Visual Privacy Control for Metaverse and the Beyond

Tatsuya Amano, *Osaka University, Suita, 565-0871, Japan*
Teruhiro Mizumoto, *Chiba Institute of Technology, Chiba, 275-0016, Japan*
Srikant Manas Kala and Hirozumi Yamaguchi, *Osaka University, Suita, 565-0871, Japan*
Tomokazu Matsui and Keiichi Yasumoto, *Nara Institute of Science and Technology, Ikoma, 630-0192, Japan*

Metaverse technologies are transforming how distant individuals interact in immersive virtual environments. These technologies, in combination with the latest developments in 3D sensing, will create a hybrid metaverse blending virtual and physical spaces. However, high-resolution 3D sensing poses privacy risks from unrestricted spatial data sharing. It may inadvertently expose the human visual appearance and private objects and spaces in the users' physical environments to the shared metaverse. To mitigate these challenges, this article surveys the visual privacy issues in the era of the metaverse. It then introduces a visual privacy control method enabling privacy protection in the current metaverse and beyond. Our goal is to empower users to control the visual privacy of spatial objects by comprehending their inherent semantics. The design allows users to define their privacy protection levels or, if preferred, delegate this responsibility to an automated control system. We present two use-case scenarios to demonstrate this concept.

METAVERSE AND THE BEYOND—APPROACH TO VIRTUAL AND PHYSICAL SPACE SHARING

The COVID-19 pandemic has increased the need for remote communication, which emerging metaverse technologies facilitate immersive virtual communication. These platforms use advanced VR/AR/XR technologies to create virtual environments that closely mimic physical settings. For example, the high degree of freedom in user actions creates an ultrarealistic experience. Moreover, AR-based communication can expand our physical surroundings into mixed reality environments, breaking the boundaries between actual and virtual spaces. This enables new communication methods in the metaverse era.

Recently, shared extended reality (XR) spaces have garnered significant attention.[1] Advancements in 3-D sensing technologies now allow real-time capture and integration of users' physical environments into these shared spaces using high-resolution 3D data. In this context, we recognize the rise of extended metaverses, where virtual and physical realities merge seamlessly.

While MR/XR overlays limited, instantaneous digital content onto the real world, the metaverse concept focuses more on capturing and streaming entire physical spaces in high-fidelity 3D into collaborative virtual worlds, facilitated by strategically placed sensors. This integration happens in real time, resulting in persistent shared environments. These environments seamlessly blend tangible real-world elements with digital realms. Such an approach enables deeper and more immersive hybrid experiences within the realm of collaborative metaverse applications. Within such a metaverse, participants located remotely could connect through immersive visual streams generated from their local spaces. For instance, family members dispersed geographically could come together in a shared visual space that renders their individual kitchen environments, enabling remote collaborative cooking.

This article discusses the challenges associated with realizing such technologies in the metaverse and beyond.

Digital Object Identifier 10.1109/MPRV.2024.3365989
Date of publication 4 March 2024; date of current version 30 May 2024.

VISUAL PRIVACY CONTROL—FROM UNALTERED TO AVATARS

Privacy preservation in AR/XR systems is a critical concern. Studies such as those by Stephenson et al.[2] and Kim et al.[3] highlight various threats to user privacy, ranging from remote secret theft by external/internal observers to unauthorized access by those with physical device access. Erebus[3] proposed an access control framework for third-party AR applications to prevent intentional and accidental data gathering and transmission. Rajaram et al.[4] focused on access control for sharing AR content, identifying regular other users on the application's as potential privacy threats.

Similarly, we identify "honest-but-curious users" as potential inadvertent privacy violators who, despite adhering to application norms, may engage in unintentional privacy breaches due to curiosity. Consequently, a different set of users, victims, could inadvertently expose sensitive visual details of their physical spaces to others within the shared metaverse. These risks emerge uniquely from the extended meterverse, where curiosity can lead to unintended visual privacy breaches. Recognizing the paramount importance of visual information, we specifically address this challenge through targeted privacy controls that protect against unintended exposure of private visual details between users' interconnected physical-virtual spaces.

Protecting visual privacy within shared spaces is a significant challenge, as discussed by Padilla-López et al.[5] The topic of visual privacy preservation has been prevalent, with several techniques presented in various domains, including AAL.[6] This issue is further amplified when high-resolution or 3-D sensors are used to capture scenes. De et al.[7] explained various components of privacy from the perspective of general security and privacy properties as information that should be protected. From the viewpoints of security and privacy, Wang et al.[8] categorized the information that should be covered when utilizing technologies related to virtual reality. The scope of Wang et al.[8] excludes object information and the decision whether or not to show one's space to others. The discussion of privacy in the traditional metaverse context typically revolves around protecting more abstract personal information and sensor data. In contrast, the privacy requirement-level discussed in this article relate to more specific elements, such as objects and the environment within one's own space.

To thoroughly discuss this issue, we categorize the physical objects in these spaces and analyze their inferred attributes. The presence of individuals and their appearances is of most importance. Sharing virtual spaces from home should be done cautiously to prevent the unintentional exposure of family members to strangers. Alternatively, sharing only the presence of occupants, without revealing their identities, may be preferable for those wishing to maintain privacy at home. Personal belongings, the arrangement and design of rooms, furniture and walls can inadvertently disclose preferences, hobbies, family dynamics, economic status, and other private information.

A simple solution is to replace real individuals with virtual avatars and tangible items with digital counterparts. However, the balance between privacy protection and realism introduces a spectrum of quality levels. Completely virtual substitutes may sacrifice realism, but can significantly reduce data traffic volume and the risk of privacy breaches. The challenge here lies in our inability to quantify privacy, necessitating intuition in striking the right balance between conflicting objectives: maintaining privacy and replicating reality. Subsequently, we will define how privacy can be modeled and evaluated.

MODELING AND EVALUATING PRIVACY IN METAVERSE

In the metaverse that incorporates real-world data, each user's physical environment including his/her visual appearance and private objects is exposed to the shared virtual space. To tackle this issue, we need a model to estimate the privacy preference for each object, including the user's appearance in the environment. In this section, we first model users' privacy preferences in the metaverse, then build and evaluate this model.

Privacy Model

The significant advantage of the metaverse systems is the highly immersive interaction, as they merge the physical space to which they belong. However, sensing and sharing everything to increase immersion raises privacy issues. For example, in an interaction between a young family household with a small boy and a household of his grandma living alone, seeing the mess in the family's room may be acceptable. Still, when a family member attends a remote cooking class, she feels the need to hide as much of the clutter in her space as possible. Moreover, she may not want to expose her face during cooking, but may want to show her face to others when she discusses with others the dishes they cooked. Even in such cases, a nervous person may want to hide his space, while a generous person feels he can expose everything, which means the preference depends on personality.

In other words, users' privacy requirements in such metaverse systems can vary depending on factors such as the type of a shared object, the relationship with the interaction partner, the activities of each user, the personality traits, and the environment of the shared space.

A crucial issue in sharing spaces is the tradeoff between the enhanced sense of immersion that comes from reflecting the entire space and the risk of unwanted places or objects being shared. In the proposed system, participants must set processing options for each object to adjust this tradeoff. If machine learning could be used to predict privacy requirement levels, it would eliminate the need to set processing options for each object, enabling easier interaction. Although there are metrics of privacy concerns such as IUIPC-10,[9] they are not suitable for use in tasks such as whether to show or not to show each object in large numbers.

Therefore, we define the privacy level estimation model as a function pl to quantify users' privacy preferences

$$pl : \mathcal{C} \times \mathcal{O} \times \mathcal{P} \rightarrow \mathcal{L} \quad (1)$$

where $\mathcal{C}$ is the set of contexts, $\mathcal{O}$ is the set of objects, $\mathcal{P}$ is the user's personality traits, and $\mathcal{L}$ is the set of privacy protection levels (such as {minimum, low, medium, high, highest}). Context can primarily be distinguished into activity, environmental, and social contexts. The activity context indicates what the user is doing, which can be acquired by extending conventional activity recognition methods. The environmental context represents the information of the space itself, including the state of the shared space and temporal information. The social context includes the relationship between the user and the person they interact with. $\mathcal{P}$ represents the user's unique personality traits or preferences, which can be defined in various ways. This article aims to realize a model that considers individual differences using indices widely used to measure personality traits. Collecting and combining this context information, personal characteristics, and object information through crowdsourcing, as discussed in the article, can infer the privacy requirement levels.

How to Build a Model

A dataset containing target object information and associated privacy levels is needed for machine learning-based inference. However, all combinations of contexts, objects, and personalities for user privacy levels are expensive and time consuming. Therefore, we devise a strategy of collecting datasets through crowdsourcing and developing a model that is effective for sparse datasets.

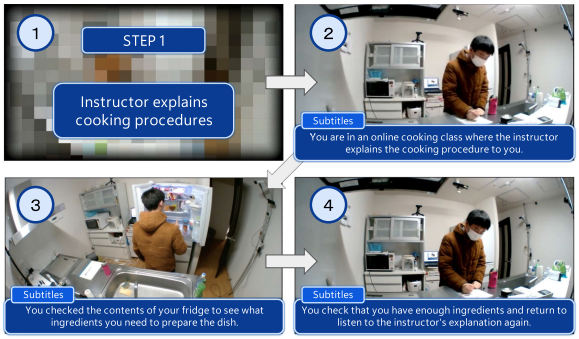

FIGURE 1. Overview of instructional video for crowdsourcing (The original videos were captured from three perspectives in a smart home at Nara Institute of Science and Technology (NAIST) and edited with subtitles to explain the context as needed.).

Collecting Data Through Crowdsourcing

Ideally, the machine learning model requires collecting the privacy requirement levels of all objects that can be shared on the metaverse system. However, collecting levels for every object is not practical regarding data collection costs. Therefore, to cover various domain and resident attributes, we gather privacy levels using crowdsourced experiments. However, crowd-sourced data are sparse, which poses challenges for inference on untrained objects. To address this problem, we describe a proposed method that utilizes the distributed representation of language to supplement these sparse data.

We conduct the questionnaire in two stages. First, personality traits are collected on a crowdsourcing platform. We adopt BIG5 personality traits, which are used in many studies and services to measure human personalities. Next, we ask participants who responded to the first stage of the questionnaire about their privacy awareness when sharing physical space to the metaverse. In this survey, we show instructional videos and images of specific living scenarios and ask about the privacy level of various objects in the scene. Crowdsourcing aims to collect their privacy requirement level while feeling fully immersed. Therefore, as instructional videos for questionnaires, research group members demonstrated activities simulating remote cooking class activities. The overview of the instructional video is shown in Figure 1. The survey question asked participants to rate their privacy requirement levels using a Likert

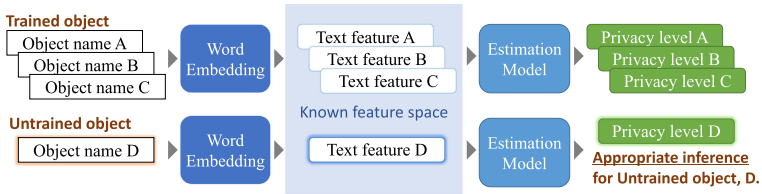

FIGURE 2. Overview of privacy estimation model with word embedding.

scale with five levels of evaluation. As an example scenario, we define a cooking class scenario with the following phases: preparing the food (Step 1), cooking (Step 2), and exchanging impressions of the finished dish (Step 3). Some objects are common to each step, while the remaining objects appear or disappear with each step, in other words, as the cooking scenario progresses. The objects in question totaled about 30, consisting of items that might appear in the context of a remote cooking class. A list of the objects questioned is listed in Table 1.

Developing a Model Effective for Sparse Datasets

An issue with privacy estimation is that when "objects whose privacy level has not been learned in the past" are found, the user must define a new privacy level for each object. To address the problem, we hypothesize that similar privacy levels would be set for similar objects. Under this hypothesis, we apply distributed representations of words to the inference of privacy levels. Words with a distributed representation can express distances in a vector space. Therefore, to effectively infer the privacy level of new objects not associated with a privacy level, we train a model to input the distributed representation of an object instead of the object name and output the privacy level.

Figure 2 shows the privacy level inference method using word embedding. The method converts the object names in the dataset into a distributed representation before training. It maps the untrained object names to the known feature space, which is expected to improve inference accuracy.

TABLE 1. List of object names questioned in the questionnaire.

	Step 1	Step 2	Step 3
Common objects between steps	microwave, kitchen paper, cupboard, kettle, electrical outlet, refrigerator, bag, scissors, tape, memo paper, sink, remote controller, cellophane tape empty pet bottle, empty can, hand soap, camera, rubber band, tablet, your body, your face, your hand		
Step-specific objects	pencil box, measuring cup	measuring cup, knife, used utensils, lid, cut board, dishes	completed meals, IH cooker

Constructing a Model

We construct a machine learning model using a dataset collected through crowdsourcing. The constructed model is a three-layer deep neural network, and the parameters are set empirically. To handle objects with unknown privacy levels, a leave-one-step-out evaluation is performed. The step is the three steps of the cooking class defined in the considered scenario and the objects that appear differently depending on the step. Therefore, the test data under evaluation always contain multiple objects with unknown levels.

Evaluation of The Model

Figure 3 shows the results of the dimensional reduction of the distributed representation of object names to two dimensions. We can group object names based on their distances from each other in the figure. For example, we found a group related to dishes. This group included finished dishes and used cooking utensils. We also found a group related to home appliances, which contained a refrigerator, a microwave oven, and a remote control. These cluster relationships improve the accuracy of privacy-level prediction compared to simply taking object names as identifiers.

As a result, the approach that uses the distributed representation has about 4% better F-measure than that without the distributed representation as shown in Table 2. This evaluation was conducted on a limited dataset. The recognition accuracy would be further improved in a natural environment with more unlearned objects.

INTEGRATION OF PRIVACY CONTROL WITH TRAFFIC MANAGEMENT

We propose a novel metaverse architecture that optimizes point cloud streaming while ensuring visual privacy through semantic communication.[10] This system separates and transmits the semantics and

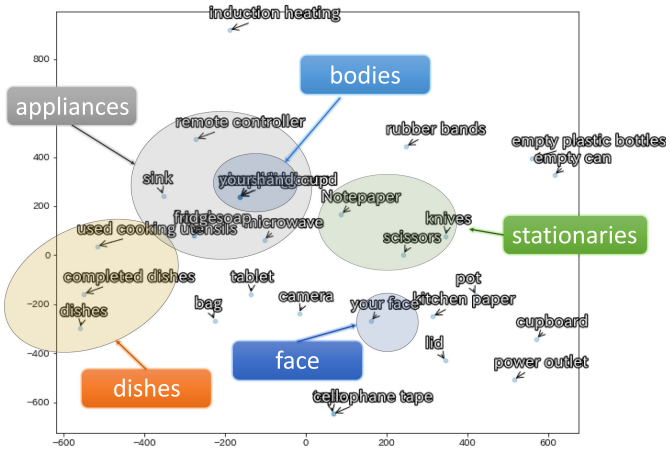

FIGURE 3. Visualization of embedded object names (translated from Japanese). We employ fastText to generate a distributed representation of object names as 300D vectors, which are then compressed into two dimensions using t-SNE.

TABLE 2. Result of leave-One-Step-Out.

Privacy Level	Baseline			Proposed			
	Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**	**Support**
1	0.330	0.452	0.381	0.413	0.511	0.457	6360
2	0.220	0.312	0.258	0.193	0.280	0.229	3547
3	0.627	0.441	0.518	0.679	0.436	0.587	12893
macro avg.	0.392	0.401	0.386	0.429	0.436	0.424	

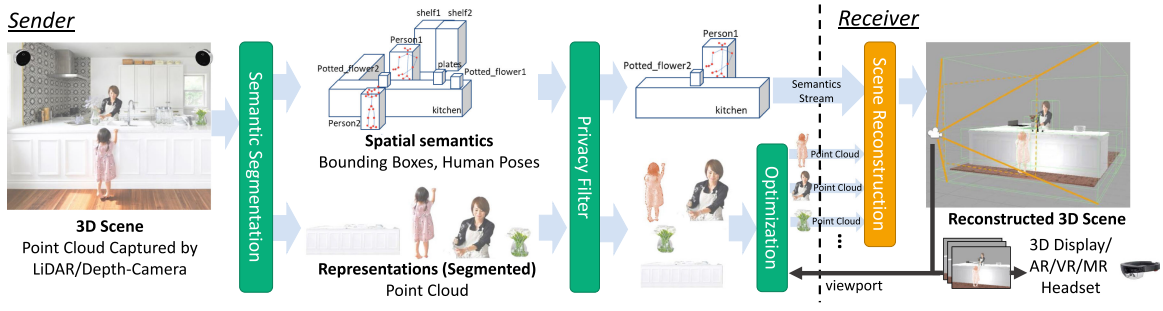

FIGURE 4. The proposed system architecture based on semantic communication (one-way). The point cloud streams are sent via the network to the receiver, who can view the scene through MR/VR headsets or regular displays.

representations of objects in the space, allowing users to control privacy at an object level.

The overview of the architecture is shown in Figure 4. The sender system captures the scene using RGB-depth cameras, creating a detailed 3D point cloud of the environment. This system is grounded in semantic communication, where the semantics of an object, such as its 3D bounding box and type (e.g., human, chair) are separated from their visual representation.

Following capture, the sender applies semantic segmentation to the point cloud. This process generates spatial semantics for each object, which, along with their point cloud representations, are subject to privacy filtering. This filtering is guided by the privacy level estimation model, as detailed in the "Modeling and

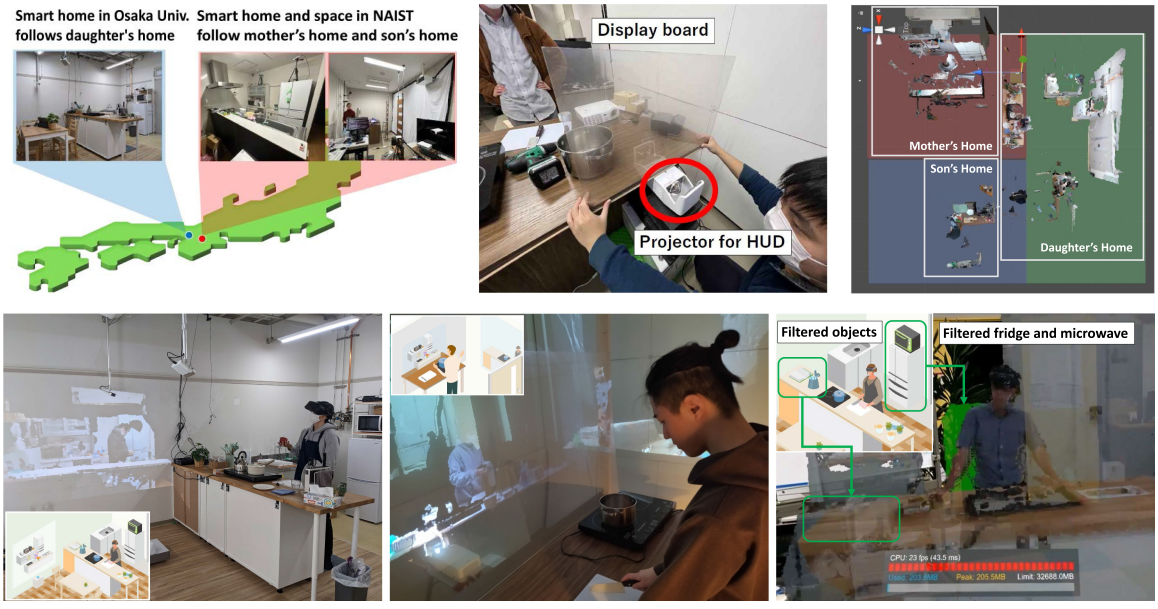

FIGURE 5. Use-case 1. (top left: experimental locations, top center: construction of a HUD prototype, top right: integrated space (bird's eye view), bottom left: projector view, bottom center: HUD view, bottom right: MR view). A notable aspect of this system is the implementation of privacy filters, as demonstrated by the daughter's use of them to manage what her mother can see. She sets up the system to hide clutter on the kitchen countertop and replace the contents of the refrigerator with a virtual green box, based on her mother's preferences. This feature highlights the system's capability for selective sharing and customization of the 3-D space, catering to individual privacy and aesthetic preferences.

Evaluating Privacy in Metaverse" section. For example, a sender may wish to hide the details of people outside the kitchen and show only their presence and behavior, by using raw point clouds without colors. After this filtering, the sender applies two types of point cloud optimization to address network and data constraints. These optimizations include spatial points downsampling and temporal downsampling efficiently manage data transfer rates and match network capacity.

The receiver reconstructs the 3D scene using received semantics and object point clouds. Each semantic information carries a URI to access the sender's corresponding point cloud stream. Point-cloud data are integrated into the scene based on the associated semantics. This setup enables users to explore the scene through various devices by choosing different viewpoints within the environment.

USE CASE AND EXPERIMENTAL SETUP

We implemented the proposed system to demonstrate the two use case scenarios: 1) family members remotely sharing cooking experiences from distributed kitchens, and 2) connecting a home environment and users at a campsite. The system captures real-time 3-D scenes using Azure Kinect. Point clouds are integrated via ROS[a] middleware and transferred over the Internet. Various devices display the integrated metaverse for each scenario.

Use Case 1—Remotely Sharing Cooking

We demonstrate a scenario of a mother living alone and her two children residing separately. Concerned for their mother, the children wish to remotely share cooking experiences with her via the proposed system. The system integrates kitchen spaces and objects into a shared metaverse. The children use AR/VR displays, while the mother uses a home theater projector to view the scene. Aware of the mess in her kitchen, the daughter presets privacy filters to conceal irrelevant objects based on her mother's preferences. The system hides the clutter on the countertop and replaces a green box for the refrigerator.

We set up experimental kitchen spaces at Osaka University and NAIST, shown in Figure 5 (top left). At Osaka University, the setup includes a projector and

[a][Online]. Available: https://wiki.ros.org

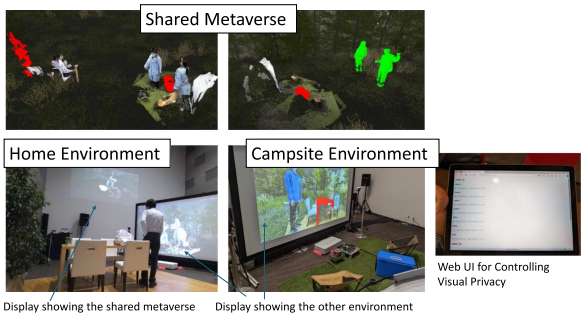

FIGURE 6. Use Case 2—Remotely Sharing Camping. The home environment contained tables, chairs, plants, and shelves. A trash bag was also intentionally placed as a representative object that users may want to hide. The campsite space included camping chairs, a tent, and a cooler box.

HoloLens2, providing a comprehensive view of the integrated kitchens. One of the NAIST locations, designed to represent the mother's home, features a smart kitchen and projector, while the other NAIST site, which depicts the son's environment, is equipped with a head-up display (HUD). These technologies facilitate a seamless and immersive experience in all locations, allowing the family to interact as if they were in the same space.

Use Case 2—Remotely Sharing Camping

We prepared two spaces within a commercial facility near Osaka Station: a home environment and a campsite setup. The setup of the experimental environment is depicted in Figure 6. This use case scenarios elderly users with limited mobility sharing space and interacting with their grandchildren at a remote campsite through the system.

Users could view the shared metaverse scenes through wall-scale screens. They were provided a tablet with a simple web-based interface to control the visibility of individual objects. For each object, users could select to show the original view, display only the geometric shape, or completely remove objects from the shared view.

SYSTEM PERFORMANCE AND USER EXPERIENCE

In our latency evaluation conducted in the first use-case experiment, we compared the original point cloud transmission, requiring about 400 Mb/s bandwidth, with our proposed architecture's optimized streaming. The sender's point cloud typically contained 276,480 points at a 5 FPS. The average PING latency from sender to receiver is $3.8\,\mathrm{ms}$ and Iperf3 testing reveals average downlink bandwidths of 229 Mb/s. The original data transmission significantly increased scene latency. However, our architecture, which adaptively optimizes point cloud data at the object level, effectively reduced the total end-to-end latency by 96%, from $456.3\,\mathrm{ms}$ to $15.0\,\mathrm{ms}$. This demonstrates a substantial improvement in maintaining low latency for real-time virtual environments, even under bandwidth constraints.

In our second use-case experiment, we linked a home environment with a campsite environment, setting the sensor frame rate to 15 FPS. By applying the proposed optimization process, we significantly reduced data traffic: in the home environment, it decreased by 89% from 116 to 12 Mb/s, and in the campsite, by 90% from 400 to 38 Mb/s.

From a privacy protection perspective, our performance evaluation focused on investigating the capabilities of object detection and tracking. This is crucial to ensure that the system can effectively and continuously conceal user-specified objects or maintain designated visual styles. In a specific scenario at the campsite environment with four persons present, we focused on whether our system could continuously track and conceal two persons selected by a user. Over a duration of 300 s, our system achieved an accuracy of 0.91 ID-Recall, 1.0 ID-Precision, and 0.95 ID-F1 in tracking accuracy. False negatives, leading to temporary privacy breaches, were mainly due to occlusions or specific activities like crouching or rapid movements. Although false negative detections were rare in our results, even a single missed detection can lead to temporary privacy breaches by inadvertently revealing individuals who need to be concealed. To mitigate this, our system is designed to default to making all newly detected objects unshared.

DISCUSSIONS

While the proposed system enables configurable visual privacy preservation within extended metaverse environments, we did not empirically evaluate whether personal user attributes could still be inferred from environment depictions even after applying privacy filters. Studies should measure how much private information can still be deduced or predicted from filtered spaces. Adaptive obfuscation techniques could then be developed to balance privacy risks with realism. Accounting for human mobility and concealed identities is also vital. Tracking people over time and space may allow re-identification or defeat concealment attempts.

ACKNOWLEDGMENTS

This work was supported by "Research and Development of Information and Communication Technologies that Contribute to Countermeasures against Infectious Diseases (222-C03)", NICT, JAPAN.

REFERENCES

1. M. Speicher, B. D. Hall, and M. Nebeling, "What is mixed reality?," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–15.
2. S. Stephenson, B. Pal, S. Fan, E. Fernandes, Y. Zhao, and R. Chatterjee, "Sok: Authentication in augmented and virtual reality," in *Proc. IEEE Symp. Secur. Privacy*, 2022, pp. 267–284.
3. Y. Kim, S. Goutam, A. Rahmati, and A. Kaufman, "Erebus: Access control for augmented reality systems," in *Proc. 32nd USENIX Conf. Secur. Symp.*, 2023, pp. 929–946.
4. S. Rajaram, C. Chen, F. Roesner, and M. Nebeling, "Eliciting security & privacy-informed sharing techniques for multi-user augmented reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–17.
5. J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "Visual privacy protection methods: A survey," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4177–4195, 2015.
6. S. Ravi, P. Climent-Perez, and F. Florez-Revuelta, "A review on visual privacy preservation techniques for active and assisted living," *Multimedia Tools Appl.*, pp. 14715–14755, 2023.
7. J. A. De Guzman, K. Thilakarathna, and A. Seneviratne, "Security and privacy approaches in mixed reality: A literature survey," *ACM Comput. Surv.*, vol. 52, no. 6, pp. 1–37, 2019.
8. Y. Wang et al., "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surv. Tut.*, vol. 25, no. 1, pp. 319–352, Firstquarter 2023.
9. N. K. Malhotra, S. S. Kim, and J. Agarwal, "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model," *Inf. Syst. Res.*, vol. 15, no. 4, pp. 336–355, 2004.
10. T. Amano, S. M. Kala, T. Mizumoto, and H. Yamaguchi, "Semantic communication for capacity-aware remote collaboration," in *Proc. 18th Int. Conf. Wireless Mobile Comput., Netw. Commun.*, 2022, pp. 381–386.

TATSUYA AMANO is currently an assistant professor with Osaka University, Suita, 565-0871, Japan. His research interests include spatial computing. Amano received his Ph.D. degree in information and computer sciences from Osaka University, Japan. He is a member of the IEEE. He is the corresponding author of this article. Contact him at t-amano@ist.osaka-u.ac.jp.

TERUHIRO MIZUMOTO is currently an associate professor with Chiba Institute of Technology, Chiba, 275-0016, Japan. His research interests include cyber-physical systems. Mizumoto received his Ph.D. degree from Nara Institute of Science and Technology, Japan. He is a member of IEEE. Contact him at eruhiro.mizumoto@p.chibakoudai.jp.

SRIKANT MANAS KALA is a specially appointed assistant professor with the Mobile Computing Lab, Osaka University, Suita, 565-0871, Japan. His interests lie in the domain of cellular networks, entrepreneurship, and venture capital. Contact him at manaskala@ist.osaka-u.ac.jp.

HIROZUMI YAMAGUCHI is currently a professor with Osaka University, Suita, 565-0871, Japan. His current research interests include pervasive, mobile and smart computing. Yamaguchi received his Ph.D. degree in information and computer sciences from Osaka University, Japan. He is a member of IEEE. Contact him at teruhiro.mizumoto@p.chibakoudai.jp.

TOMOKAZU MATSUI is currently an assistant professor with Nara Institute of Science and Technology, Ikoma, 630-0192, Japan. His current research interests include activity recognition. Matsui received his Ph.D. degree from Nara Institute of Science and Technology, Japan. He is a member of IEEE. Contact him at m.tomokazu@is.naist.jp.

KEIICHI YASUMOTO is currently a professor with the Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, 630-0192, Japan. His research interests include systems and applications in smart homes, smart cities, and smart life. Yasumoto received his Ph.D. degree in information and computer sciences from Osaka University, Japan. Contact him at yasumoto@is.naist.jp.