Junxiao Shen shawn.shenjx@yahoo.com

Research Statement

In the dynamic realm of technological advancement, my research operates at the pivotal intersection of **machine learning** and **mixed reality (MR)**, primarily aimed at augmenting human abilities. This work is grounded in the conviction that these two unique yet harmonious technological fields, when optimally integrated, can amplify **human action**, **cognition**, and **perception**.

Machine learning, as the software facet of this paradigm, holds the promise of enabling intelligent decision-making and adaptive learning processes. Simultaneously, **mixed reality**, acting as the hardware platform, offers an immersive, interactive environment that effortlessly merges physical and digital components.

By harnessing these technologies, my research endeavors to **enhance human abilities** in a way that is both groundbreaking and unparalleled. The overarching aim of m y s research is to surpass traditional human capability limits, thus empowering individuals to perceive, comprehend, and engage with their surroundings with increased precision and efficiency. This aim is not solely about advancing technology, but fundamentally about amplifying human potential.

1 Current Research

The subsequent section outlines my ongoing research endeavors, concentrated on human augmentation through machine learning and mixed reality.

1.1 Action Augmentation

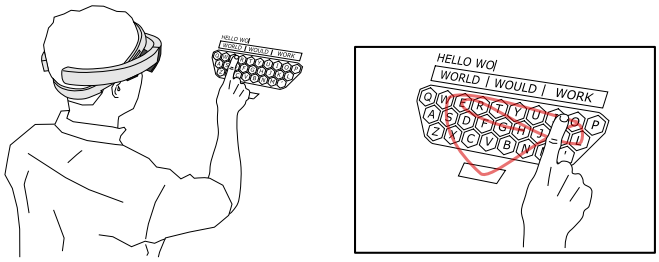
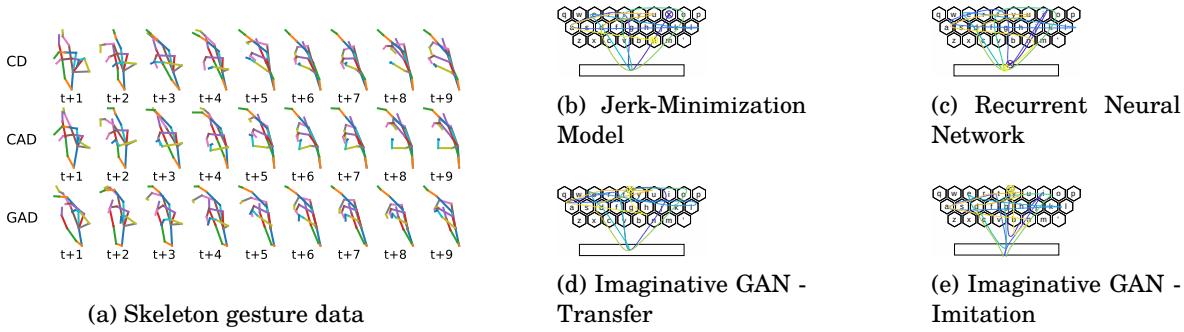


Figure 1: Illustration of a user gesturing on a mid-air gesture keyboard in AR (left), and illustration of a user's gesture trajectory (right).

In the sphere of action augmentation, my current research is explicitly centered on the use of bare hands as a conduit for interaction within mixed reality systems. Leveraging machine learning, I aim to optimize this interaction, creating a frictionless interface that enhances human actions in a blended environment of the virtual and physical worlds. My focus is honed on two key applications: First, I explore the concept of a mid-air gesture keyboard (see Figure 1), which allows users to input text via intuitive hand movements. This approach seeks to redefine traditional methods of text entry, making it more efficient and natural within a mixed reality context. Secondly, I delve into the broader domain of interaction through gesture recognition. This involves the machine learning algorithms learning, predicting, and responding to a wide array of user gestures, thereby facilitating a more dynamic and responsive interaction within the mixed reality system. Through these dual avenues of research, I aim to augment human action capabilities, fostering a future where the seamless integration of the virtual and physical worlds is not just possible, but commonplace.

However, I face several challenges, including data sparsity due to limited MR headset use, the complexities of user interface (UI) and user experience (UX) design in a three-dimensional MR environment, and the difficulty for designers unfamiliar with machine learning to apply it effectively. To address these



(a) Skeleton gesture data

(b) Jerk-Minimization Model

(c) Recurrent Neural Network

(d) Imaginative GAN - Transfer

(e) Imaginative GAN - Imitation

Figure 2: **(a)**: Visualized gesture examples illustrating the difference between the clean data (CD) which is raw data; and the conventional augmented data (CAD) generated from conventional approaches such as adding noise, shifting and scaling etc; and GAN augmented data (GAD) synthesized from Imaginative GAN. **(b)-(e)**: Visualized gesture trajectories illustrating the different samples synthesized from four different generative models.

issues, I tackle the central research hypothesis from five directions: mitigating data sparsity, optimizing UI and UX in mid-air gesture keyboard, recognizing gestures robustly, and democratizing machine learning for designing gestures.

I addressed the data sparsity issue by introducing a generative model, **Imaginative GAN**, based on the Generative Adversarial Network (GAN) for synthesizing skeleton gestural data [2]. The performance of this model surpasses traditional data augmentation methods, as demonstrated by comparative analysis and Figure 2a. Additionally, the proposed model is employed to generate synthetic trajectory data on a mid-air gesture keyboard [3]. It is compared with other proposed methods as shown in Figure 2b to Figure 2e, not with the intention to establish superiority, but to highlight different characteristics that make it suitable for various scenarios where synthetic data is utilized.

I then respond to UI optimization by unveiling **AdaptiKeyboard** [7], a mid-air gesture keyboard that employs multi-objective Bayesian optimization (MOBO) to adjust its layout size, optimizing both speed and precision. The MOBO process is illustrated by Figure 3. By tailoring to individual user habits and preferences, AdaptiKeyboard method boosts text entry speed and accuracy by **14.4%** and **13.8%** respectively, compared to a baseline design on the HoloLens 2 with constant size.

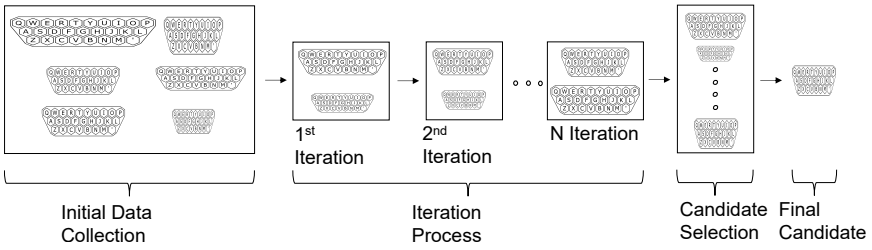


Figure 3: Six different layout candidates with varying sizes are generated and evaluated by the user, then the process will introduce another two candidates for evaluation and this forms one iteration. Such an iteration needs to be repeated four to five times. Finally, a selection of candidates exhibiting optimal speed-accuracy trade-offs are generated. The user can then select the final optimal candidate based on their own preferences.

The UX optimization challenge is addressed by presenting a unique decoding technique, a **neural motion model**, that converts 3-D gesture paths drawn on a mid-air gesture keyboard directly into the desired text [4]. The proposed decoder can consequently support novel designs, including the **removal of visual feedback** and **relaxed delimitation threshold**, which could lead to noisy input and subsequently result in the failure of traditional decoders using shape-matching algorithms. The

efficiency of the decoder is validated through studies demonstrating significant enhancements in entry speeds and minimized error rates, even in the absence of visual feedback. We found that, with adequate practice, this proposed decoder can facilitate a mid-air gesture keyboard with the aforementioned novel designs to achieve a typing speed of over 30 words per minute which is comparable to the typing speed on a physical keyboard. The demonstration of text entry on the mid-air gesture keyboard using the proposed decoder can be viewed with the provided link[1].

I address the challenge of gesture recognition by introducing a novel **key gesture spotting algorithm**, designed to be robust and low-latency. This algorithm has shown high precision and the ability for early detection across four diverse hand skeleton gesture datasets [6]. Furthermore, we contribute to the democratization of machine learning by integrating this algorithm, along with a variety of models and unique data processing and augmentation techniques, into an intuitive Graphical User Interface (GUI) and an Application Programming Interface (API). Collectively, we refer to this as **Gesture Spotter**. User studies have confirmed that these tools enable developers to efficiently build custom gesture recognition models, thereby accelerating the gesture design process. The demonstration of using the Gesture Spotter GUI to create a gesture recognition model that can recognize rock, paper, scissors, and other background activities (null gestures) in just 10 minutes can be seen in the provided link[2].

During my internship at Meta Reality Lab, I delved deeper into the subject of gesture recognition, proposing a more advanced concept known as **open-world gesture recognition**. This approach encourages models to adaptively learn new data from new gestures, different users, and novel situations. In the open-world scenario, gesture data can deviate from lab-collected data, leading to the potential failure of traditional recognition models. Therefore, I utilized **continual learning** to allow the model to persistently learn new data patterns in the open-world [1]. Moreover, I proposed an **automatic gesture annotation framework** capable of concurrently recognizing gesture classes and determining gesture nucleus positions [8]. This enables open-world gesture data collection without requiring users to explicitly annotate the gestures. Both the continual learning framework on gesture recognition and the automatic gesture annotation framework have found extensive application within Meta.

1.2 Cognition Augmentation

In my research on cognition augmentation, I focus on the development of large language models that can aid in memory augmentation and predictive text entry.

In the realm of memory augmentation, I put forward a novel **memory augmentation** system that employs an **encode-store-retrieve** process, capitalizing on augmented reality head-mounted displays for life logging, capturing, and preserving egocentric videos [5]. Given the substantial volume of video data produced, the system utilizes natural language encoding for video data, storing them in a vector database. This method fine-tunes a vision language model on egocentric data for the encoding process and enables natural language querying via vector database retrieval. When evaluated using the QA-Ego4D dataset, this system surpassed traditional machine learning models and garnered higher response scores than the user's own memory in real-life episodic memory tasks during a user study.

For predictive text entry, I propose **KWickChat** [9], a multi-turn augmentative and alternative communication (AAC) dialogue system fine-tuned from **Generative Pretrained Transformer 2 (GPT-2)** designed for nonspeaking individuals with motor disabilities. KWickChat aims to bridge the communication gap by providing a sentence-based text entry system that generates suitable sentences based on keyword entry. Underpinned by the GPT-2 language model, KWickChat leverages dialogue history and persona tags to improve response quality. The system showed a keystroke savings of around 71%

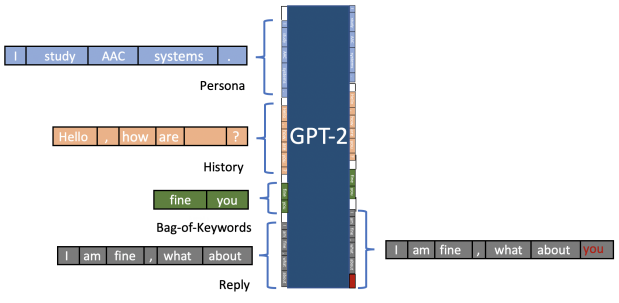


Figure 5: Structure of the KWickChat model.

[1]Mid-Air Gesture Keyboard: https://youtu.be/yGVWpzkL5BE
[2]Gesture Spotter GUI: https://youtu.be/ctY8-VT6MDY

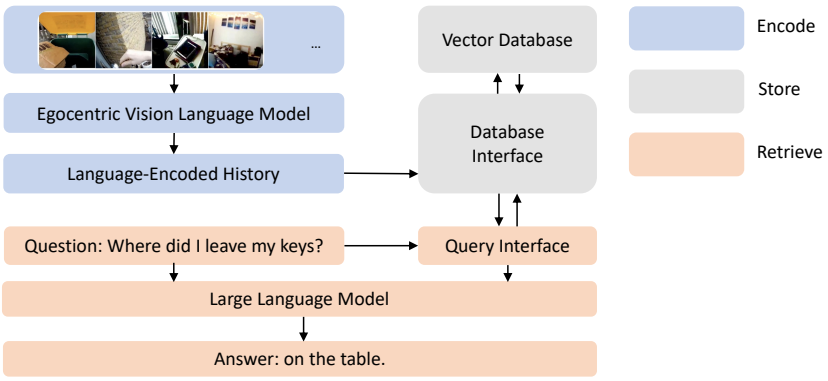


Figure 4: Workflow of the Memory Augmentation System

with a word prediction error rate of 0.65. Furthermore, semantic consistency between generated and reference sentences was rated highly by human judges, indicating the system's effectiveness.

These two applications of large language models serve as powerful **artificial intelligence (AI) agents** for cognition augmentation, enhancing human memory and communication capabilities.

2 Future Research

Expanding upon my current research in cognition and action augmentation, I plan to delve into the realm of art and design. I aim to use gesture recognition algorithms to transform bare hands into **art tools**, a novel form of action augmentation. Additionally, I'm interested in leveraging generative models for **AI-assisted art design** and **AI-generated art**, constituting a form of cognition augmentation.

Furthermore, my future research will focus on **perception augmentation**, specifically visual and **auditory enhancement**, using mixed reality and machine learning. These advancements hold immense potential for improving **accessibility** for individuals with sensory impairments, a key aspect of my future research endeavors.

For **visual augmentation**, I aim to explore how mixed reality devices equipped with advanced computer vision algorithms can be used to augment human vision. This could involve developing technologies that enhance our ability to perceive fine details, improve depth perception, or even enable us to see in low-light conditions. For instance, by integrating machine learning algorithms that can process and enhance real-time video feeds, we could potentially create mixed reality systems that augment our natural vision, allowing us to see with greater clarity and precision.

In terms of **auditory augmentation**, I am interested in how we can use mixed reality and machine learning to enhance our hearing capabilities. This could involve developing context-aware algorithms that can filter and amplify specific sounds based on the user's environment and preferences. For instance, a mixed reality system could be designed to recognize and amplify the voices of specific individuals in a crowded room, or to filter out background noise in a busy street.

Lastly, **explainable-AI** plays a pivotal role in human augmentation research. It ensures the transparency and understandability of AI systems, which is crucial when these technologies are used to augment human abilities [10]. By making AI's decision-making processes interpretable, users can trust and effectively utilize these enhancements. Furthermore, it allows for continuous improvement and adaptation, ensuring that AI-driven human augmentation is safe, efficient, and beneficial for all users.

To summarize, my academic and professional journey has endowed me with a distinct mix of skills, encompassing advanced research, funding procurement, and industry partnership.

References

[1] **Junxiao Shen**, Matthias De Lange, Xuhai Xu, Ran Tan, Naveen Suda, and Evan Strasnick. Quantitative exploration of continual learning methods for wrist-worn open-world gesture recognition using a design engineering approach. Under Submission.

[2] **Junxiao Shen**, John Dudley, and Per Ola Kristensson. The imaginative generative adversarial network: Automatic data augmentation for dynamic skeleton-based hand gesture and human action recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2021.

[3] **Junxiao Shen**, John Dudley, and Per Ola Kristensson. Simulating realistic human motion trajectories of mid-air gesture typing. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 393–402. IEEE, 2021.

[4] **Junxiao Shen**, John Dudley, and Per Ola Kristensson. Fast and robust mid-air gesture typing for ar headsets using 3d trajectory decoding. *IEEE Transactions on Visualization and Computer Graphics (TVCG), forthcoming*, 2023.

[5] **Junxiao Shen**, John Dudley, and Per Ola Kristensson. Encode-store-retrieve: Enhancing memory augmentation through language-encoded egocentric perception. Under Submission.

[6] **Junxiao Shen**, John Dudley, George Mo, and Per Ola Kristensson. Gesture spotter: A rapid prototyping tool for key gesture spotting in virtual and augmented reality applications. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 28(11):3618–3628, 2022.

[7] **Junxiao Shen**, Jinghui Hu, John J Dudley, and Per Ola Kristensson. Personalization of a mid-air gesture keyboard using multi-objective bayesian optimization. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 702–710. IEEE, 2022.

[8] **Junxiao Shen**, Xuhai Xu, Ran Tan, and Evan Strasnick. Towards simultaneous gesture classification and localization with an automatic gesture annotation model. Under Submission.

[9] **Junxiao Shen**, Boyin Yang, John J Dudley, and Per Ola Kristensson. Kwickchat: A multi-turn dialogue system for aac using context-aware sentence generation by bag-of-keywords. In *27th International Conference on Intelligent User Interfaces (IUI)*, pages 853–867, 2022.

[10] Xuhai Xu, Anna Yu, Tanya R Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, **Junxiao Shen**, et al. Xair: A framework of explainable ai in augmented reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–30, 2023.