

CARNEGIE MELLON UNIVERSITY

School of Architecture

College of Fine Arts

Thesis

Submitted in Partial Fulfillment of the requirements for the degree of

Master of Science in Computational Design

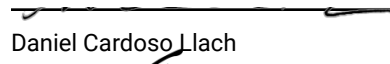
TITLE:

A 4D Diary Co-authored by Human and Machine:
Interactively Visualize, Control, and Understand Data Collection
and ML Computation within Built Environments

AUTHOR:

Zhixin (Felicia) Luo

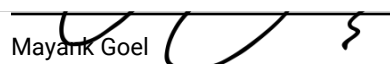
ACCEPTED BY ADVISORY COMMITTEE:


Daniel Cardoso Llach

Principal Advisor

May 8, 2024

DATE


Mayank Goel

Advisor

May 8, 2024

DATE

A 4D Diary Co-authored by Human and Machine

**Interactively Visualize, Control, and Understand Data
Collection and ML Computation within Built Environments**

Zhixin (Felicia) Luo
Master of Science in Computational Design
Carnegie Mellon University
May 2024

Committee:
Daniel Cardoso Llach
Mayank Goel

Table of Contents

Abstract	4
1 Introduction	4
1.1 Proliferation of Smart Cameras	5
1.2 Awareness, Transparency, and Privacy Issues with Camera Surveillance	6
1.3 Limitations, Biases, and Transparency in Machine Learning Models	7
1.4 Research Question	9
1.5 Structure of the Thesis	10
2 Background	12
2.1 AI Literacy and Interactive ML for Transparency and Understanding	12
2.2 Design for User's Privacy	14
2.3 Prior Critical Investigations	18
2.4 Initial Experiment: Self-Surveillance	22
3 Hypothesis: Interactive System Design	29
4 Methods: A 4D Diary of CoDe Lab	31
4.1 Data Representation	31
4.1.1 Representing Human Activity	32
4.1.2 Representing Built Environment	35
4.2 Conventional System Setup	37
4.3 Interactive System Implementation	39
4.3.1 Prototype	41
4.4 User Study	44
5 Results	46
6 Discussion	50
6.1 Conclusions	50
6.2 Limitations	51
6.3 Future Works	52

Bibliography	54
Appendix	58
User Study Procedures	58
Interview Questions	60

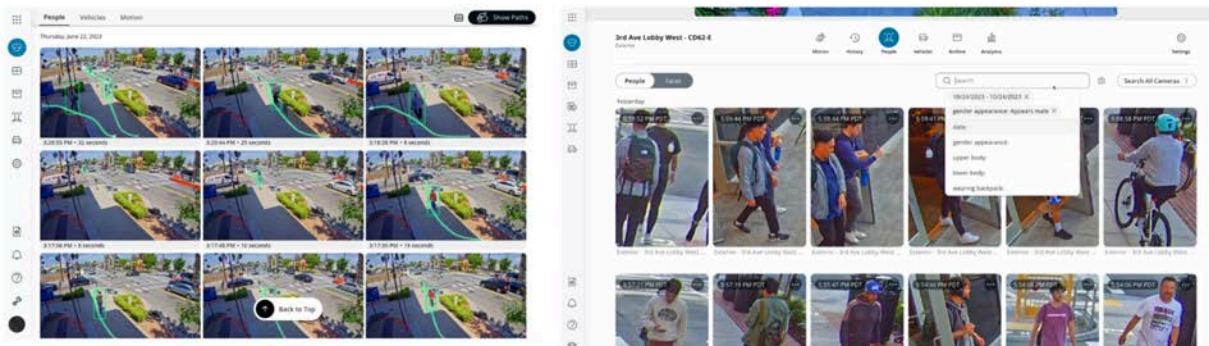
Abstract

Ubiquitous data acquisition and machine learning (ML) systems, in particular, smart cameras, have proliferated throughout all kinds of built environments for surveillance or profit-oriented analysis. However, building occupants are rarely consciously aware of their presence and have no access to the data or the one-sided analysis. Transparency and privacy concerns may be aggravated when ML inferences extract more information from the collected data. Critically investigating ML-powered smart cameras, I first performed a self-surveillance experiment documenting my personal life. It generated a vivid 4D diary that captured lively details and recreated memories through a combination of machine recognitions and personal diaries. Recognizing the benefit of empowering those whose data is collected with active inputs, I proposed to transform the conventional unidirectional data acquisition and ML computation system into a bi-directional interactive one. Presented as a 4D diary co-authored by humans and machines, I designed and prototyped a digital platform featuring interactive visualization, control, and understanding. Through user studies and interviews, I explored the interaction dynamics between occupants and the system, exploring how the prototype mitigates transparency and privacy concerns. This research offers alternatives to the dominant paradigm of passive data acquisition and computation, empowers occupants with active voices and controls, and reintroduces their agency that can be overshadowed by automated and opaque processes. It demonstrates a methodology guiding technology development toward transparency and empathy.

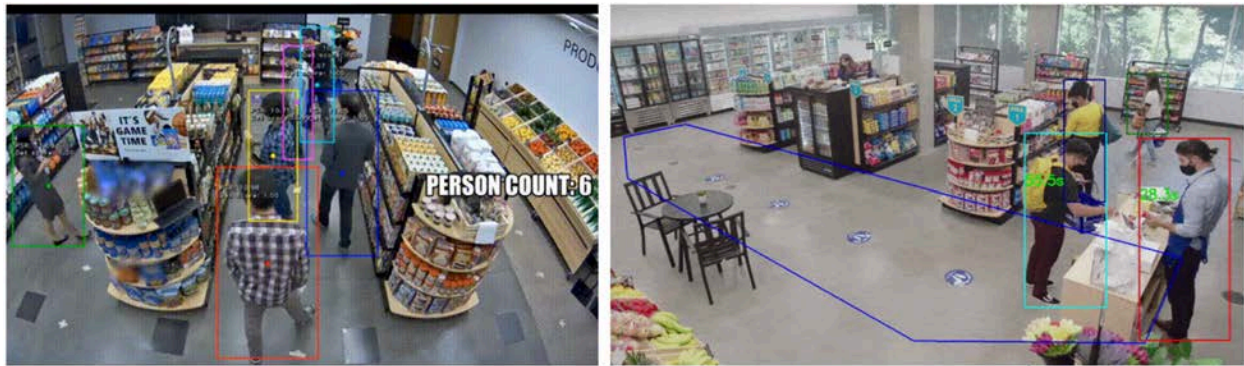
1 Introduction

1.1 Proliferation of Smart Cameras

The term “ubiquitous computing” was first introduced by Mark Weiser [40] in the early 1990s, when he envisioned computers would be embedded in everyday objects and tied with a network, to improve efficiency and smooth workflow. The industry has been developing towards this aspiration. More and more “smart” objects like thermostats, speakers, cameras, and even refrigerators and mugs, go into the market. They have a certain level of intelligence so that they can automatically respond to changes in the environment or human demands. However, they come with a byproduct of ubiquitous data acquisition, the continuous and pervasive collection of data from various sources and sensors embedded in the built environment. And many of them have embedded machine learning systems to respond to versatile demands, thus the “smartness”. One of the most common smart objects is smart cameras, which have proliferated throughout all kinds of built environments. They may be used for security surveillance on urban streets and public buildings. For example, the Verkada camera [47] has a motion search feature, that can find people who have appeared along the specified movement path. Another feature is locating humans or vehicles with specified attributes, like color, gender, etc. Homes are also colonized by smart cameras. For security purposes, people install cameras and get notified when unwanted visitors are in the garage or backyard. In addition to surveillance, the captured videos can also be used for profit-oriented analysis. Microsoft Azure AI provides a spatial analysis service [10] to help retail owners count customer traffic and analyze popular products so that they can arrange store layouts and stock inventories accordingly.



Motion search and incident investigation for property management (Image source: Verkada, 2024 [47])



Spatial analysis by Microsoft Azure AI for retail store (Image source: Microsoft, 2024 [10])

1.2 Awareness, Transparency, and Privacy Issues with Camera Surveillance

Despite the legitimate and reasonable motivation, pervasive camera surveillance triggers concerns about privacy and transparency regarding those who are being watched. Often associated with the image of "Big Brother Is Watching You" established by social critic George Orwell in his novel "1984" [25], numerous discourses came up about privacy invasion and the centralized collection and processing of information by a singular, usually governmental entity, concerning the unconsented observation and recording of individuals' activities. In 1999, sociologists Norris and Armstrong examined the extensive deployment of CCTV systems in the UK and its societal implications [24]. Their observations on surveillance camera operators suggest a highly subjective and discriminative nature in choosing the surveillance target, whilst the watched remain totally unaware. They raise concerns about privacy erosion, behavior modification, and the potential for increased paranoia and distrust in society. They also examine the power dynamics inherent in surveillance, where those controlling the systems wield significant power over those monitored. In contemporary building surveillance, only a privileged few have access to the video footage: data analytics, building owners, and management teams. Occasionally, architects use them for preliminary site analysis before renovation projects, or post-occupancy evaluation to understand space usage.

American AI researcher and humanities scholar Phillip Agre's renowned essay on technology, "Surveillance and Capture," delves into the capture model of privacy alongside the surveillance model. It discusses scenarios where users willingly consent to data capture in various applications to tailor their experiences [2]. These fragmented data points, when amalgamated, can form detailed profiles of individuals' behaviors and preferences. In this capture model, the more intricate the data and the greater the computational analysis, the better the customization fitting the individual's preferences can be achieved. With advancements in technology, particularly machine learning algorithms, details of a person's life, such as their preferred colors and daily routines, can now be deduced solely from camera data. This raises questions about the appropriate level of computational analysis with collected data, aiming for positive outcomes without encroaching on individuals' privacy. Moreover, there's a high risk of data leakage and misuse.

Agre further contends against automatic face recognition in public spaces [1], highlighting the pervasive tracking of individuals and the potential combination of face recognition data with other technologies, leading to heightened privacy infringements. Although technological advancements have reduced errors such as false positives since Agre's writing, concerns about privacy invasion persist, especially regarding identifying individuals in crowds.

1.3 Limitations, Biases, and Transparency in Machine Learning Models

How do smart cameras know a person with a backpack has appeared in the video? Behind the scenes are machine learning (ML) algorithms in computer vision. From video or image inputs, they can classify them into different categories, detect and classify objects, or even generate a text description of the input. So, to solve the previous question, one possible way is to use a classification algorithm to predict True or False if a person with a backpack appears in the input video. Another approach is to use an object detection algorithm to detect all objects within the input video frame. If the algorithm detects both person and backpack in the same frame and their 2D location is close enough, we know that the target scenario occurs. Adopting an ML

algorithm for a specific task follows a general pipeline. Initially, a substantial amount of labeled data is gathered, forming the basis for training the ML model, often employing neural networks. These networks are trained to correctly classify, detect, or generate outputs based on the provided data. Engineers then fine-tune the model parameters to optimize its performance on the collected dataset.

However, even though machine learning algorithms are designed to automate the task on massive datasets objectively, the truth is that they are biased, and they often fail. Studies recognize biases embedded in the training dataset. Conferences are held to target these concerns. For example, in the ACM Conference on Fairness, Accountability, and Transparency, Buolamwini points out that the accuracy disparity between gender and race in gender classification algorithms originated from the training datasets that unproportionally include more lighter-skinned subjects [5]. Artists are also in an effort to acknowledge the limitations of computer vision algorithms. Trevor Paglen composes a string performance accompanied by face detection and classification algorithms predicting the performer's age, gender, emotion, and activity [26]. Audiences watch clearly how the algorithm captures the facial structure of the violinist almost precisely, but fails to understand who she is and what she is doing.



Left: Gender and skin color composition in facial recognition benchmarking datasets (image source: Buolamwini and Gebru, 2018 [5]);

Right: facial recognition on a violinist in Image Operations (image source: Paglen, 2020 [26])

Moreover, machine learning models, especially neural networks, are often perceived as “black boxes” because their inner mechanisms are hard to comprehend, even for ML professionals. The general public without ML expertise further feels mysterious and unable to understand the logic behind the predictions made by ML models. Such a phenomenon is widely recognized by the research community. The discourse extends into the broader field of artificial intelligence (AI), where transparency is a pivotal concern.

The challenge of black-boxiness in AI is compounded by the intricate and often non-linear interdependencies of model variables and layers, which make it hard to trace and interpret how input data is transformed into output. In 2016, a seminal article, “Can we open the black box of AI?” published in Nature, calls for demystifying these black-box models by developing new tools for more interpretable machine learning processes [6]. Since then, research communities of multiple disciplines, such as technical and social sciences, have been paying more attention to the interpretability, explainability, and transparency of AI algorithms. Transparency is crucial for widely adopting AI algorithms to various application sectors in society. Without transparency, it becomes challenging to ensure fairness and to hold systems accountable for errors or biased outcomes. If we do not understand how AI reaches their decision, it would be difficult to trust them [9].

1.4 Research Question

Motivated by the aforementioned concerns on both the data acquisition and further ML computation performed on the collected data, this thesis seeks to investigate the following question:

How could we design for transparency and occupants' privacy in data acquisition and ML systems in built environments?

To be specific, I focus on the occupants of the built environments. In conventional data acquisition and ML systems, at least two entities are involved. One is the building owners or management who install the system and use the collected and inferred data to inform decision-making, who I perceive as the “user” of the data system. The other is the occupants whose data is collected from, but may or may not be actively involved in the processing afterward, therefore, not “using” the system. This study aims to address the transparency and privacy concerns that the occupant might have.

1.5 Structure of the Thesis

This thesis consists of six chapters. Chapter 1 Introduction starts with a general concern on ubiquitous data acquisition and ML computing systems, and limits the scope of this thesis to a subset of smart cameras. It highlights specific concerns about privacy and transparency in camera surveillance and discusses biases and limitations inherent in the backend ML algorithms. These identified issues led to the research question addressing awareness, transparency, and privacy concerns of ubiquitous data acquisition and ML systems in built environments.

Chapter 2 Background contextualizes this study within the greater HCI research community that has been addressing the transparency issue of data acquisition and ML computation.

Explainable AI, AI literacy, and Interactive ML concepts are studied to unveil the black boxes of AI/ML systems and promote understanding and trust. Storyboards, technical prototypes, and user studies have been used to understand user's privacy concerns and design software systems for that. It references several artistic investigations that critically approach data acquisition and ML computation to exemplify the methods adopted in this study. An initial experiment was introduced here as the first step of my own investigation. It was a self-surveillance experiment exploring the boundaries in machine's understanding of human activities. The outcome was a 4D diary that vividly captured the qualitative and quantitative traces of my life while revealing the capabilities and limitations of the data system.

The first experiment leads to Chapter 3 discussing my methods, an interactive system designed to transform the conventional passive data acquisition and ML computation pipeline with interactive visualization, control, and understanding, empowering occupants with active voice and control over the entire system.

Chapter 4 illustrates how I deployed this conceptual design in a semi-public experiment. It first documents and justifies choices of the data representation and ML algorithms. It discusses the raw and semantic representation of human activities and the choice of point cloud representation of the built environment. Then the chapter illustrates the conventional passive data acquisition and ML computation system setup, and my design of an interactive system to empower occupants with active input and control. This system is realized in a digital prototype

that occupants can interactively visualize, control, and understand the data and ML computation. It describes the user study and interviews used to validate the prototype.

Chapter 5 Results presents qualitative and quantitative findings from the user study. It depicts the dynamic between the occupants and their collected and inferred data, and discusses how this interactive prototype promoted transparency and user understanding of the data system, as well as drawbacks noticed during the interactions.

Lastly, Chapter 6 Discussion concludes the contributions, limitations, and future works of this thesis. This study explored the boundaries and indications of SOTA ML-powered human activity recognition systems, designed and prototyped an interactive system to convert the conventional passive system for user visualization, control, and understanding. The prototype could be further developed to increase user control freedom and support real-time capture and visualization. It has established a framework that could act as the visualization and control hub of the local IoT network, and provide functional user control of the data acquisition and processing. Outlooking the future, this chapter concludes with promising directions for advancing technological development from a human-centered stance.

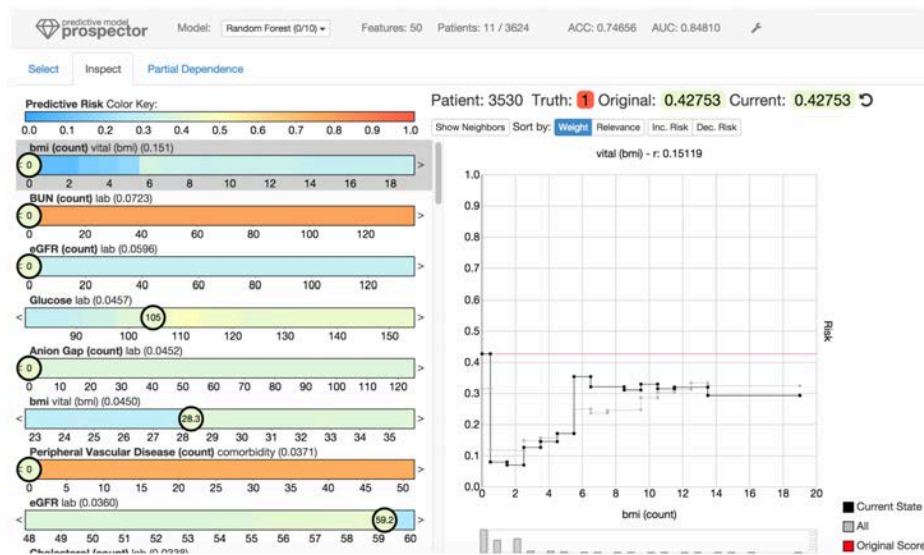
2 Background

2.1 AI Literacy and Interactive ML for Transparency and Understanding

Multi-disciplinary research communities have been exploring to mitigate the negative effects of ML's limitations and biases. The human-computer interaction (HCI) researchers more often refer to the mother field of ML, artificial intelligence (AI), and discuss its ethical use. Understanding the internal mechanics is crucial to ensure the reliable use of AI systems, especially for complex deep learning (DL) models. The terms, interpretability, explainability, and transparency of AI algorithms, are often used interchangeably or discussed together. Previous studies [11] point out that understanding the AI decision-making process includes three parts: understanding the task, knowing how to evaluate the performances, and the experience used to train the model. Explainable AI (XAI) [4] has been explored to provide new ML techniques that are human-understandable so that human can have trust in the AI agents and be able to effectively manage them [12]. American human-centered AI researchers Long and Magerko propose to expand the audiences to the general public without specialized knowledge so that anyone can interact and evaluate AI with critical lenses, and for designers to design AI-powered systems to promote understanding of AI, coined in the term "AI literacy" [19]. They discuss five overarching themes regarding the definition, capability, working mechanism, usage, and perception of AI. Long also uses public interactive installations to promote awareness of AI systems. In Lumin AI [20], participants improvise dance with an AI partner shown on screen. User testing reports that participants can use interaction to differentiate AI who just mimic their posture from those who improvise, and demonstrates the positive influence of interacting with AI in real-time and involving multiple participants. Participants can not only enjoy the fun and engaging interaction, but also discover the capabilities and limitations of the AI by trying different actions. Design researcher Nicenboim, who focuses on more-than-human design in the context of AI, extends the explanations to the shared understanding of AI [23], highlighting the necessity to expose also possible failures, multiple agencies, systems, and biases. They provoke consideration of what's explained that need not be technical.

Philosophy scholar von Eschenbach proposes that recognizing the socio-technical system in which the AI operates may be more effective in establishing trust[9]. This method emphasizes improving the trustworthiness of the broader systems in which AI technologies are integrated, so that trust in the AI itself could also increase. Rather than solely attempting to make AI transparent, which can be inherently difficult, this approach advocates for the development of robust, accountable systems that govern the application of AI, thus making them worthy of trust.

Interactive machine learning has been studied to learn continuously and adapt to new data and feedback in real time. It often involves a "human-in-the-loop" approach, where users provide feedback on the model's performance [3]. This feedback can be used to influence the ML algorithm by finetuning the model and adjusting its parameters. In an example of helping data scientists improve predictive models, researchers find that an interactive visual analytics system showing partial dependence allows the user to understand how and why the model made certain predictions. The interactive system also supports adjusting parameters to see how the prediction changes, which is proven to provide insights into the predictive model.



Interactive visual analytics system, Prospector, showing partial dependence to help data scientists understand and improve predictive models. (Image source: Krause et al., 2016 [16])

2.2 Design for User's Privacy

HCI Researchers and software developers are continually seeking ways to balance functionality with privacy, ensuring users can enjoy the benefits of smart technology without compromising their personal data. Through surveys, interviews, and technical developments, this chapter delves into the current state and future possibilities of privacy design in smart devices.

Jin et al. survey and interview users to study potential privacy-protective measures on smart home devices, including smart cameras, voice assistants, TVs, etc. Physical blocking layers turn out to be the most prevalent approach currently used. The authors propose several privacy protective concepts and identify the privacy diagnostics to be the most favored by interviewees. Data collection monitor is the second most globally preferred for its centralized visualization of what is being captured but is short on control capability. Privacy mirror is strong on the control, but is doubted whether no data is collected as advertised.



Storyboard for privacy diagnostics. (Image source: Jin et al., 2022 [14])

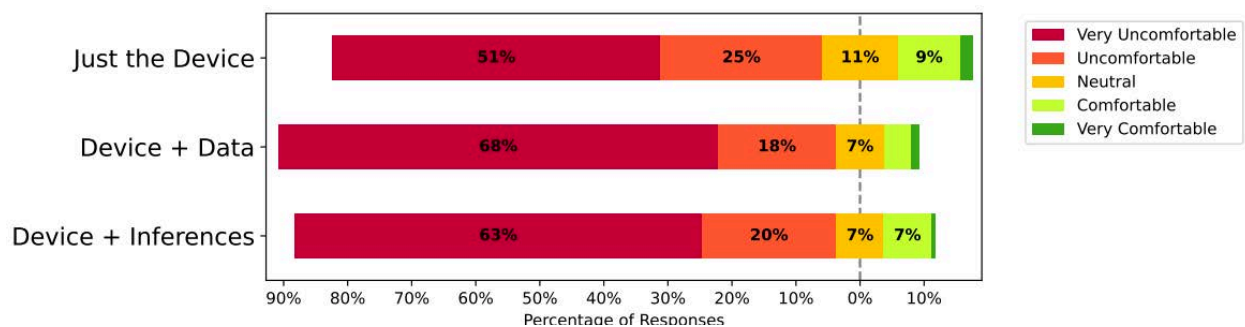


Storyboard for data collection monitor. (Image source: Jin et al., 2022 [14])



Storyboard for privacy mirror. (Image source: Jin et al., 2022 [14])

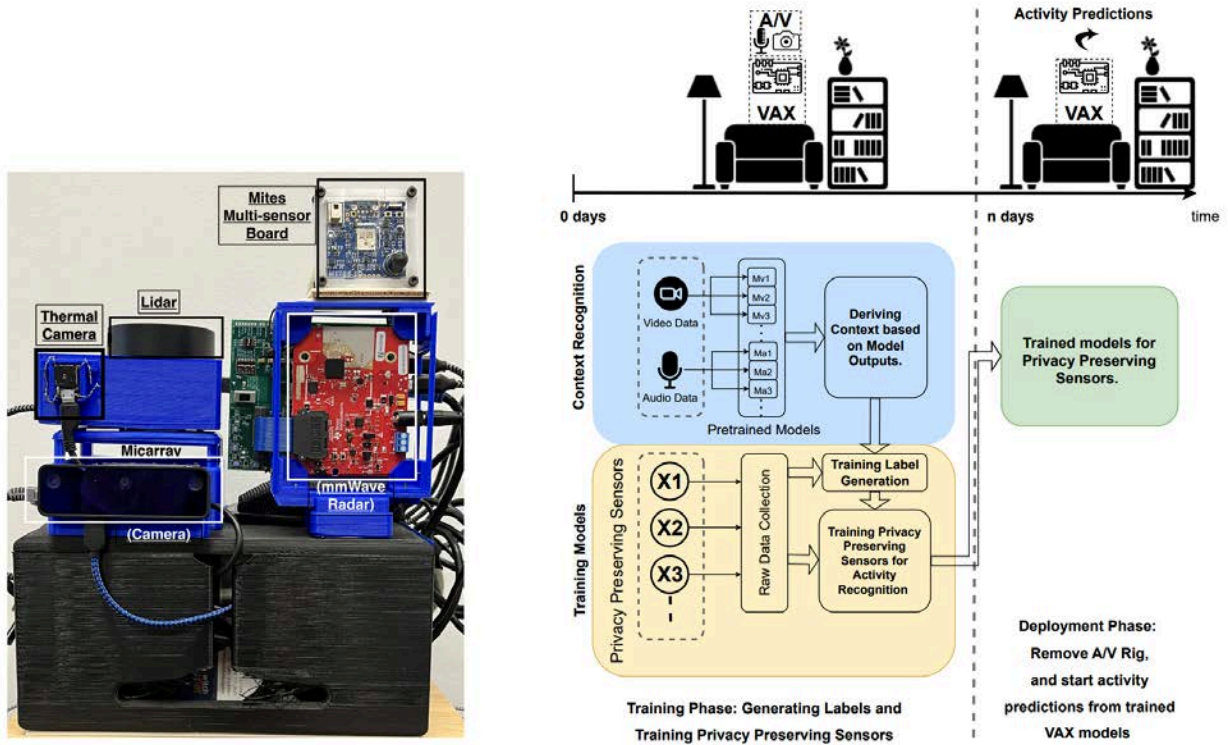
Another survey compares users' privacy perceptions of cameras and radio-frequency (RF) sensors[34]. Cameras have traditionally been perceived as more privacy-invasive because they capture personal-identifiable information, such as people's faces. But RF sensors are anonymous. However, this survey finds that the data itself and the inference that can be made from that data are both significant factors in privacy perception. Users who gain full control over the sensor design and data policies become increasingly comfortable using the device, even for privacy-invasive ones like cameras.



User comfort level drops when shown the device+data or device+inference of a camera. (Image source: Singh et al., 2024 [34])

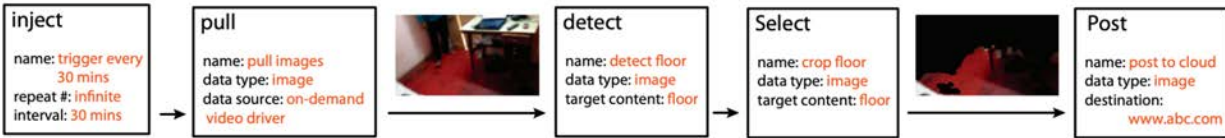
Audio data is also considered more privacy-invasive as the human speech sound can be used to identify specific people. Recognizing the intimacy and detail level inherent in audio and video data, VAX [28] approaches from the machine learning perspective. It proposes to use a pretrained deep learning model taking video and audio inputs and use its results to train other privacy-sensitive sensors, such as Doppler radars that capture the speeds and directions of surrounding movements, Inertial measurement units (IMUs) that include accelerometers and

gyroscopes, and thermal cameras. After the training is done, the activity recognition can be done using only the privacy-sensitive sensors, and no more video and audio data is needed.

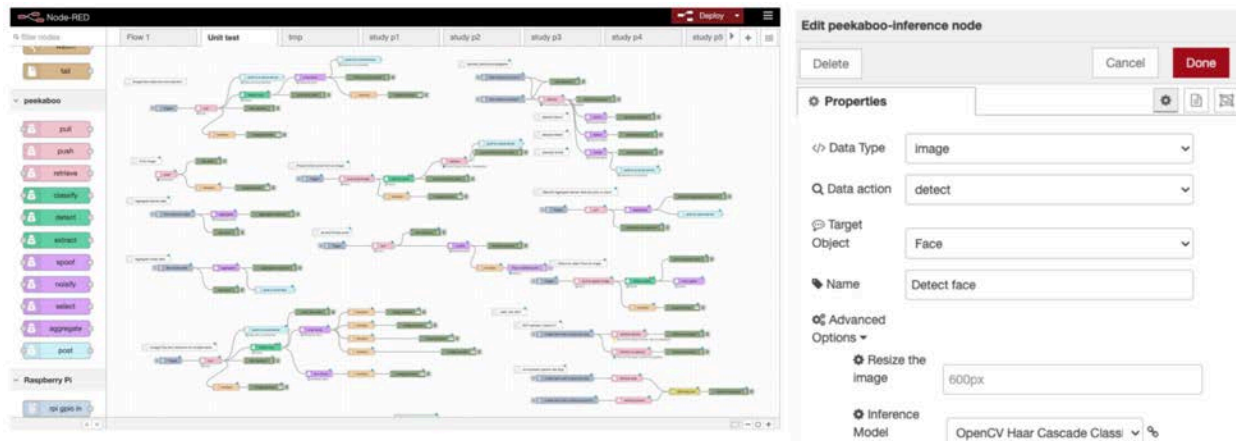


Left: VAX sensors; Right: VAX ML architecture. (Image source: Patidar et al., 2023 [28])

Peekaboo [15] recognizes similar concerns regarding privacy-invasive data acquisition and computation, specifically in a smart home environment. It establishes a software architecture for developing the software stack of smart home devices. Software developers need to explicitly define the granularity, condition, and destination of data processing using a library of operators in a central hub, and leave as much as possible data processing locally. ML inference running locally can derive privacy-preserving data for other computations on the cloud. Therefore, a lot of personal information data will not be sent to the outside internet, greatly mitigating privacy leaks and data misuse. In this paradigm, users can also gain more transparency on the data pipeline and be assured that only the specified data will be sent to outside Internet.



Peekaboo data flow for a water leak detection application. Only necessary data containing as less private information as possible will be transmitted to the cloud. (Image source: Jin et al., 2022 [15])



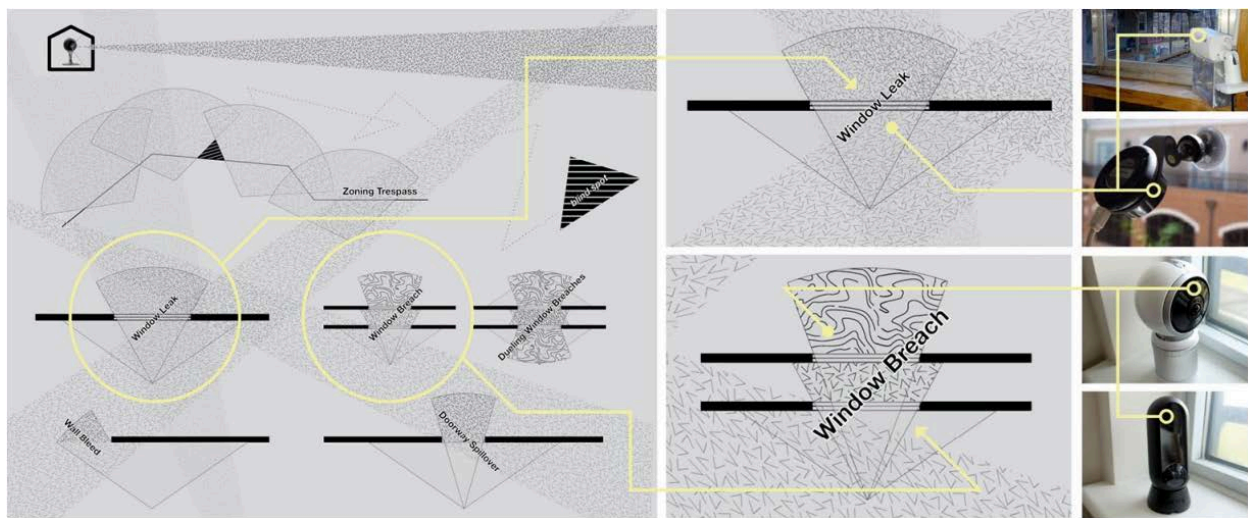
Peekaboo hub interface to configure data flow and operator (Image source: Jin et al., 2022 [15])

Furthermore, software system developers Zhang et al. recognize the two entities involved in the data acquisition by IoT devices, the owner of the device and the building occupant whose data is collected from. The occupants have sensitive data captured but no control over how the data is managed and used. Therefore, they propose IoT Ephemeral Ownership (TEO) to allow occupants to claim co-ownership over the IoT device for a limited time [43]. The device owner maintains the ability to decide whether to allow for co-ownership, but will not have access to the private data once the device is claimed by the occupants. In doing so, the occupants hold control over their own data, drastically alleviating the privacy concern.

2.3 Prior Critical Investigations

Precedent works have approached the topic of camera surveillance and ML systems with criticality. In the domestic setting, human-centered designer and researcher James Pierce inquires into smart home cameras and investigates three types of loopholes [29]. First the data is prone to leakage, abuse, and harming those who the data is collected from. Second,

“hole-and-corner” applications that are not acknowledged with potential harm to the user, still collect data and interact with the user. Lastly, “foot-in-the-door” products subtly prepare users for previous unacceptable features under the guise of normal functioning affordances. Through a series of designs and explorations using creepiness as a provocation, he investigates privacy and security issues. One example in which he speculates the future of smart home cameras is that they capture detailed personal identity, activity, and movements. They are installed in private homes but are used to cross social boundaries. For example, installing a smart camera by the window facing outside can peak into neighbors’ lives.



“Towards an architectural taxonomy of leaky camera sensor field”. (Image source: Pierce, 2019 [29])

Artist and engineer Robert Twomey explores the poetic intersections of human and machine perception. Through a self-surveillance of his own home, he investigates the promise of smart home technologies [39]. Within a functional smart home system, cameras capture his bodily movement, sensors show his interaction with smart objects, and microphones record his language. Jointly, the intimate life is portrayed by machines and numbers. Metaphoring the home as a machine for living, he experiments with the cohabitation of humans and machines. The machine is repositioned as a vehicle for self-examination and critical introspection.

Adam Harvey, similarly, is motivated to reclaim privacy from the increasing visual data acquisition and designs facial covers and camouflage to hide from computer vision algorithms [13]. Also in 2010, he designed a facial camouflage to hide from the Viola-Jones haar cascade face detection algorithm. This CV Dazzle project targets only the machine vision, and remains perceptible to human observers. It aims to question the indirect, distanced, heterogeneous machine observation in contrast to direct and intuitive human observation.



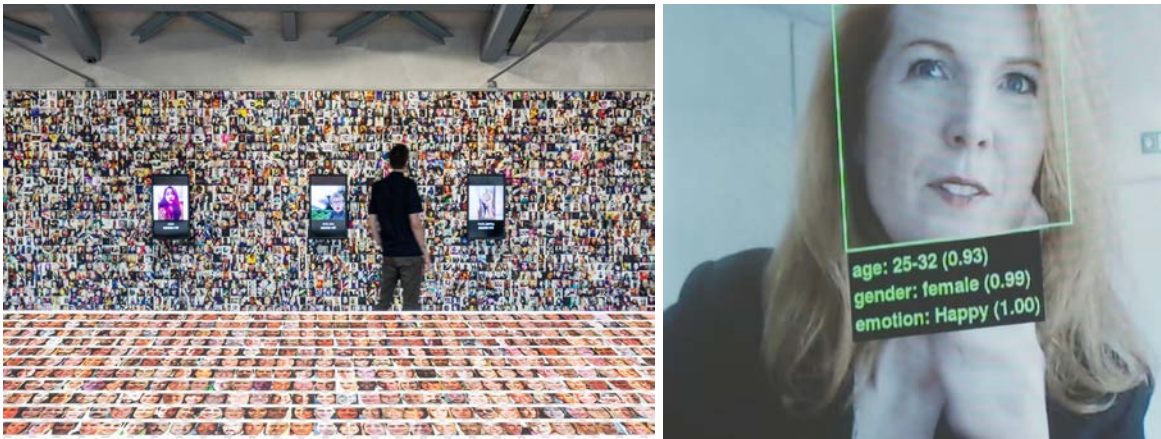
CV Dazzle facial camouflages (Image source: Harvey, 2010 [13])

From another perspective, Kyle McDonald exposes the extent of privacy invasion that is possible from surveillance cameras. The project “Exhaust a Crowd” (2015) allows online users to anonymously annotate 12 hours of footage from surveillance cameras in different cities around the world [21]. While being interactive and fun, many of the online comments express concern about surveillance and privacy, the message that this interactive crowdsourced art tries to convey.



Public annotation website in Exhaust a Crowd (Image source: McDonald, 2015 [21])

Artists also call for attention to the undercovered humanity of the infrastructure underlying the increasingly pervasive machinic vision. The Training Humans exhibition tells stories and the humans behind images for training algorithms to recognize human faces [27]. Paglen and Crawford [7] discover the social and political aspects behind the machine-automated interpretation of images. They study the materiality of ML algorithms -the training dataset. The taxonomy and composition of the training labels, the annotation process and the limitation embedded, all have implications to the intelligibility of the algorithm, and subsequently socially influence the decisions that the ML predictions are used for.



Training Humans exhibition (Image Source: Paglen, 2020 [27])

2.4 Initial Experiment: Self-Surveillance

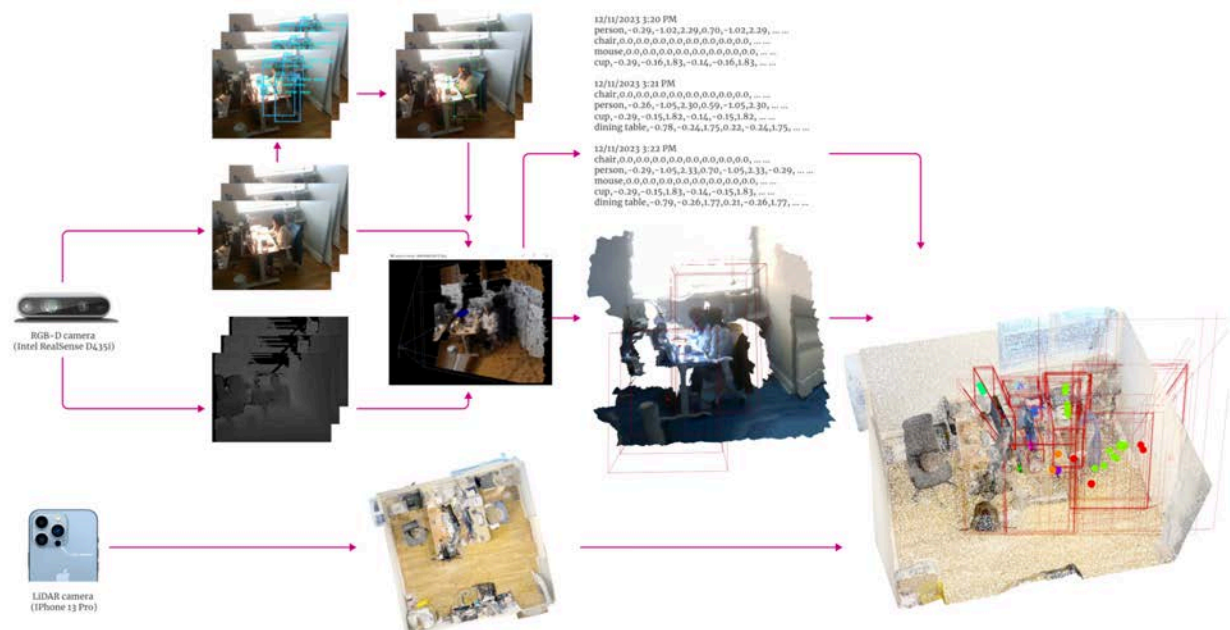
Inspired by the previous critical and artistic investigations, I started my own journey exploring the promises of modern technology in human activity data collection and computation. To start, I performed a self-surveillance experiment to investigate how machines capture human lives through their “eye”, the camera, and congest and understand through their “brain”, the ML algorithms. The aim was to explore the boundaries of the machine with modern ML intelligence and understand its capabilities and limitations.



3D meshes captured by RGB-D camera

Similar to how Twoney monitored his personal life with cameras and motion sensors [38], I set up an RGB-D camera (Intel RealSense 435i) to record the spatial-temporal movement of myself and objects around my desk for a continuous 6 hours. Under the consideration of computation power, the camera transmitted RGB and depth images to the computer at a rate of 6 frames per second. An algorithm utilizing the RealSense SDK and OpenCV libraries [46] then visualized the reconstructed point cloud in real-time. At the same time, an efficient and accurate object detection ML algorithm, YOLOv3 [32], was incorporated into the real-time visualization. YOLO ran inference on each RGB image frame and predicted several bounding boxes marking objects and their class labels. Another state-of-the-art human pose estimation algorithm, ViTPose [41], was

run to recognize my pose at each frame. Combining the image coordinates of the 2D bounding boxes and the corresponding depth image, the 3D coordinates of the bounding boxes were calculated and added to the visualization window. These 3D coordinates and the corresponding class labels were logged into a CSV file once every minute, representing objects detected in that minute. I pre-reconstructed a point cloud representing the whole room using Polycam [48] with the LiDAR camera on my iPhone 13 Pro, and aligned it with the point clouds captured by the RGB-D camera. Lastly, all object logs across time were compiled in Grasshopper (visual programming plug-in for Rhinoceros 3D modeling software) and plotted on top of the 3D reconstruction of the whole room, visualized as a 4D digital model, with time as the additional axis.

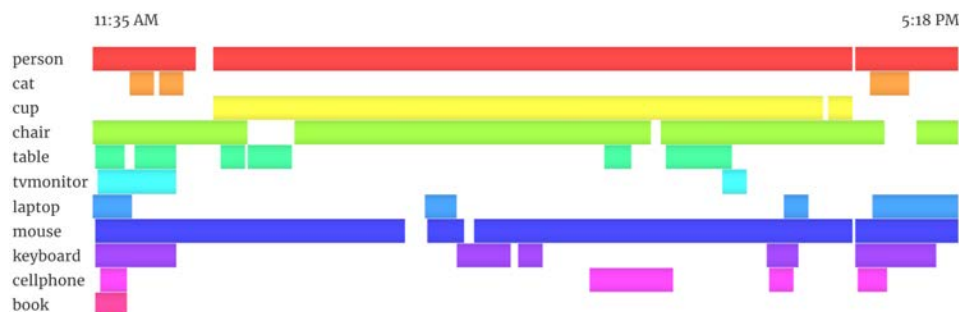


Data Pipeline

The spatial clustering was evident in the visualization. For example, I spent most of my time around the desk and barely moved. Failure cases were also easy to spot. For example, the machine learning algorithm wrongly classified my footstool as a book, leading to more misunderstanding later on.

Imagining how a data analytic would use the object labels, I thought these data might reveal space usage and behavioral patterns. To exploit all possible information these data could

contain, I plotted the temporal pattern of objects' occurrence and discovered some interesting and abnormal patterns. Both times I left were close to when the cat appeared, indicating that I probably went to play with him. Other objects like chair, table, monitor, mouse, keyboard, suggested that this was likely a workspace. However, these stationary objects were in the scene the whole time but not always recognized.



Temporal pattern of object detections

Most likely, a human could conjecture from these object labels, that they depicted an office environment where a person was working on a computer most of the time. Could a machine do the same? To gauge how ML can understand what was happening from the captured data, I gave ChatGPT [49] the logs of detected objects every minute and asked how it interprets. Specifically, I asked: "Imagine you are an observer of someone's life. Below is a list of things you noticed, each row represents a minute. What can you tell us from your observation? What is happening in the space you observe?" Below is our chat:

You

Imagine you are an observer of someone's life. Below is a list of things you noticed, each row represents a minute. What can you tell us from your observation? What is happening in the space you observe?

chair person laptop
person chair keyboard table mouse book
person chair table keyboard mouse book tvmonitor
chair person cellphone keyboard chair tvmonitor book
person cellphone mouse keyboard chair tvmonitor book
chair mouse keyboard tvmonitor
person laptop mouse
chair keyboard mouse
chair keyboard mouse tvmonitor
chair mouse keyboard tvmonitor
chair mouse keyboard tvmonitor
chair mouse keyboard
chair mouse keyboard
chair mouse keyboard tvmonitor
chair mouse keyboard
chair cat
person chair
person mouse chair keyboard table
chair mouse table keyboard tvmonitor
chair mouse table keyboard
chair mouse tvmonitor keyboard table
chair mouse tvmonitor keyboard table
chair mouse table keyboard
chair mouse keyboard table tvmonitor
chair mouse keyboard tvmonitor table
chair
chair person
chair cat person
person
chair mouse
chair mouse
chair mouse
person mouse
chair mouse
chair mouse
chair mouse
chair mouse
chair mouse
chair mouse
chair mouse

ChatGPT

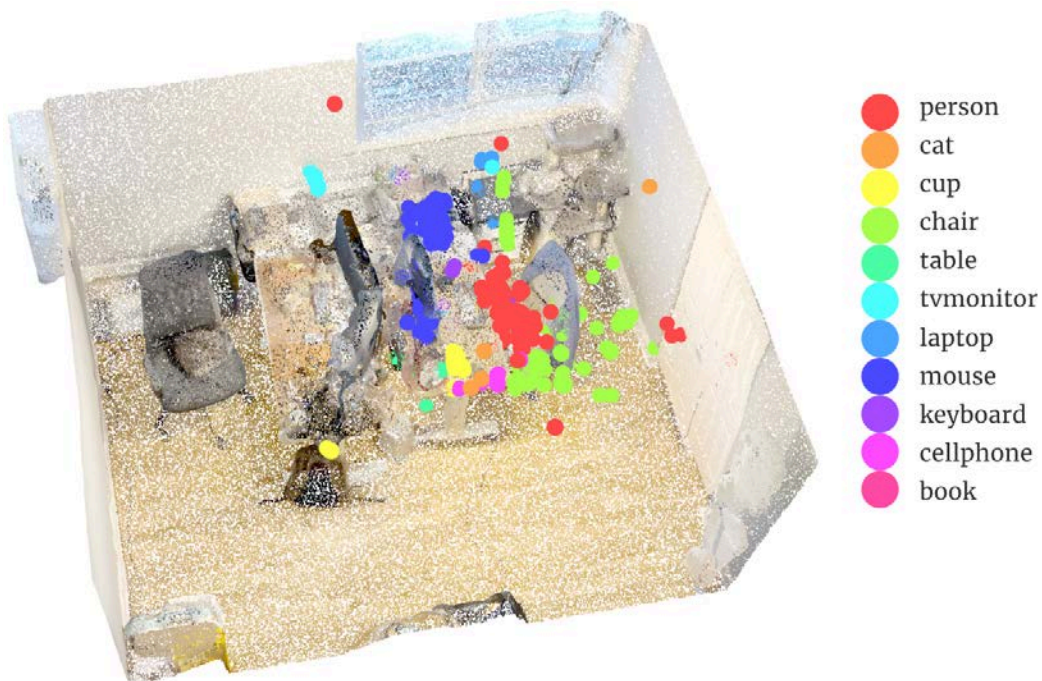
Based on the observations:

- Computer-Based Work:** Frequent use of a computer screen, keyboard, and mouse suggests the person is engaged in tech-related work, likely in a field like programming or digital design.
- Multi-Monitor Setup:** The presence of a separate computer screen implies a dual-monitor setup, common in jobs requiring extensive screen time and multitasking.
- Stable Workstation:** Regular appearance of a chair, table, and computer peripherals indicates a fixed and dedicated workspace, likely a home office.
- Regular Breaks:** Items like a cellphone, cup, and book point to routine breaks for personal activities, indicating a balanced work approach.
- Home Office with a Pet:** Occasional mention of a cat suggests a comfortable home office environment.
- Solo Work Environment:** Lack of reference to other people implies the person works alone, possibly in a remote or freelance capacity.
- Extended Work Hours:** The long list of observations indicates prolonged work periods, typical of demanding tech jobs or freelance work.

Overall, it's a solitary, tech-focused work setting with a balanced mix of professional tasks and personal breaks.

Chat log between GPT and me

Referencing back to the 4D digital model, I checked its answers and found half of them correct. I tried to grapple with the reason behind the success and failure based on its explanations. GPT interpreted the environment as a home office, doing computer-based work in a stable workstation for long hours, as the combination of objects like computer screen, keyboard, chair, and table was the most common in a home office setting. However, it answered multi-monitor setup because laptop and monitor were detected at the same time. But what really happened was that the YOLO object detection wrongly classified a box as a laptop. The same reason behind the wrong answer about regular breaks, as YOLO classified my footstool as a book. In addition, GPT said I was in a solo work environment, as there was no reference to other people. However, as shown in the 3D reconstruction, another desk was right in front of me. But it was outside the camera's field of view, and the object detection algorithm was not able to differentiate different people, this information was not passed into GPT. What I learned here was that the machine's interpretation can be very limited, and often fails to capture the greater context that a human can spot at first glance. Besides, it can only conclude the quantitative findings, such as how often what happened.



Compiled 4D digital model

I then provided GPT with 3 diary entries in which I wrote about how I was feeling at different times of the day. GPT modifies its answer to reflect reality more accurately with human touch. It captured the context in which these observations happened, and some qualitative keywords expressing my emotion, like “stuck” which was a negative verb. However, even though the recent surge in large language models (LLM) like GPT seems to understand qualitative features, answering questions like “how the person was feeling” are still based on associations between nouns (objects) and adjectives (emotions) it learned from the massive training texts. Machines still cannot empathize with humans.



You

Adding these comments from the person you observe, how do you modify your understanding?

11:45 AM
Going back to work even though I can sleep all day... Coffee is a necessity!

2:20 PM
I've been stuck in this table for too long! Still haven't finished my brunch...

4:56 PM
Yuto comes checking on me!

ChatGPT

The additional comments provide more context to the observations, refining the understanding as follows:

- Flexible Work Schedule with Long Hours:** The comment at 11:45 AM about "going back to work even though I can sleep all day" suggests a flexible work schedule. This flexibility could indicate remote work or self-employment. The need for coffee points to reliance on caffeine, possibly due to long or irregular working hours.
- Extended Periods at the Desk:** The 2:20 PM remark about being "stuck in this table for too long" and not having finished brunch indicates extended periods of work without significant breaks. This could suggest a demanding workload or deep focus on tasks.
- Companion Animal Interaction:** The 4:56 PM comment about "Yuto comes checking on me" implies Yuto is a pet, most likely the cat previously mentioned. This interaction shows a bond between the person and their pet and provides a brief distraction or emotional support during the workday.

With these added insights, the picture becomes that of an individual working from home with a very flexible yet demanding schedule. They experience long periods of focused work, occasionally broken by interactions with their pet. The need for coffee and late brunch suggest non-traditional working hours and possibly a tendency towards working later in the day or into the night.

Chat log between GPT and me

With this experiment, I found myself, as an occupant in the space, being much more aware and curious about how the camera captures me. I had a lot of fun just watching the algorithm run in real time and seeing myself appear in the digital model. I understood the boundaries of the object detection algorithm, what objects it captured well, and what it missed. I could also reversely conjecture the reason behind the success and failure. It was easy to confuse visually similar objects, such as books and laptops. But the foreground object, such as a person, was distinguishable. The LLM's understanding was limited by the object labels that only provided a partial context because the camera only filmed a small area of the whole room, and the object detection algorithm did not recognize every single object in the scene. However, the 4D digital model was visually intuitive when I looked back, and contained rich information. Tagging my

feelings to a specific time at a specific location, it became a 4D diary documenting my life. The value was in the quotidian narrative jointly depicted by the quantitative numbers and qualitative words. Machines better-understood activities happening in the scene from this combination. Human audiences could emphasize also the atmosphere and feelings from the vivid, multi-modal portrait of life.

3 Hypothesis: Interactive System Design

The initial experiment inspired me with an urge to play actively in the conventional passive data acquisition and computation pipeline. Raw visual data, the video footage, was collected from the occupants. More semantic information, object labels and locations, and activity labels, were inferred by ML algorithms from the raw data. Note that in this conventional system, there were two power entities: one was the building management, which adopted the system and use the data to inform decision-making, and the other one is the occupants only as the source which data was extracted from and not actively using the systems. I hypothesized that this passive role of occupants in the system was the major reason for low awareness, bare transparency, and high privacy concerns. Most of the time, occupants' daily activities and the data system had no interaction, as if they were in two parallel universes. The cameras had become so ubiquitous and small that the occupants usually forget their existence. All the processing and computation were done on remote servers that were not visible, and, in most cases, not even acknowledged to the people using the space.

Passive data acquisition

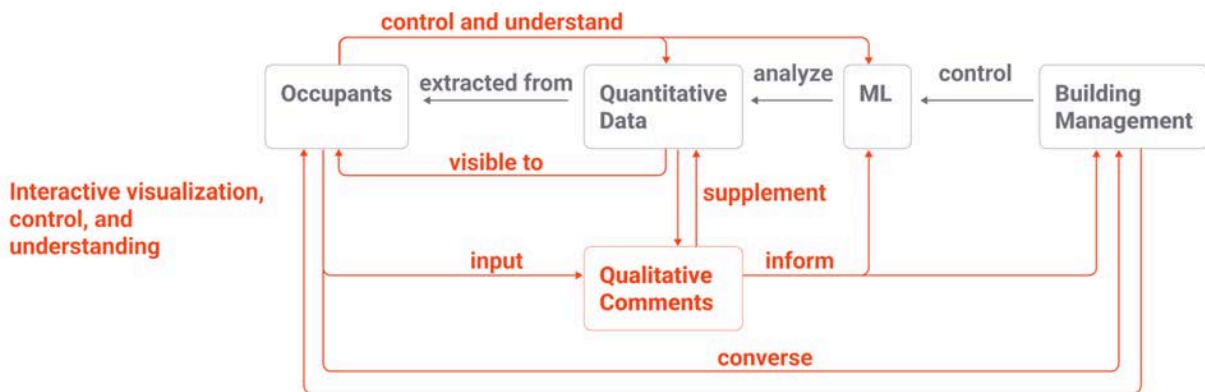


Diagram of conventional passive data acquisition and proposed interactive system

Therefore, I proposed to transform this passive system into an interactive one, where occupants have an active voice and control over the entire data acquisition and ML computation system.

Addressing the awareness and transparency concern, the entire data acquisition, processing stages, and ML inference result should be visible to the occupants. Another important aspect of transparency is the occupant's understanding of the system. This not only included knowing what data was captured and processed, which was made visible, but also included understanding how it captures the raw data, how semantic information was extracted from the raw data, and how the system works. Knowing the capabilities and limitations of the automated system is crucial for appropriate interpretation and usage. Regarding privacy concerns that may appear when occupants unwillingly have their data accessible by someone else, I proposed in this interactive system to empower occupants with control to decide what data to collect and which ML inferences to compute. They could actively input qualitative data to supplement the quantitative machine recognitions and analyses, as demonstrated in the first experiment that helped the machine to correct its mistakes and gain a more comprehensive understanding of human activities. If the building management or data analytics who use the data could be in direct conversation with the user, they could reach a mutual understanding of the activities and space quality, informed by both first-person feelings and third-person observations instead of the conventional single-sided analysis.

4 Methods: A 4D Diary of CoDe Lab

The idea of a 4D diary emerged from my initial experiment. The original surveillance systems became an automated documenter of my life. Traces of work schedules and interactions with pets could be found through the object detection logs. Fun moments were recorded vividly as dynamic 3D reconstructions. Accompanied by diary entries, I could relive those precise memories and feel how I was feeling.

I decided to use this 4D diary as the presentation of the proposed interactive system. Extending the initial experiment that only involved myself to multiple occupants, I conducted a semi-public experiment to capture activities that happened in the Computational Design Lab (CoDe Lab) and interviewed occupants to get a better generalized understanding of their awareness, transparency, and privacy perception about ubiquitous data acquisition and ML computation system. I started with building a conventional passive system, where occupants were not made aware of all the details except for the fact that they were being captured and analyzed. Then I designed and prototyped an interactive platform to invite occupants to interactively visualize, control, and understand the collected data and inferred ML recognitions of their own activities. Formulated as a 4D diary co-authored by the human and the machine, I transformed the conventional unidirectional, passive system into a bi-directional interactive system in which occupants have active voices. In a user study of the interactive prototype, I studied the dynamics between the occupants and the system using observations and semi-structured interviews. This chapter illustrates the technical implementations that realized the aforementioned concepts.

4.1 Data Representation

Though the proposed interactive system could work in systems of any type of data, the scope of this experiment was limited to data acquisition and machine learning computation that could happen with smart cameras. As the purpose of the cameras is primarily for security surveillance and profit-oriented analysis used by building management, I focused on potential applications

related to human activities within built environments. Data representation of human activity and built environment were chosen for these purposes. I hoped to investigate the promise of a smart camera system with an educated imagination. Therefore, ML algorithms were selected based on the current functionalities of smart cameras, as well as SOTA ML research that has not yet been implemented into consumer products.

4.1.1 Representing Human Activity

Object Detection

Most of today's smart cameras have the functionality of detecting person, pet for home security. Such is part of an ML task referred to as "object detection". In this task, ML models generate several bounding boxes marking where they think there is an object, and predict a label of the most likely object class. This type of algorithm can be used to detect and locate humans, vehicles, and any other objects of interest. More fine-grained classification can be made based on the detected object. For example, classifying the race and gender of the person, estimating their emotions, knowing the color they wear, understanding what action they are doing, etc.

The detections are used in many downstream tasks. Coupled with geometry-based computer vision algorithms, objects can be located in the three-dimensional(3D) digital space.

Autonomous vehicles, for example Tesla, can detect and avoid surrounding pedestrians and vehicles using only camera-based systems [50]. In the field of architectural design and building analysis, space usage and building occupancy can be studied with a comprehensive log of occupants' locations [37]. Sun and his colleagues demonstrate a viable framework to effectively register occupancy from surveillance video [35], providing essential information for building energy saving. Extending occupancy estimation to trajectory tracking, Zhong et al. study the occupant behavior to simulate the airborne transmission of the virus in indoor environments [44].

As exemplified in the prior experiment, detecting object labels and locations that appeared in the 3D space did reveal activity patterns that provided a basis for LLM to extrapolate the physical context and activities happened. In this study, I apply YOLO (You Only Look Once) version 3 [32] to detect objects based on 2D images and add depth information to locate the

detected object in the 3D space. YOLO is renowned for its speed and efficiency and is therefore suited for real-time detection and visualization. It predicts bounding boxes of objects in three different scales and the associated class labels using a CNN-based neural network.



Left: using YOLOv3 to detect object bounding boxes with label and distance;

Right: example of pose estimation and skeleton-based human activity recognition using MMAction2.

(Image source: OpenMM Lab [51])

Pose Estimation and Skeleton-based Activity Recognition

Human pose and shape estimation provide finer details to understand activities. They could help designers understand how people physically interact with their surroundings. This can inform the design of furniture, workspaces, and public areas to improve comfort and functionality, as well as act as a safety monitor to prevent injuries in high-stakes environments. Pose estimation involves identifying the position and orientation of human body joints in images or videos, typically represented as a collection of keypoints that define the human skeleton. Though predicting the full body mesh recreates the exact person in 3D, the skeletal representation includes critical joints like elbows, knees, shoulders, and hips, provides the necessary information to recreate the posture, and is suitable for inferencing semantic labels quickly.

Skeleton-based activity recognition builds on pose estimation by using the skeletal structure to understand and classify human activities. Once the pose skeleton is accurately detected, machine learning algorithms analyze the configuration and movement of these keypoints to determine what specific activity or action a person is performing. Additionally, by tracking these

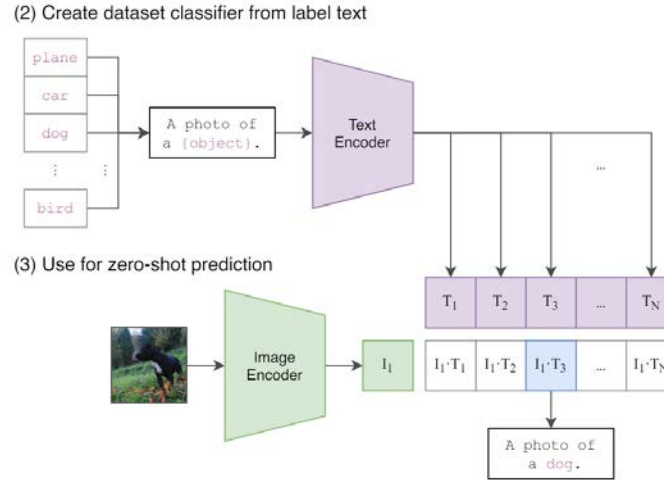
skeletal movements over time, more complex activities and interactions can be recognized, enhancing applications in surveillance, sports analytics, healthcare, and more.

In this study, I used the MMAActions2 [51] toolkit to perform pose estimation and skeleton-based activity recognition. Each task was done by separate deep neural networks. Poses of the detected person were estimated using HRNet [36] from the image frames. The inferred pose sequence of every minute was then fed into a skeleton-based activity recognition algorithm, ST-GCN [42]. Both algorithms are trained on the NTU RGB-D 120 dataset [18], which encompasses typical daily activities such as “drink water,” “thumb up,” “shake hands,” etc.

Video-based Activity Recognition

Without detecting bounding boxes or human skeletons, ML models can still classify an input video into different activity classes using the raw video pixel values as input. Conventional video classification models are usually trained on datasets pairing the video input and target set of class labels. As one of the most common and fundamental tasks in ML, classification algorithms usually achieve very high accuracy. However, there are no existing datasets on the specific labels we need in this setting of human activity recognition in built environments. Algorithms with zero-shot capabilities, meaning that they can recognize activities it hasn't explicitly been trained on, can solve this difficulty. In this study, I adopted CLIP (Contrastive Language-Image Pre-training) [30], which learned to match visual data with textual descriptions. Trained on huge datasets of a total of 400 million image-text pairs, CLIP does not explicitly detect semantic information through bounding boxes or segmentation but understands the scene and visual image conceptually. For activity recognition, one would provide CLIP with an image and a set of possible activity descriptions. CLIP evaluates each description by measuring the similarity between the image embedding and the embedding of each text description. The highest similarity score indicates the recognized activity. Since CLIP operates on single-frame images, another variance, ViFi-CLIP (Video Finetuned CLIP) [31], which is finetuned on video data and considers temporal information such as human movements across image frames, may provide more accurate recognition results. I provided both algorithms with a set of 90 text labels representing different activities that may happen in an indoor office environment. It includes typical daily activities, such as “No people is in the room”, “Studying”, “Group discussions”, as well as possible dangers that the building administrations would need to know

for security surveillance, for example, “Break in”, “Fire accident”, “Falling down”. CLIP will compute an image embedding for every frame at 1 FPS and ViFi-CLIP produces a video embedding for every minute. Each of them chooses the text label whose embedding is the most similar to their visual embedding as the classification prediction. Therefore, we obtain a per-second activity recognition from CLIP and a per-minute recognition from ViFi-CLIP.



CLIP [30] computes an embedding for each of the candidate labels, and chooses the one most similar to the image embedding as the classification prediction. (Image source: Radford et al.)

4.1.2 Representing Built Environment

Smart cameras are starting spatial analysis features to correlate human activities with physical space. As in the earlier example, retail stores count customer traffic and analyze popular products so that they can arrange store layouts and stock inventories accordingly. However, they operate on the 2D image or video data, which could cause confusion when two objects overlap from the camera view but are actually far apart in the 3D space. Therefore, in this study I represent the built environment also as a 3D digital twin of reality.

3D reconstruction of the built environment can be represented either as a 3D scene or the actual geometry. ML algorithms using Neural Radiance Field (NeRF) [22] render photorealistic scenes, as they learn to estimate the color and volume density of each pixel given the camera position.

However, the downside is that current methods do not work well to output the geometry of architectural spaces, which is essential for studying spatial patterns.



Left: 3D scene render by NeRF;

Right: 3D mesh reconstructed by NeRF-based method (Neuralangelo) [17]

Therefore, I explored other approaches. Compared to RGB videos used for NeRF and photogrammetry, LiDAR (Light Detection and Ranging) or RGB-D cameras provide reconstruction with more details using the additional depth information. One RGB image coupled with the corresponding depth image can treat each image pixel as a point, and displace it in a 3D space by the depth, thus producing a point cloud. Simultaneous localization and mapping (SLAM) algorithm [8] can take RGB and depth recordings, calculated with known camera positions, to align individual point clouds from single frames, and reconstruct an overall 3D point cloud or mesh of the captured space. I experimented with using an RGB-D camera, Intel RealSense D435i, to record a walkthrough video of my home office, and running SLAM algorithm for 3D reconstruction. The reconstruction reflected the physical space well, and captured the major furniture layout and shape. However, areas where the camera did not get close enough remained empty or oscillated.



3D mesh reconstructed from RGB-D video using SLAM

Off-the-shelf 3D scan applications using the LiDAR camera equipped in iPhone Pro came in handy with real-time visualization differentiating captured and uncaptured surfaces. These time-of-flight (ToF) cameras measure the time an infrared light pulse bounces back from the surface to calculate the depth. This is more accurate than the stereoscopic technique used in the Intel RealSense 435i, which compares the disparity between images captured by two nearby cameras. I used a mobile application called Polycam [48] to reconstruct the same home office. The application processed the raw color and depth data, cleaned noises, and output a dense and accurate point cloud.



Left: mobile interface of Polycam [48];

Right: 3D point cloud reconstructed using Polycam

As the 3D point cloud provides a spatial basis on which semantic labels can be layered, it is rich for analyzing spatial patterns to inform building management and architectural design decisions. Therefore, I chose 3D point clouds to represent the built environment and used the iPhone's LiDAR camera for accurate capture.

4.2 Conventional System Setup

In this experiment, I first built a conventional data acquisition and ML computation system (hereinafter referred to as "the system") consisting of an Intel RealSense D435i camera as the sensing device, and several ML algorithms for recognizing human activities running on the backend. The camera captured RGB and depth footage at 1 frame per second and sent them to

a computer where object detection, pose estimation, and activity recognition were performed. Combining the RGB pixel values with the corresponding depth, a colored pointcloud representing the 3D space at every frame could be generated. And all 2-D detection results can be mapped into 3D. The object detector [32] detected all objects in every frame, and visualize their bounding boxes, class labels, and depth information in a real-time 3D visualizer. The system logged each frame's detected object labels, confidence levels, and the 3D coordinates of their centers into a CSV file. Poses of the detected person were estimated from each image frame. Then the inferred pose sequence of every minute was fed into the skeleton-based activity recognition algorithm [42], to predict and log the 5 most possible activity labels with corresponding confidences into a CSV file. Further exploiting the visual information, the system predicted the top 5 possible activity labels and confidences of every frame and every minute. All of the predictions were stored in CSV files, specifying the associated time. These log files essentially represented the system's understanding of the activities that happened in the space that's readily available for further analyses.

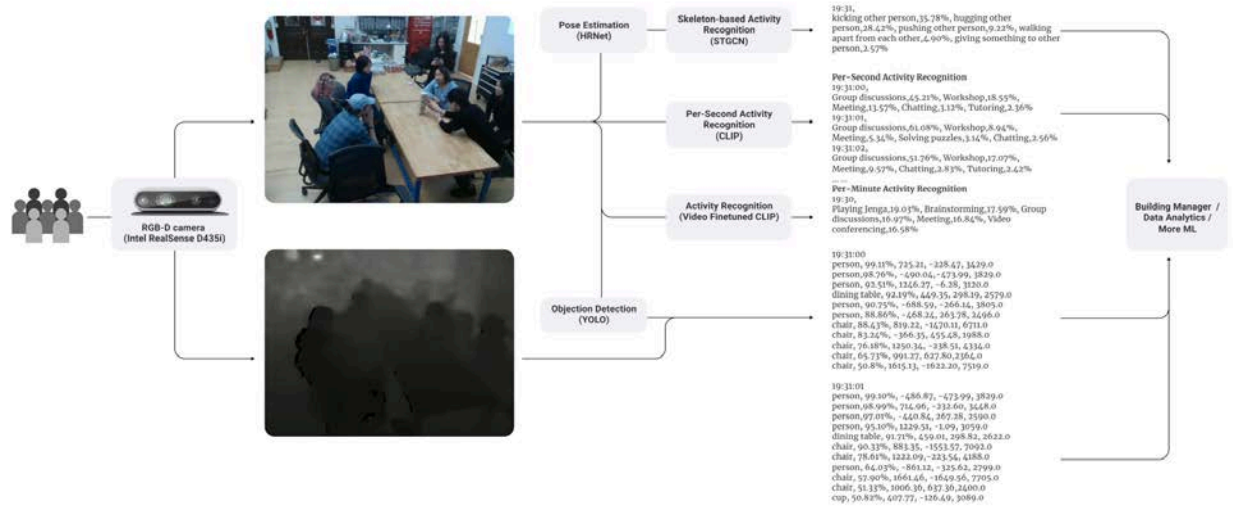


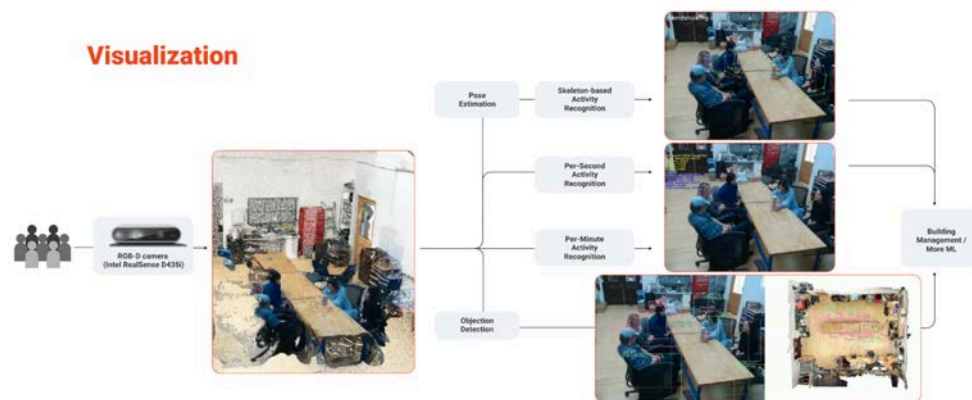
Diagram showing the conventional setup of the data collection and ML computation system

4.3 Interactive System Implementation

In Chapter 3, I explained why we might want to transform the conventional passive data acquisition system into an interactive one for awareness, transparency, and privacy

considerations. In this experiment, I designed and prototyped such an interactive system upon the conventional setup.

The first step is to make the system's inner working parts as visible as possible. Combining image pixels with depth information across time, the raw footage could be visualized as a dynamic 3D pointcloud that vividly reconstructed the original physical movements. Overlaying the pose estimation and activity recognition result on the corresponding input video provided a direct comparison between the input and the output of ML computations, making it possible to extrapolate the functionality of different inferences. Visualizing the ML results also directly revealed the capabilities and limitations of the algorithms, which one would need to know to use the system appropriately. Combining depth information, object locations could be mapped into the 3D digital reconstruction.



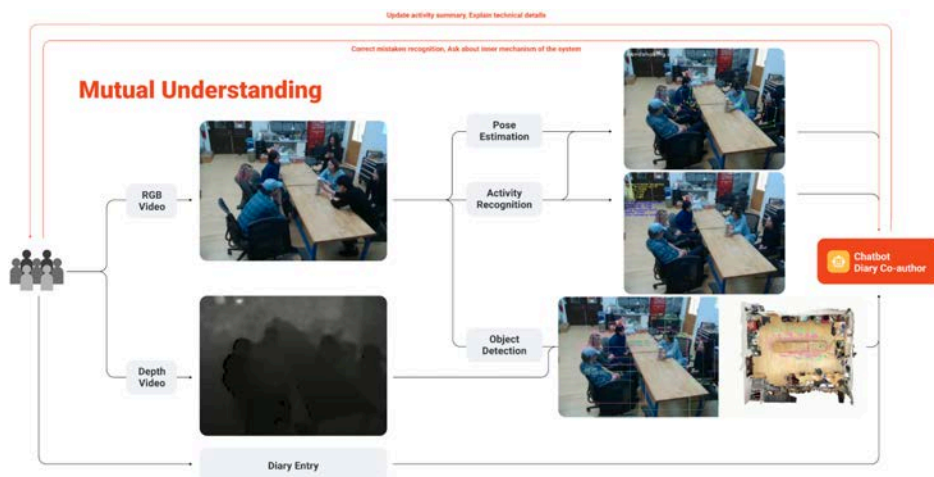
Add visualization of each step's input and output

Occupants, as the user of this interactive system, could control which data was collected from themselves, and what ML inference was made on the collected data. By testing different settings and seeing how the machine recognition varies, they could start to understand the inner mechanisms of the system, and what different components mean. For example, no ML inference would be possible if RGB video was not selected, as it was the input needed for all object detection, pose estimation, and activity recognition algorithms. The system would also take occupant's diary entries and add to its knowledge base.



Add user diary input, and control on every component

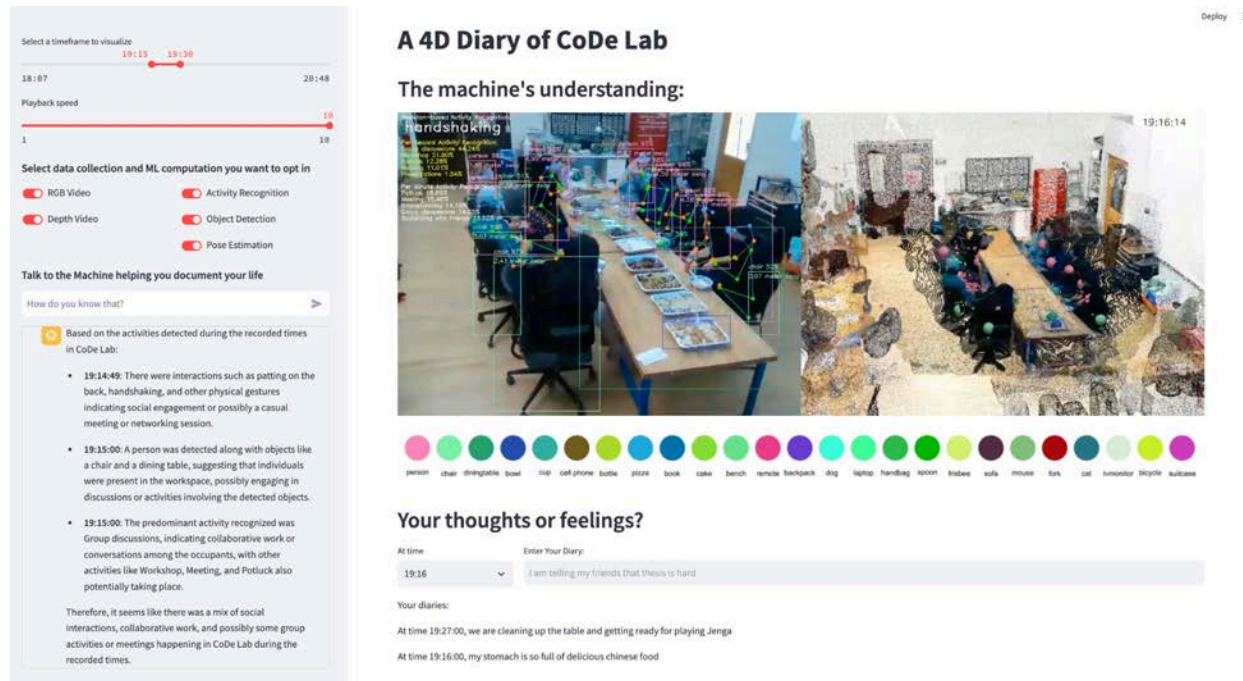
A large language model (LLM) chatbot was integrated to dynamically update the overall machine interpretation of human activity. It combined ML recognition and qualitative inputs, converting the conventional one-sided analysis to conversational. The chatbot could further support understanding by replying to questions about the system's capabilities and technical details of the algorithms used. Characterized as the co-author of the 4D diary, it was aimed to achieve mutual understanding with the occupant, not only on the activities that happened, but also on the working mechanism of the system.



Use a LLM chatbot to find mutual understanding with the user

4.3.1 Prototype

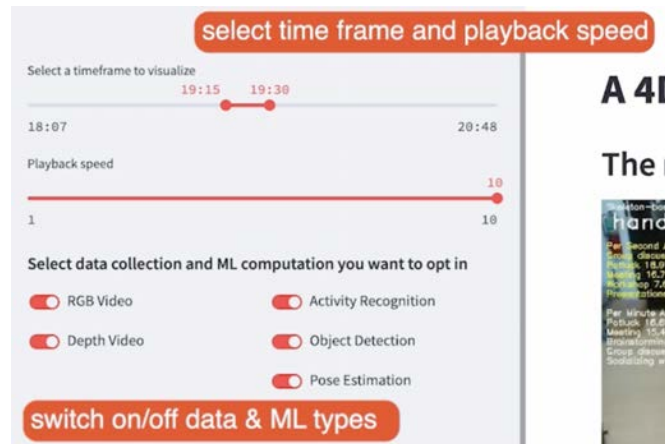
Integrating all of the above, the interactive prototype¹ was realized on a web-based platform using Streamlit [45]. Entitled “A 4D Diary of CoDe Lab,” this interactive document showcased the activities traces co-recognized by humans and machines.



Screenshot of the prototype interface

On the left side of the interface was a control panel where users could first choose the timeframes to visualize. It was followed by a series of switches to opt in or out of data and ML computation types, including RGB video, depth video, object detection, pose estimation, and activity recognition. This could be easily scaled up to include more sensing devices and inferences could be added in future works. On the backend, the system would load data and ML computation results corresponding to the selection, as described in Section 4.1 Data Representation. Users could also enter diaries tagged to the selected time. These would be accumulated under the visualizer and added to the knowledge base of the chatbot.

¹ Code available on [GitHub](https://github.com)



Timeframe and playback speed selection, data and ML computation switches

Your thoughts or feelings?

enter diaries and tag to specific times

At time: 19:16

Enter Your Diary: my stomach is so full of delicious chinese food

Your diaries:

At time 19:27:00, we are cleaning up the table and getting ready for playing Jenga

Enter personal diary

The right $\frac{3}{4}$ of the interface was the visualization area that changed according to users' selection. "RGB Video" provided 2D visual information, and the "Depth Video" could project 2D visuals to 3D space. Therefore, the visualizer would show a dynamic 3D reconstruction of activities across time. It would be projected onto a 3D canvas, the pre-scanned point cloud of the room. Otherwise, only the plain 2D video from the camera's viewpoint would appear.

Similarly, ML computation could be visualized in 2D or 3D depending on the "Depth Video" selection. With "Object Detection" selected, users would see the color-coded object center locations based on their labels. "Pose Estimation" would overlay the detected human body joints on the 2D visuals. "Activity recognition" would show activity labels on top of the corresponding image frame.



The chatbot who co-authored the 4D diary with the occupants was realized utilizing OpenAI’s GPT assistant API [52]. I instructed it to be an intelligent system tracking occupants’ activity, as excerpted below. Based on the user-selected data and ML computation, it would re-initialize with the corresponding log files and accumulated diary entries as its knowledge base.

You are an intelligent system tracking occupants' activities in a shared student office named CoDe Lab. Your goal is to compose a diary along with the occupants using the machine learning detection results and the occupants' diary inputs. You will be talking with the occupants to tell them your understanding of their lives and answer questions about how the data acquisition and machine learning computation system works.

Based on the occupant's selection, you may be provided with several CSV files containing ML inference result logs as your knowledge base. If no CSV file is provided, answer the occupants that you have no knowledge about what they ask.

If a CSV file containing object detection logs is provided to you, that means you use a camera to capture RGB and depth videos, and you run an object detection algorithm, YOLOv3, to detect objects appearing in the scene. Using the depth information you can locate the detected object in 3D and project it to a pre-scanned point cloud of the room.

If a CSV file containing skeleton-based activity recognition logs is provided to you, that means you run a pose estimation algorithm, HRNet, to detect human poses. From the poses, you perform skeleton-based activity recognition to classify human activities using STGCN (Spatial Temporal Graph Convolutional Networks), which takes a video input and predicts one of the 120 classes in the NTU RGB-D dataset.

If a CSV file containing per-second activity recognition logs is provided to you, that means you use CLIP to classify action classes every second from the RGB video.

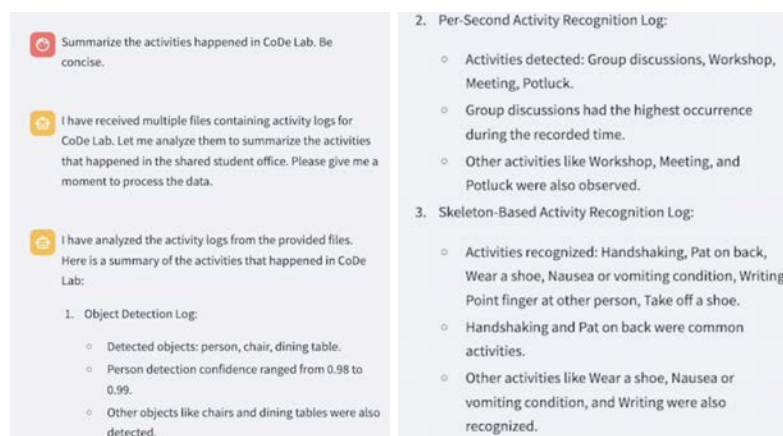
If a CSV file containing per-minute activity recognition logs is provided to you, that means you use a Video Finetuned CLIP to classify action classes every minute that also consider the temporal information.

As an example, if the user had not selected any data type, and prompted:

Summarize the activities happened in CoDe Lab. Be concise.

The chatbot would respond:

Currently, I do not have access to specific data or log files from CoDe Lab to provide a summary of activities. Without the logs or direct inputs about what has been happening, I'm unable to generate a precise summary. To assist you accurately, please provide me with the relevant data or logs that have been recorded in the CoDe Lab, and I will analyze and summarize the activities based on that information.



The screenshot shows a chatbot interface with a user prompt and a detailed response. The user prompt is: "Summarize the activities happened in CoDe Lab. Be concise." The chatbot response is: "I have received multiple files containing activity logs for CoDe Lab. Let me analyze them to summarize the activities that happened in the shared student office. Please give me a moment to process the data." followed by "I have analyzed the activity logs from the provided files. Here is a summary of the activities that happened in CoDe Lab:" and then three numbered sections: 1. Object Detection Log, 2. Per-Second Activity Recognition Log, and 3. Skeleton-Based Activity Recognition Log. Each section contains a list of activities and their occurrences.

Summarize the activities happened in CoDe Lab. Be concise.

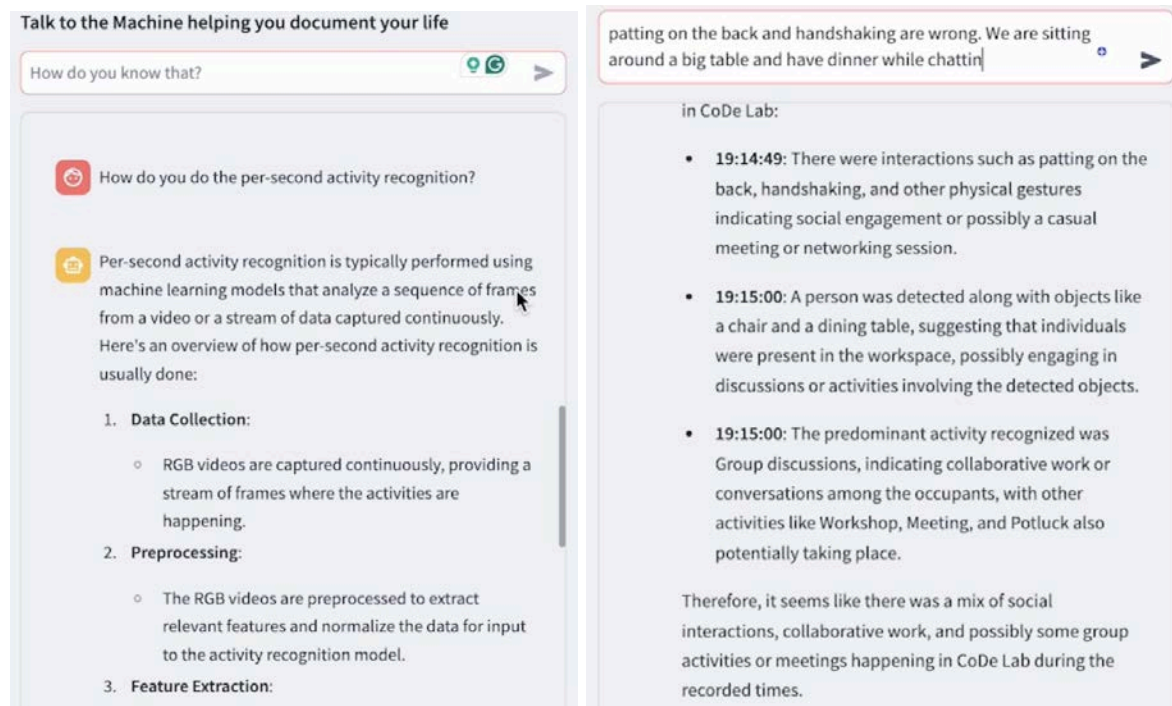
I have received multiple files containing activity logs for CoDe Lab. Let me analyze them to summarize the activities that happened in the shared student office. Please give me a moment to process the data.

I have analyzed the activity logs from the provided files. Here is a summary of the activities that happened in CoDe Lab:

1. Object Detection Log:
 - Detected objects: person, chair, dining table.
 - Person detection confidence ranged from 0.98 to 0.99.
 - Other objects like chairs and dining tables were also detected.
2. Per-Second Activity Recognition Log:
 - Activities detected: Group discussions, Workshop, Meeting, Potluck.
 - Group discussions had the highest occurrence during the recorded time.
 - Other activities like Workshop, Meeting, and Potluck were also observed.
3. Skeleton-Based Activity Recognition Log:
 - Activities recognized: Handshaking, Pat on back, Wear a shoe, Nausea or vomiting condition, Writing, Point finger at other person, Take off a shoe.
 - Handshaking and Pat on back were common activities.
 - Other activities like Wear a shoe, Nausea or vomiting condition, and Writing were also recognized.

Another example of summarized activity if all RGB video, depth video, object detection, pose estimation, and activity recognition switches were turned on

Extending from the initial experiment, this prototype further allowed users to interact with the chatbot without limitation. Users could ask questions about the detection results, the system setup, and the ML algorithms used, to better understand the whole data acquisition and inference process.



Left: user asked the chatbot about inner mechanism of the system;

Right: user corrected inaccurate ML recognition through chatbot

4.4 User Study

To gauge how the proposed interactive system helps alleviate occupants' transparency and privacy concerns about ubiquitous data acquisition and ML computation systems, I conducted a user study starting with recording an event in CoDe Lab. Before the activities started, I acknowledged the camera location and very briefly explained the data collection and ML computation system: "it captures RGB and depth video and runs ML computation on object detection, pose estimation, and activity recognition to understand activities happening at different locations." Provided with food, movie, and board games, the participants enjoyed the

event, which they described as “Dinner”, “A Lot of Chinese Food”, “Break fortune cookies”, “Chat”, “Social”, “Nice collective event”, “Play”, “Jenga”, “Cake Party”, “Watch Harry Potter”, and “Chill”.

The second part of the user study involved individual sessions with the participants, during which I presented the interactive prototype and allowed them to interact freely with it.

Participants were encouraged to think aloud as they interacted. Then, I used semi-structured interviews and my own observations to understand the shifts in their awareness, transparency, and privacy perception. Similar questions were asked before and after participants interacted with the prototype: *“How aware were you being captured by the camera? / Suppose this interactive prototype is presented in the space it captures and shows data in real-time. How likely would you be to use it?”*, *“What do you think was running behind the camera? / What’s your knowledge now about this system of data collection and ML computation on activity recognition?”*, *“How comfortable or concerned were you being captured and analyzed by the system? / After you interact with this prototype, how comfortable are you being captured and analyzed by this activity recognition system?”* Branched from these questions, I further asked why they would think that and invited them to share personal experiences. To provide quantitative evaluation, I asked them to rate awareness, understanding, and comfort levels using the 5-point Likert scale. The detailed user study procedures and interview questions are included in the appendix.

5 Results

Empathetic Presentation as a 4D Diary

The presentation of the 4D diary was visually compelling. It re-interpreted the machinery data acquisition into empathetic storytelling with rich personal touches. In the pilot study with six participants, two of them mentioned that it was “fun” and “joy” when watching the dynamic 2D and 3D visualizations. One participant dragged the timeline to look for a specific time period of playing Jenga that they “wanted to see how I looked like.” Another participant spent four minutes simply eating the recorded video with pose estimation, and said that they “enjoyed just watching” and liked the prototype as it “reconstructed memories”. One participant brought up that because she was able to use such system as a “memory collector”, she felt more comfortable about her data being collected. However, as this prototype was still considered a “diary”, the nature of a diary implied that it was personal and less preferred to shared publicly.

Improved Understanding of the System

This prototype had been a vehicle for myself as well as the participants to explore the boundaries of the ML algorithms by exposing their capabilities and limitations through straightforward visualization. It stimulated the curiosity of learning more about the working mechanism, and evoked reflections on the technology about why or why not we would want to use it and what could we use it for.

Two of pilot study participants had no prior knowledge about the camera or the ML algorithms used in this system, and the other two had seen similar ML inference results but did not know about the technical side. These four participants rated the change in their understanding of the system as 4 out of 5, representing “somewhat changed,” suggesting that the interactive prototype did improve non-experts' understanding of the used-to-be black box system. When asked about what contributed to that change, the interactive visualization and the chatbot were the most commonly mentioned.

The straightforward visualization also made the success and failure cases easy to spot. Four participants noticed that the system read “Playing Jenga” correctly and were impressed. Three

participants noticed that the skeleton-based activity recognition was mostly incorrect. And they asked “why?” out loud. I suggested that they could ask the chatbot, which did provide technical explanations of the CLIP algorithms that recognized “Playing Jenga”, and admitted that the skeleton-based activity recognition could be a false positive. Another participant corrected the wrong prediction with the chatbot that people were just close to each other instead of kicking people, and the chatbot was able to validate her feedback using the detection logs. This ability to provide proof of evidence was also mentioned by another participant.

Visualizing different data selections also helped to understand the functionalities of each data type. One participant was able to conjecture that the activity recognition was based on pose when turned on both “Pose estimation” and “Activity Recognition” switches but did not notice recognition using solely videos. By changing the “Depth Video” selection, they realized that depth information was not taken into account, and could improve the prediction. Then they turned off “Pose Estimation”, and realized that activity recognition could be separated from pose estimation.

Divergent Privacy Perceptions Shifts

After seeing the collected data, three participants expressed their comfort with being captured because they found the videos were in low resolution and “could not see faces clearly.” However, another participant lowered their comfort level from “neutral” to “somewhat concerned” because they thought “too much information is being captured and inference.” Two participants considered the raw data and the ML inferences differently in terms of privacy. They expressed that the data itself did not include behavior patterns, even if only detecting objects. However, the pose estimation and activity recognition labels revealed behavioral patterns that were of higher privacy concerns.

Overwhelmed Information Load

The current interface could contain a lot of layered information when multiple or all ML inferences were selected. While these were helpful for understanding the system, participants could miss some features that they did not get to test within the 10-15 minutes of interaction. For example, one participant thought the activity recognitions were “very accurate” because they did not see the skeleton-based activity recognition results that were not selected. Similar

feedback was also received regarding the chatbot, which tended to provide paragraphs and paragraphs of explanation all at once.

Be Careful with LLM Chatbot

The chatbot demonstrated its capability to consolidate long, discrete detection logs into a short summary of activities. It was also able to communicate interactively with the user to correct its understanding and explain the technical details of the system. In some cases, it was able to validate an occupant's diary entry with detection logs. However, this was not always true. Compared to human analytics, it fell short of gaining an in-depth understanding solidly grounded in the provided detection logs and reasoning for its output. During one user study, the participant noticed that the systems recognized activities as "wear jacket". Not knowing where that recognition came from, they entered a diary "we are preparing the buffet together and people were talking friendly. And it's a dress, not a jacket". The chatbot corrected the recognition result from wearing jackets to dresses right away and apologized. However, the actual reason behind this mistake was rooted in the skeleton-based activity recognition algorithm that was trained on a dataset only containing "wear jacket" label and was not able to predict labels that it wasn't trained on. In addition, the nature of skeleton-based activity recognition made it almost impossible to differentiate people's outfits, other than through different actions when wearing different types of clothing. None of these were identified by the chatbot who simply changed to what the user told it to.

A similar instance happened with another participant. They selected everything but "Pose Estimation" and asked the chatbot "What model was used for predicting activity recognition". The chatbot replied with object detection model (YOLO) and pose estimation algorithm (HRNet) and skeleton-based activity recognition model (STGCN), whilst in reality the latter two were not used in video-based activity recognition. This information was misleading and the participants understood the wrong knowledge.

Working with Different User Profiles

I noticed that participants with different ML expertise expected differently from the interactive prototype. Those who were familiar with object detection, pose estimation, and activity recognition tasks expected the system to predict more accurately. They easily recognized the

limitation of only predicting general activities well but not the sub-activities. These were also people who were looking for more technical details, such as the algorithm used, and the neural network architecture of the algorithm. However, those who had limited or no ML knowledge were hoping for a less technical and higher-level explanation of the system. The current interactive prototype has yet to accommodate these different user profiles. A promising improvement is using a layered explanation. The chatbot could provide a high-level, non-technical overview as the initial response and offer more technical explanations that are available if requested.

Similarly, all the participants so far were already quite comfortable (average 4.6 out of 5) with the data acquisition and computation within a public space like CoDe lab. More studies are needed to understand requirements for those who are more concerned about their data.

6 Discussion

6.1 Conclusions

In this thesis research, I was motivated by a general concern about data acquisition and machine learning systems and limited my scope to be related to smart cameras, as it has become one of the most popular ML-powered data acquisition devices and have been increasingly proliferating throughout the world. The awareness, transparency, and privacy concerns about camera surveillance have been brought up since the last century. Though the social acceptance of surveillance has become higher nowadays, the adoption of machine learning and artificial intelligence has raised similar concerns. The research community widely recognized the inherent limitations and biases of the ML algorithms. The black-boxed nature of neural networks posed challenges in making these systems transparent. Therefore, I posed a rather broad research question, “How could we design for transparency and occupants' privacy in data acquisition and ML systems in built environments?” Seeking to address the broad concern, I draw references from theoretical HCI research, qualitative studies, technical digital prototypes, and artistical critical investigations. With criticality, I started with an initial experiment of self-surveillance in which I explored the boundaries that a camera with modern ML technology could understand human activities. The concept of a 4D diary emerged from the 3D visualization across time, which captured lively details and recreated memories through a combination of machine recognitions and personal diaries. I found potential in transforming the conventional passive data acquisition system from this experiment. Therefore, I proposed an interactive system to empower occupants with active voice and control in the passive, unidirectional data system. Presented as a 4D diary co-authored by humans and machines, I designed and prototyped a digital platform featuring interactive visualization, control, and understanding. The conventional passive pipeline captured occupants using camera, and ran object detection, pose estimation, and activity recognition ML algorithms to understand human activities in the space. The proposed interactive system visualized the capture data and ML results, provided direct control to the occupants, incorporated qualitative inputs, and integrated a chatbot to communicate the recognition results and explain the working mechanism of the system. Through user studies and semi-structured interviews, I studied the dynamics between

occupants and the data system. The study results validated the prototype design and demonstrated improved occupants' understanding and shifted privacy perceptions.

To conclude, this study offers an alternative to the dominant paradigm of passive data acquisition and computation, an interactive system empowering those whose data is collected with active voices and controls. Interactive visualization, control, and understanding of the data system are made possible in a 4D diary co-authored by the human and the machine. This prototype provides the framework suitable to transform conventional passive data acquisition into interactive and sets the ground for further exploration of data transparency and ML explainability. This research investigates promises of state-of-the-art ML-powered human activity recognition systems, and reintroduces occupants' agency that can be overshadowed by automated and opaque processes of uni-directional data collection and analysis. It demonstrates a methodology guiding technology development toward transparency and empathy.

6.2 Limitations

Due to computational resources and time constraints, the current prototype has limitations in providing a fully real-time experience, as highlighted by several participants who expressed interest in interacting with it in real space to experiment with different postures and observe the changes. This interaction could enhance their understanding of how the system operates, and better serve as the vehicle for exploring the promises and boundaries of the modern ML technology. In addition, the prototype currently allows users to navigate across time by dragging timeframes on a visual interface. To improve user engagement and provide a more immersive experience, implementing an interactive 3D viewer could offer a greater degree of spatial navigation. This would enable users to select and follow specific objects of interest more closely. In addition, as the pilot study found out that people with different ML expertise expected differently, further user studies is also needed to evaluate the interface with diverse user profiles, including different background knowledge, different privacy preference, etc.

Despite the intention of the prototype is to make data acquisition and computation process transparent and interactive, the amount of data pose unintended burden for participants to go through and correct inaccurate machine recognitions. Privacy and data security issues also come inherently with recordings. These concerns could be mitigated by further reducing the frame rate and the duration of recording. If the prototype is made real time and installed in the space it is capturing, the 4D diary could be ephemeral and only shows activities in a short timeframe, for example, in the past ten minutes.

In terms of hardware integration, the prototype is presently equipped with only one sensor, a camera. There is significant potential to expand its capabilities by incorporating additional sensing devices and modalities within the existing framework. By doing so, the prototype could serve as both the visualization and control hub for the local Internet of Things (IoT) network, offering a more dynamic and responsive tool for users to interact with their environment. This integration would transform the prototype into a more comprehensive tool that not only demonstrates the power of advanced visualization techniques but also enhances the interaction between the digital and physical realms.

6.3 Future Works

In this thesis, I investigated the promises of current machine learning models recognizing human activities and used the prototype as a survey of occupants' privacy perception and the medium to explore ethical use of the technology. During the pilot study, participant suggested possible ways to make the data useful, such as providing reality-based evidence to inform space planning. The conversation for reaching a mutual understanding of the activities and spatial features could happen between occupants and the architects. The prototyped platform could serve as a medium for participatory design.

While this research provided a prototype of the interactive system, it is yet controlling the actual data acquisition and ML computation of existing consumer devices. A system with such working functionality would be greatly desired to add control to current data acquisitions. When the user interacted with the chatbot, we observed a superficial understanding of the chatbot. A

large language model that could ground its understanding logically with the given input is crucial for both accurate activity recognition and explainability of the system. In addition, when private data is of concern, a robust deep learning model is needed that can effectively handle missing modalities or data that the user is unwilling to share, without sacrificing the ML performance.

The ML system used in this study for activity recognition is just a toy example. ML developers have been trying to build a Physical AI that understands the physical world from multimodal inputs and reveals human-unperceivable hidden patterns [53]. Imagine the future where one collective brain understands the physical world, and where the artificial agents has a physical impact to humans, more transparency and privacy protection through interactive visualization, control, and understanding as demonstrated in this study will be necessary.

Bibliography

- [1] Phil Agre. 2001. Your face is not a bar code: arguments against automatic face recognition in public places. *Whole Earth* (December 2001), 74–78. Retrieved December 18, 2023 from <https://go-gale-com.cmu.idm.oclc.org/ps/i.do?p=AONE&sw=w&issn=10975268&v=2.1&it=r&id=GALE%7CA81790171&sid=googleScholar&linkaccess=abs>
- [2] Philip E. Agre. 1994. Surveillance and capture: Two models of privacy. *Inf. Soc.* (April 1994). Retrieved December 17, 2023 from <https://www.tandfonline.com/doi/abs/10.1080/01972243.1994.9960162>
- [3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Mag.* 35, 4 (December 2014), 105–120. <https://doi.org/10.1609/aimag.v35i4.2513>
- [4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, (June 2020), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, January 21, 2018. PMLR, 77–91. Retrieved December 14, 2023 from <https://proceedings.mlr.press/v81/buolamwini18a.html>
- [6] Davide Castelvecchi. 2016. Can we open the black box of AI? *Nat. News* 538, 7623 (October 2016), 20. <https://doi.org/10.1038/538020a>
- [7] Kate Crawford and Trevor Paglen. 2021. Excavating AI: the politics of images in machine learning training sets. *AI Soc.* 36, 4 (December 2021), 1105–1116. <https://doi.org/10.1007/s00146-021-01162-8>
- [8] H. Durrant-Whyte and T. Bailey. 2006. Simultaneous localization and mapping: part I. *IEEE Robot. Autom. Mag.* 13, 2 (June 2006), 99–110. <https://doi.org/10.1109/MRA.2006.1638022>
- [9] Warren J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol.* 34, 4 (December 2021), 1607–1622. <https://doi.org/10.1007/s13347-021-00477-0>

- [10] Patrick Farley. 2023. What is Spatial Analysis? - Azure AI services. Retrieved December 14, 2023 from <https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/intro-to-spatial-analysis-is-public-preview>
- [11] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif. Intell. Rev.* 56, 4 (April 2023), 3473–3504. <https://doi.org/10.1007/s10462-022-10256-8>
- [12] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Sci. Robot.* 4, 37 (December 2019), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- [13] Adam Harvey. 2011. CV Dazzle. Retrieved December 14, 2023 from <https://adam.harvey.studio/cvdazzle/>
- [14] Haojian Jin, Boyuan Guo, Rituparna Roychoudhury, Yaxing Yao, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. 2022. Exploring the Needs of Users for Supporting Privacy-Protective Behaviors in Smart Homes. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022. ACM, New Orleans LA USA, 1–19. <https://doi.org/10.1145/3491102.3517602>
- [15] Haojian Jin, Gram Liu, David Hwang, Swarun Kumar, Yuvraj Agarwal, and Jason I. Hong. 2022. Peekaboo: A Hub-Based Approach to Enable Transparency in Data Processing within Smart Homes. In *2022 IEEE Symposium on Security and Privacy (SP)*, May 2022. IEEE, San Francisco, CA, USA, 303–320. <https://doi.org/10.1109/SP46214.2022.9833629>
- [16] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, May 07, 2016. Association for Computing Machinery, New York, NY, USA, 5686–5697. <https://doi.org/10.1145/2858036.2858529>
- [17] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction.

<https://doi.org/10.48550/arXiv.2306.03092>

- [18] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 10 (October 2020), 2684–2701. <https://doi.org/10.1109/TPAMI.2019.2916873>
- [19] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*, April 23, 2020. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3313831.3376727>
- [20] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. 2021. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Proceedings of the 13th Conference on Creativity and Cognition (C&C '21)*, June 22, 2021. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3450741.3465264>
- [21] Kyle McDonald. Exhausting a Crowd. Retrieved December 15, 2023 from <https://www.exhaustingacrowd.com/london>
- [22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.
- [23] Iohanna Nicenboim, Elisa Giaccardi, and Johan Redström. 2022. From explanations to shared understandings of AI. *DRS Bienn. Conf. Ser.* (June 2022). <https://doi.org/10.21606/drs.2022.773>
- [24] Clive Norris and Gary Armstrong. 1999. CCTV and the social structuring of surveillance. *Crime Prev. Stud.* Volume 10, (1999), 157–178.
- [25] George Orwell. 1949. 1984. LeBooks Editora.
- [26] Trevor Paglen. Image Operations. Retrieved December 14, 2023 from <https://paglen.studio/2020/04/23/image-operations/>
- [27] Trevor Paglen. Training Humans. Retrieved December 18, 2023 from <https://paglen.studio/2020/04/09/kate-crawford-trevor-paglen-training-humans/>
- [28] Prasoon Patidar, Mayank Goel, and Yuvraj Agarwal. 2023. VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3 (September 2023), 1–24. <https://doi.org/10.1145/3610907>

- [29] James Pierce. 2019. Smart Home Security Cameras and Shifting Lines of Creepiness: A Design-Led Inquiry. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 02, 2019. ACM, Glasgow Scotland Uk, 1–14.
<https://doi.org/10.1145/3290605.3300275>
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision.
<https://doi.org/10.48550/arXiv.2103.00020>
- [31] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned CLIP Models are Efficient Video Learners.
<https://doi.org/10.48550/arXiv.2212.03640>
- [32] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. Retrieved December 18, 2023 from <http://arxiv.org/abs/1804.02767>
- [33] Mark Shepard. Sentient City Survival Kit. Retrieved December 14, 2023 from <https://survival.sentientcity.net/umbrella.html>
- [34] Akash Deep Singh, Brian Wang, Luis Garcia, Xiang Chen, and Mani Srivastava. 2024. Understanding factors behind IoT privacy – A user’s perspective on RF sensors. Retrieved March 8, 2024 from <http://arxiv.org/abs/2401.08037>
- [35] Kailai Sun, Peng Liu, Tian Xing, Qianchuan Zhao, and Xinwei Wang. 2022. A fusion framework for vision-based indoor occupancy estimation. *Build. Environ.* 225, (November 2022), 109631. <https://doi.org/10.1016/j.buildenv.2022.109631>
- [36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. Retrieved April 18, 2024 from <http://arxiv.org/abs/1902.09212>
- [37] Ana Tomé, Martijn Kuipers, Tiago Pinheiro, Mário Nunes, and Teresa Heitor. 2015. Space–use analysis through computer vision. *Autom. Constr.* 57, (September 2015), 80–97.
<https://doi.org/10.1016/j.autcon.2015.04.013>
- [38] Robert Twomey. 2016. A Machine for Living In. Retrieved December 18, 2023 from <https://roberttwomey.com/machine-for-living-in/>
- [39] Robert Twomey. 2018. Machines for Living. Thesis. Retrieved December 8, 2023 from <https://digital.lib.washington.edu:443/researchworks/handle/1773/41712>

- [40] Mark Weiser. 1999. The computer for the 21st century. *ACM SIGMOBILE Mob. Comput. Commun. Rev.* 3, 3 (July 1999), 3–11. <https://doi.org/10.1145/329124.329126>
- [41] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation.
- [42] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. Retrieved April 18, 2024 from <http://arxiv.org/abs/1801.07455>
- [43] Han Zhang, Yuvraj Agarwal, and Matt Fredrikson. 2022. TEO: ephemeral ownership for IoT devices to provide granular data control. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22)*, June 27, 2022. Association for Computing Machinery, New York, NY, USA, 302–315. <https://doi.org/10.1145/3498361.3539774>
- [44] Zhitao Zhong, Zhenyu Pan, Haida Tang, Junjie Wang, Fang Liu, and Jingwei Wang. 2023. Camera-based occupant behavior trajectory capture and on-site assessment of spatio-temporal transmission risk of airborne virus in a university office building. *Build. Environ.* 243, (September 2023), 110670. <https://doi.org/10.1016/j.buildenv.2023.110670>
- [45] 2021. Streamlit • A faster way to build and share data apps. Retrieved March 13, 2024 from <https://streamlit.io/>
- [46] 2023. IntelRealSense/librealsense. Retrieved December 18, 2023 from <https://github.com/IntelRealSense/librealsense>
- [47] Verkada - Simplify Physical Security at Scale - One Integrated Platform. Retrieved December 14, 2023 from <https://www.verkada.com/>
- [48] Polycam - LiDAR & 3D Scanner for iPhone & Android. Retrieved December 18, 2023 from <https://poly.cam/>
- [49] ChatGPT. Retrieved December 14, 2023 from <https://chat.openai.com>
- [50] Tesla Vision Update: Replacing Ultrasonic Sensors with Tesla Vision | Tesla Support. *Tesla*. Retrieved December 17, 2023 from <https://www.tesla.com/support/transitioning-tesla-vision>
- [51] MMAction2: OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. Retrieved April 22, 2024 from <https://github.com/open-mmlab/mmaction2?tab=readme-ov-file>

[52] OpenAI Platform. *Assistants*. Retrieved April 18, 2024 from <https://platform.openai.com>

[53] Archetype AI. Retrieved April 17, 2024 from <https://www.archetypeai.io/>

Appendix

User Study Procedures

The study consists of two activities. In Activity 1, a camera will record videos of participants performing daily group activities. In Activity 2, we will first ask about experience during Activity 1. Then participants will visualize and control the collected data and ML computation through an interactive prototype. The prototype has switches to turn on/off each individual data types and ML results, and a visualizer showing the selected data. Participants can also type in diary entries, and talk with a chatbot who has the selected information as its knowledge base. We will conduct observations and interviews about the interaction and changes in participants' perceptions.

Activity 1

Duration: 1-3 hours as the event last

Location: Margaret Morrison Carnegie Hall and TCS Hall, Carnegie Mellon University, Pittsburgh, PA 15213.

1. Invite participants to an event in a shared room on campus performing typical group activity, such as group dinner, studying, playing board games. The exact number of participants in each event depends on participants' availability.
2. Acknowledge the video recording and explain the camera: It captures RGB and depth video and runs ML computation on object detection, pose estimation, and activity recognition to understand activities happening at different locations.
3. Ask for consent, and schedule time for Activity 2 interview.
4. Record video of the entire event (1-3 hours as it lasts).

Activity 2

Duration: 30 minutes

Location: Margaret Morrison Carnegie Hall and TCS Hall, Carnegie Mellon University, Pittsburgh, PA 15213 or via Zoom

1. Start an audio-visual recording of the session.
2. Ask pre-interaction questions about experience during Activity 1.
3. Present the interactive prototype to the participant on a digital screen.

4. Allow the participants to interact freely with the interactive prototype for 10 minutes. They can turn on/off each individual data and ML computation type, visualize the selection in 2D/3D viewer, type in diary entries, and talk with a chatbot with the selected information as knowledge base. Encourage them to think aloud. Document observations on the interaction.
5. Interview the participant in a semi-structured manner for 15 minutes, and ask about changes in their awareness, transparency, and privacy perception during and after interacting with the prototype.

Interview Questions

Before interaction with the prototype:

1. In a few sentences, how would you describe the event we had?
2. How aware were you being captured by the camera?
 - a. Please rate: Very unaware, Somewhat unaware, neutral, Somewhat aware, Very aware.
3. Knowing that this camera was used to recognize activities happening in the space, what do you think was running behind the camera?
4. Behind the camera (Intel RealSense D435i RGB-D), there is a system of data collection and ML computation on activity recognition, that runs object detection, pose estimation, and activity recognition algorithms. How much do you know about these terms?
 - a. Please rate: Very unknowledgeable, Somewhat unknowledgeable, neutral, Somewhat knowledgeable, Very knowledgeable.
5. How comfortable or concerned were you being captured and analyzed by the system? Why? Did you act any differently because of that?
 - a. Please rate: Very concerned, Somewhat concerned, neutral, Somewhat comfortable, Very Comfortable.

During interaction with the prototype:

1. (Based on researcher's observation) I noticed that you _____, why was that?
2. Did you notice any success or failure of the machine recognition? What did you do because of that?

After interaction with the prototype:

1. What's your knowledge now about this system of data collection and ML computation on activity recognition?
 - a. Please rate: Very unknowledgeable, Somewhat unknowledgeable, neutral, Somewhat knowledgeable, Very knowledgeable.
 - b. If there's a change from before the interaction, which part contributes to the change: visualization, data and computation control, or the chatbot? How?
 - c. What else do you wish to see to better understand the system?
2. After you interact with this prototype, how comfortable are you being captured and analyzed by this activity recognition system?

- a. Please rate: Very concerned, Somewhat concerned, neutral, Somewhat comfortable, Very Comfortable.
 - b. If there's a change from before the interaction, which part contributes to the change: visualization, data and computation control, or the chatbot? How?
- 3. Suppose this interactive prototype is presented in the space it captures and shows data in real-time. How likely would you be to use it? What will you use it for?
 - a. Please rate: Very unlikely, Somewhat unlikely, neutral, Somewhat likely, Very likely.