Desenvolvimento e Avaliação de um Sistema de Escrita e Desenho no Ar com Caneta Eletrônica e VLMs

Luma T. L. de Souza[1], **Rafael A. D. Caldeira**[1], **Sérgio D. C. Leal**[1], **Maria Clara P. de Souza**[1], **Thiago M. Paixão**[1], **Richard J. M. G. Tello**[1]

[1]Instituto Federal do Espírito Santo (IFES) - Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, Serra - ES - Brasil

{lumatavares501,rafaeldeps15,sergiodanielcamponezleal}@gmail.com,

maria.peterle.souza@gmail.com,

{thiago.paixao, richard}@ifes.edu.br

Abstract. *The need to explore new forms of human–computer interaction and to expand resources for writing and drawing in digital environments motivated the development of this project. To this end, a pen equipped with a bluish LED at its tip was designed to generate a luminous point. The structure was produced using 3D printing and incorporated an ESP32 microcontroller with Bluetooth technology, enabling integration with the computer. The system was designed to capture, through a camera, the movements performed in the air with the pen and, with the support of Vision–Language models, to recognize both written words and drawn images. In the case of images, the system can also generate an enhanced version of the drawing by using the image description as a reference; however, this functionality will not be explored in the present study. Finally, a comparison among different models was carried out, using 12 words for testing — 6 in Portuguese and 6 in English — and 5 drawings from different classes. The models with the best performance achieved 84% accuracy with Gemini 2.5 Flash in image detection and 88.3% accuracy with Perplexity.ai's model, based on GPT-4.1, in word detection.*

Resumo. *A necessidade de explorar novas formas de interação homem–máquina e de ampliar os recursos para escrita e desenho em ambientes digitais motivou o desenvolvimento deste projeto. Para isso, foi concebida uma caneta equipada com um LED de tom azulado em sua extremidade, responsável por gerar um ponto luminoso. A estrutura foi produzida por meio de impressão 3D e incorporou um microcontrolador ESP32 com tecnologia Bluetooth, possibilitando a integração com o computador. O sistema foi projetado para capturar, por meio de uma câmera, os movimentos realizados no ar com a caneta e, com o auxílio de modelos de Visão–Linguagem, reconhecer tanto palavras escritas quanto imagens desenhadas. No caso das imagens, o sistema também gera uma versão aprimorada do desenho, utilizando a descrição da imagem como referência; entretanto, essa funcionalidade não será explorada no presente estudo. Por fim, foi realizada uma comparação entre diferentes modelos, utilizando 12 palavras — sendo 6 em português e 6 em inglês — e 5 desenhos de distintas classes. Os modelos com melhor desempenho apresentaram 84% de acurácia com o Gemini 2.5 Flash na detecção de imagens e 88,3% de acurácia com o modelo da Perplexity.ai, baseado no GPT-4.1, na detecção de palavras.*

1. Introdução

Air writing (AW), ou escrita no ar, é uma forma de interação humano–máquina em que gestos realizados no espaço livre são convertidos em caracteres ou palavras exibidos na tela [Chen et al. 2016]. Trata-se de um campo promissor para o desenvolvimento de novas tecnologias, com potencial de ampliar a acessibilidade digital e oferecer novas formas de interação.

Entre suas aplicações, destaca-se a capacidade de fornecer a usuários com necessidades específicas uma alternativa de comunicação e apoio, como, por exemplo, para pessoas com dislexia [Vaidya et al. 2022]. De modo geral, o *air writing* pode se tornar uma interface de entrada mais universal em comparação com métodos tradicionais, como o teclado e o mouse.

Diversas tecnologias já foram empregadas para registrar esses gestos, incluindo ondas de rádio, dispositivos dedicados, sensores vestíveis e visão computacional [Elshenaway and Guirguis 2021]. Um exemplo notável é o Microsoft Kinect, que utilizava uma câmera infravermelha para rastrear movimentos corporais e possibilitar a interação humano–máquina em aplicações como jogos.

Neste trabalho, propõe-se o desenvolvimento e a avaliação de um sistema de *air writing* que integra um dispositivo externo sem-fio e técnicas de visão computacional. A solução consiste em uma caneta eletrônica com terminal luminoso, desenvolvida por meio de impressão 3D, que utiliza a plataforma ESP32 e comunicação Bluetooth. A luz emitida pela caneta é captada por uma câmera web, permitindo o reconhecimento da escrita ou de desenhos realizados pelo usuário. Além disso, a caneta possui botoeiras que possibilitam funcionalidades adicionais, como a mudança de cores e ativação/desativação da escrita.

O sistema possui duas funções principais: detecção de palavras e geração de descrições de imagens, sendo esta última responsável pela função de aprimoramento do desenho. Cada função — reconhecimento de palavras e descrição/imagem — é processada de forma independente, utilizando modelos específicos para cada tarefa. Avaliou-se a acurácia de diferentes modelos em cada função, com o objetivo de identificar aqueles mais adequados para implementação no sistema.

2. Trabalhos Correlatos

Para embasar o desenvolvimento deste projeto, realizou-se uma análise de trabalhos publicados sobre o tema da escrita no ar (Air Writing) e de tecnologias relacionadas.

Em [Barbosa et al. 2024], principal referência para este estudo, foi desenvolvido um sistema Air Writing capaz de seguir a movimentação dos dedos e reconhecer gestos. Além disso, no trabalho também foram utilizados modelos para a realização da reconhecimento de palavras, sendo eles: HTR (Handwritten Text Recognition) [Vloison and Xiwei 2021], que obteve uma acurácia de 80,90% em situações ideais, e TrOCR [Li et al. 2021], que obteve 95% de acurácia nas mesmas condições.

Na pesquisa de [Alam et al. 2019], foi proposto um modelo para reconhecimento de palavras escritas no ar utilizando uma rede neural convolucional (CNN, do inglês *Convolutional Neural Network*). O modelo foi treinado e validado com um conjunto de 26 mil caracteres. A arquitetura da CNN é composta por camadas de entrada, convolução, camadas densas e de saída. Para prevenir *overfitting*, foi aplicada a técnica de dropout

nas camadas de convolução 1 e 2. Esta é uma técnica de regularização em que neurônios são selecionados aleatoriamente e ignorados durante o treinamento, ajudando a melhorar a generalização do modelo. O trabalho reportou uma acurácia de 97,29% no reconhecimento das palavras.

No trabalho de [Chen et al. 2019], foi projetado um sistema que utiliza óculos inteligentes para a captação de imagens RGB e emprega um modelo — uma Rede Neural Convolucional Regional de Máscara modificada (Mask R-CNN) — capaz de detectar a ponta do dedo de um indivíduo, permitindo a realização de escrita no ar em tempo real.

Finalmente, em [Wang et al. 2017], foi desenvolvido um mouse sem fio voltado para baixo consumo de energia. Para isso, foi utilizado o microcontrolador LPC54100, em conjunto com o sensor de movimento MPU6050. Além disso, foram empregados componentes de baixo consumo com conectividade Bluetooth, responsáveis pela transmissão dos dados. O projeto contemplou tanto o controle do cursor quanto as funcionalidades de clique e comunicação.

3. Materiais e Métodos

3.1. Construção da Caneta

Visando o desenvolvimento de um sistema sem fio, foi realizada uma análise técnica de comparativa entre modelos existentes no mercado. O modelo escolhido foi o ESP32 DEVKit v1 de 30 pinos, que apresenta dimensões reduzidas e pode ser instalado em diversos ambientes, além de contar as tecnologias de comunicação Bluetooth e Wi-Fi.

Para a construção da caneta, foram utilizados resistores, uma bateria de 9V, uma placa ESP32, um diodo LED de 10 mm (luz branca com tonalidade azul) e botoeiras não retentivas nas cores azul, vermelho, preto e verde, que representam as cores disponíveis para escrita no sistema.

O corpo da caneta foi confeccionada com o auxílio de uma impressora 3D. A integração da caneta com o sistema é exemplificada na Figura 1.

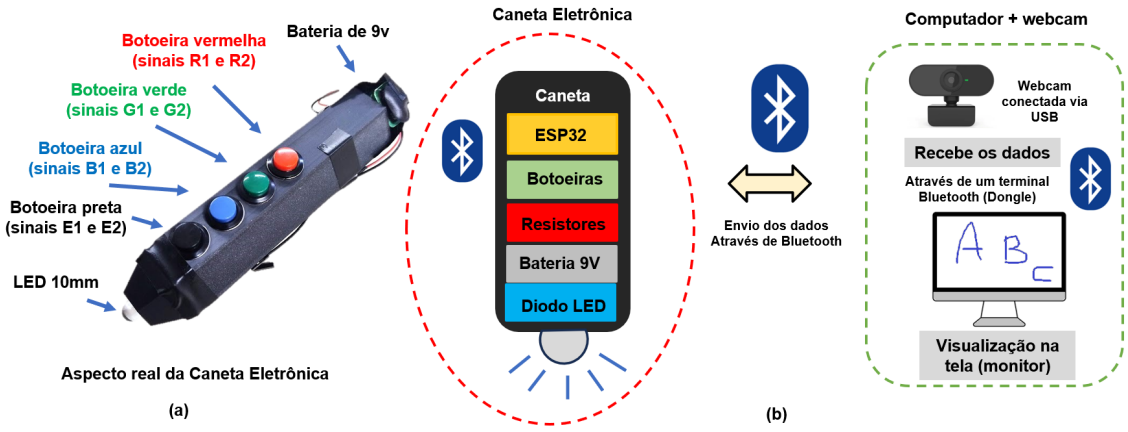


Figura 1. (a) Aspecto real da caneta eletrônica; (b) Componentes da comunicação caneta-computador.

No sistema proposto, o microcontrolador através da ativação/desativação das botoeiras (botões táteis de pulso não retentivos) envia os seguintes comandos ou sinais:

- R1: se a botoeira vermelha for apertada, a cor vermelha inicia a escrita;
- R2: caso a botoeira vermelha for apertada novamente, a cor vermelha de escrita é interrompida;
- G1: se a botoeira verde for apertada, a cor verde inicia a escrita;
- G2: caso a botoeira verde for apertada novamente, a cor verde de escrita é interrompida;
- E1: se a botoeira preta for apertada, a cor preta inicia a escrita;
- E2: caso a botoeira preta for apertada novamente, a cor preta de escrita é interrompida;
- B1: se a botoeira azul for apertada, a cor azul inicia a escrita;
- B2: caso a botoeira azul for apertada novamente, a cor azul de escrita é interrompida.

Cada um desses comandos indica a inicialização ou a interrupção da escrita, mantendo a luz acesa para identificação pelos algoritmos de visão computacional. Essa luz, emitida pela caneta, é captada por uma câmera web HP modelo W300, compatível com diversos sistemas operacionais e capaz de operar em diferentes resoluções de imagem. O dispositivo utiliza conexão USB 2.0 e atinge uma taxa máxima de captura de 30 quadros por segundo (fps). Para o processamento, foi utilizado um computador equipado com processador Intel® Core™ i7-9700 CPU @ 3.00GHz e 24 GB de memória RAM.

3.2. Localização do Ponto Luminoso baseada em Visão Computacional

Para realizar a detecção do ponto luminoso, é necessário identificar inicialmente, por meio da webcam, a luz específica emitida pelo LED incorporado na caneta (diodo LED 10 mm). Para isso, a imagem capturada é convertida para o espaço de cores HSV, que permite separar de forma mais robusta as informações de tonalidade, saturação e intensidade luminosa. Em seguida, define-se um intervalo de valores (*lower bound* e *upper bound*) que representa a tonalidade azul característica do LED.

A partir desse intervalo, é criada uma máscara binária, na qual cada pixel que se encontra dentro da faixa de cor especificada recebe o valor branco (255), indicando que pertence à região de interesse, enquanto os demais recebem o valor preto (0), indicando que não fazem parte da luz do LED. Essa máscara isola visualmente e computacionalmente o ponto luminoso, permitindo aplicar operações morfológicas para reduzir ruídos e, posteriormente, detectar o contorno e a posição exata do LED na cena, mesmo em ambientes com iluminação ambiente mais intensa. Na Figura 2 ilustra o funcionamento da máscara.

Após essa etapa, foram utilizadas funções da biblioteca `OpenCV`, como a findContours, cujo objetivo é identificar o contorno da máscara e viabilizar o seu reconhecimento. A partir disso, tornou-se possível detectar e acompanhar o ponto luminoso, definido como ponto de interesse. Esse ponto serve como referência para o traçado das linhas que representam os movimentos realizados. A Figura 3, apresenta-se um dos desenhos e uma das palavras produzidas com o sistema da caneta, a partir da localização do ponto luminoso.

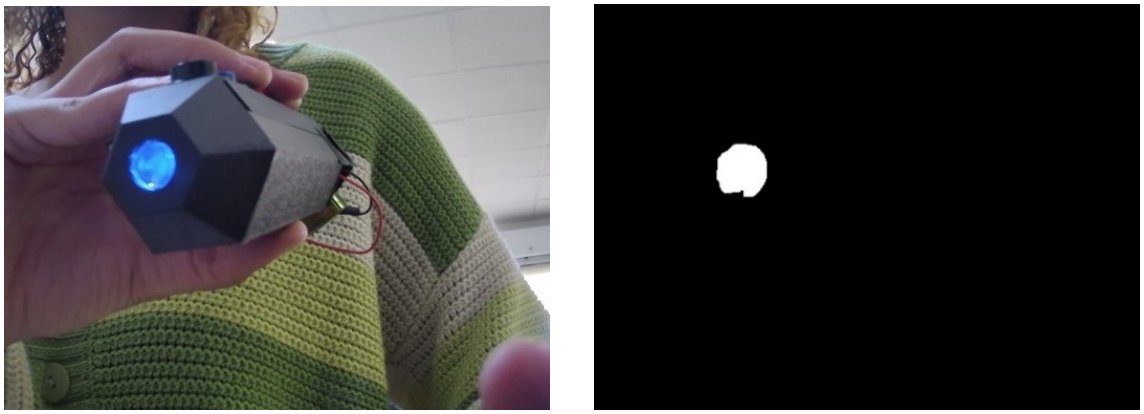
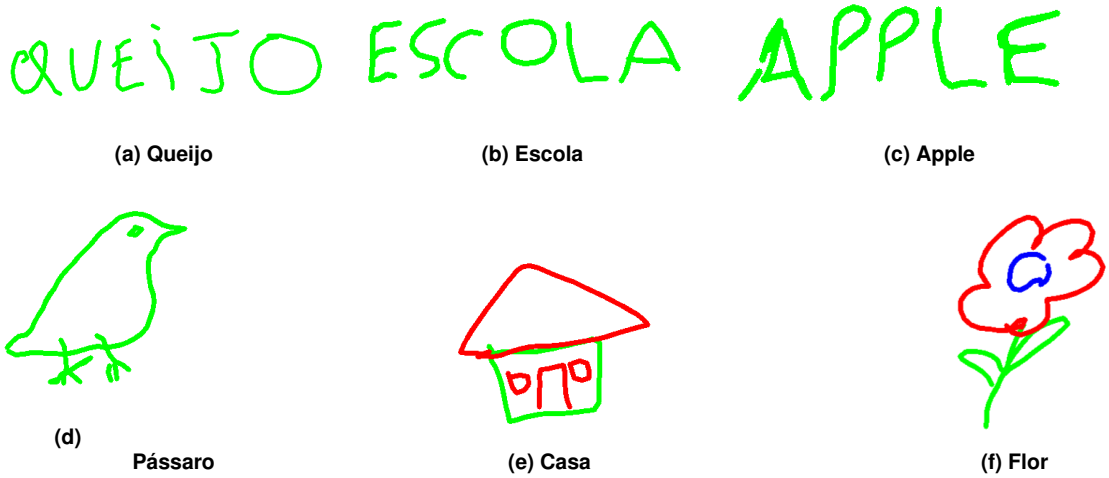


Figura 2. Imagem com e sem máscara para detecção do LED.



(a) Queijo **(b) Escola** **(c) Apple**

(d) Pássaro **(e) Casa** **(f) Flor**

Figura 3. Exemplos de desenhos e palavras produzidos com o sistema da caneta.

3.3. Reconhecimento de palavras e imagens

Neste trabalho, tanto palavras quanto desenhos são inicialmente tratados como imagens, permitindo a utilização de Modelos Visão-Linguagem (VLMs - *Vision-Language Models*). Diferentemente dos Modelos de Linguagem de Grande Porte (LLMs - *Large Language Models*), os VLMs são capazes de identificar e descrever conteúdos visuais, integrando informações visuais e textuais para um reconhecimento mais completo.

Esses modelos permitem o reconhecimento de palavras e a descrição de desenhos. Dessa forma, em nosso sistema foram utilizados modelos distintos para cada tarefa — reconhecimento de palavras e reconhecimento e descrição de imagens — possibilitando a diferenciação entre eles e atribuindo funções específicas a cada tipo dentro do sistema, conforme ilustrado na Figura 4. Adicionalmente, a descrição gerada pelo sistema é empregada para criar uma versão aprimorada da imagem, como mostrado na Figura 5.

Dessa forma, o sistema não apenas realiza a identificação e descrição de palavras e desenhos, mas também permite a criação de versões aprimoradas das imagens.

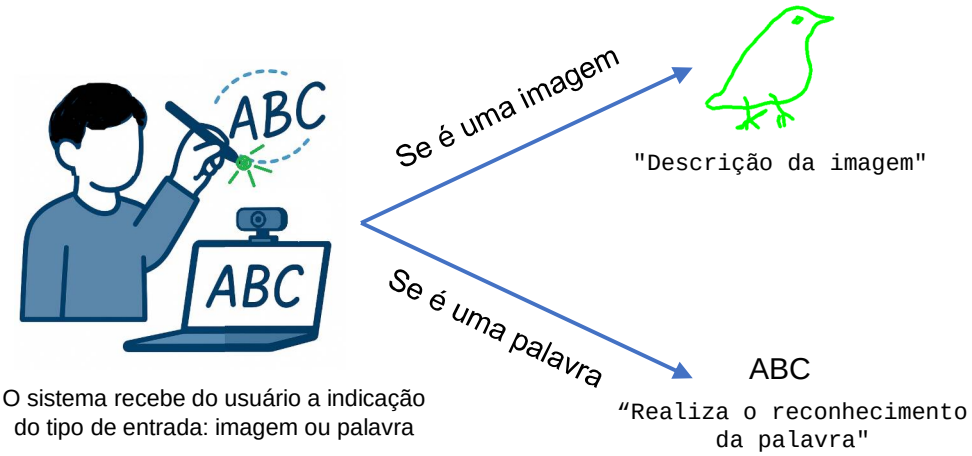


Figura 4. Representação do funcionamento do sistema para identificação de palavras e imagens.

(a) Desenho produzido utilizando o sistema da caneta eletrônica.

(b) Desenho aprimorado gerado a partir da descrição do desenho anterior.

Figura 5. Comparação entre o desenho realizado com o sistema da caneta eletrônica (a) e a versão aprimorada gerada a partir de sua descrição (b).

4. Experimento

Neste trabalho, as métricas de desempenho foram baseadas na capacidade dos modelos em detectar palavras e descrever os desenhos realizados pelos usuários. Para isso, utilizamos os seguintes modelos para descrição de imagens: Gemini-2.5-flash, VIT GPT2 image captioning, GIT base coco, Kosmos-2 e BLIP image captioning base; e, para o reconhecimento de palavras, TrOCR e Perplexity.ai com GPT-4. Cinco participantes contribuíram para a construção de uma base de dados escrevendo as palavras listadas na Tabela 1, sem um tempo pré-estabelecido, iniciando as palavras em português e, posteriormente, as palavras em inglês. Em seguida, cada participante realizou os desenhos indicados na Tabela 2, correspondentes às figuras de cada classe descrita, também sem limite de tempo. Cada desenho e cada palavra foram salvos em arquivos individuais para testes posteriores com os modelos.

Tabela 1. Palavras usadas para detecção de escrita

PT	EN
Queijo	Cheese
Maçã	Apple
Parabéns	Congratulations
Mundo	World
Laranja	Orange
Escola	School

Tabela 2. Desenhos usados para a detecção de imagens

Imagens
casa
pássaro
sol
flor
carro

Foi analisada a capacidade de reconhecimento das inteligências artificiais, calculando-se a acurácia desses modelos utilizando a seguinte fórmula:

$$\text{Acurácia} = \frac{\text{Número de outputs corretos}}{\text{Número total de inputs}} \times 100\% \quad (1)$$

Para o domínio de imagens, atribui-se 1 (um) ponto ao modelo quando a descrição ou a classe da imagem estiver completamente correta; 0,5 (meio) ponto quando houver acerto parcial (correspondente a aproximadamente metade da resposta esperada); e 0 (zero) ponto quando não houver acerto algum. Utilizamos legenda aberta (*open caption*) em vez de fechada (por exemplo, "carro", "ave"), pois a descrição da imagem é posteriormente utilizada para a geração de novas imagens. Ademais, a Figura 6 a seguir ilustra o critério de pontuação.

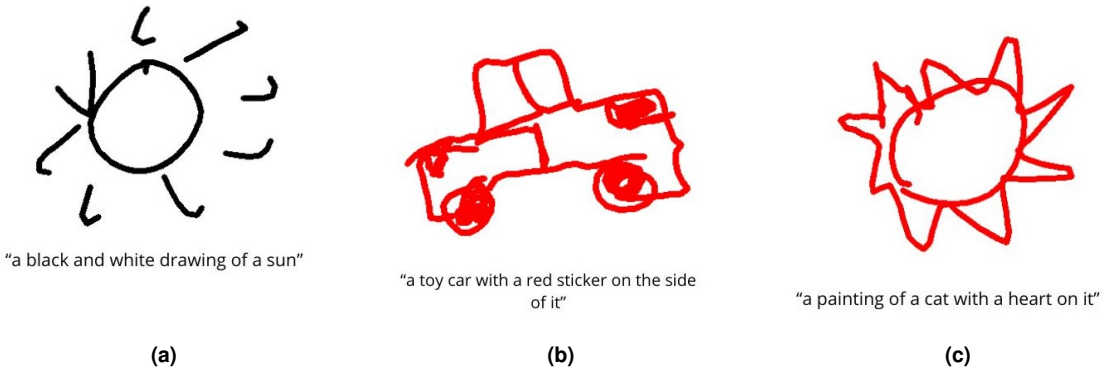


Figura 6. Exemplo de pontuação para o modelo de imagens: (a) acerto completo; (b) acerto parcial; (c) sem acerto.

No âmbito da escrita, foram conduzidas duas análises distintas. A primeira utilizou a ferramenta Perplexity.ai, configurada com o modelo ChatGPT 4.1, considerando a palavra como um todo para o cálculo da acurácia. Já a segunda análise empregou o modelo TrOCR e, de acordo com o trabalho de [Barbosa et al. 2024], a acurácia foi obtida por meio da métrica Character Error Rate (CER), conforme exemplificado na Figura 7.

5. Resultados e Discussão

Os resultados obtidos no teste de reconhecimento de escrita, considerando os dois modelos avaliados, estão apresentados na Tabela 3. Nesta etapa, buscou-se comparar o de-

Figura 7. Exemplo de saída do sistema Character Error Rate (CER).

sempenho do Perplexity.ai com GPT-4 e do modelo TrOCR, ambos aplicados às amostras geradas pelos participantes do experimento.

Tabela 3. Performance of the models for handwriting detection

Model	Acurácia
Perplexity.ai	$88.33 \pm 17,28\%$
TrOCR	$64,12 \pm 13,01\%$

Nos testes com o Perplexity.ai, optou-se pela versão web em vez da API, devido à maior possibilidade de personalização e à ausência de limites de uso. A acurácia geral dos participantes no experimento foi de 88,3%, enquanto a do modelo TrOCR foi de apenas 64,1%. Essa diferença significativa deve-se, em grande parte, ao fato de o modelo TrOCR ter sido utilizado em sua versão disponibilizada no Hugging Face, sem qualquer tipo de treinamento adicional. Como resultado, o modelo apresentou dificuldades na interpretação de determinadas palavras em português, como "maçã", que contém caracteres não utilizados na língua inglesa e, portanto, não reconhecidos pelo modelo, cujo pré-treinamento foi realizado exclusivamente com escrita manual em inglês.

Ambos os testes revelaram que a principal dificuldade encontrada pelos modelos foi o fator humano, especificamente a caligrafia, que frequentemente se mostrava ilegível até mesmo para humanos. Adicionalmente, observou-se um padrão em que os participantes apresentaram um maior número de acertos nas palavras finais do experimento, sugerindo uma adaptação progressiva ao manuseio da caneta. Complementarmente, os resultados do reconhecimento de imagens podem ser observados na Tabela 4.

Tabela 4. Desempenho dos modelos para detecção de imagens

Modelo	Acurácia
Gemini-2.5-flash	$84,00 \pm 20.8\%$
GIT base coco	$74,00 \pm 33.5\%$
BLIP image captioning base [Li et al. 2022]	$72,00 \pm 37.7\%$
Kosmos-2 [Peng et al. 2023]	$64,00 \pm 32.4\%$
VIT GPT2 image captioning	$14,00 \pm 14.2\%$

Como mostrado pelos resultados obtidos, o modelo do Gemini alcançou o melhor

desempenho na detecção e descrição de imagens, superando os outros modelos. Esse resultado demonstra maior capacidade do modelo em gerar descrições mais precisas e coerentes em relação ao conteúdo visual apresentado. Portanto, esse modelo mostra-se como a opção mais adequada para ser implementado em nosso sistema para essa tarefa. Ademais, observa-se aqui uma limitação relacionada ao uso por parte dos usuários, uma vez que os desenhos podem ser feitos com base em suas interpretações individuais e habilidades artísticas, que nem sempre são suficientemente desenvolvidas.

6. Conclusões

Neste trabalho, foi desenvolvido um sistema hardware–software capaz de detectar um ponto luminoso emitido por uma caneta eletrônica, construída por meio de impressão 3D e composta por componentes como botoeiras, resistores e a plataforma ESP32. Além disso, o sistema realiza a geração de imagens a partir de descrições. Trata-se de um sistema intuitivo, com funcionalidades como troca de cores, implementadas por meio de técnicas de visão computacional.

Além disso, realizamos o *benchmarking* de diferentes modelos de inteligência artificial para as tarefas de detecção de imagens e reconhecimento de escrita. utilizando dados obtidos a partir de 5 participantes, que escreveram 12 palavras — sendo 6 em inglês e 6 em português —, além de realizarem 5 desenhos. Os resultados demonstraram que o modelo Gemini-2.5-Flash apresentou o melhor desempenho na descrição de imagens, com acurácia de 84%, enquanto o modelo Perplexity.ia com GPT 4.1 atingiu 88,3% de acerto na detecção de palavras. Para além dos fatores técnicos, observamos que o desempenho do sistema também foi influenciado por aspectos humanos, como a dificuldade dos participantes em escrever ou desenhar com precisão utilizando a caneta — possivelmente devido ao seu tamanho ou ergonomia. Outro fator limitante identificado foi a dependência da qualidade da iluminação: quando a luz emitida pela caneta não é corretamente detectada, a captação do movimento se torna falha ou cessa completamente, comprometendo o funcionamento do sistema. Dessa forma, como proposta para trabalhos futuros, sugere-se o desenvolvimento de uma caneta mais ergonômica e confortável, assim como a adoção de métodos de detecção mais robustos, que não dependam exclusivamente da luz e que funcionem adequadamente em diferentes cenários.

7. Agradecimentos

Todos os autores agradecem o apoio institucional para o desenvolvimento deste projeto. Em especial, as autores Luma T. L. de Souza, Sérgio D. C. Leal e Maria Clara P. de Souza agradecem ao Instituto Federal do Espírito Santo (IFES) pela concessão das bolsas de Iniciação Científica vinculadas ao Programa PICTI (Edital PRPPG nº 03/2024).

Referências

Alam, M. S., Kwon, K.-C., and Kim, N. (2019). Trajectory-based air-writing character recognition using convolutional neural network. In *2019 4th International Conference on Control, Robotics and Cybernetics (CRC)*, pages 86–90.

Barbosa, C. E., Pereira, T. B., do Carmo, I. M., Tello, R. J., Boldt, F. A., and Paixao, T. M. (2024). Reconhecimento de texto para sistemas air writing: Um estudo experimental. In *Escola Regional de Informática do Espírito Santo (ERI-ES)*, pages 21–30. SBC.

Chen, M., AlRegib, G., and Juang, B.-H. (2016). Air-writing recognition—part i: Modeling and recognition of characters, words, and connecting motions. *IEEE Transactions on Human-Machine Systems*, 46(3):403–413.

Chen, Y.-H., Su, P.-C., and Chien, F.-T. (2019). Air-writing for smart glasses by effective fingertip detection. In *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pages 381–382.

Elshenaway, A. R. and Guirguis, S. K. (2021). On-air hand-drawn doodles for iot devices authentication during covid-19. *IEEE Access*, 9:161723–161744.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2021). Trocr: Transformer-based optical character recognition with pre-trained models. `https://github.com/microsoft/unilm/tree/master/trocr`.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306.

Vaidya, V., Pravanth, T., and Viji, D. (2022). Air writing recognition application for dyslexic people. In *2022 International Mobile and Embedded Technology Conference (MECON)*, pages 553–558.

Vloison, V. and Xiwei, H. (2021). Deep learning framework for line-level handwritten text recognition. `https://github.com/vloison/Handwritten_Text_Recognition`.

Wang, K., Zeng, W., Ma, C., Cheng, C., Sun, P., Wang, L., and Cai, W. (2017). The design of wireless air mouse based on lpc54100. In *2017 36th Chinese Control Conference (CCC)*, pages 6409–6413.