support personalized interactions, we needed to build a system that combines both fine-grained tracking and reliable identity assignment and correction - the focus is not on indoor positioning per se. We present a method that combines two input technologies, commercial camera-based motion tracking and Radio Frequency Identification (RFID) technology. Our system prototype (Figure 1) is a combination of a shared display, plus two components: (1) a Microsoft *Kinect*™; and (2) a single RFID reader. The *Kinect*™ is used to track visitors within the exhibition space, which supports fine-grained location resolution (less than ±1 cm in a 2.4x2.4m space [14]) and is highly reactive (responsive to user actions in less than 10ms). The use of inexpensive RFID tags (credit-card sized tokens carried by visitors) allows the system to identify who is in the exhibition room. We developed a probabilistic Bayesian model that combines data from each of the two components of our system to: (1) preserve the identity of visitors who enters a new exhibit and (2) to resolve identity ambiguities that occur when visitors occlude one another or temporarily step outside of the camera's view frustum.

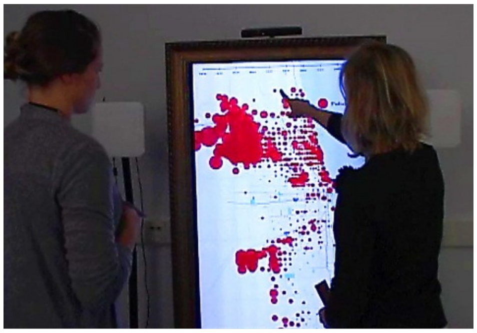


Figure 1. Two people interacting with our system prototype. A 65" display is used in combination with two RFID antennas (left and right side) and one *Kinect*™ camera (on top)

We tested our system during in-laboratory experiments and with museum visitors at a local history museum. The lab tests were designed to test the performance of the system using repeated pre-scripted user motions known to be problematic for the system, while the *in situ* tests were done to assess the performance under naturalistic use conditions.

BACKGROUND AND RELATED WORK

While we are hardly the first to explore embodied interaction with museum exhibits, our problem space (multi-user exploration of rich data sets [5]) demands a system that supports both location awareness (identifying *where* visitors are) and context awareness (identifying *who* each visitor is). Here we review prior work on location awareness and unique identification in indoor spaces.

Overview of Common Indoor Tracking Technologies

Camera-Based Tracking

Camera based systems use computer vision techniques to detect the outlines of visitor bodies or special printed codes (called fiducial symbols) from live video feeds. They provide fast (in the order of milliseconds) and accurate (on the order of millimeters) tracking, but preserving the identity of users is still a challenge. Single-camera systems are limited by the view frustum of the camera, so unless a camera can be mounted on a high ceiling to cover the entire exhibit space [8], or unless smart strategies are used to manage camera movements (e.g., [7]), the camera can easily "lose" individuals. A room can be outfitted with multiple cameras to completely cover space, but apart from the expense, this also involves significant alterations to the space (mounting cameras, power drops, and data transmission channels within the space of interaction), and a careful placement of cameras to prevent occlusion-related misidentification [16]. Unfortunately, calibrating these systems is non-trivial and this approach still doesn't support the *unique* identification of visitors – which is needed if the interaction design for the exhibit calls for associating a visitor with a specific profile, as would be needed for personalized exhibit experiences.

Fiducial symbol-based strategies could be used to establish identity (visitors could carry or wear a token or sticker bearing a fiducial symbol), but these have limitations when the symbol is moved too far away from the camera (an exception is presented in [30]), or when it is tilted too far.

Alternatively, users could wear special-purpose colored or reflective clothing visible from multiple angles (e.g., hats), but this may be impractical (some museums avoid wearables owing to the risk of transmitting head lice). Some systems, including the *Kinect*™, can perform limited face recognition to disambiguate amongst visitors, but these approaches suffer the problem of false positives or false negatives when identifying multiple users [32]. Furthermore, the recognition rate decreases in poor lighting conditions [14] and its performance may be poorer with different ethnic groups [24]. When the interaction design for an exhibit is continuous (e.g., if the display changes along with visitor proximity), such mis-assignments could entirely derail the interactive experience.

Infrared Tracking

Infrared (IR) distance sensors provide fine-grained location detection like camera systems (on the order of millimeters). They can be constructed in different ways, either by assembling a grid of IR beams and sensors that gets interrupted by visitor bodies [28] (often used to support input to multi-touch displays) or by using IR cameras to detect the presence of visitor bodies or of Infrared beacons [11]. Beam-based IR systems require a fair amount of work to build, are only as accurate as the grid spacing (which can get expensive if a large area is to be covered), and are not able to uniquely identify users (they can distinguish amongst individuals, but there is no way to link an identified body with a specific user profile). IR camera systems require a line of sight and, unless visitors wear specialized beacons (which can suffer from occlusion problems like fiducials), such systems cannot uniquely identify approaching visitors.

Radio-Frequency Tracking

Radio Frequency ID (RFID) systems work by transmitting radio signals from an antenna (or array of antennae) and "listening" for the signal's "reflection" as it bounces off specialized printed circuits (usually embedded in credit-card-like devices). The main advantage of RFID systems is that the radio waves are not line-of-sight only – they can penetrate most non-metallic substances. RFID systems are extremely reliable when identifying tags in the read range of the reader, which is why they are so commonly used by manufacturers and shippers to track the location of items in warehouses. Parts and parcels are not prone to moving about of their own volition, however – so most RFID tracking systems are configured to detect only when a card enters or exits a space, like the Intellibadge [10] system, not to triangulate the precise location of a card. Likewise, the RFID technology employed in [21] [22] is suitable to identify the users of the display; but its limited localization is insufficient for supporting embodied interaction. There have been several attempts at expanding the capacity of RFID for localization for tracking items in motion. [3] introduces a location fingerprinting method to localize RFID tags, using a map of the received signal strength indicator (RSSI). However, the granularity of the localization that can be achieved by these systems is generally at the level of one room, which is insufficient for highly-interactive applications.

Zigbee-based systems (such as [6] and [17]) have an accuracy of 1.5-2.0m, which may not be enough for controlling a highly-interactive system within a small interaction space. Among the most accurate and responsive RFID tracking systems are Ultra Wide Band systems, such as the commercial Ubisense, which still have properties unsuitable for highly-interactive applications (an accuracy of ±30cm along with a requirement of 5-6 seconds to get a stable reading after a tag has been moved [31]).

Ultrasonic Tracking

Ultrasonic sensors [20] can supply nuanced localization combined with reliable identification, but these systems can be difficult to use and calibrate. Like multi-camera systems, they require many receivers to be installed in the location where tracking is to be performed to locate emitters [15].

Hybrid Tracking Systems

Given the limitations of many single-technology methods for both localizing and identifying individuals, quite a few researchers have explored combining different methods to support both of these needs. Concurrent radio and ultrasonic signals [26], can get 2-dimensional positioning sufficient for room-based embodied interaction, but the ~10x10cm precision is not sufficient for fine-grained interactions. Hello.Wall [25] uses one long-range and one short-range RFID reader to define zones of interaction, with fine-grained interactivity supported with a specialized hand-held device. Another approach which requires a device (in this case, accelerometer-equipped cell phones) carried by the users is [29], which combines the accelerometer data with CCTV video inputs, but this system can still inadvertently swap user identities. Another system achieved finer control by combining passive RFID tags (to recognize the user) with an ultrasonic sensor, where distance from the display would zoom pictures associated with the user in or out according to the direction of movement [27]. This approach still requires a custom installation and extensive calibration owing to the narrow line-of-sight of many ultrasonic sensors.

Combining RFID Identification with Low-Cost Commercial Localization

Our system differs from existing approaches for its ability to overcome the trade-off between fast, accurate tracking and the ability to preserve the unique identity of users, thus allowing exhibit designers to create personalized embodied interaction experiences. Compared to many commercial tracking systems such as Nikon Metrology iGPS our system does not require a direct line of sight and requires the user to carry only an inexpensive credit-card sized tag. Compared to Ultra Wide Band RFID system such as Ubisense, our system is more fine-grained (±1cm vs ±30cm with Ubisense) and faster (see [31]) when the tagged object (a person) is walking within the exhibition space (a tracking update takes less than 10ms). Furthermore, our research is different in that: (1) we exploit low-cost, easily-obtainable technology (a Microsoft *Kinect*™ and semi-passive RFID); (2) installation requires little to no modification of the exhibit space; (3) we develop a Bayesian probabilistic model that supports robust identification for applications that require reliable recognition of individuals.

EMBODIED INTERACTION WITH PERSONALIZED CENSUS DATA

Our pilot exhibit, CoCensus, is situated in a history museum, and encourages visitors to explore the US Census data. Data sets, even when visualized attractively, are typically not all that engaging, but there is evidence that interactive data visualizations like Hans Rosling's Gapminder might change this. To help visitors explore a visualization of Census data, we allow them to "role play" as the data subsets associated with their own self-identified ethnicities (see Figure 2). We use embodied interaction to help visitors "feel" the connection to *their* data as they explore them: when a visitor approaches the display, his subset of data becomes more prominent on a map (via opacity and z-ordering). Our design metaphor is a mirror, reflecting the users' own data in response to their body movements within the interaction space, in order to engage visitors in joint explorations of rich data sets. When two visitors move within the exhibition space, they collectively conceal and reveal patterns of where their respective data subsets do or do not intersect, sparking conversations such as comparing the settlement patterns of different immigrant groups. These patterns indicate whole hosts of interesting phenomena from segregations to migrations, and help visitors connect their own personal family stories to the larger trends visible on the map.

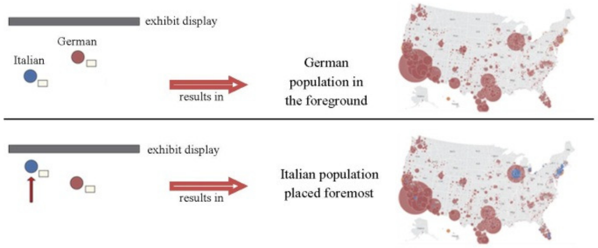


exhibit display
German
Italian
results in
German population in the foreground
exhibit display
results in
Italian population placed foremost

Figure 2. The metaphor for our collaborative data exploration design is a data mirror. Visitors approaching the exhibit should see "their" data reflect their body movement.

Indeed, part of the learning that occurs at the exhibit is in the emergent discussions among visitors about their respective data sets, which are highly personalized, so it is critical that visitors are able to identify which portions of the data display are under each person's control.

ILLUSTRATION OF TRACKING CHALLENGES

The example scenario highlights the functional requirements of an indoor tracking system that guarantees a fully personalized and embodied interaction experience. More specifically, the system should be: (1) *Reactive*; the system must be fast enough to allow continuous and transparent interaction with the user. When a visitor is moving within the exhibition space, the system should be able to continuously track its current position. When a visitor leaves one exhibit and moves closer to another, the new one should proactively respond to the visitor's approach. (2) *Fine-grained*; so that even subtle motions can be detected, to support more nuance in interaction even in a small room or exhibit. For instance, if we want to emphasize the data of the visitor who is closer to the screen, the system should be able to determine if one person is just one step ahead of the other. (3) *Not invasive*; both towards the visitors and the exhibition space itself. On the one hand, in a public space, it may be impractical to ask people to wear special purpose clothing or to carry heavy or expensive devices. Nor do we wish to unduly alter the exhibition space – many museums (such as the one we tested in) are housed in historical buildings where changes to the walls and ceiling (e.g., mounting receivers or power drops) are impossible. (4) *Able to mirror users*; interaction with a system is fully embodied only if it is able to react to the full bodily behaviors of each user, so that each visitor understands *who* is producing the observed feedback.

These four functional requirements (reactive, fine-grained, not invasive, able to mirror users) suggest the use of a compact camera tracking system, such as the *Kinect*™. We performed a formative study in a museum, using only the *Kinect*™ and Primesense OpenNI. With solo visitors, 5 out of the 9 visitors lost their identity at some point during their 15-minutes interaction with the system. For instance, when a guided-tour group walked through the exhibition space, a new identity was assigned to the visitor.

PROBLEM STATEMENT

Without any other supporting technology, such as face recognition or fiducial symbols (which suffers from the problems described in the related work section), the *Kinect*™ can track a shape in the exhibition space, but is not able to associate a previously-defined profile to that shape, e.g. to preserve the identity of a visitor moving through exhibits, possibly in different rooms. The use of the RFID technology is a good fit for supporting the camera tracking system in preserving identities. For instance, when one visitor enters an exhibition, the new shape seen by the *Kinect*™ can be associated with the new RFID tag and the system can *reactively* display the data related with that user's profile (associated to the ID of the RFID tag) and *mirror* that user's movements, as the *Kinect*™ is able to detect human shapes within its field of view and track them with *fine-grained* resolution in the three-dimensional space. However, this process becomes challenging when two or more people enter the room at the same time, a common occurrence. The camera tracking system detects two or more new shapes, the RFID reader two or more new IDs, but who is who? There are potentially $n!$ associations between n new shapes and n new tags. One trivial solution would be to make visitor A always enter the exhibit room before visitor B, who needs to enter before visitor C and so on. Of course, this would not be feasible in a museum, when multiple people are supposed to move freely from one room (and one exhibit) to the other. This motivates the need for a more intelligent merging of visitor detection sensors, such as the probabilistic Bayesian model here described.

Formally, let us denote as n the number of subjects being tracked, and define the set $R = \{1, \dots, n\}$ of RFID identifiers, and the set $K = \{1, \dots, n\}$ of *Kinect*™ shapes. At time t, a tuple $r_t = \langle (a_{t,i}, rssi_{t,i}) \rangle_{i \in R}$ is received, where $a_{t,i} \in \{0,1\}$ identifies which of the two antennae attached to the RFID reader received the signal, and $rssi_{t,i} \in N_{\geq 0}$ is the strength of that signal. A tuple $s_t = \langle s_{t,i} \rangle_{i \in R}$ of $(x_{t,i}, y_{t,i})$ coordinates is also received from the *Kinect*™, representing the position of the i-th shape. We define the identification function as a bijection $\rho: R \to K$, that corresponds to a unique mapping of RFID identifiers to skeletons, that can assume $n!$ different values. The problem is to estimate the correct identity assignment ρ_t^* at time t given the observation histories $r_{0:t}$ and $s_{0:t}$.

PROPOSED SOLUTION

Our methodology is based on exploiting the statistical dependence between the data coming from the camera tracking system and from the RFID antennas. In order to do so, we developed a probabilistic Bayesian filter where the identification function ρ is treated as a random variable representing the hidden state of the system.

One assumption sometimes made when tracking indoor radio signals is that, for a given location, the RSSI is normally distributed [3]. However, as observed in similar work with Wi-Fi signals [23], the RSSI registered for any given location exhibits more complex distributions, e.g. bimodal. For this reason, we use histograms to provide a richer, nonparametric characterization of the probability distribution of signal strengths. Since RSSIs assume integer values only, it is natural to assign each integer level to one bin of the histogram. During execution, a window of W RSSI observations is kept for each RFID tag in range, and the derived histogram is compared to the *fingerprints* that have been stored for specific, known locations using the Kullback-Leibler (KL) divergence. Upon this measure, a conditional probability is built to carry out a Bayesian update and obtain the posterior probability distribution for the identification function ρ. At that point, the *Maximum A Posteriori* (MAP) value of ρ is chosen as the actual one. Details of the methodology follow.

Kullback-Leibler (KL) Divergence

The KL divergence [18] is an information theoretical measure of the difference between two probability distributions p and q. For distributions over a finite, discrete set X, the KL divergence is:

$$KL(p||q) = \sum_{x\in X} p(x)\frac{\log(p(x))}{\log(q(x))}$$

which is clearly non-symmetric. When a symmetric measure is preferable, a symmetrized version is often used, simply defined as $KL_S(p,q) = KL(p||q) + KL(q||p)$.

Localization Based on KL-kernel Regression

In order to guess the correct association function ρ^*_t, the system first computes an estimation $s^R_{t,i}$ of the position of each RFID tag $i \in R$ detected by the antennae, using the information r_t. To perform this, we follow an approach similar to the one presented in [23] .

For each tag and each antenna, the system maintains a histogram $H^a_{t,i}$ of the RSSI values received from tag i on antenna a in the time window $[t-W,t]$ for some specified window length W. In addition, a collection of RSSI histograms (fingerprints) $F = \{H^a_l\}_{l\in L, a\in\{0,1\}}$ is pre-computed for a set of known locations L. It is reasonable to assume that the signals received on the two antennas are conditionally independent, given the position of the RFID tag. Under these circumstances, it can be shown that the symmetrized KL divergence for the joint distribution RSSI for the two antennas can be computed as the sum of the two components, i.e. $KL_S(H_{t,i},H_l) = \sum_{a\in\{0,1\}} KL_S(H^a_{t,i},H^a_l)$.

The above formula defines the symmetrized KL divergence between the joint histogram of RSSI collected during localization and the one corresponding to the fingerprint location l. Upon this measure, a *kernel function*[1] is computed as $k(H_{t,i},H_l) = e^{-\alpha\, KL_S(H_{t,i},H_l)}$, where α is a parameter of the model that needs to be tuned appropriately. The kernels $\{k(H_{t,i},H_l)\}_{l\in L}$ are treated as weights for estimating the current position of a tag i being tracked. In particular, the K larger kernels are selected and the position is estimated through the following weighted sum:

$$s^R_{t,i} = (x^R_{t,i}, y^R_{t,i}) = \frac{\sum_{l\in L_K}(x_l,y_l)k(H_{t,i},H_l)}{\sum_{l\in L_K} k(H_{t,i},H_l)},$$

where L_K are the locations associated with the K largest kernels. K is a parameter whose optimal value is estimated empirically.

Probabilistic Filtering for the Association Function

The estimation at time t of the position of the RFID tags is combined with the information s_t coming from the *Kinect*™ to perform the probabilistic Bayesian update described in the following. The *posterior* probability distribution over the values of the association function ρ is computed by multiplying the *prior* probability (i.e. the one obtained at time $t-1$) by the *likelihood* of the received observation. In our case, we have:

$$p(\rho_t|s_t,s^R_t) = Z^{-1}p(s_t,s^R_t|\rho_t)p(\rho_{t-1}),$$

where Z is a normalization factor. The likelihood $p(s_t,s^R_t|\rho)$ is built by considering the accuracy of the of the received observation. In our case, we have:

$$D_\rho(s_t,s^R_t) = \frac{1}{n}\sum_{i\in R} d(s_{t,\rho(i)}, s^R_{t,i}),$$

where $d(\cdot,\cdot)$ is the Euclidean distance between two points.

Intuitively, this distance is small when the positions estimated through the RSSIs accurately describe the real location of the subjects, assuming that tag i is carried by subject $\rho(i)$. This quantity is therefore inversely correlated with the likelihood of the equation above. Experimentally, we observed that a negative exponential relation is a good fit for this correspondence. Therefore, the overall update is:

$$p(\rho_t|s_t,s^R_t) \propto e^{-\beta D_{\rho_t}(s_t,s^R_t)}p(\rho_{t-1})$$

The parameter β controls the magnitude of the update: a small value makes the system more robust to local perturbations, while a large one makes it more responsive. At each time step, the system sets the current identity assignment to the configuration assuming the maximum value (MAP). Mathematically, we have:

$$\rho^*_t = \text{argmax}_{\rho_t} p(\rho_t|s_t,s^R_t)$$

SYSTEM EVALUATION

We tested the system in our research lab to assess its reactivity and stability under different scenarios. We then evaluated it during two sessions at a small history museum in Chicago, to determine if the level of performance we achieved was sufficient for actual visitors to find it usable. 60 museum visitors participated to the *in situ* evaluation. Both in-lab and *in situ*, users were divided in groups of two people and interacted with the exhibit prototype described

[1] A *kernel* is a symmetric function assuming value one if the arguments are equal and decaying to zero when their dissimilarity grows.

earlier. An RFID tag was given to each person before entering the exhibition space.

Experimental set-up

The system was deployed by mounting a 65” screen, the two RFID antennas, and the *Kinect*™ on a movable TV stand on wheels, hence providing high flexibility with respect to the venue configuration. The RFID infrastructure is based on a Thingmagic *Astra* reader, which is connected to an additional 6dB antenna using a 1m long coaxial cable. The tags were semi-passive ISO 18000-6C PowerID PowerP-E704 tags. We used *Primesense OpenNI* to obtain tracking data from the *Kinect*™ for all our tests but the second user-study at the museum, in which we used the *Microsoft* SDK.

During the tests reported in the following section, we used a 2.4x2.4m interaction space divided for calibration into a 4x4 grid. Our system configuration is shown in Figure 3.

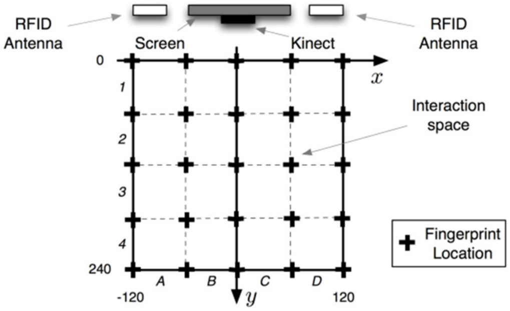


Figure 3. System configuration: RFID antennae, camera tracking system, display, 4x4 grid. The measures are in cm.

Parameter tuning

In order to gather the fingerprint information, we performed a reading of 1000 RSSI values in each of the 25 corner points of a 4x4 grid, which took less than one hour. To train the values of K and α, we collected 1000 RSSI values from 20 randomly selected points within the 2.4x2.4m interaction space, and chose the parameters providing the best tracking accuracy. Noticeable discrepancies in the optimal value of K were observed across the interaction space. Acknowledging them, we set $K = 5$ for those points with y <80cm or $y > 180$cm, and $K = 20$ for all the other points. The optimal value of α was estimated at 0.1 in all cases. Note that the overall setup phase took only a few hours to complete, and it is mostly automatized. Setting β equal to 0.01 guarantees a good trade-off between robustness and responsiveness. A window length (W) of 20 was used during the experiments. This is in line with our original intention to create a system that requires only a short calibration that can be performed without any knowledge of how the system works.

Preliminary In-lab Evaluation

The in-lab evaluation was performed with members of our research group. Each test case was repeated 10 times (for a total of 80 instances). We marked the 4x4 calibration grid on the floor, to make sure that each test was repeatable.

Test Cases

We analyzed different instances of three main test cases: Initialization, Tag Exchange, Stress Tests.

Initialization: Two people enter the exhibition space together. Under *Far x*, *Far y*, and *Close*, two users enter the exhibition space and remain still for 60 seconds, within the average 2 minute linger time observed at most museum exhibits. In the first case, they stand at variable distance from the screen, but farther than 1 m from each other on the x-axis of the grid; in the second case, they stand farther than 1 m from each other on the y-axis of the grid; in the third case, they both stand within a radius of 1 m. Under *Moving*, two people enter the interaction space, stand for 10 seconds facing the screen (which simulates two visitors looking for the first time at the data that appear on the screen), and then start “interacting with the system”, i.e. they move towards and away from the shared screen to change the opacity of the bubbles displayed. We expected this case to be the most challenging *Initialization* test, as RSSI is notoriously sensitive to tag movements; however, we also expected it to be the more realistic scenario in a museum setting,

Tag Exchange: Two people exchange their tags. This may happen if two visitors decide to swap the set of data that each of them is controlling. In our scenario, this means exchanging the nationality; if each visitor was controlling different features (for instance, timeline and zoom) with their motions, it would mean swapping “embodied functionality”. It is worth noting that these tests serve double-duty: they also replicate how the system reacts when an erroneous identity assignment occurs. In both test cases, two people enter the space, stand for 5 seconds, interact with the system for 40 seconds; at $t_{ex} = 45$ seconds, they exchange tags. Under *Keep pose*, nothing else happens; under *Swap pose*, the two users also swap their position on the grid: for instance, if user A was in cell A3 and user B in cell C1 when the swap occurs, A gives his tag to B, wears B’s tag and moves to cell C1, while B moves to cell A3, then resume their interaction with the system.

Stress Tests: we tested our system under some un-realistic scenarios, which are however known to be critical for the RFID infrastructure. Under *Free movement*, two users enter the interactive area, stand facing the screen for 5 seconds, and then start walking, running, and jumping in the space, frequently occluding each other and facing different directions. Under *Circles*, two people face the screen for 5 seconds and starts to walking in circles in two different halves of the interactive space. We designed this last case because the RSSI is notoriously influenced by the angle between the tag and the antenna.

Evaluation Metrics

Ground truth (which user had possession of which tag and when) was recorded manually using a stopwatch. We used several metrics to compare system performance:

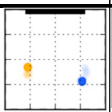
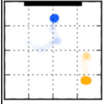
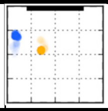
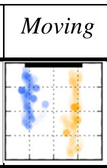
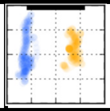
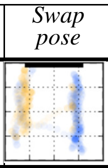
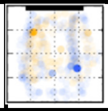
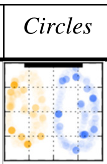
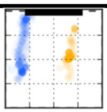
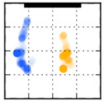
		Initialization				**Tag Exchange**		**Stress Tests**		***in situ* Museum**	
		Far x	*Far y*	*Close*	*Moving*	*Keep pose*	*Swap pose*	*Free mov.*	*Circles*	*OpenNI*	*MS sdk*
Example											
$\Delta T_{adj,1}$ (s)	$\bar{X}$ SD	2.433 3.159	6.029 4.364	1.574 2.640	5.818 5.290	4.133 4.007	2.985 2.423	2.147 2.175	6.695 6.942	10.129 10.077	5.088 5.392
$MOTA_1$	$\bar{X}$ SD	0.991 0.026	0.937 0.134	1.0 0.0	1.0 0.0	1.0 0.0	1.0 0.0	1.0 0.0	0.905 0.101	0.903 0.133	0.902 0.193
$\Delta T_{adj,2}$ (s)	$\bar{X}$ SD	-	-	-	-	11.879 4.579	15.106 7.117	-	-	-	-
$MOTA_2$	$\bar{X}$ SD	-	-	-	-	1.0 0.0	1.0 0.0	-	-	-	-
$MOTA_{tot}$	$\bar{X}$ SD	0.959 0.044	0.790 0.268	0.915 0.203	0.932 0.077	0.853 0.078	0.832 0.090	0.967 0.033	0.839 0.141	0.815 0.133	0.847 0.213

Table 1. Evaluation of the system. Initialization, Tag Exchange and Stress Tests were performed in-lab. The last column presents the results of the in situ evaluation at a museum. For each evaluation metric, we report mean $\bar{X}$ and standard deviation SD

- $\Delta T_{adj,1}$: time (seconds) required by the system to stabilize after two new users enter in the interaction space. It is defined as the time elapsed until the system performs a swap to an identity assignment that: (1) afterwards is kept for at least 20 seconds; (2) is coherent with the ground truth. $\Delta T_{adj,1}$ is a measure of the *reactivity* of the system (one of our design goals): if the initial identity assignment is slow, people may not realize that the system is interactive and pass it by.
- Multiple-Object Tracking Accuracy ($MOTA_{t_1,t_2}$) [4]: measures the *accuracy* of the system and is defined as the fraction of $[t_1, t_2]$ during which the system assigns the correct identities. High values of accuracy allow the system to *mirror* the action of each user on her/his own data.
- $\Delta T_{adj,2}$: time (seconds) required by the system to stabilize after two users exchange their tags. It is defined as the difference between $t_{adj,2}$ (when the system performs a swap to an identity assignment that is coherent with the new ground truth and is stable for at least 20 seconds afterwards) and t_{ex} (when the tag exchange occurs). This value measures the *reactivity* of the system in re-assigning identities.

Figure 4 illustrates the meaning of these parameters in one instance of the *Tag Exchange* test case. We denoted as $MOTA_1$ the value of $MOTA_{t_1,t_2}$in the interval $[t_1 = t_{adj,1}, t_2 = t_{ex}]$, i.e. after the system stabilizes, as $MOTA_2$ the value in $[t_{ex}, t_{end}]$, i.e. when the system stabilizes again after the tag exchange, and as $MOTA_{tot}$ the value in $[0, t_{end}]$, i.e. also when the system is not stable yet. In all the other test cases (when a tag exchange does not occur), $MOTA_1$ is the value in the interval $[t_{adj,1}, t_{end}]$, $MOTA_2$ has no meaning and $MOTA_{tot}$ is the same.

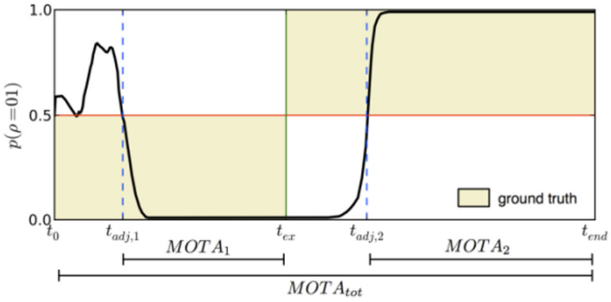


Figure 4. Parameters used for the system evaluation.

Results after the In-lab Evaluation

The results of the in-lab evaluation are reported in Table 1. In the first row, we report one instance of the users' paths showing how they interacted with the system, generated from the tracking data collected with the *Kinect*™. The first user is shown in green, the second in red. The thick black line represents the screen location, and the darkest areas are those where the user spent the most time.

Initialization. As we expected, the system is faster and more accurate initially identifying two users when they are far from each other along the x-axis (left-right) rather than when they are far on the y-axis, as the two antennae are placed to the left and right of the screen. This result will change depending on different antennae configurations. Also, as we expected, the average time required by the system to stabilize (under 6 seconds, which is always a one-time cost, as then people are tracked with the *Kinect*™ when they move) was higher when the users were moving, than when they were standing far from each other on the x-axis (under 2.5 seconds). Surprisingly, though, the accuracy ($MOTA_1$ and $MOTA_{tot}$) of these two test cases is comparable, even though the RSSI is affected by tag movement. This can be explained by: (1) the KL-divergence is computed on a window of RSSIs (20 values), not on a single instance of the RSSI; (2) the probabilistic model mediates the inaccuracy and instability of the data coming from the RFID. Furthermore, *Moving* proved to be a better scenario than *Far y*, probably because receiving data from different locations (people were moving in the space) reduced the risk of standing in spots that were more difficult for the RSSI-based system to disambiguate.

Once under *Far y* and once under *Close* the system failed to stabilize to the correct identity assignment during the 60 seconds interaction. The cause of this problem and whether or not a wrong identity assignment may persist during a longer interaction should be investigated in future work. For our analysis, we incorporated this information in the average $MOTA_{tot}$ for each of the two test cases; we decided to consider these two instances as outliers for $T_{adj,1}$and $MOTA_1$, as the system never reached a stable identity assignment (those two evaluation metrics refer to stable identity assignments).

Tag Exchange. We did not observe a significant difference between the two tests cases. As we expected, $T_{adj,2}$ is considerably higher than $T_{adj,1}$, as the system has to adjust itself and switch from one stable configuration to the opposite one. It is worth noting that $T_{adj,2}$ includes the time the two users require to give the tag to each other, wear it and, under *Swap pose*, to move in the grid (without necessarily facing the screen); this additional time generally varied between 3 and 6 seconds.

Stress Tests. The results of the two stress tests were surprising. As we expected, the accuracy ($MOTA_1$) of *Circles* is the lowest of the in-lab test cases; however, this value is still high (the identity assignment was preserved in 90% of the interaction time), considering that the RSSI is known to be dependent on the tag angle and on tag movements. Under *Free movement*, we even obtained values of both $T_{adj,1}$ and $MOTA_1$ that outperform many of the other in-lab test cases. This tendency is very promising for highly interactive scenarios, given the range of interactions this would support.

In Situ Evaluation

The *in situ* evaluation took place at a small history museum and cultural center, in two separate 2-weeks sessions (the second one took place after four months). 30 museum visitors participated in each session, in groups of 2 people, for a total count of 60 people. Our *in situ* trials, one in a main hallway and one in a regular gallery space, were both located where many non-interacting visitors walked through the interaction space.

Methods

The moderator explained how the system worked and asked each visitor to pick up one RFID tag and to specify the ethnicity whose data she/he wanted to explore. After that, the moderator invited the two participants to enter the room with the shared screen. No constraints were given on the interaction. Visitors were free to explore the data, to interact with the system for about one minute, and to ask questions to the moderator and to each other about the data displayed on the shared screen. We also removed the grid that we used for the calibration, to prevent people from being influenced by the cell limits. We marked a 1cm-thick blue line on the floor 90 cm from the screen (the end of the interaction area) but did not point it out to visitors.

Results

The results ($T_{adj,1}$, $MOTA_1$, and $MOTA_{tot}$) of the first *in situ* evaluation are presented in Table 1 (Museum - OpenNI), while the interaction patterns exhibited by visitors are shown in Figure 5.

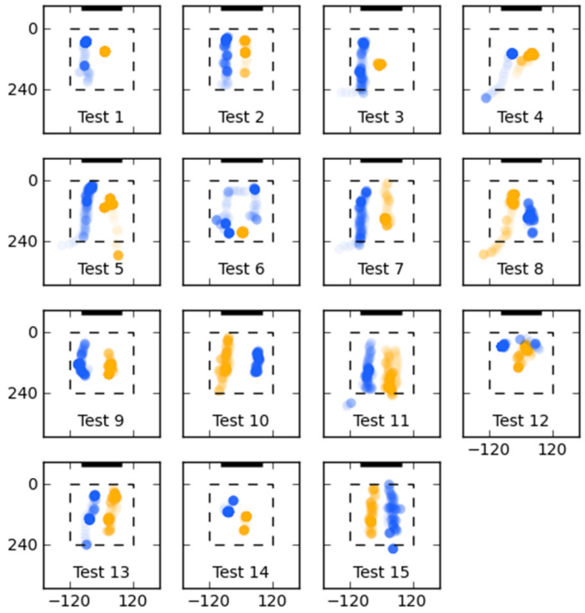


Figure 5. Interaction patterns during the *in situ* evaluation. First user in blue, second in yellow. Darker traces corresponds to the locations were the user spent more time. Measures are in cm. The interaction space that we calibrated is dashed.

Both the moderator and the two participants knew which “nationality” (i.e. which tag) each visitor had chosen. In all cases the system stabilized to the correct identity assignment (ground truth). As we expected, these results were less good than in-lab: the average time needed by the system to stabilize was 10s, and the accuracy after the initialization phase ($MOTA_1$) was 90.3% on average.

We performed a second museum session in order to validate the previously obtained results with a different *Kinect*™ API (Microsoft’s official SDK). The results of this second evaluation are presented in Table 1 (Museum – MS sdk). The accuracy of the system ($MOTA$) is coherent with the values observed with OpenNI. However, the system was now faster in disambiguating the identity of people (ΔT_{adj}). A possible explanation for this improvement is that the Microsoft SDK seems to be faster and more robust in recognizing human shapes at the beginning of the interaction. It is worth noting that switching from OpenNI to Microsoft SDK was transparent to the Bayesian model and to the user.

Discussion

During our *in situ* evaluation, the system worked as desired to promote data exploration. We observed users engaged in conversations related to data interpretation, such as "*I'd never realized that there were so many Germans and Italians in Chicago!*" or remarks that indicated that the personalization of the experience was prompting comparison (such as "[BRITISH] *It looks like I'm along the Lake*" - "[POLISH] *And I'm not.*") We also observed some design parameters that should be considered when constructing personalized embodied interaction experiences in a museum setting.

Passersby. Non-participating visitors passing by the exhibit are a challenge that cannot be avoided, certainly during exhibit use and possibly also during calibration. Despite our best efforts to perform calibration at "dead times" at the museum, people crossed the interactive area several times, which we suspect slightly affected the *in situ* system performance. Blocking off the space during calibration, or if that is impossible, calibrating when the museum is closed is recommended. During use, we had a number of passersby walk behind the users or sometimes even between the users and the screen. The identity resolution process is accelerated if we take the simple measure of discarding any Kinect-detected shapes outside the marked interaction boundaries, and then relying on the algorithm to assign valid RFIDs to shapes within the interaction boundaries.

Stepping outside the interaction space. During the first *in situ* evaluation, visitors stepped outside the interaction space that we had previously calibrated in 11 cases (73%). The second *in situ* evaluation also confirmed the users' tendency to stand really close to the screen (between 10 and 90 cm from the 65" display) when talking about the data. The *Kinect*™ camera is not able to track people when they are close to the sensor. Even though our system was successful in recovering the identity of those users, the overall accuracy ($MOTA_{tot}$) was affected. Also, visitors cannot actively interact with the system when they are not tracked. In order to minimize this recurrent behavior, a more tangible delimiter (like a stanchion) could be installed to mark this interaction boundary.

Dealing with Identity Ambiguities. Even though the system takes few seconds to adjust at the beginning of the interaction, visitors never noticed this initialization time. In all the test cases, they stood for longer than 10s after entering, just looking at the screen. Also, people never made any remarks about identity mismatches in 28 cases (93%). This is probably due to the fact that: (1) people were engaging in conversation about the data; (2) most problems occurred when people were standing close to each other, so they were not able to really see any effect on the screen. However, as ambiguous identity assignments may derail the interaction experience, we plan on damping the system responsiveness when the probability level of the current identity assignment drops below a certain threshold.

How people carry the tag. At the beginning of 4 tests people entered the exhibition space holding the tag in their hands, until the moderator asked them to pin it to their shirt. This behavior might have influenced the accuracy of the system. During the second *in situ* evaluation, we inserted the tag into a badge holder that people naturally wore around their necks, which resolved the problem.

CONCLUSION AND FUTURE WORK

In this paper we presented and validated a model that combines input from a mid-range RFID reader with input from a commercial camera-based tracking system: we used the Kinect™ as a source of tracking data and we augmented it with an RFID-based identity assignment monitor to attain tracking precision and identity preservation not possible with any existing system. Visitors can actively control and explore their own data with their bodies, in a shared space.

Some guidelines should be considered when this system is incorporated in the design of a museum exhibit, in order to optimize its performance. (1) Given the initial delay in reconciling identities, and that accuracy was higher when people were moving in the space ("Moving" vs. "Far y"), this system is better used in spaces that are large enough to allow visitors to walk towards the display for a few seconds (before they would expect it to respond to them); (2) Users should be prevented from stepping too close to the camera, possibly with small stanchions; (3) As the system might temporary produce an erroneous identity assignment (even though for a very short time interval), it should enter an uncertainty state when the probability of an identity assignment is below a threshold – either intentionally "damping" the system responsiveness until the identity is resolved, or even generating prompts (e.g. asking visitors to walk around) to automatically correct the problem.

Even though we installed and tested one single prototype, our vision is that a museum might add multiple such exhibits throughout their galleries – while our system would not track visitors in-between these locations, it would allow for continuity of experience from exhibit to exhibit. If many interactive exhibits are installed across the museum, they can support an experience that is sensitive to not just on who you are but also to what you have/haven't seen already in the museum. Future work should investigate the performance of this system in a bigger interaction space, which may allow the presence of a greater number of users.

The personalized perspective leads visitors to noticing and questioning patterns in the data, and making comparisons across each other's data sets - this phenomenon will be further investigated in future work. Personalized interactions have the potential to fundamentally change museum experiences. In museums, visitors often tend to linger in a peripheral zone before engaging with an exhibit, especially if strangers are present. If this system can issue personalized "invitations" to visitors to "step right up", it can help break down these barriers and possibly even get visitors to speak to each other (a known hard problem). In cultural and historical institutions this can be particularly valuable, as part of the experience is about coming to understand the lives and histories of others. Modern technologies, including handhelds, may tend to isolate us from one another. On the contrary, the system that we presented might be able to bring us together.

ACKNOWLEDGMENTS

This work is funded by the NEH Digital Humanities Start-Up Grant HD 51357. We would like to thank the staff of the Jane Addams Hull House Museum.

REFERENCES

1. Allen, S. Designs for learning: Studying science museum exhibits that do more than entertain. *Science Education 88*, S1 (2004), S17-S33.
2. Antle, A.N., et al. Towards Utopia : Designing Tangibles for Learning. *Design*, (2011), 11-20.
3. Bahl, P. and Padmanabhan, V.N. RADAR: an in-building RF-based user location and tracking system. *Proceedings IEEE INFOCOM 2000*, 775-784.
4. Bernardin, K., Elbs, A., and Stiefelhagen, R. Multiple object tracking performance metrics and evaluation in a smart room environment. *VS'06*.
5. Cafaro, F., Lyons, L., Radinsky, J., and Roberts, J. RFID localization for tangible and embodied multi-user interaction with museum exhibits. *Ubicomp'10*.
6. Callaway, C., Stock, O., Dekoven, E., et al. Mobile Drama in an Instrumented Museum : Inducing Group Conversation via Coordinated Narratives. *IUI'11*, 73-82.
7. Camp, F.V.D., Voit, M., and Stiefelhagen, R. Efficient Person Identification using Active Cameras in a Smartroom. *MM Workshop*, (2010), 17-22.
8. Carreras, A. and Pares, N. Designing an Interactive Installation for Children in a Museum to Learn Abstract Concepts. *EdMedia 2007*, 3454-3459.
9. Correia, N., at al. A multi-touch tabletop for robust multimedia interaction in museums. *ITS 2010*, 117-120.
10. Cox, D., at al. IntelliBadge^{TM}: Towards Providing Location-Aware Value-Added Services at Academic Conferences. *Ubicomp 2003*, (2003), 264-280.
11. Hallaway, D., Hollerer, T., and Feiner, S. Coarse, inexpensive, infrared tracking for wearable computing. *Seventh ISWC 2003*, (2003), 69-78.
12. Hornecker, E. Interactions around a contextually embedded system. *TEI 10*, (2010), 169.
13. Jimenez Pazmino, P. and Lyons, L. An exploratory study of input modalities for mobile devices used with museum exhibits. *CHI 2011,* ACM (2011), 895-904.
14. Khoshelham, K. Accuracy Analysis of Kinect Depth Data. *GeoInformation Science 38*, 1.
15. Khoury, H.M. and Kamat, V.R. Evaluation of position tracking technologies for user localization in indoor construction environments. *Autom. in Construction 18*, 4
16. Kim, K. and Davis, L.S. Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering. *ECCV 2006*.
17. Kuflik, T., Lanir, J., Dim, E., et al. Indoor positioning: challenges and solutions for indoor cultural heritage sites. *IUI 2011,* ACM (2011), 375-378.
18. Kullback, S. and Leibler, R.A. On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1.
19. Lehn, D. Vom, Heath, C., and Hindmarsh, J. Conduct and Collaboration in Museums and Galleries. *Symbolic Interaction 24*, 2 (2001), 189-216.
20. Lonsdale, P., et al. Context awareness for MOBIlearn: Creating an engaging learning experience in an art museum. *MLEARN 2004*.
21. McCarthy, J.F., et al. Augmenting the Social Space of an Academic Conference. *CSCW 2004*.
22. McDonald, D.W., at al. Proactive displays: Supporting awareness in fluid social environments. *Transactions on Computer-Human Interaction 14*, 4 (2008), 1-31.
23. Mirowski, P., Steck, H., Whiting, P., et al. KL-Divergence Kernel Regression for Non-Gaussian Fingerprint Based Localization. *Ratio 10*, (2011), 21-23.
24. Phillips, P.J., Jiang, F., Narvekar, A., Ayyad, J., and O'Toole, A.J. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept. 8*, 2 (2011).
25. Prante, T., et al. Ambient Agoras. *CHI 04*, (2004), 763.
26. Priyantha, N.B., Chakraborty, A., and Balakrishnan, H. The Cricket location-support system. *MobiCom 2000*.
27. Ryu, H.-S., Yoon, Y.-J., Lim, M.-E., Park, C.-Y., Park, S.-J., and Choi, S.-M. Picture navigation using an ambient display and implicit interactions. *OZCHI 07*.
28. Shankar, M., et al. Human-tracking systems using pyroelectric infrared detectors. *Optical Eng. 45*, 10 ('06).
29. Teixeira, T., et al. Tasking Networked CCTV Cameras and Mobile Phones to Identify and Localize Multiple People. *Main*, (2010), 213-222.
30. Xu, A. and Dudek, G. Fourier Tag: A Smoothly Degradable Fiducial Marker System with Configurable Payload Capacity. *CRV '11*, (2011).
31. Zhang, Y., Partridge, K., and Reich, J. Localizing tags using mobile infrastructure. *Lo'CA07*, (2007), 279-296.
32. Zhao, W., et al. Face recognition: A literature survey. *ACM Comput. Surv. 35*, 4 (2003), 399-458.