“Juntos Venceremos” – Um *exergame* para teste de jogo *multiplayer* utilizando visão computacional

André Bonetto Trindade[1], Joice Luiz Jeronimo[1], Marcelo da Silva Hounsell[2], Rafael Stubs Parpinelli[2]

[1]Departamento de Engenharia Elétrica – Instituto Federal de Santa Catarina (IFSC) – Joinville – SC – Brasil

[2]Departamento de Ciência da Computação – Universidade do Estado do Santa Catarina (UDESC) – Joinville – SC – Brasil

{andre.bonetto, joice.jeronimo}@ifsc.edu.br, {rafael.parpinelli, marcelo.hounsell}@udesc.br

Abstract. *This paper describes the development of a multiplayer exergame that uses a machine learning tool to track people's poses in real time. The system consists of conventional equipment: computer, webcam and video projector. After an exploratory search of open source models, the Movenet tool was selected, which uses a convolutional neural network as the basis of the algorithm that can detect up to 6 people at a response speed greater than 30 frames per second, allowing the game to run without noticeable latency to the user.*

Keywords – *Computer Vision, Exergame, Autism*

Resumo. *Este artigo descreve o desenvolvimento de um jogo multijogadores do tipo exergame que utiliza ferramenta de aprendizado de máquina para o rastreamento de poses de pessoas em tempo real. O sistema é composto por equipamentos convencionais: computador, webcam e projetor de vídeo. Após uma busca exploratória de modelos de código aberto, foi selecionada a ferramenta Movenet que utiliza rede neural convolucional como base do algoritmo e que pode detectar até 6 pessoas em velocidade de resposta superior a 30 quadros por segundo, possibilitando a execução do jogo sem latência perceptível ao usuário.*

Palavras-chave *– Visão Computacional, Exergame, Autismo*

1. Introdução

A inserção de tecnologia digitais como ferramenta auxiliar em diversas áreas do conhecimento está cada vez mais consolidada. Um exemplo disso é utilização de jogos digitais como ferramenta educacional e até mesmo em sessões terapêuticas. Em [Trindade, 2022] foi descrito o desenvolvimento e avaliação de uma plataforma interativa de projeção de jogos sérios ativos (*exergames*) voltados para o público com transtorno do espectro autista (TEA) que, através da utilização de visão computacional e equipamentos convencionais (projetor de vídeo, computador e *webcam*), detecta a interação do jogador com o jogo sem a necessidade de sensores ou dispositivos extras de controle. Especialistas na área de TEA avaliaram a plataforma e apontaram

melhoramentos, entre eles, a importância da utilização dos jogos por mais de um jogador (jogos *multiplayer*) pela razão de que um dos déficits comuns na população autista é a dificuldade de socialização, e desta forma, jogos sérios *multiplayers* poderiam ser uma forma de incentivo e oportunidade de treino de socialização para esta população.

Visão computacional é uma vertente de estudo e aplicação da inteligência artificial que utiliza a capacidade dos computadores em interpretar e entender informações visuais ou imagens digitalizadas [Szeliski, 2022]. Ela utiliza uma variedade de métodos, incluindo processamento de imagens, reconhecimento de padrões, aprendizado de máquina e redes neurais artificiais para executar tarefas como detecção de objetos, reconhecimento facial e rastreamento de movimento. As redes neurais artificiais são modelos computacionais inspirados no funcionamento do cérebro humano. Essas redes são compostas por unidades de processamento interconectadas e são conhecidas como neurônios artificiais. As redes neurais são capazes de aprender a partir de exemplos e ajustar os pesos das conexões neurais de diversas camadas para tarefas de classificação, regressão, processamento de linguagem natural e outras aplicações [Goodfellow *et al*, 2016].

Jogos digitais sérios são jogos desenvolvidos com o envolvimento de especialistas na área tema do jogo e que aproveitam as características de entretenimento de um jogo digital convencional e inserem um objetivo sério, com o intuito de aumentar o engajamento do jogador naquele tema sério escolhido [Alvarez e Djaouti, 2011]. *Exergame* é um tipo de jogo que necessita que o jogador faça algum tipo de esforço físico ou movimentação para executar o jogo e, neste contexto, este trabalho teve como objetivo principal a busca e teste de um modelo de detecção de poses humanas por visão computacional capaz de identificar mais de uma pessoa em um quadro de imagem e com velocidade de inferência superior a 30 quadros por segundo (30 FPS) a fim de poder ser inserido em um *exergame multiplayer* utilizando projeção interativa e, ao mesmo tempo, manter a fluidez e conforto visual.

2. Trabalhos Relacionados

Através de busca exploratória em repositórios de artigos e motores de busca na *internet*, foram encontradas e selecionadas as ferramentas listadas na Tabela 1 para serem analisadas quanto a características básicas de reconhecimento de poses humanas. A seleção destas ferramentas baseou-se nas seguintes premissas:

- Ter código aberto;
- Ser capaz de identificar mais de uma pessoa no quadro e;
- Relação de tempo de processamento e qualidade mínima para poder ser utilizado em jogos (>30 FPS).

Os dados da Tabela 1 foram recolhidos do artigo ou da documentação disponibilizada pelos autores de cada ferramenta. Desta maneira, percebe-se que das ferramentas selecionadas, apenas 40% explicitaram que a ferramenta é capaz de detectar pessoas em velocidade igual ou superior a 30 quadros por segundos. Apenas a ferramenta *DirectPose* não explicita no artigo e na documentação como obter acesso a ferramenta. Excetuando a *DensePose*, as ferramentas têm como saída pontos chaves específico do corpo humano e, na maioria das vezes, das articulações dos membros

inferiores, superiores, tronco e cabeça. Algumas delas (30%) dão a possibilidade de identificar as extremidades dos membros (mãos e pés), e por fim, 60% permitem o rastreamento de pontos específicos da face (nariz, olhos, boca e nariz).

Tabela 1. Ferramentas de detecção de pessoas.

#	Ferramenta	Ano	Referência	Código Aberto	MultiPessoa	>30FPS	Pontos de Detecção
1	AlphaPose	2022	Fang, H. S. *et al*.	Sim	Sim	Sim	17
2	Movenet	2021	TensorFlow	Sim	Sim	Sim	17
3	BlazePose	2020	Bazarevsky, V. *et al.*	Sim	Sim	Sim	33
4	Xnet	2020	Mehta D. , *et al.*	Sim	Sim	Sim	17
5	DirectPose	2019	Tian, Z., Chen, H. e Shen, C.	Não Explicita	Sim	Não explicita	17
6	OpenPose	2019	Jo, B., Kim, S.	Sim	Sim	Não (22FPS)	15
7	Dense Pose	2018	Ɜüler, R. A., Neverova, N., & Kokkinos, I	Sim	Sim	Não explicita	Malha de Pontos
8	MultiPoseNet	2018	Kocabas, M., Karagoz, S. e Akbas, E.	Sim	Sim	Não (23FPS)	17
9	ArtTRack	2017	Insafutdinov, E. *et al.*	Sim	Sim	Não explicita	12
10	DeepCut	2016	Pishchulin, L. *et al.*	Sim	Sim	Não	14

A Tabela 1 auxiliou na tarefa de escolha de uma ferramenta para construir o jogo teste, sendo o primeiro fator relevante o tempo de processamento, logo restringindo a escolha para os modelos que explicitaram serem capazes de executar com taxa maior de 30 quadros por segundo. O segundo ponto determinante na escolha foi a qualidade do material e documentação disponível e da acessibilidade da ferramenta. Deste modo, a ferramenta escolhida para a construção do jogo *multiplayer* neste trabalho foi a *Movenet*, por atender os requisitos anteriores e ser uma ferramenta recente. A seguir, será detalhado o funcionamento da *Movenet*.

3. Movenet

Movenet[1] é uma ferramenta que utiliza uma rede neural convolucional que é executada em imagens baseadas no *Red/Green/Blue* (RGB) e prevê localizações de articulações humanas de até 6 pessoas no quadro da imagem de entrada em taxas superiores a 30 quadros por segundo e capaz de entregar até 17 pontos de referência do corpo humano, mesmo em condições de oclusão.

A entrada de dados deve ser um quadro de vídeo ou imagem, representado como um tensor em forma dinâmica, o qual a altura e largura do quadro precisam ser múltiplos de 32. A escolha do tamanho da imagem de entrada controla a compensação entre velocidade *versus* precisão, então é importante determinar um formato de entrada que melhor se adapta à aplicação desejada.

A saída do modelo é composta por um tensor da forma [1, 6, 56], que correspondem:

- A primeira dimensão é a dimensão do lote, que é sempre igual a 1;
- A segunda dimensão corresponde ao número máximo de detecções de instância, o qual o modelo pode detectar até 6 pessoas no quadro da imagem simultaneamente;
- A terceira dimensão representa as coordenadas dos locais previstos da caixa delimitadora que limita cada pessoa (*bounding box*), coordenadas dos pontos-chaves (*keypoints*) de cada articulação de cada pessoa e pontuações de confiança do modelo (*score*).

[1] www.tensorflow.org/hub/tutorials/movenet

Os primeiros 17 elementos da terceira dimensão citada anteriormente são os locais dos *keypoints* e os *scores* no formato: [y_0, x_0, s_0, y_1, x_1, s_1, …, y_16, x_16, s_16], onde y_i, x_i, s_i são os coordenadas yx (normalizadas para o quadro da imagem, por exemplo, intervalo em [0.0, 1.0]) e a pontuação da confiança da i-ésima articulação correspondentemente. A ordem dos 17 *keypoints* é: [nariz, olho esquerdo, olho direito, orelha esquerda, orelha direita, ombro esquerdo, ombro direito, cotovelo esquerdo, cotovelo direito, pulso esquerdo, pulso direito, quadril esquerdo, quadril direito, joelho esquerdo, joelho direito, tornozelo esquerdo, tornozelo direito]. Os 5 elementos restantes [ymin, xmin, ymax, xmax, *score*] representam a região da caixa delimitadora (*bounding box*) em coordenadas normalizadas e a pontuação de confiança da instância.

Neste trabalho a ferramenta *Movenet* foi utilizada para a detecção de pessoas em uma área de jogo a fim de determinar a interação entre os jogadores e o jogo executado. Ela foi aplicada em conjunto com a linguagem de programação *Python*[2] e os módulos: (i) *Pygame*[3] para a criação de jogos em *Python* e; (ii) *OpenCV*[4] que é uma biblioteca específica para aplicações de visão computacional. A seguir será detalhada a construção do jogo "Juntos Venceremos", um jogo colaborativo que utiliza a visão computacional.

4. Juntos Venceremos

O jogo "Juntos Venceremos" foi desenvolvido para fazer testes de funcionalidades da ferramenta *Movenet* com a participação de dois jogadores e tem como objetivo principal a colaboração dos jogadores para encontrar a localização da resposta certa. O estilo do jogo é semelhante a um *Quiz*. Foram adicionados ao jogo um conjunto de 10 perguntas a fim de teste de funcionalidades, contudo esta quantidade e temas de perguntas podem ser alterados se necessário. A interface do jogo consiste em duas telas:

- Tela de controle: nesta tela é realizada a calibração da superfície de jogo e na qual podem ser monitorados os esqueletos detectados pela visão computacional em tempo real (Figura 1a). Esta tela fica no computador;
- Tela de jogo: é a tela na qual o jogo será executado e ela deve ser projetada na área física do jogo (Figura 1b). Esta tela é enviada ao projetor de vídeo.

4.1. Calibração da área de jogo

A primeira parte da utilização do jogo se destina a identificação da área de jogo na captura de vídeo pela *webcam* que deverá ser considerada pelo algoritmo de visão computacional, como visto na Figura 1a que mostra os pontos de calibração utilizados (vértices do polígono em vermelho). Esta informação será necessária para determinar a posição do jogador que estará sobreposta à projeção. O jogo foi desenvolvido para projeções verticais (paredes, por exemplo), ou seja, em superfícies que estejam perpendiculares ao plano de terra, e devido a possibilidade da *webcam* não estar alinhada à projeção, esta calibração ajudará na correção de perspectiva e assim, reduzirá erros na determinação de posicionamento dos jogadores sobre a área de jogo definida.

---

[2] www.python.org

[3] www.pygame.org

[4] www.opencv.org

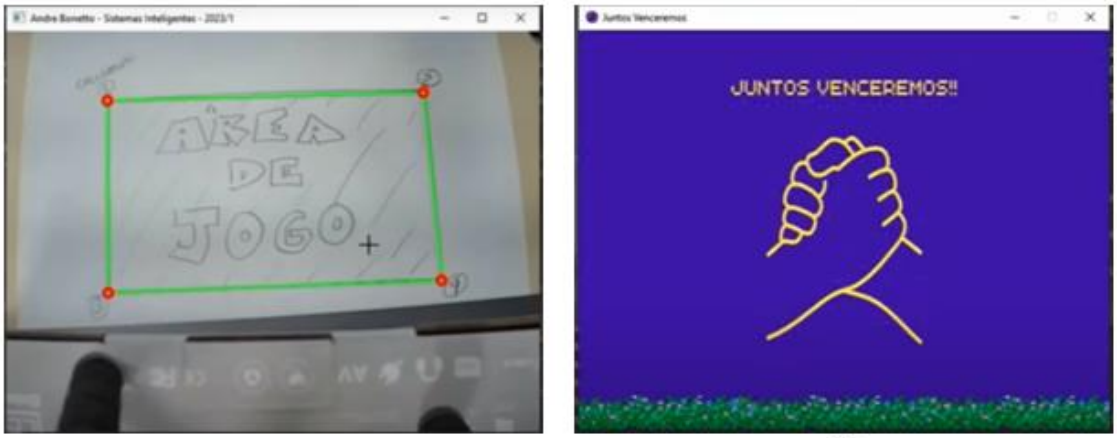

Figura 1. (a) Tela de controle e (b) tela de projeção do jogo.

O processo de calibração consiste em determinar os 4 vértices da projeção captada pela *webcam* utilizando cliques de *mouse* sobre a tela de controle na seguinte sequência: vértice superior esquerdo, vértice superior direito, vértice inferior esquerdo e vértice inferior direito, conforme mostra a Figura 1a, na qual os quatros vértices da área de jogo já foram determinados pelo usuário (vértices em vermelho e linhas verde). Estes 4 pontos são enviados ao módulo OpenCV que realiza uma função para corrigir a perspectiva.

4.2. Jogando o "Juntos Venceremos"

O jogo "Juntos Venceremos" é liberado após a execução do processo de calibração e, por se tratar de um jogo colaborativo, é necessário que dois jogadores sejam detectados na área de jogo selecionada. Para iniciar a partida os dois jogadores devem se cumprimentar usando as mãos conforme mostra a Figura 2. O sistema reconhece como um "aperto de mãos" quando as coordenadas X e Y dos pulsos estiverem dentro de certos limites de distância entre si.

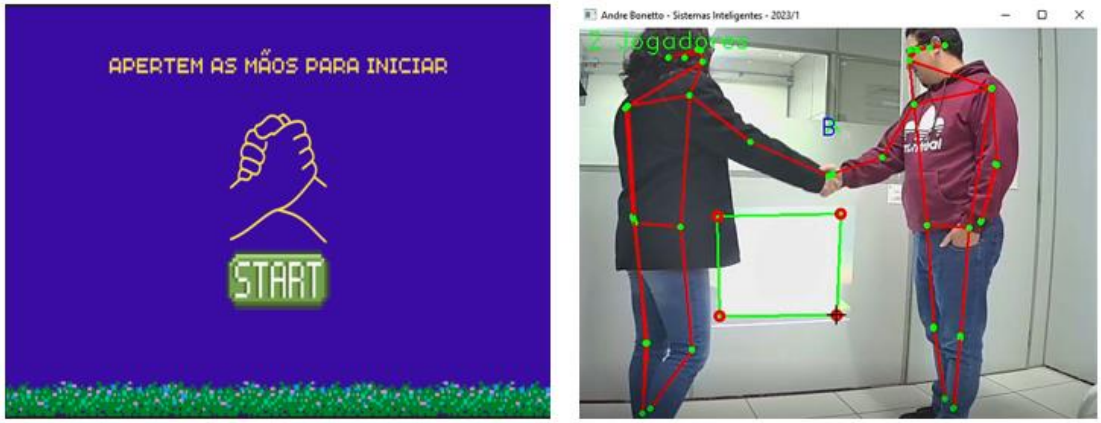

Figura 2. Jogadores se cumprimentando.

Ao iniciar a partida, cada jogador deve se posicionar em uma das laterais da projeção (Figura 3a) e, quando isto acontecer, uma pergunta sorteada é projetada na tela

de jogo com duas alternativas de resposta à esquerda e à direita da projeção, como exemplifica a Figura 3b.

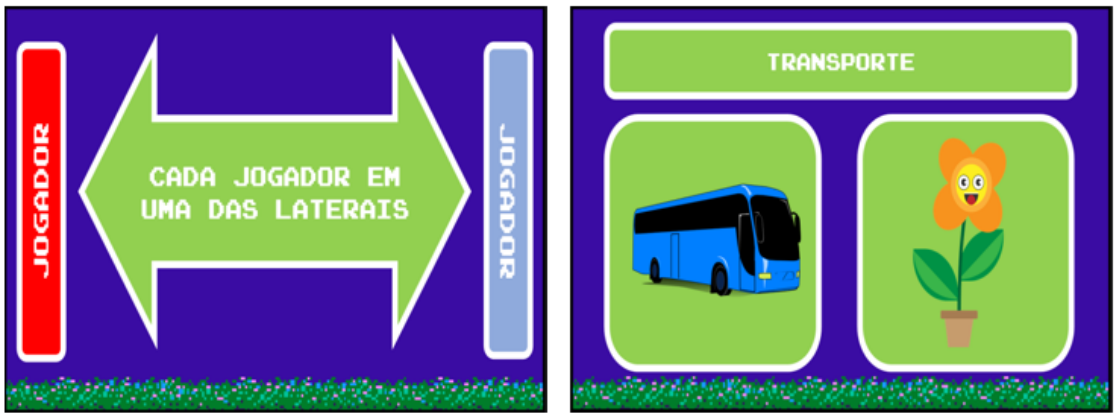

Figura 3. (a) Instrução para os jogadores e (b) exemplo de pergunta.

A partir desta etapa o jogo espera a resposta escolhida de forma colaborativa pelos jogadores, e para isso ambos devem se posicionar sobre a resposta certa para dar continuidade ao jogo. Como o objetivo principal do jogo é testar a visão computacional na situação de *multiplayers*, o jogo foi simplificado para esperar até que os jogadores se posicionem sobre a resposta certa, sem limite de tempo.

Quando ambos os jogadores se posicionam sobre a resposta certa, o jogo emite um *feedback* sonoro e visual demonstrando que eles acertaram a questão (Figura 4a) e em seguida pede para que os jogadores se posicionem novamente nas laterais (Figura 3a) para dar continuidade ao jogo. Este ciclo se repete por 10 perguntas, quando então eles serão recompensados com um troféu, como mostra a Figura 4b.

Figura 4. (a) *Feedback* entre perguntas e (b) *Feedback* final.

Após a tela de premiação ser extinta, outra tela com as opções de continuar a jogar ou sair do jogo (Figura 5a) é exibida. Caso os jogadores queiram continuar a jogar, eles devem se cumprimentar novamente, mas se eles quiserem sair do jogo, ambos devem levantar as mãos e o sistema fechará automaticamente (Figura 5b).

Figura 5. (a) Instrução para os jogadores e (b) jogadores saindo do jogo.

5. Resultados

A utilização da ferramenta *Movenet* mostrou-se robusta e capaz de identificar mais de um jogador em tempo suficiente baixo a ponto de ser utilizado como motor de rastreamento dos jogadores em uma projeção de jogos, sem uma latência que pudesse causar dificuldades na mecânica do jogo proposto. A ferramenta se mostrou capaz de identificar as pessoas em situações de baixa luminosidade e em poses variadas. Não houve problemas de rastreamento mesmo com jogadores de diferentes estaturas utilizando vestuário de diversos tipos e cores. O fundo de cenário em que os jogadores se posicionaram não teve influência nos testes realizados. O modelo se manteve estável permitindo identificar a pessoa mesmo se estivesse de costas na imagem (Figura 6a) e também em momentos de oclusão parcial de partes do corpo (Figura 6b).

Figura 6. (a) Jogador de costas para câmera e (b) oclusão de jogador.

Houve dificuldades de rastreamento quando o jogador se posicionava lateralmente na imagem conforme mostra a Figura 7a, em que a jogadora à esquerda parou de ser identificada. Se houver algum detalhe na projeção do jogo que possa se assemelhar à silhueta humana (Figura 7b), pode gerar uma falsa detecção.

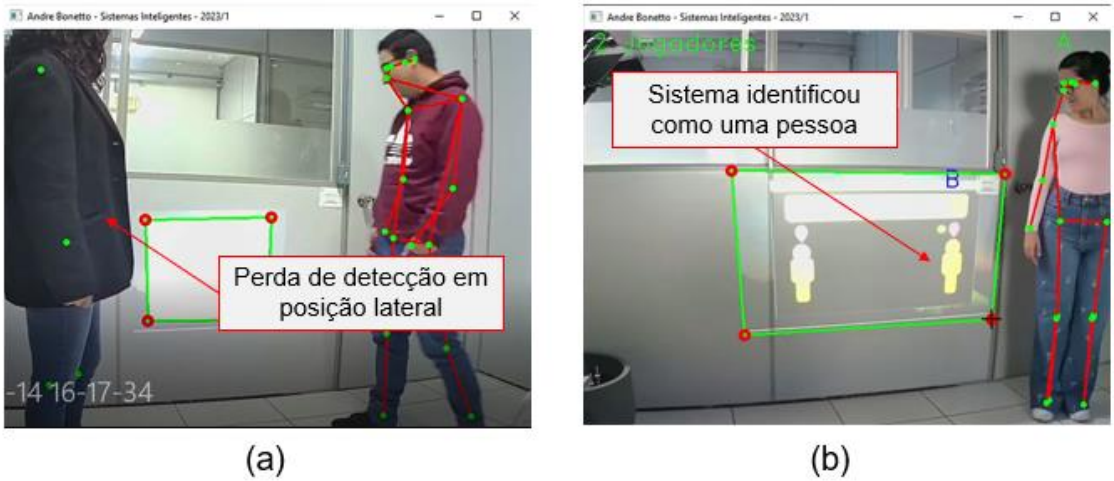

Figura 7. (a) Jogador posicionado lateralmente e (b) falsa detecção.

Outra particularidade da ferramenta é que a terceira dimensão do tensor de saída pode variar a sequência de identificação de pessoas, por exemplo, em um quadro de imagem uma pessoa identificada pode ser a primeira do tensor, e no quadro posterior a mesma pessoa pode ser identificada como segunda do tensor, como mostra a Figura 8, em que os personagens identificados por A e B, são alternados pelo sistema em dois quadros consecutivos. Isto gera dificuldades na hora de criação do jogo, pois o algoritmo deverá contornar esta situação, em caso que, por exemplo, o jogo crie uma pontuação para os jogadores especificamente, logo ele terá que rastrear e de alguma forma fazer um tipo de memorização de cada jogador que estiver na área de jogo.

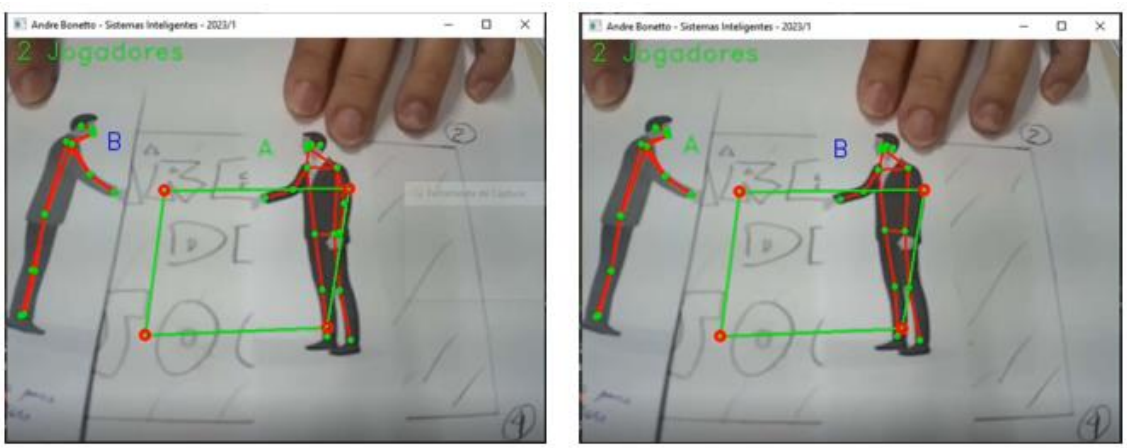

Figura 8. Alternância entre quadros na identificação dos jogadores.

A ferramenta provê 17 pontos de referência do corpo, contudo as extremidades de mebros, mãos e pés, não são contempladas por estes pontos. Logo, isto também cria dificuldades, por exemplo, caso queira-se construir botões de seleção através dos dedos da mão, isto não será possível executar de forma direta. Outra situação é o caso em que a projeção do jogo seja direcionada para o solo e o jogo precise tomar como referência a posição dos pés dos jogadores, da mesma forma, isto não será possível fazer a detecção de forma direta.

6. Conclusões e trabalhos futuros

A inserção de ferramentas de tecnologia em diversos campos do conhecimento mostra-se útil e capaz de automatizar diversos processos e, por consequência, deixando mais tempo para os profissionais que se utilizam destes recursos para aprimorar seus conhecimentos e melhorar seus resultados em seu campo de atuação. O avanço nas novas tecnologias relacionadas ao campo da inteligência artificial vem abrindo novas oportunidades em diversas outras áreas, como pôde ser visto neste trabalho, no qual se utilizou da visão computacional e conseguiu-se criar um jogo colaborativo *multiplayer* com equipamentos convencionais e de fácil acesso, a fim de atender uma necessidade apontada por especialistas em TEA em relação à construção de jogos digitais que objetivam a socialização dos jogadores.

Através da busca exploratória foi possível encontrar diversas ferramentas de detecção de poses de código aberto e que, de modo geral, a ferramenta *Movenet* escolhida para os testes, atendeu de forma satisfatória os objetivos determinados, pois foi possível criar um jogo de *multiplayer* com baixo tempo de latência, utilizando linguagem de programação em código aberto e em que dois jogadores precisaram colaborar para avançar no jogo.

Como trabalho futuro pode-se elencar os seguintes melhoramentos:

- Utilização de outros modelos para realizar uma comparação em termos de desempenho;
- Através da própria visão computacional, desenvolver um algoritmo de calibração automática da área de projeção;
- Desenvolver jogos com outras mecânicas/temas e quantidade de jogadores.

Agradecimentos

Os autores gostariam de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brasil) pela bolsa de produtividade DT2 (processo 306613/2022-0) e, à Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) pelo financiamento parcial ao laboratório LARVA, T.O. No.: 2023TR284 e GASR - T.O No.: 2021TR930.

Referências

Alvarez, J. e Djaouti, D. (2011) “An introduction to Serious Game Definitions and Concepts”, Anais do Serious Games & Simulation for Risks Management Workshop, pp. 11–15.

Bazarevsky, V. *et al*. (2020). “Blazepose: On-device real-time body pose tracking”, In: arXiv preprint arXiv:2006.10204.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). “Realtime multi-person 2d pose estimation using part affinity fields”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291-7299.

Fang, H. S. *et al.* (2022). “Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time”, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Goodfellow I., Bengio Y., e Courville A. (2016). “Deep Learning”, Disponível em: www.deeplearningbook.org. Acesso em: 16-Junho-2023.

Güler, R. A., Neverova, N. e Kokkinos, I. (2018). “Densepose: Dense human pose estimation in the wild”, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7297-7306.

Insafutdinov, E. *et al.* (2017). “Artrack: Articulated multi-person tracking in the wild”, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6457-6465.

Jo, B. e Kim, S. (2022). “Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices”, Traitement du Signal, Vol. 39, No. 1, pp. 119-124.

Kocabas, M., Karagoz, S. e Akbas, E. (2018). “Multiposenet: Fast multi-person pose estimation using pose residual network”, In: Proceedings of the European conference on computer vision, pp. 417-433.

Mehta, D. *et al.* (2020). “XNect: Real-time multi-person 3D motion capture with a single RGB câmera”. ACM Transactions On Graphics, Vol. 39, No. 4, pp. 82-1.

Pishchulin, L. *et al*. (2016). “Deepcut: Joint subset partition and labeling for multi person pose estimation”, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4929-4937.

Szeliski, R. (2021). “Computer Vision: Algorithms and Applications 2nd Edition”, In: Final draft, September 30-2022, The University of Washington, Springer.

TensorFlow (2023). “MoveNet: Modelo de detecção de pose ultrarrápido e preciso”. Disponível em: www.tensorflow.org/hub/tutorials/movenet. [Acesso em: 01-maio-2023].

Tian, Z., Chen, H. e Shen, C. (2019). “Directpose: Direct end-to-end multi-person pose estimation”, In: arXiv preprint arXiv:1911.07451.

Trindade, A. B., Pereira, G. B. e Hounsell, M. da S. (2022). “Chão Interativo e Jogos Sérios Ativos para Autistas: A Plataforma T-TEA e o Jogo RepeTEA”. In Anais Estendidos do XXI Simpósio Brasileiro de Jogos e Entretenimento Digital, pp. 512-521.