# My Being to Your Place, Your Being to My Place: Co-present Robotic Avatars Create Illusion of Living Together

Bumsoo Kang[1*], Inseok Hwang[2†], Jinho Lee[2], Seungchul Lee[1],
Taegyeong Lee[1], Youngjae Chang[1], Min Kyung Lee[3]
[1]KAIST, [2]IBM Research, [3]Carnegie Mellon University
[1]{bumsoo, seungchul, tglee, yjchang}@nclab.kaist.ac.kr, [2]{ihwang, leejinho}@us.ibm.com, [3]mklee@cs.cmu.edu

## ABSTRACT

People in work-separated families have been heavily relying on cutting-edge face-to-face communication services. Despite their ease of use and ubiquitous availability, experiences in living together are still far incomparable to those through remote face-to-face communication. We envision that enabling a remote person to be spatially superposed in one's living space would be a breakthrough to catalyze pseudo living-together interactivity. We propose HomeMeld, a zero-hassle self-mobile robotic system serving as a co-present avatar to create a persistent illusion of living together for those who are involuntarily living apart. The key challenges are 1) continuous spatial mapping between two heterogeneous floor plans and 2) navigating the robotic avatar to reflect the other's presence in real time under the limited maneuverability of the robot. We devise a notion of functionally equivalent location and orientation to translate a person's presence into another in a heterogeneous floor plan. We also develop predictive path warping to seamlessly synchronize the presence of the other. We conducted extensive experiments and deployment studies with real participants.

## CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; **Ubiquitous and mobile computing systems and tools**; • **Computer systems organization → Robotics**;

## KEYWORDS

Robotic avatar; telepresence; co-presesnce; co-located interaction

**ACM Reference Format:**
Bumsoo Kang, Inseok Hwang, Jinho Lee, Seungchul Lee, Taegyeong Lee, Youngjae Chang, Min Kyung Lee. 2018. My Being to Your Place, Your Being to My Place: Co-present Robotic Avatars Create Illusion of Living Together. In *MobiSys '18: The 16th Annual International Conference on Mobile Systems, Applications, and Services, June 10–15, 2018, Munich, Germany.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3210240.3210348

*This work was done in part while he was on an internship at IBM Research - Austin.
†The corresponding author.


*MobiSys '18, June 10–15, 2018, Munich, Germany*

ACM ISBN 978-1-4503-5720-3/18/06…$15.00
https://doi.org/10.1145/3210240.3210348

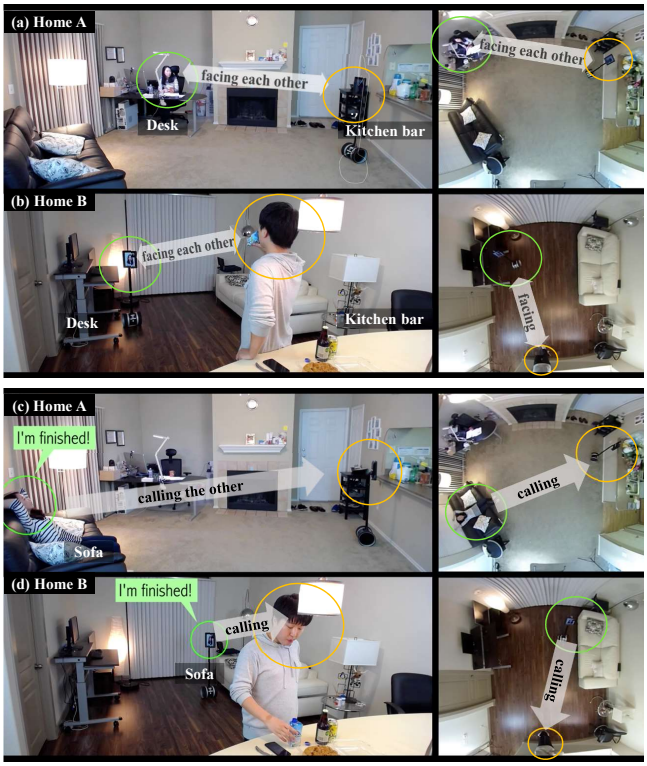



**Figure 1: HomeMeld [1] enables family members to naturally present at two heterogeneous homes at the same time.**

## 1 INTRODUCTION

Think of a work-separated family. Two partners who have been appointed to different schools, or a military officer serving at a base far from her husband and children. It is unfortunate but yet they can manage to see and talk to each other every day, thanks to cutting-edge mobile, networking, and multimedia technologies.

Imagine a fictional system that superposes two distant homes and the people therein. It provides real-time self-mobile avatars directly in the physical space of both homes; e.g., the person at Home A is present in Home B through her avatar, and vice versa. The avatar could be a hologram or a humanoid robot looking, moving around, and acting exactly in the same way as she is doing at her remote home, despite all the spatial and interior differences between those homes. For example, a military officer enters her quarters after work; she sees his young son's avatar running to her at the porch,

[1]Our video is available at: https://goo.gl/7xJwJj

welcoming him with open arms. At the same time, the little boy sees his mother's avatar entering his porch and runs to her. They dine together; each of them is dining at one's own table with the other's avatar dining at the same table. They sit on the couch together, talk and laugh, and she sees him falling asleep next to her.

This fictional system highlights an avatar *intelligently* mimicking the remote person in real-time. The avatar does not simply copy every inch of one's activity and location as-is; it relocates and adjusts itself to make it look perfectly natural in different home environments. This system enables a distant couple to be mutually *co-present* in each other's living space, with all activities and movements intelligently mirrored to each other. Such co-presence brings many pseudo living-together experiences incomparable to those of today's telecommunication. They naturally perceive each other's live presence and context directly in one's living space. It could be even a peripheral perception [48] with little cognitive cost. Making an interaction is hassle-free, with no interference. Also, small interactions may naturally blend with other on-site activities such as cooking or household chores, which are often discouraged in remote conversation. In essence, they may communicate over a distance, but they commune together when being co-present.

Realizing such a system opens up a vast spectrum of technical and design challenges. In this paper, we build an initial prototype, *HomeMeld*, a zero-hassle self-mobile robotic system serving as a real-time, co-present avatar to create a persistent illusion of living together. HomeMeld is built on top of a commercial telepresence robot hardware [2] and a CNN-based computer vision technique, letting the person be device-free at all times.

Among many questions along the path towards this ambitious dream, HomeMeld focuses on intelligent simultaneous positioning of the avatar from a human point-of-view. Figure 1 shows two scenes of HomeMeld operating at two distant homes. The first scene shows the wife's home (Figure 1a) and the husband's home (Figure 1b) at the same time. The wife is sitting at her desk and talking to her husband's avatar standing against her kitchen bar and facing her. Conversely, the husband is standing against his kitchen bar and talking to his wife's avatar located at his desk and oriented towards him. The second scene shows those two homes a few moments later (Figure 1c and 1d, respectively). The wife has moved to her couch, calling her husband's avatar that has turned around towards her kitchen bar. In the husband's home, her avatar has moved autonomously to his couch, as she has, and calls him from behind. Having the other person's presence at such an equivalent location is important because it is an intuitive indicator to his/her context, for example, she may be busy with her work and he may feel like drinking something.

As illustrated above, HomeMeld synchronizes the avatar's presence with the remote person in terms of both space and time. We highlight that, given an arbitrary location/orientation in Home A, defining the human-perceived equivalent location/orientation in Home B is a fuzzy problem susceptible to uncertainty, subjectivity, and dependency on individual home environments.

We address this problem through a two-tiered approach. We begin with establishing human insights; we collect extensive, human-labeled anecdotal point-to-point mappings between heterogeneous floor plans. This data set revealed interesting trends of varying human-perceived equivalence between in-home locations, depending on a location's proximity to household objects with inherent functions. Then we leverage those trends to derive a generalized computational model that evaluates the *functionally equivalent location* in a home for a given location in another home.

Our model of functionally equivalent location synchronizes the avatar's presence in terms of space, but not necessarily in terms of time. Our model translates a person's real-time path in Home A into a sequence of functionally equivalent locations in Home B. It does not necessarily ensure the same traveling distance, resulting in the avatar lagging behind. We devise *predictive path warping* to ensure the user experience of same-time co-presence.

To continuously detect a device-free person's in-home location/orientation, HomeMeld takes a single-camera vision approach. We instrument a room with a ceiling-mounted 180° camera in favor of obstruction-free, accurate localization. This design choice led us to train a custom CNN model localizing a person and detecting her orientation viewed from a very unusual angle, at every possible location and orientation in various homes. We devise unique strategies to collect, augment, and synthesize the data set to mitigate the data-hungry challenges of training a custom-purposed CNN model.

Although our initial prototype is based on a simplified robotic hardware that is unable to mimic every possible human motion, the key features of HomeMeld could be seamlessly extended towards higher fidelity mimicry as more sophisticated hardware with higher degrees-of-freedom such as a humanoid robot, as well as higher-precision indoor human activity sensing [37] become available.

Our contributions are threefold. First, we envision a physical co-presence service superposing two distant, heterogeneous living spaces through self-mobile avatars. Second, we build a device-free working prototype, through which we address self-navigating telepresence robots featuring human-perceived spatio-temporal equivalence across heterogeneous homes. Third, we conduct small-scale deployments with real participants suffering from involuntary distant living and discuss their experiences with HomeMeld.

## 2 PRESENCE OVER A LONG DISTANCE

Needless to say, it is widely agreed that living together is crucial for people in a family-like relation. Formally, a phenomenological study [54] revealed the prominent elements representing the experience of family-like intimacy. Those include *presence*—noticeable existence of a person with another, *time continuity*—shared experiences persisting over time such as a whole day, *boundary-free*—no boundaries between them in both physical and psychological ways, and *nonverbal communications*. Not surprisingly, those are the privileges naturally granted when living together.

The growing trends of globalization and work-force mobility are impacting the traditional living-together models. In 2006, a census result reports that 3.6 million married Americans live in a different city from their spouses due to work, which is a 53% increase since 2003 [5]. Those involuntarily living apart adopt computer-mediated communication technologies to see and talk to each other, share a moment, and feel the other's existence, yet which offers partial, limited coverage of the aforementioned elements.

Real-time face-to-face communication services, such as a video chat, are a predominant tool that conveys not only words but also

some nonverbal cues [34, 44]. Although some couples use a video chat to improvise a narrow-band life-sharing channel at home [29], those are generally not compatible with long-lasting, relaxed experience sharing as at home; a video chat is often presumed attention-demanding and preemptive of other in-home activities.

Experimental technologies have been proposed to enrich family interaction by sharing a task-specific tangible activity [30, 62] or augment a conversation channel with additional sensory stimuli [28, 46]. Those exhibit a specific form of presence at a certain moment, but are not designed to provide a continuous, boundary-free remote presence. Alternatively, a large body of work has explored providing a feeling of ambient presence of the other person [17, 18, 31, 32], which allows continuous signaling of one's presence, yet through a very low-fidelity channel hardly helping meaningful communication.

*Telepresence*, first coined by Marvin Minsky [41], aims for a more holistic presence of a remote person. In light of this, a life-sized real-time projection provides a long-lasting, boundary-free experience of telepresence in a device-free augmented reality style [49]. But the stationary instrumentation of a projector renders the other person immobile. While wearable augmented reality devices could let free the remote person, keeping the wearable device on the face hampers long duration usage, making the user feel not as comfortable as being at home. A fundamental shortcoming is that a virtual remote person and a physical local person may pass through each other. Such events, which are implausible in the real world, cause disruption in the immersive presence experience [33].

Telepresence through a robotic agent was originally explored for situated collaboration between remote workplaces [47]. Such robotic hardware has evolved and has become commercially available at an affordable cost [1, 2], which are essentially a video chat terminal on a movable platform piloted by a remote user. Experiments are in progress to broaden its penetration, e.g., in academic conferences [45], companies [36], and very recently in homes [63].

Robotic telepresence has been found to be useful for strengthening a sense of presence between remote people, mainly because of its ability to move around [36]. But we stress that the perception of a remote person's presence is not mutual. A robot must be controlled by a remote pilot, who stays on the interface. While the other person who is mingling with the robot is free to move anywhere and do anything in his space, the piloting person is left effectively immobile under high cognitive loads of both piloting and video-chatting tasks. Furthermore, if we imagine a symmetric setup where each person pilots a robot in the other person's home and vice versa, it would be pointless as both become immobile and their experiences would be no better than a stationary video chat.

We argue that such asymmetric experiences and requirements of humans in the loop are fundamental hurdles against embedding natural, hassle-free, long-lasting co-presence using a robotic agent within distant people's unconstrained life in their living spaces.

## 3 DESIGN STUDY

In this section, we explain the current practices and challenges found from real work-separated families, through which we have established the key design requirements of HomeMeld. We conducted a preliminary user study that included one-on-one semi-structured interviews and a design workshop. First, we investigate current ways and limitations of communication between remote family members, then distill essential design considerations to realize the fictional system for providing a sense of living together.

**Table 1: Participant demographics**

| Gender | Age | Occupation | Period | Remark |
|---|---|---|---|---|
| M | 27 | Graduate student | 1 year | |
| M<br>F | 29<br>31 | Army surgeon<br>Dentist | 6 months | Married |
| M | 26 | Graduate student | 1 year | |
| F | 29 | Teacher | 10 months | |
| M<br>F | 33<br>35 | Researcher<br>Post-doc | 4 years | Married |
| M | 30 | Post-doc | 2 years | |
| M | 32 | Graduate student | 3 years | |
| M | 36 | Researcher | 3 months | |
| M<br>F | 29<br>28 | Graduate student<br>Engineer | 3 years | Married |
| F | 38 | Professor | 3 years | |
| M | 31 | Researcher | 6 months | |

### 3.1 One-on-one Semi-structured Interviews

We were interested in those who are work-separated from family or distant couples. We recruited 14 participants for one-hour one-on-one semi-structured interviews (see Table 1) from the authors' online social networks in both the United States and South Korea, and from online communities in a public research university in South Korea. Each participant was rewarded a $10 gift certificate. Questions explored in the interviews were: 1) communication channels that they use to interact with remote family members, 2) limitations of current ways of remote communication, 3) moments that they feel like being together with the other family member, and 4) potential ways to provide an illusion of living together.

**Voice/video chatting:** Most participants used voice/video chatting as a main communication channel, however, they responded that pre-arrangement is a significant inconvenience for two reasons. First, it is hard to communicate upon an unexpected event at the agreed time (e.g., delayed schedule, connection problem, etc.) Second, it requires waiting until both sides are available. It often hampers vivid interaction, resulting in a smaller amount of time for interaction, and losing chances to share the experience just-in-time. Such hardship in arranging daily conversation often makes them feel it obligatory, even on a day that is tricky to talk.

Participants also reported that they sometimes feel even more lonely after the end of the communication. It deeply reminds them of needs for a feeling of continuous togetherness.

**Face on the camera, eyes on the screen:** Participants find it hard to video chat and do other activities at the same time because they consider it a norm that they should keep eyes on the other person over the camera. All participants agreed that voice/video chatting requires a higher attention than co-located conversation. Even they feel pressured to keep talking because a momentary pause during remote conversation feels awkward (also supported in literature [39]) or might be a signal to terminate the conversation.

**Serendipitious interaction when living together:** Participants reported advantages of co-located family conversation. Notably, it

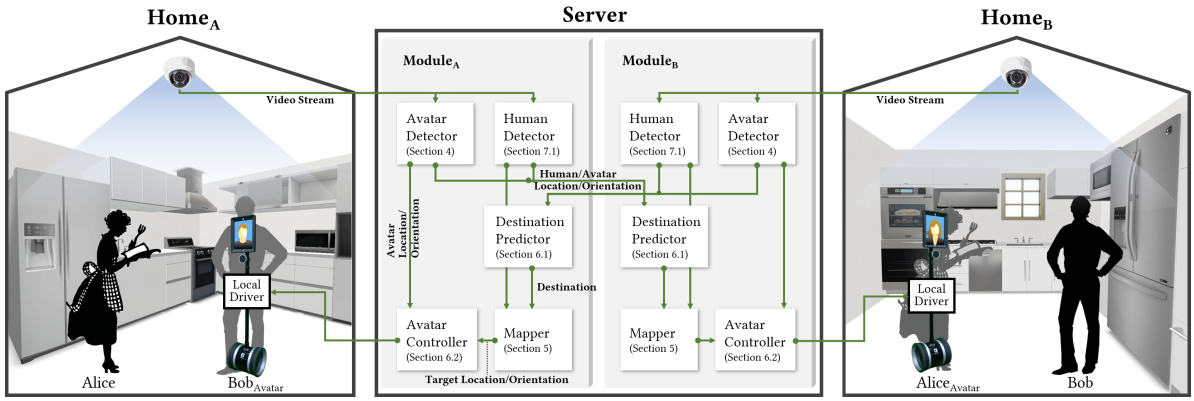



**Figure 2: HomeMeld system overview, instrumenting each home with a 180° camera, a telepresence robot, and a server module**

is so easy to perceive even a subtle context of the other and initiate a relevant chat. For example, a wife starts a casual conversation with asking her husband what he is going to cook when she sees him going to the kitchen, hears a simmering sound, or smells a spicy flavor. Some participants said that knowing the other's context helps to blend their chat while doing routine activities (e.g., dining or doing household chores), which catalyzes a serendipitous interaction to naturally stem from what they are doing now. Also, they feel it acceptable to look elsewhere from time to time while casually chatting, such at a boiling pan or floor being mopped.

## 3.2 Design Workshop

We conducted a two-hour focus group discussion to clarify key design considerations to realize our target system. We recruited six participants who experienced both living together with their families and living apart from them. Three of them are married couples who live apart, and the others are separated from their parents and live alone. Each participant received a $20 gift certificate. Below we list the key desired properties that we identified.

**Device free:** They discussed the inconvenience of voice and video calls. In particular, they pointed out that the preparation process involving a computer or a smartphone for remote chatting was cumbersome, e.g., setting up the equipment, and difficulties to go along with housework such as washing dishes. Most participants agreed that using or wearing the devices every time is a major inconvenience and hinders intimate sharing of everyday lives.

**Physical occupancy:** Most participants said the importance of a substantial embodiment to realize such a system. With holograms or VR/AR, the avatars pass through the user's body and untouched, which may be a factor of diminishing presence. Previous research claims that the illusion of reality with a virtual object can be shattered when their hands pass through it [9].

They stressed that *physically occupying space* is important in giving a sense of co-presence. When real people are together, they have to wait or move around to avoid bumping into the other people in their way. Once they know the other person is a penetrable simulation, they will get used to passing through the other person, developing thoughts that the other person is ignorable.

**Location/orientation:** Some participants claimed a key feature of knowing what a remote family member is doing is often related to where she is at that home. For example, if the system indicates the other person's location in front of the refrigerator, a user easily infers that the other person may be looking for food or drinks in the refrigerator. Interestingly, some participants expected that a location where the other person is staying even for a moment may be a stronger indicator of her activity, than the passing-by locations along a continuous path when the other person is moving.

They also discussed the orientation of the other person. They responded that whether the other person is looking at them or not is quite important because it indicates her willingness to interact. They also said that what the other person is looking at can be an additional clue indicating her current or near-future activity.

Participants demanded to be able to recognize the other family member's context in a most intuitive form; for example, to recognize the other person's location, the most intuitive form would be an in-situ visual indicator, rather than textual or auditory description.

# 4 SYSTEM OVERVIEW

We designed HomeMeld, a self-mobile robotic mutual co-presence system requiring zero control from users and zero mobile devices. Figure 2 shows an architectural overview. It consists of a ceiling mounted camera, a telepresence robot, and a server module for each home, which all cooperate to realize the system.

A ceiling mounted 360° camera is installed at the center of the target living space. One hemisphere of the camera is facing downward, obtaining a 180° field of view (FoV) of the room. It continuously streams the video of the living space to the server. The server is comprised of two modules, each includes multiple components, which are responsible for a home and the person living therein. The two modules work almost independently, except for the exchange of human/avatar information between two homes.

For ease of explanation, we name the two people Alice and Bob respectively for the rest of the paper. In $Home_A$, Alice lives with $Bob_{Avatar}$ and Bob lives in $Home_B$ with $Alice_{Avatar}$. The system settings and operations described herein are symmetric to both Alice and Bob, although we explain only one-sided operations.

In $Home_A$, the avatar detector finds the location and orientation of $Bob_{Avatar}$ from the video feed, by placing a ArUco vision marker [22] on top of the avatars and applying trigonometry. The marker is sometimes not detected at peripheral area of the FoV due to spherical distortion. We interpolate the location using local

encoders of the robot's wheel. Once the marker is detected again, we correct accumulated errors which usually do not exceed 15 cm.

From the same video feed, the human detector extracts the location and orientation of Alice. Unlike the avatar, Alice, who is supposed to be device-free, has no vision marker. Our design choice is to use a convolutional neural network (CNN) model to obtain the location and orientation of humans. We trained a custom CNN model with our own data set because, to our knowledge, a data set of labeled top-view images of people and a model fit for it were not publicly available. We will discuss this issue in §7.

Once the location/orientation of the $Bob_{avatar}$ and Alice are captured, the system is ready to decide how to move the $Alice_{Avatar}$. Translating the location/orientation from a floor plan to another is not trivial. We introduced the notion of *functionally equivalent location/orientation* to tackle this problem. The household objects in $Home_A$ and the location of $Bob_{Avatar}$ are combined to form *functional objects* of $Home_A$, and the mapper translates the location/orientation of Alice in relation to those functional objects. For this, the location/orientation information of Alice and $Bob_{Avatar}$ in $Home_A$ is sent to $Module_B$, where the controller for $Alice_{Avatar}$ is located at. This is the only communication between the two modules of the server. §5 describes the mapper in more details.

The server module applies an optimization to predict a moving person's likely destination upon the changes of her location and orientation. With enough confidence about Alice's destination, it alters Alice's location to the predicted destination so that $Alice_{Avatar}$ can move more promptly. §6 discusses the destination predictor.

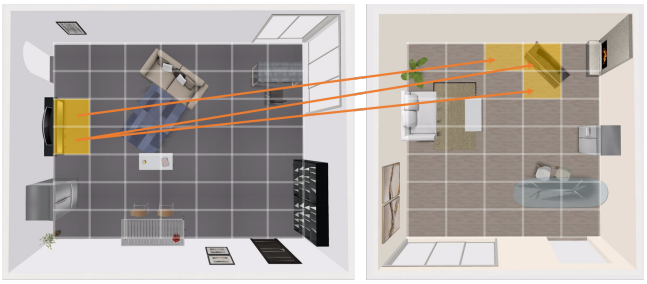
Lastly, the avatar controller controls the robotic avatar in two ways: 1) It plans a path for $Alice_{Avatar}$ in $Home_B$ towards the functionally equivalent location of Alice's predicted destination. The path gets updated on new coordinates of $Alice_{Avatar}$. 2) Importantly, the path planning does not necessarily respect every interim location of Alice to be mapped. It respects only a few locations of Alice of higher significance in terms of functional location, and makes $Alice_{Avatar}$ take a shortest-time curve between such locations to minimize the travel time lag from Alice. The local driver running at $Alice_{Avatar}$ interpolates the intermittent route updates from the avatar controller by dead-reckoning. We discuss more details of the avatar controller in §6.

## 5 MAPPING HETEROGENEOUS HOMES

Our first question is how to reproduce individual's location and orientation at another home in a contextually equivalent way. A challenge is the spatial heterogeneity between the two homes, e.g., different floor plans and furniture arrangements. Obviously, reproducing individual's location and orientation as-is (i.e., same coordinates) will not make sense. To find out human-perceived equivalence about the avatar's locations/orientations, we conducted an empirical study with 20 participants, which is described below.

### 5.1 Functionally Equivalent Location and Orientation

We collected an empirical reference dataset regarding the spatial equivalence between heterogeneous floor plans. From public floor plan repositories such as [4], we collected 20 pairs of heterogeneous



**(a) Source Floor Plan** **(b) Target Floor Plan**

**Figure 3: Human-labeled floor plan mapping**

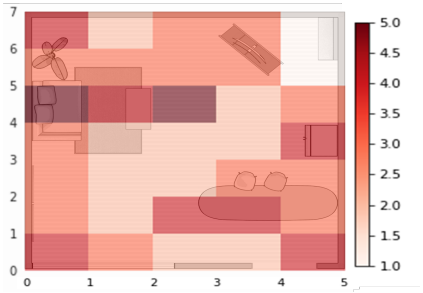



**Figure 4: Heatmap showing participants' agreement rates**

floor plans with various room sizes and numbers of objects. We divided each floor plan into one-square-meter grids. We recruited and asked 20 volunteers to map grids in $Home_A$ to semantically equivalent ones in $Home_B$. Each pair is covered by randomly assigned four or five participants. They were allowed to make one-to-many mappings because a grid in a small home may have multiple semantically equivalent grids in a larger home (see Figure 3). We collected human-labeled ground truth data for 2921 grids. We then conducted short interviews on how they mapped those grids.

We found that most participants started with grids close to in-home objects, and then moved on to the rest of grids by imagining the path from an object to another. Figure 4 shows a heatmap sampled from our results set. It represents how much the participants agree to the mapping results. This implies that the location equivalence is correlated to the proximity to certain in-home objects.

The findings naturally led us to devise a rationale of *functionally equivalent location & orientation* defined based on the in-home objects. When Alice is at home ($Home_A$), her context is likely related to an in-home *functional object* that she is using, such as a refrigerator or a dining table. Being close to the object and looking at the object are strong indicators that Alice is likely interacting with the object, which is consistent with the findings from [50]. For a given location in $Home_A$, a functionally equivalent location in $Home_B$ would be roughly defined as the locations in $Home_B$ related to the same or similar purposes, or near such objects. Similarly, functionally equivalent orientation would mean the orientation that functions the same in two homes, e.g., facing the TV. In HomeMeld, we focus on making $Alice_{Avatar}$ have functionally equivalent location and orientation of Alice as much as possible. The detailed method is described in the succeeding subsections. Note that we include the $Bob_{Avatar}$ in $Home_A$ (not to be confused with $Alice_{Avatar}$ in $Home_B$) as one of the functional objects in $Home_A$, because Alice may be looking at or near Bob when they are together.

**Table 2: Description of the symbols used in Section 5.2**

| Symbol | Description |
|---|---|
| $\mathbf{a}$ | Location vector of Alice. |
| $\mathbf{a}'$ | Location vector of $Alice_{Avatar}$. |
| $\mathbf{o}_k$ | Location vector of functional object $k$ in $Home_A$. |
| $\mathbf{o'}_k$ | Location vector of functional object $k$ in $Home_B$. |
| $\mathbf{D}$ | Distance vector from $\mathbf{a}$ to $N$ functional objects. i.e., $\mathbf{D} = (d_k \mid d_k = \|\mathbf{a} - \mathbf{o}_k\|, k = 1 \ldots N)$ |
| $\mathbf{D'}$ | Distance vector from $\mathbf{a}'$ to $N$ functional objects. i.e., $\mathbf{D'} = (d'_k \mid d'_k = \|\mathbf{a}' - \mathbf{o}'_k\|, k = 1 \ldots N)$. |
| $r$ | Size ratio between $Home_A$ and $Home_B$. |
| $\hat{\mathbf{u}}_k$ | Unit vector of object $k$'s orientation in $Home_A$. |
| $\hat{\mathbf{u}'}_k$ | Unit vector of object $k$'s orientation in $Home_B$. |
| $\theta_k$ | Angular difference between $\mathbf{o}_k$ and $\hat{\mathbf{u}}_k$. |

## 5.2 Mapping Policy

From the findings above, we conclude that the most important aspect of the mapping is to connect the functionally equivalent locations. While it is straightforward to make one-on-one mappings between the same class of functional objects, the question is how to make mappings between arbitrary points in inter-object space. Imagine Alice is moving from a sink to a table, which are next to each other in $Home_A$ but far apart in $Home_B$. The path is not obvious, and possibly not even straight due to other objects.

The common idea underlying our mapping strategies is to use the functional objects as *anchors*. For a given location in $Home_A$, we find an equivalent location in $Home_B$ by using the distances to the same set of functional objects as the similarity metric. In this light, we first built two preliminary mapping strategies; both take a tuple of distances to multiple functional objects into account, e.g., $(d_{couch}, d_{tv}, d_{table})$. Based on real operations, we devised the third mapping strategy considering only a single functional object.

**1) Minimizing distance-ratio-differences** is inspired from trilateration. This strategy is to locate the robot to the point where the ratios among the distances in the tuple are closest across homes. From a reference point where Alice is at, a distance vector to the functional objects is built, and the mapped point at $Home_B$ is the location minimizing the ratio of the vectors. Mathematically, the function takes Alice's location vector $\mathbf{a} = (x, y)$, the sets of functional objects' location vectors $\{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_k\}$ and $\{\mathbf{o'}_1, \mathbf{o'}_2, ..., \mathbf{o'}_k\}$ in $Home_A$ and $Home_B$, respectively, and the floor plan size ratio $r$ and outputs the location of $Alice_{Avatar}$ $\mathbf{a}' = (x', y')$. Define the distance vectors of Alice and $Alice_{Avatar}$ as follows:

$$\mathbf{D} = (\|\mathbf{a} - \mathbf{o}_1\|, \|\mathbf{a} - \mathbf{o}_2\|, ..., \|\mathbf{a} - \mathbf{o}_k\|) \quad (1)$$

$$\mathbf{D'} = (\|\mathbf{a}' - \mathbf{o}'_1\|, \|\mathbf{a}' - \mathbf{o}'_2\|, ..., \|\mathbf{a}' - \mathbf{o}'_k\|) \quad (2)$$

Then, the optimal avatar location is the location that minimizes the $n$-th norm of the vector $\mathbf{D} - r\mathbf{D'}$. We chose $n = 2$ empirically. We used `fmincon` solver to perform the optimization [3].

**2) Maximizing cosine similarity** is fundamentally similar to 1) but different in the objective function. The output avatar location is where it maximizes the cosine similarity between $\mathbf{D}$ and $\mathbf{D'}$.

To see how those two strategies represent individual's intuitive floor plan translation, we developed a simulator. It displays two home spaces $Home_A$ and $Home_B$, and updates the mapped point in $Home_B$ as the user moves the reference point in $Home_A$. We encouraged users to reproduce their frequent in-home activities in the simulator (e.g., heading to the couch after stopping by the refrigerator). Our major observation of both strategies was occasional discontinuity of the mapped points' trajectory in $Home_B$, given continuously moving the reference point in $Home_A$. Such discontinuity might deliver a wrong context to Bob seeing the robot's movement at $Home_B$. When the arrangements of functional objects are very different across homes (e.g., two functional objects are closely located in $Home_A$, while they are far away in $Home_B$), the discontinuity near the object happens more frequently.

**Table 3: Evaluation of mapping algorithms**

| Distance Error | Distance-ratio Differences | Cosine Similarity | Object-oriented Mapping |
|---|---|---|---|
| Near-object error | 3.41 | 4.45 | 3.15 |
| Far-object error | 3.12 | 4.16 | 3.10 |

In our design workshop in §3.2, participants agreed that the equivalence of in-route path is relatively less significant compared to the desination equivalence. To mitigate such discontinuities and put emphasis on the near-object equivalence, we came up with a new strategy considering the single nearest functional object.

**3) Object-oriented mapping** tries to relate the person to a single functional object that she might be interacting with. Each object is assumed to have an orientation unit vector $\hat{\mathbf{u}}_k$ (e.g., direction the TV is facing at). It partitions the given home space into several regions each of which encloses a functional object (e.g., Voronoi diagram [8]). Then a point in $Home_A$ is mapped into a point in $Home_B$ by finding an equivalent relative polar coordinate of the point with respect to the functional object being considered (e.g., scaling the distance from the functional object at the equivalent angle with respect to the object's $\hat{\mathbf{u}}_k$). The location of Alice at $Home_A$ in the $k$-th region enclosing $k$-th object is represented as

$$\mathbf{a} = \mathbf{o}_k + \|\mathbf{a} - \mathbf{o}_k\| \cdot rotate(\hat{\mathbf{u}}_k, \theta_k) \quad (3)$$

The output location for $Alice_{Avatar}$ is given by

$$\mathbf{a}' = \mathbf{o'}_k + r_k \|\mathbf{a} - \mathbf{o}_k\| \cdot rotate(\hat{\mathbf{u}'}_k, \theta_k) \quad (4)$$

where $r_k$ is the size ratio between the $k$-th regions in two homes.

Even though object-oriented mapping also turned out to have a problem of discontinuity, those mostly appear at points far from either functional object. This approach was shown to be effective for generating continuous trajectory of the mapped points within a region enclosing a functional object. However, when Alice in $Home_A$ moves from a functional object to another, we often observe discontinuity in the $Alice_{Avatar}$'s trajectory in $Home_B$ at a mid point while it traverses from a functional object to another, i.e., at the border between regions. We revisit this issue in §6.

We evaluated the mapping strategies in terms of average distance errors by comparing the algorithm-generated output with human-labeled results ($N = 2921$) collected in §5.1. We divided the entire home into near-object areas and far-object areas. For each functional object, its near-object area is defined as a circular region whose radius is half of the distance to the nearest other functional object. We separately evaluated the average distance errors for the grids within the near-object areas and those within the far-object areas, which cover 50.1% and 49.9% of entire floor plans, respectively. Table 3 shows the results; for near-object area, the distance error of the object-oriented mapping was significantly

lower than distance-ratio differences ($t = -5.30, p < 0.01$) and cosine similarity ($t = -13.67, p < 0.01$). For far-object area, the distance error of the object-oriented mapping was also significantly lower than distance-ratio differences ($t = -2.63, p = 0.01$) and cosine similarity ($t = -11.75, p < 0.01$). As a result, object-oriented mapping exhibits the lowest distance errors in both areas. The main reason why the near-object error is larger than the far-object error in object-oriented mapping is that participants have different object matching strategy when two floor plans have different functional objects (discussed in §9.3) Accordingly, we adopted the object-oriented mapping to implement the server modules.

### 5.3 Translating the Orientation

The orientation of $Alice_{Avatar}$ also needs to be translated. Like the location mapping, it is not trivial to decide which direction $Alice_{Avatar}$ should be facing, since it is hard to tell what Alice is actually looking at along her line of sight. To the same extent of the object-oriented mapping, we designed the orientation translation in the following way. In $Home_A$, we draw a straight line from Alice toward her orientation and find the nearest functional object from the line. We then calculate the angle between the line and a new line from Alice to the functional object. In $Home_B$, this angle is added to the azimuth of the unit vector from $Alice_{Avatar}$'s translated location towards the matched functional object. The resulting angle is the translated orientation of $Alice_{Avatar}$.

### 5.4 Object Matching in the Two Homes

Every home has different objects. Many of them may be common objects, but some may not be. If we use common objects as functional objects for mapping purposes, it may not always represent a person's equivalent location/orientation well due to the absence of a unique object in either home. To handle such cases of objects existing only at either home, we used fastText [13] to find the most semantically similar object and map to it. The detailed process is as follows. 1) The functional objects of a home are listed up. This step may be replaced with automatic object recognition. 2) We map common objects of those lists. For unmapped objects, we compute cosine similarities between those objects' word vectors, and map each object to the most similar object above a certain threshold.

## 6 LETTING AVATARS BE SELF-MOBILE

Another challenge in developing HomeMeld is driving the avatar to synchronize its location/orientation with those of the person in real-time. The difficulty comes from the fact that the robot is much slower than a human, especially in rotations and accelerations.

A naive way is taking a shortest-distance path, which will result in obstacle-free connected straight lines. We name it *rotate-and-forward* approach, as the avatar has to pre-rotate towards the next waypoint or the destination (provided by the mapping policy in §5), and move forward. This method is certainly slow and the human-perception is even worse, because the avatar stays at the initial location until it completes pre-rotation. This naive approach aggravates the problem to the already-slow avatar robot.

We need a way to optimize the avatar's movements to help it keep up with the human. We devise two optimization techniques, so called *predictive path warping* and *navigation through hyperspace*.

### 6.1 Predictive Path Warping

Synchronizing the arrival time at the destination is an important issue to timely convey one's possible update of activity upon arrival. However, as mentioned in §5, the heterogeneous mapping might cause the path for the $Alice_{Avatar}$ to be much longer than that of Alice and worsen the speed problem of the robot.

We add destination prediction on top of our navigation algorithm. Based on the information of Alice's real-time location and orientation, we attempt to predict Alice's destination. If there is only a single functional object in her moving direction, we can almost safely say that is her destination. If we are confident enough, we find the optimal path from the current $Alice_{Avatar}$'s position to the final destination, rather than to the current functional equivalent location of Alice. This approach will help the avatar arrive at the destination faster because it navigates along the optimized path to the destination rather than the tracing every single functionally equivalent location of Alice, which may get long and crooked.

**Confidence-aware Speed Adaptation:** What if there are multiple functional objects towards Alice's moving direction? For example, $Home_A$ has a couch and a tea table next to it. But at $Home_B$ those objects are quite apart from each other. Suppose it initially predicts the couch to be a slightly more likely destination than the tea table at modest confidence, $Alice_{Avatar}$ moves to the equivalent object in $Home_B$. However as Alice approaches closer, the prediction has been updated to be the tea table. Now, $Alice_{Avatar}$ rapidly changes its moving direction to the new destination in $Home_B$, which may convey a wrong context to Bob and result in delayed arrival. To avoid this, we find the functionally equivalent locations of those two likely destinations in $Home_B$, and find a centroid in between. HomeMeld drives $Alice_{Avatar}$ slowly towards the centroid while the prediction confidence remains low. As soon as the confidence becomes good enough, HomeMeld drives $Alice_{Avatar}$ towards the destination at a higher speed. The reason why HomeMeld drives the robot slowly at low confidence rather than holding it is to reproduce the context that Alice is moving to somewhere right now.

### 6.2 Navigation through Hyperspace

Synchronizing the time of the departure is important to provide an instant cue that the counterpart has started moving for a new activity. It generally arouses the person's attention and may help initiate a new interaction. However, in the *rotate-and-forward* approach, the pre-rotation takes a while due to the robot's limited angular speed, delaying both departure and arrival times.

We highlight that an optimal path is a shortest-time path, not necessarily a shortest-distance path as in the rotate-and-forward. The shortest-time path is a sequence of connected lines and curves of various curvature that are *maneuverable* within the robot's mechanical limits, e.g., maximum acceleration and angular speed, etc.

We devised an algorithm finding a minimum-hop navigation sequence in discretized hyperspace. For brevity, this algorithm is hereinafter referred to as *hyperspace*. We build a 5-dimensional space in which a point describes the robot's state in terms of its 2-dimensional location ($x$, $y$), linear speed ($v$), orientation ($\theta$), and angular speed ($\omega$). We discretize each dimension to pre-load every possible state and feasible transition in between. An edge connecting two 5-D points means that the robot can move from one to the

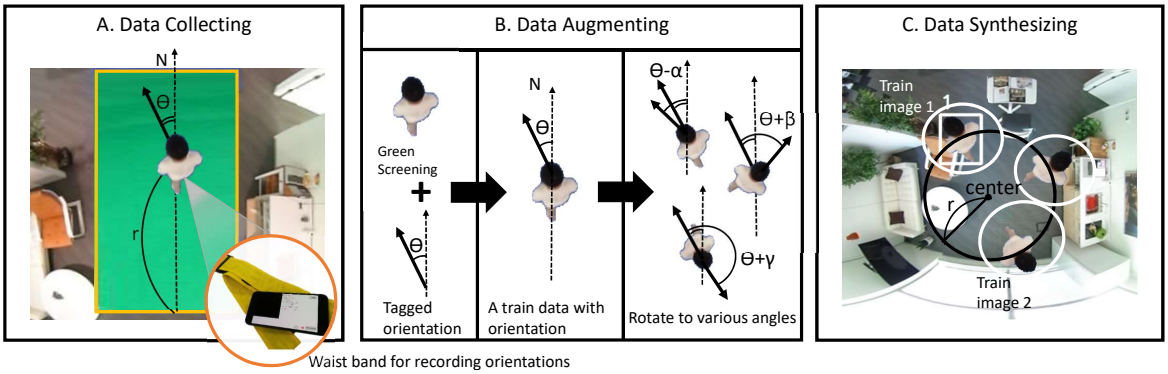



**Figure 5: Data collection and augmentation with green screening, rotation transformation, and background synthesis**

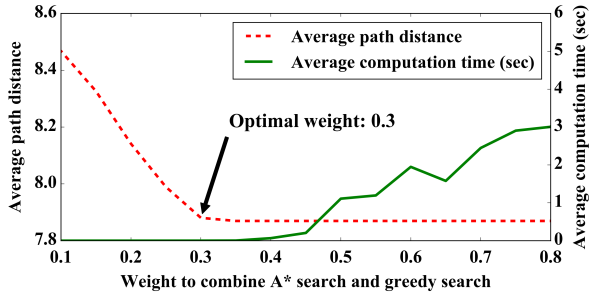



**Figure 6: Optimal weight to combine A*- and greedy-search**

other in a unit time under the robot's mechanical limits. Say a robot can accelerate up to 50 cm/s$^2$ and the unit time is 1 second. For an arbitrary point $P$ at $(x, y, v, \theta, \omega)$=(0 cm, 0 cm, 0 cm/s, 0 rad, 0 rad/s), P has edges to all points that the robot at P is reachable in a unit time by proper acc/deceleration or rotation. For example, P has an edge to (12.5 cm, 0 cm, 25 cm/s, 0 rad, 0 rad/s) as it is reachable in one second by accelerating at 25 cm/s$^2$ and keeping its orientation.

As an initial attempt, we adopted A* search [24] to hyperspace. Unlike conventional A* search, we define the weight of an edge to represent the time cost to travel along it, which is all 1 in hyperspace. With given origin and destination states, the path obtained by hyperspace describes the *navigable shortest-time path*. This path itself embeds the required navigational controls along the path, e.g., acceleration, breaking, or steering, because heading to the next waypoint represents not only a spatial displacement but also a change in navigational parameters.

A* search expands nodes based on $f(n) = h(n) + g(n)$. $h(n)$ denotes the estimated cost from the node $n$ to the goal, and $g(n)$ denotes the cost so far until the node $n$. It finds the optimal path which minimizes $f(n)$, where $n$ is the last node on the path. A* search requires extensive computation because every node with the same $g(n)$ has to be analyzed every time, which is more serious in 5-D space. A* search is equivalent to greedy search if $f(n) = h(n)$. However, greedy search does not guarantee the global optimal path. We adopted the notion of weighted A* [51] to combined those algorithms as $f(n) = h(n) + w * g(n)$, where $0 < w < 1$, to find the path to the goal while reducing the number of nodes to be computed. As shown in Figure 6, we select 0.3 as an optimal $w$ which minimizes both average path distance and computation time.

Once the avatar starts moving according to the navigation controls retrieved from the hyperspace-generated path, the path is recalculated repeatedly using the newly received information. We used 0.1 second for the recalculation interval, which equals to the command update interval in the telepresence robot we used.

# 7 HUMAN DETECTION

## 7.1 Custom Trained CNN Model

**Need for a custom trained model:** We use a ceiling-mounted 180° camera in favor of obstruction-free human detection. Tracking a person from a top view has several benefits than using an eye-level view: 1) no blind area, and 2) tracking the person's location without knowing his height. However, detecting locations and orientations of the person from a top view has not been frequently addressed in existing models, e.g., pre-defined OpenCV classifiers or pre-trained CNN models. This led us to create a custom dataset and model.

**System Implementation:** We implemented our model using a CNN-based real-time object detection framework, YOLO [53] and trained a custom model to detect a person from the top-view. We define 8 classes indicating a person's discretized orientations: north, north-east, east, etc. This model returns a class of the person's orientation and the bounding box coordinates enclosing the detected person. We compute the person's location from the bounding box coordinates. In a top view, a person's feet are always the closest part of the body to the center of camera. Using this property, we infer the person's ground-level location by finding an intersection of the bounding box and a line from the center of FoV to the center of a bounding box. We applied moving average in time domain of person's orientation to smooth the changes of discrete orientations.

## 7.2 Building Training Dataset

CNNs require a huge training dataset. To our best, we could not find a public dataset of people seen from the top view. We also need the orientations labeled on each picture, making the requirements far less common. Thus we built our own dataset as below.

**Green-screening and augmenting:** Figure 5 shows our data collection process. Fortunately, from the top view, a person's appearance is symmetric with respect to the center of FoV. We benefit from this property and a green-screening technique. A participant moves along the 1-D straight line on the green fabric with his orientations labeled automatically using a smartphone's gyroscope sensor for ground truth. We screen participants by a green-screening

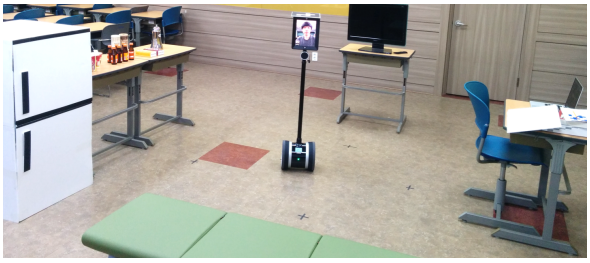


**Figure 7: Laboratory experimental setting**

technique. We then augment our dataset by rotating screened participants around the center of FoV as a pivot point. To balance our training dataset for each orientation class, we rotated screen participants with random angles but at equal frequencies per class.

**Robustness and Generality:** Artificially enlarging the dataset is a common method to reduce overfitting [16, 35]. To attain sufficient robustness and generality, we need numerous data samples of a large variety in terms of people, locations/orientations, and floor plans. We superpose our screened participants with various background images. We collected 85K original images from 14 participants who often changed the clothes. Then we finally generated 665.7K training images on 20 different background images with different furniture arrangements with arbitrary lighting conditions.

## 8 EVALUATION

In this section, we evaluate HomeMeld with respect to 1) system performance in controlled environments and 2) potential usefulness from small-scale exploratory deployments.

### 8.1 System Performance Evaluation

**Experimental setup:** We prototyped HomeMeld using commercial telepresence robots, Double 2 [2]. Its driving hardware is a bi-wheeled self-stabilizing base, on which a 9.7-inch iPad Pro is mounted in an inverted pendulum style. The iPad and the driving hardware is connected via Bluetooth. While a common way to pilot Double 2 is using its standard piloting interfaces on web and mobile platforms, it provides an iOS SDK to enable third-party piloting applications. We developed a custom pilot software to automatically drive Double 2 upon server-side commands and local dead-reckoning. The iOS SDK provides real-time encoder values individually from both wheels, allowing short-distance dead-reckoning.

The server modules are running on our GPU cluster (NVIDIA TITAN X). We mounted Ricoh Theta S cameras [6] on the ceiling, which are connected to a laptop and to the server module.

We conducted extensive measurements to evaluate HomeMeld's human recognition and avatar navigation performance. To represent a casual living space, we set up two experimental rooms and placed four kinds of common furniture, i.e., a refrigerator, a working desk, a couch, and a television. Figure 7 illustrates structures of our rooms. We then defined four types of casual in-home movements: normal, paused, serial, and turned movement (see Figure 8).

To compose the workload, we recruited 4 graduate students in South Korea (age: 22–30). They were asked to perform every movement type once at a random starting point and destination. They repeated the task with various walking speeds, clothes, and rooms. At the same time, we collected a top-view video stream as well as

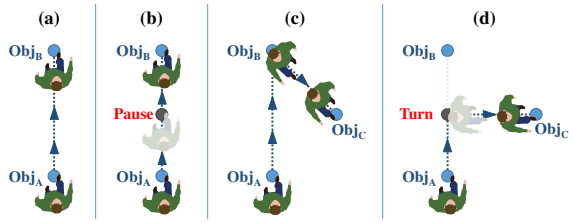



**Figure 8: Four types of casual in-home movements: (a) Normal, (b) Paused, (c) Serial, (d) Turned**

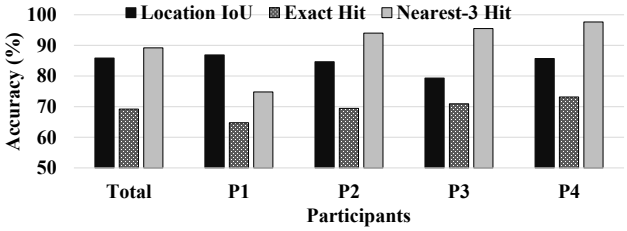



**Figure 9: Per-participant location and orientation accuracy**

the participants' body orientation by using the 180° camera and a smartphone-mounted waistband, respectively. After the collection, we manually labeled their locations for ground truth. We collected a total of 32 movement traces with 2,863 labeled video frames.

**Human recognition accuracy:** We evaluated the accuracy of the recognition model with respect to the person's location and orientation. We used three metrics: 1) detection recall, 2) location accuracy, and 3) orientation accuracy. The detection recall measures whether a system detects a person's presence, regardless of location and orientation. It is a prerequisite for the system to continuously respond to the person's movements at minimal latency. The location accuracy is measured in Intersection over Union (IoU, %) between the recognized region and the actual region. For the orientation accuracy, we used the 8 discrete classes, then used two hit ratio metrics (%): 1) *exact hit* returns 'hit' only when the recognized orientation class matches the actual one, 2) *nearest-3 hit* returns 'hit' when the recognized class falls in the range of [counterclockwise nearest class, ground-truth class, clockwise nearest class].

Overall, the model exhibits 86.1% recall, denoting that our system is practical enough to track a person's presence in real-time. The worst-case, i.e., the longest number of consecutive frames missing detection, was 11 frames, indicating a little delay of 0.9 seconds at 12 fps on average. Our model precisely recognized the person's location and orientation. The average accuracy of location detection was 85.9% and the orientation detection was 69.2% and 89.2% for exact hit and nearest-3 hit, respectively. Figure 9 shows the details.

**Latency in navigating avatars:** We measured the performance of HomeMeld in synchronizing the location/orientation of a person with those of the avatar. Latency of the CNN model was 72ms on our server hardware, which is the time taken to find the location and orientation of the user from a video frame. To measure the prominent human-perceived latency, we took two metrics, *departure latency* and *arrival latency*. Departure latency is the delay until an avatar starts moving after a person departed. Similarly, arrival latency is defined as the delay at arrival. To evaluate the effect of the navigation through *hyperspace* and with *predictive path warping*,

**Table 4: Evaluation of departure latency (sec)**

| Navigation Strategy | Normal | Pause | Serial | Turned |
|---|---|---|---|---|
| Naive | 2.74 | 2.58 | 2.49 | 2.92 |
| Hyperspace-only | 0.59 | 0.55 | 0.43 | 0.44 |
| Path-warping-only | 2.62 | 2.78 | 2.19 | 3.08 |
| HomeMeld | 0.77 | 0.81 | 0.64 | 0.66 |

**Table 5: Evaluation of arrival latency (sec)**

| Navigation Strategy | Normal | Pause | Serial | Turned |
|---|---|---|---|---|
| Naive | 8.67 | 8.03 | 11.57 | 9.14 |
| Hyperspace-only | 7.53 | 7.38 | 8.12 | 7.79 |
| Path-warping-only | 11.10 | 13.44 | 11.43 | 11.46 |
| HomeMeld | 4.87 | 4.16 | 6.33 | 6.11 |

we compared all four combinations: *naive* which navigates through rotate-and-forward without predictive path warping, *hyperspace-only*, *path-warping-only*, and HomeMeld's strategy that uses both techniques. We used the recorded walking traces in §8.1 as the input and observed the corresponding avatar's movement.

Overall, HomeMeld outperformed the other strategies in both latency metrics as shown in Table 4 and 5. Specifically, it was more than five times faster than *naive* and *path-warping-only* in arrival, showing that hyperspace achieves significantly time-efficient navigation. Although the *hyperspace-only* performed the best *departure latency*, the difference between *hyperspace-only* and HomeMeld is almost negligible, i.e., only 0.2 seconds. HomeMeld demonstrated much shorter *arrival latency* by more than a second.

One may raise concerns that the arrival latency of HomeMeld, an average of 5.37 seconds, is too high. However, our observations reveal that much of the latency is due to the final deceleration to precisely stop at the destination. We see that, when decelerating, the location is already close enough to be a meaningful cue to a human. We also measured the 'approach latency' between the person's arrival and the avatar approaching within 30 cm from the object. It turned out to be an average latency of 3.33 seconds.

## 8.2 Small-scale Deployment in the Lab-setting

We conducted a small-scale deployment to understand the feasibility and usefulness of HomeMeld in a real environment. It consisted of three steps. First, we briefly described HomeMeld to the participants. They signed an informed consent form approved by the university IRB. They watched and controlled our robotic avatar for about 10 minutes to mitigate novelty effects. Second, we requested participating couples to naturally interact with different interaction methods, i.e., video chat, telepresence robot, and HomeMeld, for 15 minutes each in a random order. After each method, they were asked to fill in a questionnaire that measures the relative level of social-presence [11], comprising 38 questions; to list a few: '*I often felt as if I was all alone*', '*My behavior was in direct response to the other's behavior*'. To encourage their interaction, we provided a list of 14 casual in-home activities such as '*look at your partner*' and '*solve a cross-word puzzle*'. Third, we conducted a 30-min semi-structured interview to obtain their feedback. The interview themes included perception of each other's presence, comparison with existing tools, and naturalness of avatars' positioning and navigation. Key questions included '*Could you imagine your partner's activity?*', '*How different was the topic of the conversation?*', and so on.

**Participants:** We recruited couples who have lived apart involuntarily for at least one month, from a public research university in South Korea. In total, we recruited eight couples. Six were unmarried couples ($C_1$-$C_6$) and two were married ($C_7$-$C_8$). Their geographic origins included South Korea, India, and Kazakhstan. They have experienced an average distant period of 15 months (min: 1, max: 60). Each participant was rewarded a $30 value gift certificate.

**Results:** The results show that the average social-presence scores in HomeMeld ($M = 5.97, SD = 0.89$) was higher than those of the video chat ($M = 4.33, SD = 0.90$) and the telepresence robot ($M = 5.53, SD = 0.25$). According to the ANOVA test, participants experienced richer social-presence from HomeMeld than video chat ($F = 3.52, p < 0.01$) and telepresence robot ($F = 9.05, p < 0.01$). Note that average social-presence score reported in the original work [11] was 5.55 for face-to-face interaction and 5.39 for teleconferencing, although their results are not directly comparable to us. It shows the potential limitation of our evaluation that the novelty effect to the robots might not be clearly removed and contributed higher scores in telepresence robots and our system.

Overall, the participants were satisfied with using HomeMeld and stated its feasibility and usefulness. All but $C_6$ expressed that the system provided a novel interaction medium while they kept perceiving each other's presence. Interestingly, although the participants knew that the avatar in the room was not the real counterpart, when they saw its existence, watched its movement, and heard voices of the counterpart from it, they could *imagine* that the counterpart was living with them. They added that such co-presence further broadened the spectrum of interaction themes like their current behavior and locations. Meanwhile, $C_6$ complained that they couldn't find each other's face immediately from the front screen of the avatar, which sometimes confused them.

We also observed that most participants liked that they could keep a conversation with little attention which was impossible in voice or video calls. However, $C_2$ explained that they were already attention-free without HomeMeld as they used to have phone calls in speakerphone mode. In terms of the robot's tangibility and mobility, two expressed their concerns about collision while they agreed that such tangibility provided fresh experiences. To further look into it, we asked their opinions about a concept of holographic avatars. Half the participants answered that it would be weird, mainly because they could pass through the counterpart and it would feel like they were facing a ghost. Meanwhile, $C_3$ and $C_6$ were not satisfied with the robot's positioning and navigation behavior. They suffered from the robots' long navigation latency mainly caused by their frequent location changes.

The participants suggested several features to improve HomeMeld, e.g., adjusting the robot's height to the counterpart's, reflecting head motions, and a 'private' mode—turning off their avatars when they do something privately. We leave them as future work.

## 8.3 Exploratory Probe in Real Living Spaces

As an exemplary case study complementing the lab-setting experiment, we conducted an exploratory user study in the wild to collect the users' experiences with HomeMeld in their real living spaces (please see Figure 10). We recruited two adult sisters ($S1$ and $S2$), aged 30 and 28, who had lived together and heavily relied on each

**Figure 10: Exploratory user study in real living spaces**

other under a single parent, but recently separated due to their jobs. We deployed HomeMeld at each home for an hour.

The study was a valuable preview arousing a few facets of user experiences and technical demands likely in real life with HomeMeld with no usage scenario given. We discuss selected findings.

**Watching TV together:** During the study, $S1$ and $S2$ turned on the same TV show for a while. $S1$ complained that the sound coming from the TV at $S2$'s home through the avatar was out of sync with her own TV sound. Since watching TV together is very common for people living together, we believe it is worth resolving the issue. As location of the TV is likely known to HomeMeld, the robot may adopt spatially selective sound capture techniques [19] and suppress the TV sound, while letting the user's voice pass through.

**Holistic smart home sharing:** $S1$ responded that, in the past, when $S2$ was already watching a TV show, she used to sit next to her sister and watched it together. $S1$ suggested that automatically turning on the same TV channel as soon as $S2$ starts watching the TV would be great to have seamless watching-together experiences. $S1$ added that synchronizing lighting systems, watering sound, and door opening of both homes would be also good for a strong sense of living together. For example, turning off the light on the dining area at $Home_A$ is automatically mirrored at the light at the equivalent space of $Home_B$. Clark et al. [17] studied synchronizing audio and lighting between two homes in a room-level. Incorporating this notion with HomeMeld would be a great addition.

**Pre-ready feature:** Our deployment commenced with the participants ready in their respective rooms and the robots staying at initial locations. $S2$ suggested that it would be even more compelling if $S2$ sees her sister's avatar already moving around when $S2$ enters her home after work, rather than seeing the robot initializing in response to her arrival. $S2$ said that she used to come home later than $S1$ when they were living together. It reminds her of past times that $S1$ welcomed her entering home. Initializing HomeMeld a few minutes before the user's arrival would be straightforward by geofencing the user's location in a close proximity to home.

**Tangible avatars:** We observed a few times that a robot bumped into a participant, mainly upon the robot's delayed movement not being able to avoid the participant's newest location. Interestingly, $S1$ responded that the collision was not entirely bad; avoiding the incoming robot was inconvenient, but at the same time, occasional bumping made her feel a strong sense of her sister's real presence.

# 9 DISCUSSION

## 9.1 Design Considerations

**Presence vs. Privacy:** A few participants discussed privacy concerns, as HomeMeld lets a person infer the other person's in-home activities at all times. However, other participants countered that; if they were living together, they are naturally aware of each other's activity and would not consider it a privacy issue. They pointed out that enjoying pseudo living-together experiences means that they should adapt their notion of privacy to what they would have if they live together. For those who want a trade off in the middle, HomeMeld would be set to operate only in a shared living area, such as the living room or the dining room which most families do not consider private when they live together. The other spaces, such as bedrooms or restrooms, may be set off limits to HomeMeld.

**Design alternatives:** We envisioned a robotic avatar which is moving, looking, and acting exactly in the same way as the family member. Another option might be a follow-me avatar always facing the local user; this would allow the user to always see the other family member. However, many participants pointed out that being watched all the time may feel unnatural and make them nervous. A follow-me mode may be an addition in future work.

**Better destination prediction:** We proposed predictive path warping in §6.1. Considering an individual's location history would increase the prediction accuracy. A possible solution may be incorporating a time-variant machine learning algorithm (e.g., HMM) into our current destination prediction.

## 9.2 How HomeMeld Scales

**Towards an entire house:** In this paper, we demonstrated use cases of HomeMeld in select areas of a house, e.g., living room and kitchen. We chose such areas as our primarily deployment space to effectively convey the core concept of HomeMeld, because those are typical shared areas where most interactions between family members likely happen. Eventually, HomeMeld should be able to cover an entire house. To do so, a possible extension would be a two-tier hierarchical mapping approach that begins with room-level mappings and then proceeds with object-level mappings. Upon the user moving to another room, a room-level mapping finds an equivalent room in which the avatar should be located. The following object-level mapping finds a precise in-room location based on functional objects within. HomeMeld could extend with multiple cameras to cover a larger home or rooms across the walls.

**Beyond two homes and two people:** While our current implementation assumes two homes with a single person located at each, we can extend the principles of HomeMeld towards more than two homes and/or more than a single person at a place, as long as the equal number of avatars are affordable. Identifying each avatar in the same place is trivial with different markers. But identifying each person in the same place may pose a challenge. We may extend the model to identify a person. Alternatively, we may leverage physical differences in household members such as height [25]; an individual's height may be opportunistically estimated in our setting by trigonometry on the slanted views when she is walking around peripheral areas. Once identified, HomeMeld would keep tracking her using inter-frame locality or feature matching [38, 55].

### 9.3 Real-world Deployment

**One-on-none, one-on-many correspondence:** For a dishwasher in Home A, suppose no matched object exists in Home B's kitchen, and even an object of a semantically close name is not found. Then HomeMeld may find one that belongs to the same group in the lexical hierarchy [40].There could be more than one matched object in Home B, e.g., Home A has a single couch in the living room whereas Home B has two. A naive method would designate a single default couch. Alternatively, our policy may dynamically choose either couch depending on the current avatar location which we treat as a functional object as well; one couch may give an equivalent location that exhibits more similar person-avatar proximity.

**Obstacle avoidance:** Obstacle avoidance is an important feature for real-world deployment, however, it was outside of the scope of this paper. We focused on how the notion of functional equivalence would facilitate natural interaction between users residing in different homes. For obstacles that mostly remain stationary, a possible approach is to register such an obstacle's location and shape. By removing the edges traversing over such obstacles from our hyperspace algorithm, it will naturally generate a shortest-time path that circumvents the obstacles. Another way is to put additional sensors on the robot and apply sensor-based obstacle avoidance techniques [12, 15, 21], and even utilize deep learning techniques on those sensor data [59, 60]. Determining dangerous area (e.g., wet areas, stairs, pool) may be done by image segmentation [10].

**Smaller face in the small screen:** The face in the screen is seen smaller than the screen when the person is far from the avatar's front camera. One possible solution is to magnify the face when the person is looking at the avatar by a face recognition technique.

**Battery life:** In the deployment at real homes, we observed that $S1$ was stationary for about 75% of the study duration. We measured battery consumption rate of the robot to estimate how long the robot could seamlessly operate. The battery consumption rate was 4.26%/hour and 7.05% when the robot remains stationary and keeps moving, respectively. We derive that HomeMeld could operate for 20.2 hours for a user of similar moving patterns to $S1$, and for 14.2 hours for an extreme user who is always moving around. Assuming daily charging while a user is sleeping or out for work, HomeMeld would be practical for daily use in terms of battery life.

**Living space with limited network bandwidth:** A family member might be assigned in a rural area, e.g., at an observatory or a remote construction site. Such living space may have network bandwidth not as abundant as a home in a developed area does. If the server modules of HomeMeld reside in cloud, we estimate its inbound traffic to be about 559 kB/s mostly due to the live camera feed. The outbound traffic is negligible, i.e., 1095 bytes/s for the command messages to the avatar. Running the human/avatar detectors locally may save much of the inbound traffic, as only the respective coordinates are sent to the cloud. Currently, a modest desktop with a mainstream GPU could execute the CNN model for real-time human detection. A mobile device may suffice in the near future, powered by recent advances in CNN frameworks optimized for mobile platforms [26, 27]. We envision more people with limited network or computing resources may benefit from HomeMeld.

## 10 RELATED WORK

**Lack of situation awareness in remote interaction:** In spite of rapid advance in communication technology, Walther et al. pointed out that the lack of nonverbal cues is the key weakness of remote interaction [61]. The lack of non-verbal cues decreases opportunities to understand a remote other's situation [23]. Brave et al. tried to overcome this issue with tangible interfaces, which create the illusion that distant users are interacting with shared physical objects [14]. Yarosh et al. designed a remote parent-child interaction by augmenting video chatting with a camera-projector system [64].

**Natural remote spontaneous interaction:** While there have been many attempts to investigate remote spontaneous interaction, most of them focus on implying simplified other-related information to the smartphone alert [57], text message [58], video chatting [20, 42, 52], household object [43], and public social agents [56]. Representing the remote other's contexts, including real-time behavior, to facilitate natural spontaneous interaction remains underexplored.

**Indoor localization:** Within a wide range of indoor positioning techniques, passive wireless indoor localization has the advantage of freeing people from wearable devices. Such localization techniques have employed WiFi base stations [65], visible light arrays [37], and a custom FMCW antenna array [7]. They exhibit various trade-offs among localization accuracy, gesture support, and complexity of infrastructure setup. While our approach using CNN-based visual recognition of user location and orientation provides reasonable accuracy with relatively simple setup, the aforementioned techniques may complement HomeMeld in further extensions that capture user contexts at higher fidelity.

## 11 CONCLUSION

We proposed HomeMeld, a zero-hassle self-mobile robotic system serving as a co-present avatar to create a persistent illusion of living together. We explored the practices and challenges for work-separated families to keep their bonds. We developed models and a system to ensure synchronous co-presence experiences with human-perceived spatial equivalence between heterogeneous living spaces. HomeMeld has been evaluated through both controlled experiments and deployments to real participants under work-separation. We envision further emergence of versatile, higher-fidelity co-presence services, and HomeMeld would serve as an early catalyst therein.

## 12 ACKNOWLEDGEMENTS

We thank our shepherd Dr. Jeremy Gummeson and the anonymous reviewers for their valuable comments. We also thank Christopher M. Durham at IBM Research - Austin for his wholehearted support of this research. The corresponding author is Inseok Hwang. This research was partly supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2017R1A2B3010504, 2017M3C4A7065955, 2017M3C4A7066473).

## REFERENCES


[1] Beam. https://suitabletech.com/products/beam. Accessed: December 7, 2017.

[2] Double Robotics. https://www.doublerobotics.com. Accessed: November 30, 2017.

[3] Find minimum of constrained nonlinear multivariable function. https://www.mathworks.com/help/optim/ug/fmincon.html. Accessed: April 9, 2018.

[4] Floorplanner. https://floorplanner.com. Accessed: April 22, 2018.

[5] Living Apart for the Paycheck. New York Times. http://www.nytimes.com/2009/01/04/fashion/04commuter.html. Accessed: December 07, 2017.

[6] Ricoh Theta S. https://theta360.com/en/about/theta/s.html. Accessed: December 07, 2017.

[7] Fadel Adib, Zachary Kabelac, and Dina Katabi. 2015. Multi-Person Localization via RF Body Reflections.. In *NSDI*. 279–292.

[8] Franz Aurenhammer. 1991. Voronoi Diagrams–a Survey of a Fundamental Geometric Data Structure. *ACM Comput. Surv.* 23, 3 (Sept. 1991), 345–405. https://doi.org/10.1145/116873.116880

[9] Mahdi Azmandian, Mark Hancock, Hrvoje Benko, Eyal Ofek, and Andrew D. Wilson. 2016. Haptic Retargeting: Dynamic Repurposing of Passive Haptics for Enhanced Virtual Reality Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1968–1979. https://doi.org/10.1145/2858036.2858226

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (Dec 2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

[11] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *4th annual international workshop on presence*. 1–9.

[12] Joydeep Biswas and Manuela Veloso. 2012. Depth camera based indoor mobile robot localization and navigation. In *2012 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1697–1702. https://doi.org/10.1109/ICRA.2012.6224766

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[14] Scott Brave, Hiroshi Ishii, and Andrew Dahley. 1998. Tangible Interfaces for Remote Collaboration and Communication. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW '98)*. ACM, New York, NY, USA, 169–178. https://doi.org/10.1145/289444.289491

[15] Norlida Buniyamin, W.A.J. Wan Ngah, Nohaidda Sariff, and Zainuddin Mohamad. 2011. A simple local path planning algorithm for autonomous mobile robots. *International journal of systems applications, Engineering & development* 5, 2 (2011), 151–159.

[16] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110

[17] Meghan Clark and Prabal Dutta. 2015. The Haunted House: Networking Smart Homes to Enable Casual Long-Distance Social Interactions. In *Proceedings of the 2015 International Workshop on Internet of Things Towards Applications (IoT-App '15)*. ACM, New York, NY, USA, 23–28. https://doi.org/10.1145/2820975.2820976

[18] Meghan Clark and Prabal Dutta. 2016. Weaving Social Fabric with a Home-to-Home Network. In *CHI '16 Future of Human-Building Interaction Workshop*.

[19] James L Flanagan, Arun C Surendran, and Ea-Ee Jan. 1993. Spatially selective sound capture for speech and audio processing. *Speech Communication* 13, 1-2 (1993), 207–222.

[20] Sean Follmer, Hayes Raffle, Janet Go, and Hiroshi Ishii. 2010. Video Play: Playful Interactions in Video Conferencing for Long-distance Families with Young Children. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 3397–3402. https://doi.org/10.1145/1753846.1753991

[21] Santiago Garrido, Luis Moreno, Mohamed Abderrahim, and Fernando Martin. 2006. Path Planning for Mobile Robot Navigation using Voronoi Diagram and Fast Marching. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2376–2381. https://doi.org/10.1109/IROS.2006.282649

[22] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47, 6 (2014), 2280–2292. https://doi.org/10.1016/j.patcog.2014.01.005

[23] Leo Gugerty, Mick Rakauskas, and Johnell Brooks. 2004. Effects of remote and in-person verbal interactions on verbalization rates and attention to dynamic spatial scenes. *Accident Analysis & Prevention* 36, 6 (2004), 1029–1043. https://doi.org/10.1016/j.aap.2003.12.002

[24] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* 4, 2 (July 1968), 100–107. https://doi.org/10.1109/TSSC.1968.300136

[25] Timothy W. Hnat, Erin Griffiths, Ray Dawson, and Kamin Whitehouse. 2012. Doorjamb: Unobtrusive Room-level Tracking of People in Homes Using Doorway Sensors. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems (SenSys '12)*. ACM, New York, NY, USA, 309–322. https://doi.org/10.1145/2426656.2426687

[26] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[27] Loc N. Huynh, Youngki Lee, and Rajesh Krishna Balan. 2017. DeepMon: Mobile GPU-based Deep Learning Framework for Continuous Vision Applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '17)*. ACM, New York, NY, USA, 82–95. https://doi.org/10.1145/3081333.3081360

[28] Inseok Hwang, Chungkuk Yoo, Chanyou Hwang, Dongsun Yim, Youngki Lee, Chulhong Min, John Kim, and Junehwa Song. 2014. TalkBetter: Family-driven Mobile Intervention Care for Children with Language Delay. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1283–1296. https://doi.org/10.1145/2531602.2531668

[29] Tejinder K. Judge and Carman Neustaedter. 2010. Sharing Conversation and Sharing Life: Video Conferencing in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 655–658. https://doi.org/10.1145/1753326.1753422

[30] Bumsoo Kang, Chulhong Min, Wonjung Kim, Inseok Hwang, Chunjong Park, Seungchul Lee, Sung-Ju Lee, and Junehwa Song. 2017. Zaturi: We Put Together the 25th Hour for You. Create a Book for Your Baby. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1850–1863. https://doi.org/10.1145/2998181.2998186

[31] Joseph 'Jofish' Kaye, Mariah K. Levitt, Jeffrey Nevins, Jessica Golden, and Vanessa Schmidt. 2005. Communicating Intimacy One Bit at a Time. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1529–1532. https://doi.org/10.1145/1056808.1056958

[32] Jina Kim, Young-Woo Park, and Tek-Jin Nam. 2015. BreathingFrame: An Inflatable Frame for Remote Breath Signal Sharing. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '15)*. ACM, New York, NY, USA, 109–112. https://doi.org/10.1145/2677199.2680606

[33] Kangsoo Kim, Divine Maloney, Gerd Bruder, Jeremy N. Bailenson, and Gregory F. Welch. 2017. The effects of virtual human's spatial and behavioral coherence with physical objects on social presence in AR. *Computer Animation and Virtual Worlds* 28, 3-4 (2017), e1771. https://doi.org/10.1002/cav.1771

[34] David S. Kirk, Abigail Sellen, and Xiang Cao. 2010. Home Video Communication: Mediating 'Closeness'. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 135–144. https://doi.org/10.1145/1718918.1718945

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Curran Associates, Inc., 1097–1105.

[36] Min Kyung Lee and Leila Takayama. 2011. "Now, I Have a Body": Uses and Social Norms for Mobile Remote Presence in the Workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/1978942.1978950

[37] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical Human Sensing in the Light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '16)*. ACM, New York, NY, USA, 71–84. https://doi.org/10.1145/2906388.2906401

[38] David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Vol. 2. IEEE, 1150–1157. https://doi.org/10.1109/ICCV.1999.790410

[39] Margaret L. McLaughlin and Michael J. Cody. 1982. Awkward silences: Behavioral antecedents and consequences of the conversational lapse. *Human communication research* 8, 4 (1982), 299–316. https://doi.org/10.1111/j.1468-2958.1982.tb00669.x

[40] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[41] Marvin Minsky. 1980. Telepresence. *Omni* 2, 9 (1980), 45–52.

[42] Paulina L. Modlitba and Christopher Schmandt. 2008. Globetoddler: Designing for Remote Interaction Between Preschoolers and Their Traveling Parents. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08)*. ACM, New York, NY, USA, 3057–3062. https://doi.org/10.1145/1358628.1358807

[43] Elizabeth D. Mynatt, Jim Rowan, Sarah Craighill, and Annie Jacobs. 2001. Digital Family Portraits: Supporting Peace of Mind for Extended Family Members. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 333–340. https://doi.org/10.1145/365024.365126

[44] Carman Neustaedter and Saul Greenberg. 2012. Intimacy in Long-distance Relationships over Video Chat. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 753–762.

https://doi.org/10.1145/2207676.2207785

[45] Carman Neustaedter, Gina Venolia, Jason Procyk, and Daniel Hawkins. 2016. To Beam or Not to Beam: A Study of Remote Telepresence Attendance at an Academic Conference. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 418–431. https://doi.org/10.1145/2818048.2819922

[46] Young-Woo Park, Kyoung-Min Baek, and Tek-Jin Nam. 2013. The Roles of Touch During Phone Conversations: Long-distance Couples' Use of POKE in Their Homes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1679–1688. https://doi.org/10.1145/2470654.2466222

[47] Eric Paulos and John Canny. 1998. PRoP: Personal Roving Presence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 296–303. https://doi.org/10.1145/274644.274686

[48] Elin Rønby Pedersen. 1998. People Presence or Room Activity Supporting Peripheral Awareness over Distance. In *CHI 98 Conference Summary on Human Factors in Computing Systems (CHI '98)*. ACM, New York, NY, USA, 283–284. https://doi.org/10.1145/286498.286763

[49] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1716–1725. https://doi.org/10.1145/2818048.2819965

[50] Matthai Philipose, Kenneth P. Fishkin, Mike Perkowitz, Donald J. Patterson, Dieter Fox, Henry Kautz, and Dirk Hahnel. 2004. Inferring activities from interactions with objects. *IEEE Pervasive Computing* 3, 4 (2004), 50–57. https://doi.org/10.1109/MPRV.2004.7

[51] Ira Pohl. 1970. First results on the effect of error in heuristic search. *Machine Intelligence* 5 (1970), 219–236.

[52] Hayes Raffle, Glenda Revelle, Koichi Mori, Rafael Ballagas, Kyle Buza, Hiroshi Horii, Joseph Kaye, Kristin Cook, Natalie Freed, Janet Go, and Mirjana Spasojevic. 2011. Hello, is Grandma There? Let's Read! StoryVisit: Family Video Chat and Connected e-Books. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1195–1204. https://doi.org/10.1145/1978942.1979121

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. https://doi.org/10.1109/CVPR.2016.91

[54] Lisa M. Register and Tracy B. Henley. 1992. The Phenomenology of Intimacy. *Journal of Social and Personal Relationships* 9, 4 (1992), 467–481. https://doi.org/10.1177/0265407592094001

[55] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision (ICCV)*. IEEE, 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

[56] Jamieson Schulte, Charles Rosenberg, and Sebastian Thrun. 1999. Spontaneous, short-term interaction with mobile robots. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, Vol. 1. IEEE, 658–663. https://doi.org/10.1109/ROBOT.1999.770050

[57] Jaemyung Shin, Bumsoo Kang, Taiwoo Park, Jina Huh, Jinhan Kim, and Junehwa Song. 2016. BeUpright: Posture Correction Using Relational Norm Intervention. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 6040–6052. https://doi.org/10.1145/2858036.2858561

[58] Frank Siegemund. 2002. Spontaneous interaction using mobile phones and short text messages. In *Workshop on Supporting Spontaneous Interaction in Ubiquitous Computing Settings, Ubicomp 2002*.

[59] Lei Tai, Shaohua Li, and Ming Liu. 2016. A deep-network solution towards model-less obstacle avoidance. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2759–2764. https://doi.org/10.1109/IROS.2016.7759428

[60] Lei Tai, Giuseppe Paolo, and Ming Liu. 2017. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 31–36. https://doi.org/10.1109/IROS.2017.8202134

[61] Joseph B. Walther. 1992. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication research* 19, 1 (1992), 52–90. https://doi.org/10.1177/009365092019001003

[62] Jun Wei, Xuan Wang, Roshan Lalintha Peiris, Yongsoon Choi, Xavier Roman Martinez, Remi Tache, Jeffrey Tzu Kwan Valino Koh, Veronica Halupka, and Adrian David Cheok. 2011. CoDine: An Interactive Multi-sensory System for Remote Dining. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/2030112.2030116

[63] Lillian Yang, Carman Neustaedter, and Thecla Schiphorst. 2017. Communicating Through A Telepresence Robot: A Study of Long Distance Relationships. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 3027–3033. https://doi.org/10.1145/3027063.3053240

[64] Svetlana Yarosh, Stephen Cuzzort, Hendrik Müller, and Gregory D. Abowd. 2009. Developing a Media Space for Remote Synchronous Parent-child Interaction. In *Proceedings of the 8th International Conference on Interaction Design and Children (IDC '09)*. ACM, New York, NY, USA, 97–105. https://doi.org/10.1145/1551788.1551806

[65] Moustafa Youssef, Matthew Mah, and Ashok Agrawala. 2007. Challenges: Device-free Passive Localization for Wireless Environments. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking (MobiCom '07)*. ACM, New York, NY, USA, 222–229. https://doi.org/10.1145/1287853.1287880