Karan Ahuja | Research Statement

My research develops novel, practical and deployable sensing systems and interaction technologies that aim to overcome challenges in high-impact application areas of health sensing, extended reality, natural user interfaces, and physical computing. To develop these solutions, I draw from my diverse set of skills, including human-computer interaction (HCI), machine learning, signal processing, computer vision, sensors, and interaction design. These efforts have led to more than 20 publications in the last 5 years, with several winning awards at top-tier computer science venues such as ACM UbiComp, UIST, and CHI.

A long-standing vision in computer science has been to evolve computing devices from passive tools to truly proactive assistants, able to boost our productivity, wellness, and many other facets of our lives. Techniques that digitize the user are crucial to achieving this, allowing computers to better understand the user, and capture their behavior routines, body pose, biomarkers, and decision patterns. Today's high-end consumer devices (Fig 1, lower right quadrant) provide coarse digital representations of a user, capturing step count, pulse, respiration, and a handful of human activities, such as running and biking. They encode sparse observations of the human body - far too little information to be holistic reactive agents. On the other end, professional, high-fidelity comprehensive user digitization systems exist. For example, motion capture suits and multi-camera rigs that digitize our full body and appearance, and scanning machines such as MRI (Fig 1, upper left quadrant) capture our detailed anatomy. However, these carry significant user practicality burdens, such as financial, privacy, ergonomic, aesthetic, and instrumentation considerations, that preclude consumer use. In general, the higher the fidelity of capture, the lower the user's practicality. Fig 1 plots these two dimensions as a high-level design space. The dashed line denotes the axis occupied by conventional approaches (Fig 1, red dots), balancing user practicality and digitization fidelity.

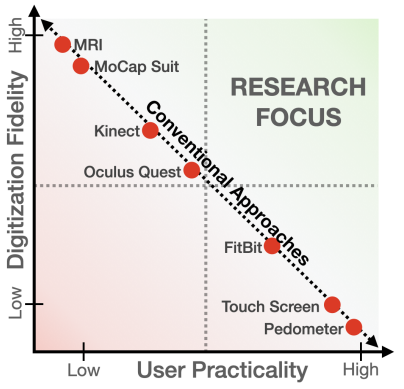

Figure 1: High-level design space of digitization fidelity vs user practicality.

My research aims to break this trend – developing sensing systems that **increase user digitization fidelity to create new and powerful computing experiences while retaining or even improving consumer practicality and accessibility,** allowing such technologies to have a societal impact. Such techniques would lie above the conventional approaches diagonal line and build towards the upper right quadrant of our design space (Fig 1, green region). In my Ph.D., I have worked across two domains (Fig 2). First, I explore techniques to **advance activity recognition** [1, 2, 3, 4, 5, 6] to create a vocabulary to digitize and express their behavior routine. Second, I develop **pose-sensing systems that are mobile** [7, 8, 9, 10, 11, 12, 13, 14] and provide a holistic representation of the user on the go. My work has received multiple awards, has been deployed in the wild, and has been licensed for commercialization by Fortune 100 companies.

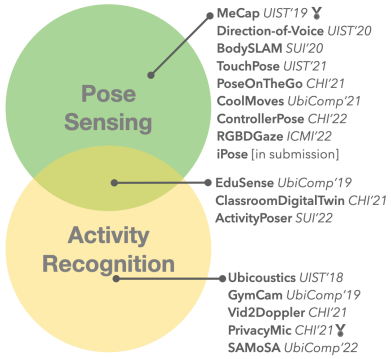

Figure 2: My research unlocks new opportunities in two domains.

Advancing Human Activity Recognition

Activities are the basic building blocks representing a user's behavior, enabling numerous applications such as real-time task assistance, context-aware computing, eldercare, personal informatics, and health sensing to name a few. However, current computing devices have limited knowledge of their user's physical and social context. For example, a high-end smartwatch can only track a few primitive activities such as walking, running, cycling, and swimming. Likewise, a smart speaker sitting on a kitchen countertop cannot figure out if it is in a kitchen, let alone know what a user is doing in the kitchen. In response, I created **Ubicoustics** [1], a novel, real-time, sound-based activity recognition system. It leverages microphones from existing everyday devices (watches, smart speakers, IoT sensors) to digitize 30 activities of interest (e.g., typing, coughing, chopping) without needing any user calibration or environment-specific training data. Ubicoustics is open-sourced and has resulted in commercial licenses to two companies for integration into their ecosystem and has been shipped as a product feature.

When developing Ubicoustics, I noticed that such an always-on acoustic activity recognition system is power-hungry due to its need to sample audio data at high rates (16 kHz and above) for accurate inferences. This increases the power burden, especially on devices such as smartwatches, where the battery cannot be made much larger. In response, I created **SAMoSA** [2] - Sensing

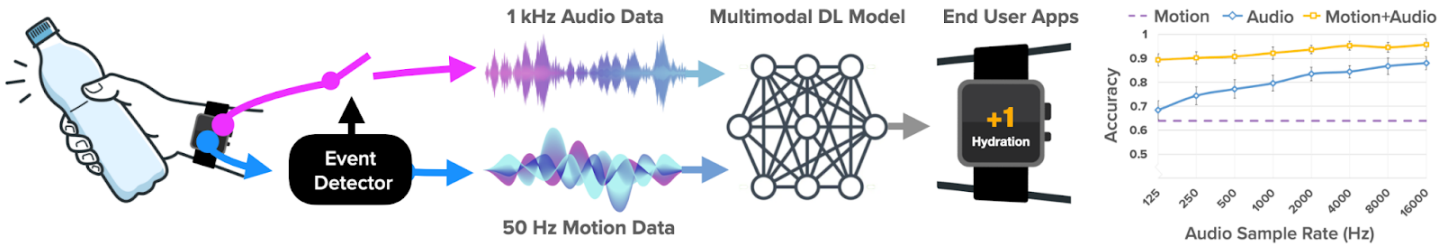

Figure 3: SAMoSA senses activity using power-efficient and privacy-sensitive (≤ 1 kHz) audio data and 50 Hz IMU motion data.

Activities with Motion and Subsampled Audio (Fig 3). To enable this, I relied on compute-optimized IMUs sampled at 50 Hz to act as a trigger for detecting activity events rather than an always-on audio-based activity classification system. Once detected, I employed a multimodal temporal deep learning model that augments the motion data with subsampled audio data (captured at ≤ 1 kHz) that reduces power consumption on mobile devices, while also rendering speech content unintelligible, thereby being more sensitive to the user's privacy. This multimodal model at its lowest audio frequency of 125 Hz is more accurate than a 16 kHz audio-only model or an IMU-only model (Fig 3, accuracy graph). This multimodal model runs locally on the smartwatch, thereby enabling **minimally invasive, real-time, fine-grained activity sensing on resource-constrained devices.**

A complementary approach to using mobile sensors is to embed sensors in the environment. This enables large-scale sensing for multiple users outside of the home and office settings, reducing the need for coordination and compliance with each user. This is typically done by upgrading each device in the environment to be "smart". However, such sensing carries a significant upgrade cost, is limited to the appliance itself, and is rarely interoperable, thus dampening adoption and defeating the purpose of a holistic view. In contrast, I strive to create **rich sensing solutions that require minimal environmental instrumentation**, thereby allowing for wide-scale deployability. For example, my work **GymCam** [3] uses only a single camera in a gym to detect, segment, and recognize various simultaneous exercises by multiple users. **EduSense** [6] is a video and audio-based scalable classroom sensing system powering a suite of data-driven instructional aids by sensing pedagogically motivated features such as hand raises, body posture, body accelerometry, aggregate student gaze, and speech acts. It is open-sourced and **deployed in over 45 classrooms across 3 universities**, providing real-time feedback to teachers. To ensure user privacy, both EduSense and GymCam process the data on the edge, discard raw video frames and only save the featurized data. As these audio and vision-based sensing paradigms migrate out of lab settings and see commercial viability, there is an aspect of trust associated with the manufacturer and developer of the sensors. For example, there can be malintent during data collection, false advertising of data recording routines, or even data leaks. To curb these, I explore **novel sensing paradigms that are inherently privacy-sensitive.**

We address this challenge by introducing **Vid2Doppler** [4]. It makes use of mmWave doppler radar to power an inherently privacy-sensitive activity recognition framework. The mmWave radar sensor returns a radial velocity across time map (Fig 4, Doppler plots) devoid of identifiable information captured by mics and cameras. The challenge, though, is that there is little existing Doppler radar data to train a human activity recognition classifier. Thus I created a software pipeline that converts videos of human activities into realistic, but synthetic doppler radar data. This involves understanding the underlying RF phenomenon in the real world, creating theoretical and mathematical models to simulate its signal profile, developing a learning framework for activity recognition, and combining all these modules to create an end-to-end sensing system. Vid2Doppler achieves an accuracy of 81.4% across 12 exemplary activities when trained on only synthetically-generated data. The accuracy further improves to 93.4% when augmented with real-world calibration Doppler data. Vid2Doppler is open-sourced offering the community an important stepping-stone towards significantly reducing the burden of training privacy-sensitive user digitization systems.

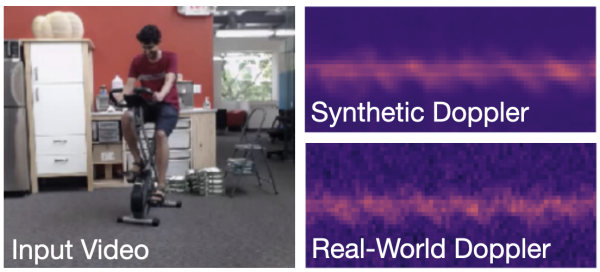

Figure 4: Vid2Doppler generates synthetic doppler from input video. Real-world doppler shown for reference.

Practical and Rich Pose Sensing

Fine-grained activities while powerful are still discrete and pre-defined. On the other hand, pose offers a continuous and more comprehensive representation of the user. It enables personalized health sensing, fitness, holoportation, social avatars, novel interactive experiences, gaming, and also has a trickle-down effect on activity recognition. A lot of efforts have focused on using

Figure 5: Full-body animoji (left) and third-person shooter game (right) enabled by Pose-on-the-Go.

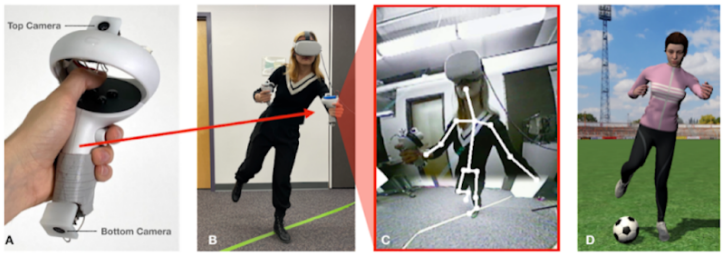

Figure 6: ControllerPose uses two fisheye cameras on each VR handheld controller to enable full-body 3D pose tracking.

external infrastructures or specialized hardware for pose sensing (e.g. Kinect, Vicon), thus constraining their capabilities to the particular location of deployment. In contrast, my research explores **full-body digitization through mobile consumer devices**, affording my approach with superior practicality and ability to work in unconstrained settings in the wild.

To this extent, I developed **Pose-on-the-Go** [8] to bring full-body pose estimation to an unmodified smartphone. Here, I use a sensor fusion-based approach, leveraging almost every sensor in smartphones - front camera to track the user's head pose, rear camera for visual odometry, user-facing depth camera for shoulder orientation, capacitive touchscreen to opportunistically track the hand touching the screen and IMU to orient the hand holding the phone. I fused data from these disparate sensors to generate a real-time, animated skeleton of the user as they operate their phone. I then use inverse kinematics to estimate and animate the probable body poses of the body key points for which I have no direct sensory data (e.g., the angle of the elbow joint). This full-body pose, while coarse, enables interesting applications such as social avatars (Fig 5 left), and immersive games (Fig 5 right). Pose-on-the-Go is **available on the** iOS App Store **as a software-only app** that can run directly on existing devices (iPhone X and above), allowing developers to build pose-enabled apps without additional hardware.

Pose-on-the-Go uses a smartphone, all the while, a user typically carries multiple mobile devices (e.g. watch, earbuds) with them. In **iPose** [13], I explore the feasibility of getting a full-body pose using solely the onboard IMUs from these devices (smartphones, smartwatches, and earbuds). Here I solve several challenges such as noisy IMU data from off-the-shelf devices, different sampling rates across devices, and a sparse and changing number of instrumentation points on the go. For example, a user can take a phone from their pocket into their hand (i.e., moving from a left hip placement to the right hand), or add to the number of sensed points by wearing their earbuds. The iPose pipeline first detects whatever subset of devices are available, tracks their on-body location, and then produces an optimal pose estimate from their IMU data (which can potentially even be from a single device). The adaptive and mobile nature of iPose enables **passive and longitudinal sensing of the user**, especially making it well-suited for health and wellness applications.

While iPose and Pose-on-the-Go use sensors already present on today's consumer devices, **ControllerPose** [7] informs the design of future virtual reality (VR) devices. It increases the digitization fidelity of current consumer VR, which only track the user's head and hand, and expands it to enable full-body pose capture. To achieve this, I integrate cameras into handheld controllers (Fig 6A), where batteries, computation, and wireless communication already exist. By virtue of the hands operating in front of the user during many VR interactions, the controller-borne cameras captured a superior view of the body for digitization (Fig 6B). The pipeline combines multiple camera views, performs 3D body pose estimation (Fig 6C), fuses this data with the tracked positions of the heads and hands to control a rigged human model, and outputs the resulting user avatar to enable novel end-user applications. For example, users can now stomp, lean, squat, lunge, balance, and perform many other leg-driven interactions in VR (e.g., kicking a soccer ball in Fig 6D).

Future Research

Through the efforts of my Ph.D., I have created powerful user digitization systems that are minimally invasive, mobile and deployable in-the-wild. However, much work needs to be done, which I look forward to investigating as a faculty member. I am also extremely excited to start new collaborations with future colleagues and explore related areas of interest, described below:

- **Health sensing and accessibility:** My experience in creating practical, rich, and deployable user digitization solutions can be directly applied to health sensing and accessibility. For example, I have started experimenting using iPose [13] to characterize the user's motion profile via pose to measure hyperactivity in children as a symptom of ADHD. I will continue and expand my collaborations with medical practitioners and domain experts to grow user digitization to

encapsulate physiological biomarkers such as heart rate and respiration rate. I further plan to build new sensors to digitize facets such as a user's muscle utilization, tissue deformation, user intent, affect, and brain activity to name a few; while keeping user practicality in mind. This would create holistic virtual representations that enable early disease detection, personalized health tracking and inform behavior-driven interventions.

- **New avenues for user digitization beyond pose:** The goal of user digitization is to create human digital twins that are an exact and faithful representation of our physical selves. While I have started working towards this by digitizing activity and pose, many other dimensions need to be captured such as facial attributes, fine-grained hand pose, skin texture, clothing, and mesh deformations to name a few. This requires the development of sensing paradigms that are rich and high in resolution, but not at the cost of user practicality and affordability. Such technologies would transform the domains of telepresence, extended reality, and remote medical assessment, among others.

- **Interactions between user and environment:** User digitization unlocks the opportunity for novel and seamless interaction between users and smart environments [11, 14]. For example, my prior work, Direction-of-Voice [11] utilizes speech as a directional communication channel to estimate the head orientation of a user for addressing smart device ecosystems. [14] explores tracking the user's alongside their immediate environment for gaze estimation. I plan to expand on this initial body of work and further investigate approaches that digitize the user and contextualize their environment to enable intuitive, rapid, and natural interaction across a multitude of computational ecosystems.

References

1. Gierad Laput, **Karan Ahuja**, Mayank Goel, and Chris Harrison. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18). Video Code
2. Vimal Mollyn, **Karan Ahuja**, Dhruv Verma, Chris Harrison, and Mayank Goel. SAMoSA: Sensing Activities with Motion and Subsampled Audio. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT) 6, 3, Article 132 (September 2022). UbiComp '22. Video
3. **Karan Ahuja***, Rushil Khurana*, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. GymCam: Detecting, Recognizing and Tracking Simultaneous Exercises in Unconstrained Scenes. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Volume 2, Issue 4, Article 185 (December 2018), UbiComp '19. (* Equal Contribution) Video
4. **Karan Ahuja**, Yue Jiang, Mayank Goel, and Chris Harrison. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems (CHI '21). Video Code
5. Yasha Iravantchi, **Karan Ahuja**, Mayank Goel, Chris Harrison, and Alanson Sample. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. In Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems (CHI '21). ***Best Paper Nomination*** Video
6. **Karan Ahuja***, Dohyun Kim*, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. EduSense: Practical Classroom Sensing at Scale. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Volume 3, Issue 3, Article 71 (September 2019). UbiComp '19. (* Equal Contribution) Code
7. **Karan Ahuja**, Vivian Shen, Cathy Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. ControllerPose: Inside-Out Body Capture with VR Controller Cameras. Proceedings of the 2022 ACM Conference on Human Factors in Computing Systems (CHI '22). Video
8. **Karan Ahuja**, Sven Mayer, Mayank Goel, and Chris Harrison. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems (CHI '21). Video
9. **Karan Ahuja**, Chris Harrison, Mayank Goel, and Robert Xiao. MeCap: Whole-Body Digitization for Low-Cost VR/AR Headsets. Proceedings of 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19). ***Best Paper Nomination*** Video
10. **Karan Ahuja**, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D. Wilson. CoolMoves: User Motion Accentuation in Virtual Reality. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Volume 5, Issue 2, Article 52 (June 2021), 23 pages. UbiComp '21. Video
11. **Karan Ahuja**, Andy Kong, Mayank Goel, and Chris Harrison. Direction-of-Voice (DoV) Estimation for Intuitive Speech Interaction with Smart Devices Ecosystems. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20). Video Code
12. **Karan Ahuja**, Paul Streli, and Christian Holz. TouchPose: Hand Pose Prediction, Depth Estimation, and Touch Classification from Capacitive Images. Proceedings of 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21). Video Code
13. Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and **Karan Ahuja**. iPose: Body Pose Estimation using Phones, Watches, and Earbuds. *Under Review*. Proceedings of 2023 ACM CHI Conference on Human Factors in Computing Systems (CHI '23).
14. **Karan Ahuja*,** Deval Shah*, Sujeath Pareddy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. 2021. Classroom Digital Twins with Instrumentation-Free Gaze Tracking. In Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems (CHI '21). (* Equal Contribution) Video