Handwritten Text Entry in Virtual Reality Using Gesture Recognition and Word Prediction

João Silva[1], **Kaique Carvalho**[1], **Luciana Cardoso**[1], **Juliana Félix**[23], **Fabrizzio Soares**[2], **Thamer Horbylon Nascimento**[12]

[1]Instituto Federal Goiano (IF Goiano) – Campus Iporá – Iporá – GO – Brazil

[2]Universidade Federal de Goiás (UFG) – Instituto de Informática – Goiânia – GO, Brazil

[3]Escola Politécnica e de Artes, Pontifícia Universidade Católica de Goiás – GO – Brazil

{joao.victor3, kaique.carvalho}@estudante.ifgoiano.edu.br,

luciana.cardoso@ifgoiano.edu.br, {julianafelix, fabrizzio}@ufg.br,

thamer.nascimento@ifgoiano.edu.br

Abstract. *This work presents an approach for text entry in virtual reality (VR) environments, using handwritten letters drawn in the air as a form of natural interaction. The system was developed for the Meta Quest 2 device and is composed of different integrated modules: real-time capture of gestures performed by the user, character recognition using a convolutional neural network (CNN) trained with the EMNIST dataset, and the construction of words through a Trie structure, which enables efficient term search based on recognized letters. Furthermore, the final selection of words is performed based on their usage frequency, which allows prioritizing more probable terms within a common linguistic context. The method allows for complete word input in VR, with consistent performance in the identification of individual letters and automatic suggestion generation, demonstrating that it can provide a fluid, intuitive experience compatible with the immersive interaction proposal in three-dimensional environments.*

1. Introduction

The growing use of wearable devices and virtual reality (VR) has driven new forms of interaction between humans and computational systems [Katona 2021]. Ubiquitous computing, as proposed by [Weiser 1991], promotes natural, continuous, and context-aware interactions. In immersive environments, there is a demand for input methods that do not interrupt the user experience.

Inspired by enactive systems [Kaipainen et al. 2011] and grounded in the theory of enactive cognition [Varela et al. 1992], this work proposes a gesture-based text entry method in which users draw cursive letters in the air using the Meta Quest 2 device. These gestures are captured and recognized by a convolutional neural network (CNN) trained on the EMNIST dataset [Cohen et al. 2017], enabling accurate letter recognition in three-dimensional environments.

This approach aims to make text entry in virtual reality more natural and fluid by removing the need for traditional graphical interfaces such as virtual keyboards. By prioritizing freehand gestural movement and intelligent interpretation of inputs, the system seeks to reduce cognitive load and foster a more immersive, intuitive, and ergonomic experience for users.

Furthermore, the combination of pattern recognition techniques, optimized data structures (such as the Trie), and linguistic frequency analysis allows for real-time word suggestions, accelerating text composition. This may make the system suitable for applications such as VR communication, note-taking in immersive environments, or contextual commands in interactive systems. Figure 1 illustrates this process, showing how words are built letter by letter with real-time suggestions from the recognized prefix.

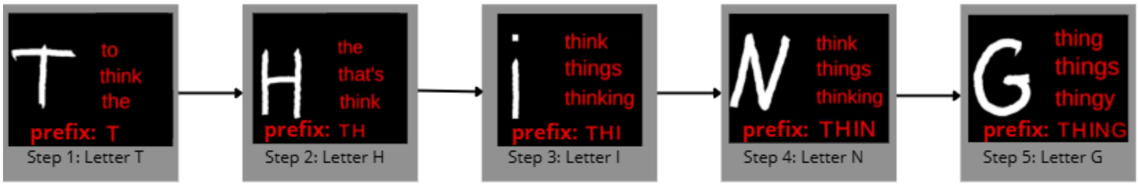

Figure 1. Illustration of the word construction process for "THING" with suggestions based on recognized prefixes, letter by letter.

2. Related Work

Several methods have been developed to facilitate text entry on devices with unconventional interfaces, such as small screens and three-dimensional environments, aiming to overcome the limitations of traditional keyboards, especially in immersive systems.

Gesture-based techniques for VR and wearable devices have emerged as natural alternatives for interaction. For example, continuous gesture recognition on smartwatches integrated with Google Cardboard was proposed in [Nascimento et al. 2017a], while the performance of gesture-based mid-air text entry with different hand postures in virtual environments was investigated in [Wang et al. 2021]. Personal authentication and recognition of aerial Hiragana input using deep neural networks was investigated in [Mimura et al. 2021], and a CNN-based air-writing recognition framework for multilinguistic characters and digits was presented in [Kumar et al. 2022]. Multimodal interfaces combining facial touch and gestures have also shown potential to improve usability [Gugenheimer et al. 2016].

Ergonomic and precision limitations of physical controllers have been discussed in [Boletsis and Kongsvik 2019a], which highlight fatigue and inaccuracy associated with prolonged raycasting. To address these, an ambiguous keyboard inspired by virtual drums was proposed in [Boletsis and Kongsvik 2019b]. Handwriting-based text entry using Leap Motion was studied in [Tsuchida et al. 2015], and a smart ring-based system for mid-air writing was introduced in [Shen et al. 2024].

Handwritten character recognition has employed techniques such as neural networks [Elmgren 2017, Lu et al. 2019], machine learning for trajectory recognition [Chen et al. 2022], association rules for handwritten character recognition [Carvalho 2000], and image preprocessing [Blanco Junior 2022]. Naïve Bayes classification based on geometric shapes was used in [Nascimento et al. 2017b] to recognize letters on smartwatches. More recent work explores immersive text editing strategies. Efficient mid-air text correction using gesture-based strategies was proposed in [Dudley et al. 2024], and probabilistic input with floating QWERTY keyboards was analyzed in [Dudley et al. 2023]. An adaptation of the flick technique for Japanese characters was implemented in [Monobe and Ohishi 2025].

These approaches reflect a convergence of ergonomics, AI, and gesture-based interaction in the evolution of VR text entry. The present work distinguishes itself by integrating these techniques into a single pipeline, combining 3D gesture capture, robust CNN-based recognition, and dynamic word suggestion using a Trie structure weighted by usage frequency—originally proposed by [Nascimento et al. 2023] for text entry on smartwatches and here adapted to immersive VR environments—offering a natural and effective text entry solution.

3. Proposed Method

This work proposes a method for text entry in virtual reality through freehand cursive gestures captured in real time using the Meta Quest 2. Users draw letters with natural hand movements, which are rendered as 3D trajectories, converted into 2D images, and sent to a remote server for recognition via a convolutional neural network (CNN). The recognized characters form a prefix buffer used to query a Trie data structure containing English words weighted by usage frequency, enabling predictive word suggestions to the user within the virtual environment. This integration aims to provide an immersive, fluid, and efficient text entry experience.

3.1. System Overview and Architecture

The system consists of four modules: (1) gesture capture, where the user draws letters in VR with the controller and the 3D trajectory is visualized and converted into 2D images for transmission; (2) image processing and character recognition on the server using a CNN trained on the EMNIST Letters dataset; (3) word generation through a Trie structure that efficiently retrieves possible completions based on the current prefix; and (4) user interaction, which displays word suggestions ranked according to frequencies derived from linguistic corpora, allowing for selection or continuation of input. Figure 2 shows the complete system architecture and communication among these modules.

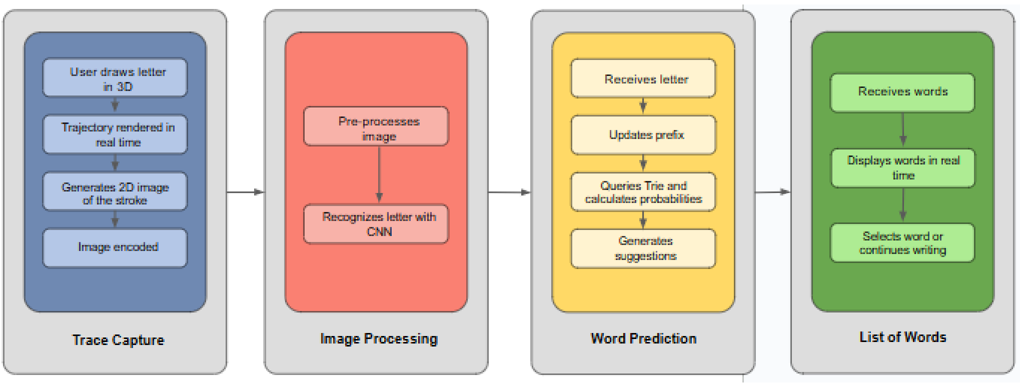
User draws letter in 3D
Trajectory rendered in real time
Generates 2D image of the stroke
Image encoded
Trace Capture
Pre-processes image
Recognizes letter with CNN
Image Processing
Receives letter
Updates prefix
Queries Trie and calculates probabilities
Generates suggestions
Word Prediction
Receives words
Displays words in real time
Selects word or continues writing
List of Words

Figure 2. System architecture showing interaction among the modules of the proposed method.

3.2. Handwriting Capture and Character Recognition

Users initiate gesture capture by pressing a controller button, recording the 3D position of the brush tip rendered by a LineRenderer. Upon completion and user confirmation, an orthographic camera captures the trace as a 2D PNG image, which is encoded and sent via HTTP to the server. The CNN classifies the image into one of the 26 English alphabet letters, adding it to the current prefix.

3.3. Word Suggestion and Prediction

As each letter is recognized and appended to the prefix under construction, the system uses a Trie data structure to efficiently retrieve all words starting with that prefix. The Trie was preloaded with a large English vocabulary associated with usage frequencies extracted from reliable linguistic corpora, allowing the determination of the likelihood of each word.

To prioritize suggestions, each candidate word w has its probability calculated by normalizing its frequency $f(w)$ relative to the sum of the frequencies of all candidate words $\{w_i\}_{i=1}^{n}$ matching the current prefix, according to the formula:

$$P(w) = \frac{f(w)}{\sum_{i=1}^{n} f(w_i)}.$$

This probabilistic approach ensures that more common and likely words are presented first, increasing typing efficiency and reducing user effort. The system displays the top three most probable words in real time within the virtual environment, allowing the user to quickly select the desired option or continue writing to refine the suggestions.

3.4. User Interaction

The user interface is integrated into the virtual environment, clearly and accessibly displaying the word suggestions generated by the server for selection. The user can navigate through the options using the VR controller or continue drawing letters to expand the prefix and refine suggestions.

This continuous interaction allows a natural and fluid input experience, combining gesture recognition with intelligent support to accelerate text composition. Immediate suggestion display and selection capability facilitate rapid word and phrase construction without traditional typing, aligning with the immersive characteristics of VR. Figure 3 shows the process of capturing 3D gestures, converting them to 2D images, and sending them to the server for recognition.

4. System Evaluation and Preliminary Results

The implemented method allows users to input complete words in VR environments by drawing cursive letters in the air using the Meta Quest 2 device. The application, developed on the Unity platform, integrates real-time gesture capture, character recognition via a CNN trained on the EMNIST dataset, and word suggestion using a Trie structure weighted by usage frequency.

Preliminary tests involved ten short sentences adapted from the set proposed by [Vertanen and Kristensson 2011], widely used in text entry studies on mobile devices.

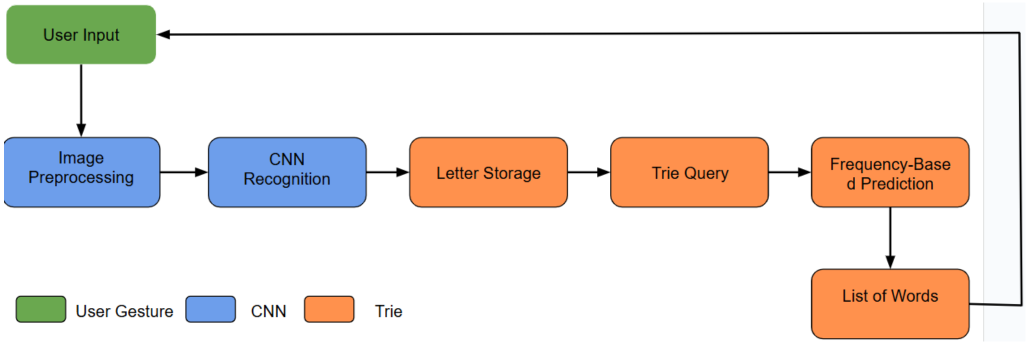

Figure 3. Illustration of the word insertion process based on letter recognition and display of suggestions in the VR environment.

Punctuation was removed due to limitations of the current system version. Table 1 presents the sentences employed during the evaluation.

Table 1. Sentences selected for preliminary evaluation.

Are you going to join us for lunch?
Are you in today?
Do you need it today?
How are you?
I am on my way.
I am walking in now.
Is it over?
OK with me.
See you soon!
Yes, I am playing.

During the evaluation, the system demonstrated stable performance: gestures were captured, letters were correctly classified, and coherent word suggestions were presented. These results confirm the feasibility of the approach for text entry in immersive environments. Although formal user testing has not yet been conducted, the system shows satisfactory results in terms of execution time, accuracy, suggestion rate, and questionnaires such as SUS and UEQ-S.

5. Final Considerations

This work introduced a method for text entry in virtual reality that allows users to draw letters in the air using the Meta Quest 2. The system combines gesture capture, character recognition through a convolutional neural network trained on EMNIST, and word formation via a frequency-weighted Trie structure. The proposed solution offers an intuitive and immersive alternative to traditional keyboards, particularly suited for environments emphasizing freedom of movement and gestural interaction.

Initial evaluations conducted by the development team confirmed the system's technical viability, with accurate character recognition and coherent word suggestions. However, challenges remain in handling less standardized gestures and improving error resilience. The next steps will involve usability testing with a broad user base using SUS and UEQ-S instruments to refine the system for practical applications in everyday, educational, and professional VR contexts.

References

Blanco Junior, M. (2022). Reconhecimento de placas de veículos utilizando redes neurais artificiais. Available at: http://repositorio.utfpr.edu.br/jspui/handle/1/35497.

Boletsis, C. and Kongsvik, S. (2019a). Controller-based text-input techniques for virtual reality: An empirical comparison. *International Journal of Virtual Reality*, 19(3):2–15. DOI: 10.20870/IJVR.2019.19.3.2917.

Boletsis, C. and Kongsvik, S. (2019b). Text input in virtual reality: A preliminary evaluation of the drum-like vr keyboard. *Technologies*, 7(2). DOI: 10.3390/technologies7020031.

Carvalho, J. V. d. (2000). Reconhecimento de caracteres manuscritos utilizando regras de associação. Technical report, Universidade Federal de Campina Grande. Available at: https://dspace.sti.ufcg.edu.br/handle/riufcg/7494.

Chen, Z., Yang, D., Liang, J., Liu, X., Wang, Y., Peng, Z., and Huang, S. (2022). Complex handwriting trajectory recovery: Evaluation metrics and algorithm. In *Asian Conference on Computer Vision (ACCV) 2022, Lecture Notes in Computer Science, vol. 13517*, pages 58–74. DOI: 10.1007/978-3-031-26284-5_4.

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE. DOI: 10.1109/IJCNN.2017.7966217.

Dudley, J. J., Karlson, A., Todi, K., Benko, H., Longest, M., Wang, R., and Kristensson, P. O. (2024). Efficient mid-air text input correction in virtual reality. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. DOI: 10.1109/ISMAR62088.2024.00105.

Dudley, J. J., Zheng, J., Gupta, A., Benko, H., Longest, M., Wang, R., and Kristensson, P. O. (2023). Evaluating the performance of hand-based probabilistic text input methods on a mid-air virtual qwerty keyboard. *IEEE Transactions on Visualization and Computer Graphics*. DOI: 10.1109/TVCG.2023.3320238.

Elmgren, R. (2017). Handwriting in vr as a text input method. Master's thesis, KTH Royal Institute of Technology. Available at: https://www.diva-portal.org/smash/get/diva2:1107665/FULLTEXT01.pdf.

Gugenheimer, J., Dobbelstein, D., Winkler, C., Haas, G., and Rukzio, E. (2016). Facetouch: Enabling touch interaction in display fixed uis for mobile virtual reality. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 49–60, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2984511.2984576.

Kaipainen, M., Ravaja, N., Tikka, P., Vuori, R., Pugliese, R., Rapino, M., and Takala, T. (2011). Enactive systems and enactive media: Embodied human-machine coupling beyond interfaces. *Leonardo*, 44(5):433–438. DOI: 10.1162/LEON_a_00244.

Katona, J. (2021). A review of human–computer interaction and virtual reality research fields in cognitive infocommunications. *Applied Sciences*, 11(6). DOI: 10.3390/app11062646.

Kumar, P., Chaudhary, A., and Sharma, A. (2022). A cnn based air-writing recognition framework for multilinguistic characters and digits. *SN Computer Science*, 3:453. DOI: 10.1007/s42979-022-01362-z.

Lu, D., Huang, D., and Rai, A. (2019). Fmhash: Deep hashing of in-air-handwriting for user identification. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7. DOI: 10.1109/ICC.2019.8761508.

Mimura, H., Ito, M., ichi Ito, S., and Fukumi, M. (2021). Personal authentication and recognition of aerial input hiragana using deep neural network. In Komuro, T. and Shimizu, T., editors, *Fifteenth International Conference on Quality Control by Artificial Vision*, volume 11794, page 1179411. International Society for Optics and Photonics, SPIE. DOI: 10.1117/12.2585333.

Monobe, K. and Ohishi, M. (2025). Research for japanese input method using flick in virtual reality. In *2025 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. IEEE. DOI: 10.1109/CACML64929.2025.11010978.

Nascimento, T. H., Felix, J. P., Santos Silva, J. L., and Soares, F. (2023). Text entry on smartwatches using continuous gesture recognition and word dictionary. In *International Conference on Human-Computer Interaction*, pages 550–562. Springer. DOI: 10.1007/978-3-031-35596-7$_3$5.

Nascimento, T. H., Nunes Soares, F. A. A. M., Vieira Oliveira, D., Lopes Salvini, R., Martins da Costa, R., and Gonçalves, C. (2017a). Method for text input with google cardboard: An approach using smartwatches and continuous gesture recognition. In *2017 19th Symposium on Virtual and Augmented Reality (SVR)*, pages 223–226. DOI: 10.1109/SVR.2017.36.

Nascimento, T. H., Soares, F. A. A. M. N., Irani, P. P., Galdino de Oliveira, L. L., and Da Silva Soares, A. (2017b). Method for text entry in smartwatches using continuous gesture recognition. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 549–554. DOI: 10.1109/COMPSAC.2017.168.

Shen, J., Boldu, R., Kalla, A., Glueck, M., Surale, H. B., and Karlson, A. (2024). Ringgesture: A ring-based mid-air gesture typing system powered by a deep-learning word prediction framework. *IEEE Transactions on Visualization and Computer Graphics*, 37(4):Article 111. DOI: 10.1109/TVCG.2024.3456179.

Tsuchida, K., Miyao, H., and Maruyama, M. (2015). Handwritten character recognition in the air by using leap motion controller. In *HCI International 2015 – Posters' Extended Abstracts*, pages 534–538. Springer. DOI: 10.1007/978-3-319-21380-4$_9$1.

Varela, F. J., Thompson, E., and Rosch, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press. DOI: 10.7551/mitpress/6730.001.0001.

Vertanen, K. and Kristensson, P. O. (2011). A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 295–298, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2037373.2037418.

Wang, Y., Wang, Y., Chen, J., Wang, Y., Yang, J., Jiang, T., and He, J. (2021). Investigating the performance of gesture-based input for mid-air text entry in a virtual environment: A comparison of hand-up versus hand-down postures. *Sensors*, 21(5). DOI: 10.3390/s21051582.

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, 265(3):94–105. Available at: http://www.jstor.org/stable/24938718.