UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Brenda Caroline Santos Mendes

Detecção de Interação usando Estimativa de Pose e Detecção de Objetos com Yolov8 em imagens de sessões de terapia com crianças com TEA

Florianópolis
2025

Brenda Caroline Santos Mendes

Detecção de Interação usando Estimativa de Pose e Detecção de Objetos com Yolov8 em imagens de sessões de terapia com crianças com TEA

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciência da Computação.
Orientador: Prof. Mateus Grellert da Silva, Dr.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Mendes, Brenda Caroline Santos
Detecção de interação usando estimativa de pose e detecção de objetos com Yolov8 em imagens de sessões de terapia com crianças com TEA / Brenda Caroline Santos Mendes ; orientador, Mateus Grellert da Silva, 2025.
55 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2025.

Inclui referências.

1. Ciência da Computação. 2. Detecção de Interação. 3. Estimativa de Pose. 4. Detecção de Objetos. I. Silva, Mateus Grellert da . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Brenda Caroline Santos Mendes

Detecção de Interação usando Estimativa de Pose e Detecção de Objetos com Yolov8 em imagens de sessões de terapia com crianças com TEA

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Profa. Carina Friedrich Dorneles, Dra.
Instituição INE/CTC/UFSC

Profa. Manuella Pinto Kaster, Dra.
Instituição BQA/CCB/UFSC

Prof. Randhall Bruce Kreismann Carteri, Dr.
Instituição UFCSPA

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciência da Computação.

ICP-Edu
Documento assinado digitalmente
Cristina Meinhardt
Data: 30/12/2025 16:33:04-0300
CPF: ***.850.380-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

GOV.BR
Documento assinado digitalmente
MATEUS GRELLERT DA SILVA
Data: 30/12/2025 20:25:33-0300
CPF: ***.664.940-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Mateus Grellert da Silva, Dr.
Orientador

Florianópolis, 2025.

Dedicado a Vanderlei, Jovana, Bruna, Bianca e Willy.
Obrigada por sempre estarem ao meu lado, com amor,
carinho e presença constante.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela vida, saúde e força que me permitiram concluir mais uma etapa da minha trajetória.

Minha gratidão ao meu pai Vanderlei, minha mãe Jovana e minhas irmãs Bruna e Bianca, pelo amor incondicional e apoio constante ao longo deste período.

Aos colegas do laboratório, agradeço pela companhia e pelas contribuições que tornaram essa jornada mais leve e enriquecedora.

Ao meu orientador, professor Mateus, sou grata pelos ensinamentos, pela paciência e pelo apoio essencial à realização deste trabalho.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), pelo apoio financeiro que viabilizou este estudo.

“Cada passo dado com coragem
aproxima você do seu objetivo.”
Geronimo Theml

RESUMO

O Transtorno do Espectro Autista (TEA) afeta muitas crianças e pode limitar habilidades de interação social, comunicação e comportamento. O acompanhamento por profissionais qualificados auxilia o desenvolvimento dos pacientes por meio de sessões terapêuticas e permite avaliar seu progresso. No entanto, esse acompanhamento costuma ser registrado manualmente pelos profissionais, o que pode ocasionar falhas nas análises, já que alguns eventos podem passar despercebidos. Para apoiar esse processo de registro, propõe-se uma automatização capaz de predizer o maior número possível de interações, utilizando técnicas de estimativa de pose e detecção de objetos da visão computacional. O objetivo é detectar interações entre os participantes presentes nas sessões de terapia com crianças com Transtorno do Espectro Autista (TEA). Para isso, foram empregados o detector de objetos YOLOv8 e o detector de pose YOLOv8-Pose. Em seguida, foram utilizadas oito heurísticas que combinam as técnicas mencionadas para predizer interações a partir das predições de detecção de objetos e da estimativa de pose. Para melhorar o desempenho, buscaram-se soluções para superar os desafios identificados nas predições desses modelos. Os resultados obtidos demonstraram que as heurísticas que utilizam a pose das pessoas presentaram desempenho superior à heurística que baseada apenas em caixas delimitadoras. Entre essas heurísticas, duas se destacaram, alcançando aumentos de até 12% a acurácia, 15.8% a precisão, 3.17% o recall e 7.34% a confiança f1, resultando em um desempenho satisfatório em relação à proposta do trabalho.

Palavras-chave: Detecção de interação, Estimativa de pose, Detecção de objetos.

ABSTRACT

The Autism Spectrum Disorder (ASD) affects many children and can limit social interaction, communication, and behavioral skills. Monitoring by qualified professionals supports the development of patients through therapeutic sessions and enables the evaluation of their progress. However, this monitoring is usually recorded manually by professionals, which can lead to analysis errors, as some events may go unnoticed. To support this recording process, an automation is proposed that is capable of predicting the largest possible number of interactions, using pose estimation and object detection techniques from computer vision. The goal is to detect interactions between participants present in therapy sessions with children diagnosed with Autism Spectrum Disorder (ASD). For this purpose, the YOLOv8 object detector and the YOLOv8-Pose estimator were employed. Subsequently, eight heuristics were used, combining the aforementioned techniques to predict interactions based on object detection and pose estimation predictions. To improve performance, solutions were sought to overcome the challenges identified in the models' predictions. The results obtained demonstrated that the heuristics using human pose achieved superior performance compared to the heuristic based solely on bounding boxes. Among these heuristics, two stood out, achieving increases of up to 12% in accuracy, 15.8% in precision, 3.17% in recall, and 7.34% in F1-score confidence, resulting in satisfactory performance in relation to the proposed work.

Keywords: Interaction detection, Pose estimation e Object detection

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FN	Falso Negativo
FP	Falso Positivo
HOI	Interação humano-objeto
TEA	Transtorno do Espectro Autista
VC	Visão Computacional
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

O Transtorno do Espectro Autista TEA, segundo (MAENNER *et al.*, 2021), é uma condição do desenvolvimento, que pode afetar habilidades de interação social, comunicação e comportamento. Existem diferentes tipos de tratamentos farmacológicos e não farmacológicos para o TEA (BOSA, 2006). O diagnóstico deste transtorno ainda é essencialmente clínico, baseado na observação do comportamento e do desenvolvimento da pessoa, em conjunto com entrevistas com pais e cuidadores, e na aplicação de instrumentos de avaliação padronizados. A realização desse diagnóstico de forma precoce é de extrema importância, pois permite que medidas de tratamento sejam aplicadas logo nos primeiros anos de desenvolvimento (BOSA, 2006). Quanto antes o diagnóstico for realizado e o tratamento iniciado, maiores são as chances de o indivíduo ter uma vida melhor, pois vai ter uma melhora significativa das características negativas do transtorno.

Uma das formas mais eficazes para tratamento é por meio de sessões de terapia que desenvolvem as habilidades afetadas pelo TEA, como linguística, social e motora. O acompanhamento destas sessões é tipicamente realizado por uma equipe multidisciplinar qualificada, a qual avalia o grau de interação do paciente, assim como seu desenvolvimento social e comunicativo (BOSA, 2006). Dessa forma, algumas análises buscam compreender as interações que ocorrem entre o paciente e o ambiente, como o que chama sua atenção, para o que ele olha, se entende o que deve ser feito em determinado jogo, entre outros aspectos (RAMÍREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018) (KOŁAKOWSKA *et al.*, 2017).

O progresso do paciente é avaliado através de uma sequência de observações e anotações feitas pelo profissional no decorrer da terapia. Dessa forma, pode ser verificado se é melhor prosseguir com o mesmo protocolo ou adaptar os procedimentos de acordo com o progresso do paciente. No entanto, esse registro manual consome tempo significativo e está sujeito a falhas decorrentes das limitações humanas na percepção de certos eventos. Nesse contexto, pode-se utilizar a estratégia de métodos automatizados com visão computacional para auxiliar o profissional.

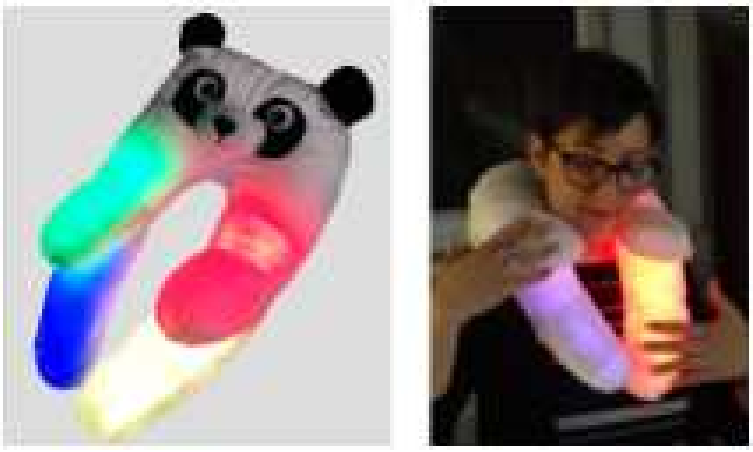
Trabalhos recentes utilizam visão computacional para automatizar diferentes aplicações relacionadas à interação entre humanos e objetos. Alguns utilizam a estimativa de pose principalmente para estimar com precisão a pose humana em tempo real (ANAND; PALANISWAMY, 2023) (ZHANG, Y. *et al.*, 2024) (DONG, C.; TANG; ZHANG, L., 2024) (LI, X.; ZENG; ZHENG, 2023), detectar comportamento de compras (LI, J. *et al.*, 2024), realizar manipulação robótica (MOU *et al.*, 2022), reconhecer linguagens de sinais através de gestos manuais, expressões faciais e pose corporal (AMRUTHA; PRABU; PAULOSE, 2021). No entanto, esses trabalhos não utilizam a estimativa de pose com o propósito de predizer interação.

Esta Dissertação busca desenvolver uma solução robusta para a detecção de interações por meio de heurísticas que combinam técnicas de estimativa de pose e de detecção de objetos. A primeira tarefa consiste em definir esqueletos para um objeto, frequentemente representado por um ser humano. Essa técnica é usada na interação humano-computador, no reconhecimento de comportamento, na detecção e rastreamento de alvos, entre outros (ZHANG, Y. *et al.*, 2024). A segunda tarefa consiste em definir caixas delimitadoras ao redor dos objetos identificados em imagens. Foram definidas 8 heurísticas, sendo que 6 delas envolvem a estimativa de pose e a detecção de objetos.

Portanto, pretende-se analisar as interações entre a criança e o Plusme e entre a terapeuta e o Plusme. Para ficar mais claro, este trabalho busca responder a seguinte pergunta: *Heurísticas que combinam as técnicas de estimativa de pose e detecção de objetos são mais eficientes do que heurísticas que utilizam apenas a técnica de detecção de objetos para detectar interação em imagens?*

Para responder à pergunta, serão utilizadas técnicas de visão computacional, campo da inteligência artificial que envolve o uso de computadores para obter uma compreensão detalhada de dados visuais (XU *et al.*, 2021).

Figura 1 – Dispositivo terapêutico Plusme

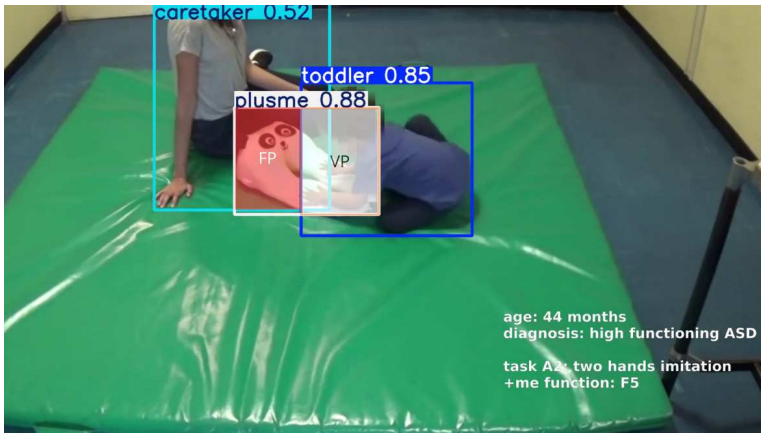


Fonte: (OZCAN *et al.*, 2022)

O presente estudo dá continuidade às investigações iniciadas no Trabalho de Conclusão de Curso do (SOARES, 2023), que realizou análise da interação entre os participantes em sessões de terapia para crianças com TEA. Dentre os participantes, estão as crianças, a terapeuta e o dispositivo terapêutico Plusme, um bichinho de pelúcia representado na Figura 1, projetado pelo Instituto de Ciências e Tecnologias Cognitivas (ISTC-CNR) com o objetivo de ser um brinquedo interativo (INSTITUTO DE CIÊNCIAS E TECNOLOGIAS COGNITIVAS, 2024) (OZCAN *et al.*, 2022). No estudo previamente citado (SOARES, 2023) utilizou a técnica de detecção de objetos e, para a detecção de interação, foi proposta uma heurística de sobreposição de caixas

delimitadoras, que são resultantes da detecção de objetos. Porém, como os participantes podem estar próximos, pode haver sobreposição das caixas mesmo sem ocorrer interação, provocando erros na detecção, conforme ilustrado na Figura 2.

Figura 2 – Representação da detecção de participantes e de suas respectivas interações entre o Plusme e as pessoas utilizando a heurística de sobreposição de caixas delimitadoras.



Fonte: Autor

A Figura 2 ilustra a previsão da interação entre a criança e o Plusme que é destacada em branco, indicando um verdadeiro positivo — no qual há uma interação real, e foi corretamente prevista. A previsão da interação entre a terapeuta e o Plusme resultou em um falso positivo, destacado em vermelho, indicando que, embora não haja interação real, a sobreposição das caixas delimitadoras gerou uma detecção incorreta.

Por fim, este estudo traz novas análises para detecção de interação. Além da heurística utilizada pelo (SOARES, 2023), foram propostas novas heurísticas, que utilizam as técnicas: de estimativa de pose e a de detecção de objetos para predizer interação. Portanto, este trabalho contribui significativamente para a área da Computação, pois explora possibilidades de soluções que une técnicas da visão computacional de forma complementar na solução proposta no campo de detecção de interação, que traz novas abordagens de como predizer essas interações.

1.1 OBJETIVOS

Este trabalho tem como objetivo detectar interações em imagens de forma automatizada, utilizando predições da detecção de objetos e de estimativa de pose para identificar as imagens que contêm interações a partir da análise das heurísticas propostas.

1.1.1 Objetivos Específicos

Para atingir o objetivo geral deste trabalho, foram definidos os seguintes objetivos específicos:

- Treinar um modelo para detectar pessoas e o Plusme nas imagens.
- Realizar predições de detecção de objetos das imagens do conjunto de teste.
- Rotular o posicionamento do Plusme em cada imagem por meio de pontos-chave, a fim de construir um conjunto de dados que possibilite o treinamento da sua estimativa de pose.
- Realizar predições de estimativa de pose das pessoas e do Plusme nas imagens do conjunto de teste.
- Desenvolver heurísticas para a detecção de interações.
- Realizar predições de detecção de interações com base nas heurísticas propostas.

1.2 ESTRUTURA DO TRABALHO

O trabalho inicia-se com a introdução, que apresenta a pergunta de pesquisa e objetivos. O capítulo 2 é composto pelo referencial teórico, no qual são discutidos os principais conceitos, incluindo as palavras-chave do trabalho. Posteriormente, o capítulo 3 traz a revisão de literatura, que aborda os trabalhos relacionados. Em seguida, o capítulo 4 descreve a metodologia utilizada no desenvolvimento do trabalho, e o capítulo 5, metodologia dos experimentos. O capítulo 6 aborda as considerações finais e por fim, as referências utilizadas.

2 REFERENCIAL TEÓRICO

Esta seção descreve os conceitos relacionados ao Transtorno do Espectro Autista, à Visão Computacional e às técnicas de Detecção de Objetos, Estimativa de Pose e Detecção de Interação. Por fim, aborda o Aprendizado de Máquina.

2.1 TRANSTORNO DO ESPECTRO AUTISTA - TEA

O TEA é um distúrbio do neurodesenvolvimento com componentes genéticos e ambientais (CAMPISI *et al.*, 2018). Trata-se de um espectro de alterações comportamentais que pode ser diagnosticado ao longo da vida, embora o diagnóstico geralmente seja realizado na infância, com sintomas perceptíveis desde os primeiros estágios do desenvolvimento (NOGAY; ADELI, 2020) (RAMÍREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018). Ao longo dos anos, houve um aumento considerável na quantidade de novos diagnósticos de TEA em crianças (KOŁAKOWSKA *et al.*, 2017)(LORD *et al.*, 2000). Dessa forma, é importante realizar esse trabalho com uma população de crianças para auxiliar os profissionais nessa demanda.

Pessoas diagnosticadas com TEA geralmente apresentam déficits nos domínios da interação social, da comunicação e de comportamentos repetitivos (SHARMA; GONDA; TARAZI, 2018). Durante a infância, a deficiência na habilidade motora é um dos primeiros sintomas que podem ser observados nas crianças com TEA (KOŁAKOWSKA *et al.*, 2017).

É importante ressaltar que o TEA constitui um espectro de condições heterogêneas, ou seja, cada paciente apresenta um perfil diferente, um caso específico (LORD *et al.*, 2000). Portanto, quando se trata de tratamentos e intervenções, os pacientes têm necessidades distintas (CAMPISI *et al.*, 2018).

2.2 VISÃO COMPUTACIONAL

A Visão Computacional (VC) é um dos pilares da inteligência artificial, responsável por estudar e desenvolver métodos computacionais capazes de analisar processos que permitam às máquinas tomar decisões a partir de elementos do seu ambiente (BARBOSA; JESUS, 2020).

As câmeras são utilizadas para capturar informações que serão posteriormente processadas pelo computador. Por meio dos métodos da VC, as imagens podem ser analisadas pelas máquinas de forma que percebam e reconheçam padrões de maneira semelhante aos humanos (CHERAPANAMJERI; RAO, 2022). Com isso, o termo VC surgiu com o intuito de imitar a visão humana de forma análoga (BARBOSA; JESUS, 2020).

Os métodos de VC possibilitam a execução automatizada de diversas tarefas,

como detecção e rastreamento de objetos, segmentação, classificação, reconhecimento de atividades, estimativa de pose, entre outras (NIGAM; SINGH, R.; MISRA, 2019).

2.2.1 Detecção de Objetos

A detecção de objetos é uma tarefa fundamental da VC (KAUR; SINGH, S., 2023) e serve como base para outras tarefas que requerem a localização de um objeto na imagem, como segmentação e estimativa de pose.

O objetivo da detecção de objetos é identificar e classificar diferentes objetos de determinadas classes presentes em uma imagem. Esse processo envolve localizar o objeto, criando uma caixa delimitadora ao seu redor, e, em seguida, classificá-lo para determinar a categoria à qual pertence (CHERAPANAMJERI; RAO, 2022)(KAUR; SINGH, S., 2023).

A utilização dessa técnica é considerada um dos componentes fundamentais de sistemas de inteligência artificial (CAZZATO *et al.*, 2020). A detecção de objetos pode ser aplicada em diversas áreas, como segurança e vigilância, varejo, agricultura, transporte e assistência médica (KAUR; SINGH, S., 2023)(CHERAPANAMJERI; RAO, 2022).

A Figura 3 apresenta um exemplo de detecção de objetos, no qual foram localizadas e classificadas as pessoas e o ônibus na imagem. As caixas delimitadoras representam cada detecção e, acima de cada uma delas, estão indicadas a classe do objeto e a precisão da predição.

Figura 3 – Exemplo de detecção de objetos

Fonte: (WANG, H.; LI, D.; ISSHIKI, 2024)

2.2.2 Estimativa de Pose

A estimativa de pose, também conhecida como esqueletização, é uma tarefa da VC responsável pela criação do esqueleto de um objeto. Geralmente, esses esqueletos são preditos para seres humanos, mas também podem ser aplicados a outros tipos de objetos, como animais.

Essa técnica utiliza como base as tarefas de detecção de objetos e classificação. A detecção de objetos localiza e desenha a caixa delimitadora ao redor do objeto, enquanto a classificação identifica a classe predefinida que ele pertence. Já a estimativa de pose define o esqueleto do objeto.

A estimativa de pose pode ser empregada para detectar comportamentos humanos servindo como base para muitas aplicações, como: segurança e vigilância, aplicações clínicas e interação humano-robô (NIGAM; SINGH, R.; MISRA, 2019).

A detecção de pose pode ser realizada tanto para uma única pessoa quanto para um grupo com várias pessoas (NIGAM; SINGH, R.; MISRA, 2019). As predições dos esqueletos são obtidas a partir de pontos-chave, compostos por vértices que representam articulações do corpo. Esses vértices são interligados por arestas, que representam os ossos do corpo (SHUJAH ISLAM, 2024).

Existem dois métodos para realizar a estimativa de pose, o método de cima para baixo (*top-down*) e de baixo para cima (*bottom-up*). No método de cima para baixo, primeiro é realizada a detecção do objeto na imagem e, em seguida, a detecção dos pontos-chave do objeto. No método de baixo para cima, primeiro são detectados os pontos-chave do objeto e, após isso, eles são conectados para formar o esqueleto. Esse último método é eficaz na detecção de uma única pessoa (ZHANG, Y. *et al.*, 2024).

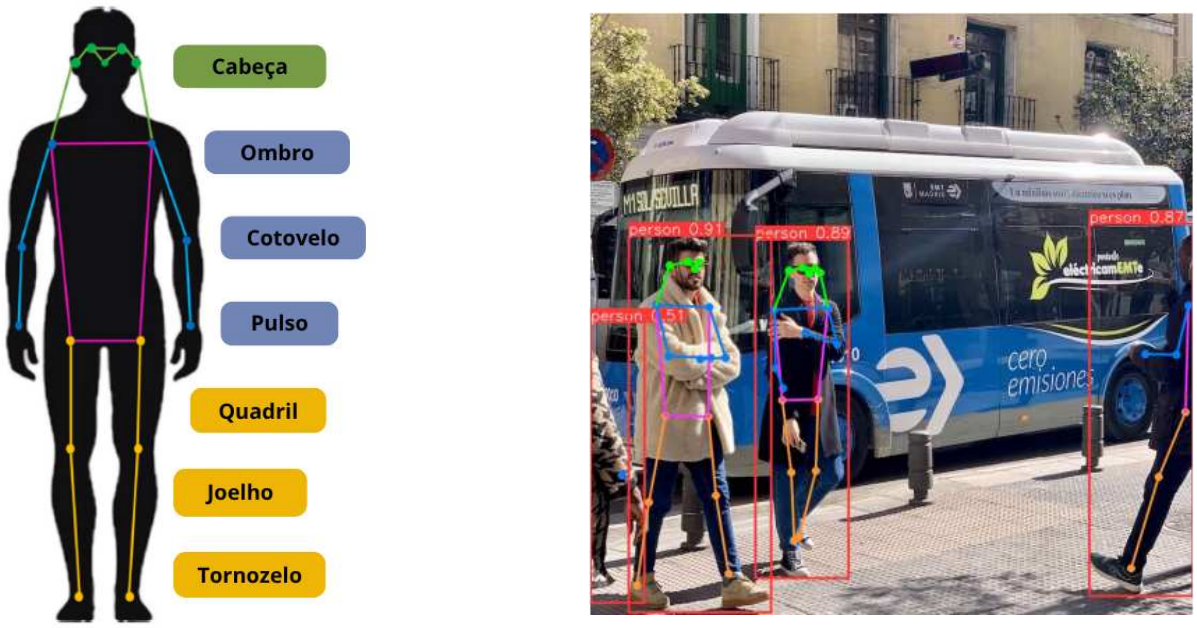
A estimativa de pose humana pode ser observada na Figura 4-esquerda, composta pela detecção de várias partes do corpo, totalizando 17 pontos-chave: orelhas, olhos, nariz, ombros, cotovelos, pulsos, quadris, joelhos e tornozelos(ZHANG, Y. *et al.*, 2024).

A Figura 4-direita mostra um exemplo de predição de estimativa de pose humana. É possível perceber que os pontos-chave de cada pessoa seguem a representação demonstrada na Figura 4-esquerda.

Para este trabalho, além da estimativa de pose das pessoas, também foi definida a pose para o dispositivo Plusme.

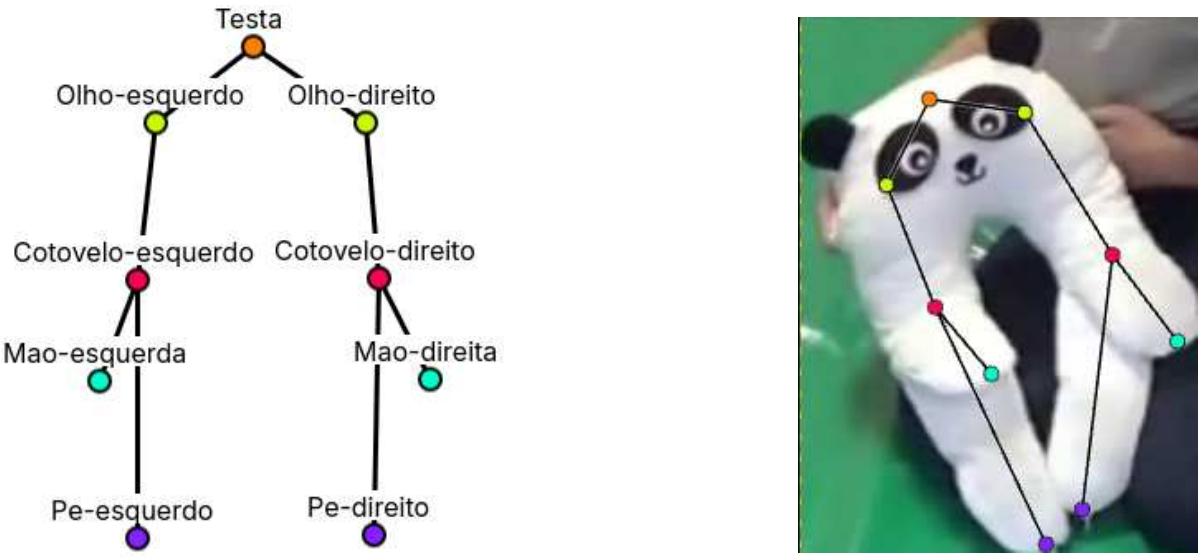
Para detectar a estimativa de pose do brinquedo, foi necessário anotar manualmente sua pose nas imagens do conjunto de dados. Foram definidos 9 pontos-chave: testa (ponto entre os olhos, um pouco acima da região dos olhos), olhos (nos limites da região preta dos olhos de ambos os lados), cotovelos (região onde tem a dobra dos braços), mãos (no término do comprimento dos braços) e pés (no término do comprimento do boneco). A Figura 5 ilustra esses pontos-chave e a anotação da pose

Figura 4 – A imagem à esquerda mostra a distribuição dos pontos-chave da estimativa de pose, enquanto a imagem à direita exemplifica a aplicação da estimativa de pose.



Fonte: (ZHANG, Y. *et al.*, 2024), (WANG, H.; LI, D.; ISSHIKI, 2024)

Figura 5 – A imagem à esquerda representa a distribuição dos pontos-chave da estimativa de pose do Plusme, enquanto a imagem à direita exemplifica a anotação da estimativa de pose realizada.



Fonte: Autor

do Plusme.

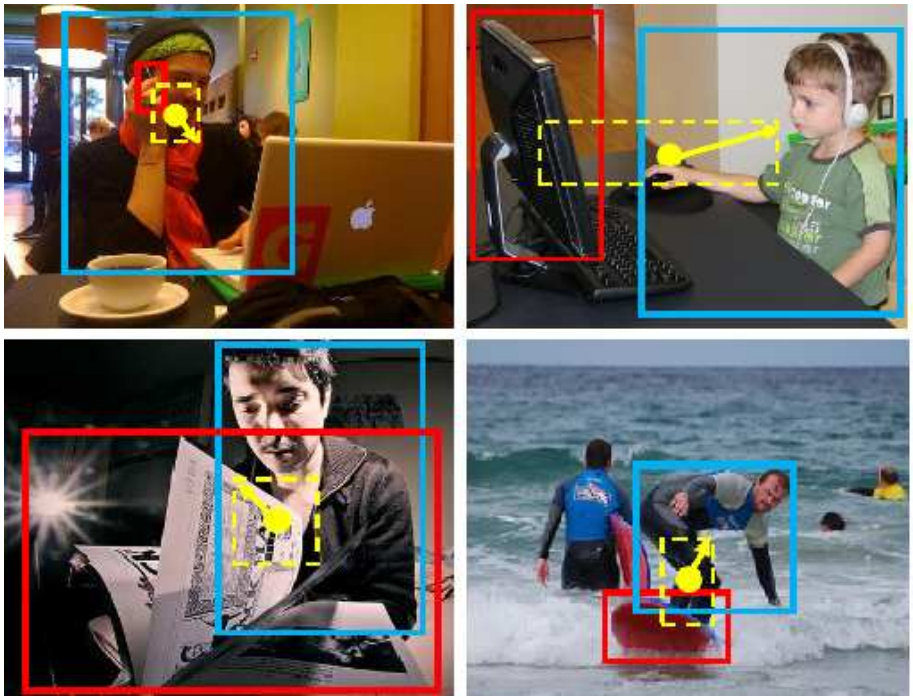
2.2.3 Detecção de Interação

A detecção de Interação humano-objeto (HOI) consiste na análise de cenas para localizar humanos, objetos e as interações entre eles (ANTOUN; ASMAR, 2023) (WANG, T. *et al.*, 2020). Essa técnica busca compreender a relação entre humanos e

objetos, ou seja, como eles interagem (WANG, J. *et al.*, 2023).

A detecção de interação pode ser realizada por abordagens de um ou dois estágios. Na abordagem de dois estágios, inicialmente é feita a detecção de objetos, localizando as caixas delimitadoras e classificando-as como humano ou objeto, em seguida, é prevista a interação entre eles (ANTOUN; ASMAR, 2023). Já na abordagem de um estágio, são extraídos da imagem todos os recursos referentes ao humano, objeto e interação simultaneamente (ANTOUN; ASMAR, 2023).

Figura 6 – Exemplo de detecção de interação humano-objeto



Fonte: (WANG, T. *et al.*, 2020)

A Figura 6 apresenta exemplos de detecção HOI entre pessoas e objetos. As pessoas foram localizadas com caixas delimitadoras na cor azul, os objetos com caixas vermelhas e a interação entre eles é demonstrada na cor amarela.

2.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é um campo da inteligência artificial responsável pela criação de modelos matemáticos capazes de detectar padrões em diversos tipos de dados (BARBOSA; JESUS, 2020).

Existem três tipos principais de aprendizado utilizados no treinamento de modelos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

O aprendizado supervisionado consiste em treinar o modelo com dados rotulados, ou seja, com dados conhecidos de entrada e saída. Após o treinamento, a

predição é comparada aos rótulos para avaliar o quanto o modelo acertou (JANIESCH; ZSCHECH; HEINRICH, 2021) (BARBOSA; JESUS, 2020).

O aprendizado não supervisionado consiste em treinar o modelo com dados não rotulados, ou seja, o modelo deve identificar padrões sem quaisquer informações prévias sobre as categorias (JANIESCH; ZSCHECH; HEINRICH, 2021)(BARBOSA; JESUS, 2020).

No aprendizado por reforço, utiliza-se um sistema de meta a ser atingida. O modelo aprende por tentativa e erro, quando ele acerta, recebe uma recompensa e quando erra, uma penalidade, ajustando seu comportamento para alcançar o objetivo (JANIESCH; ZSCHECH; HEINRICH, 2021).

3 REVISÃO DA LITERATURA

Essa seção trata da literatura dos trabalhos relacionados. Primeiro é abordado sobre a abordagem YOLO, na qual muitos trabalhos utilizaram. Posteriormente são apresentados os trabalhos relacionados à estimativa de pose, detecção de interação e TEA.

Foram realizadas diversas pesquisas ao longo do desenvolvimento do trabalho, entre os anos de 2023 e 2025, utilizando as bases: *SCOPUS*, *IEEE*, Google acadêmico e *Springer*. As palavras chave utilizadas na pesquisa foram: *pose estimation, detection interaction, object detection e yolov8*. Os artigos foram selecionados a partir do título e resumo.

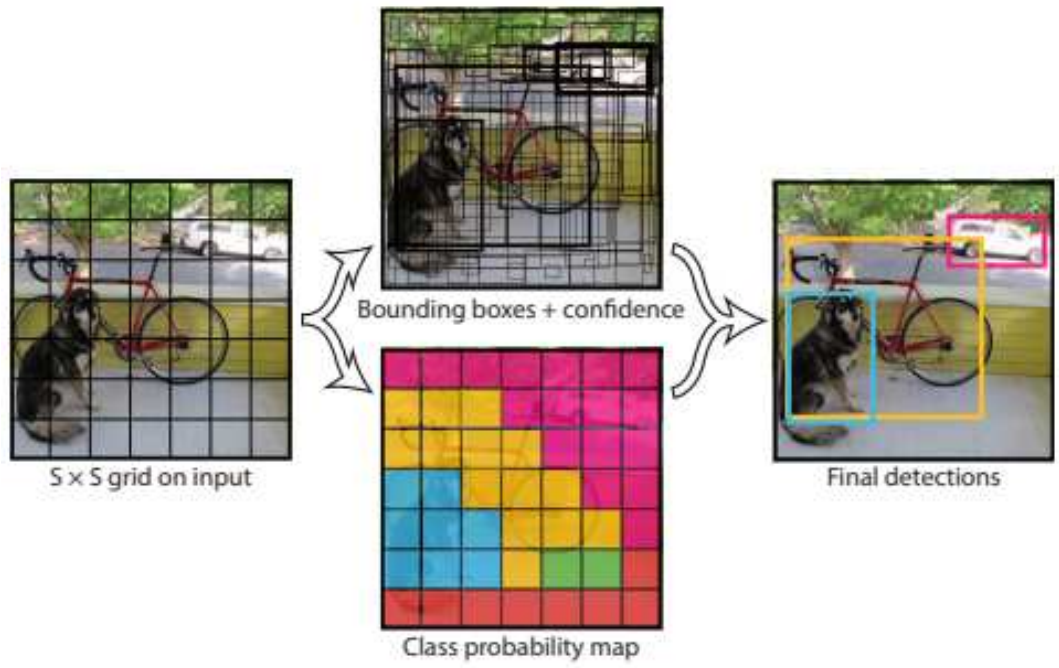
3.1 YOLO

YOLO (*You Only Look Once*), é uma abordagem para detecção de objetos em tempo real. Utiliza uma rede convolucional que prevê simultaneamente múltiplas caixas delimitadoras e probabilidade de classe para essas caixas (REDMON *et al.*, 2016).

O YOLO é o mais popular detector de objetos em tempo real, pela sua arquitetura de rede leve; métodos eficazes de fusão de recursos e resultados mais precisos de detecção (LOU *et al.*, 2023).

A Figura 7 representa como são feitas as detecções dos objetos.

Figura 7 – Representação da detecção do YOLO



Fonte: (REDMON *et al.*, 2016)

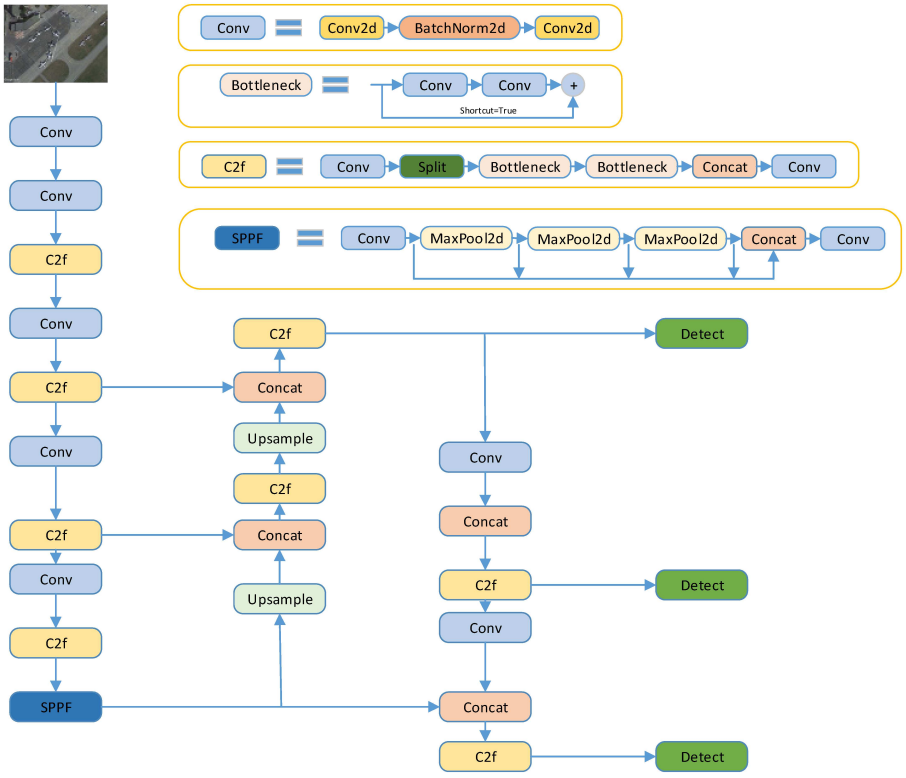
A imagem original é dividida em *grids*. São previstas múltiplas caixas delimi-

tadoras em cada célula do *grid*. Para definir qual a caixa delimitadora mais próxima da detecção correta que abrange todo o objeto, é analisado a confiança da predição, dessa forma, vai prevalecer a caixa delimitadora que possui a maior confiança em relação as outras de cada classe identificada (REDMON *et al.*, 2016).

A abordagem do YOLO iniciou com o objetivo da detecção de objetos, mas ao longo do tempo foram surgindo novas versões com uma abordagem capaz de resolver novos problemas.

A versão YOLOv8 foi desenvolvida pela Ultralytics. Ela suporta diversas tarefas da visão computacional, como: detecção, segmentação, estimativa de pose, rastreamento e classificação (ULTRALYTICS, 2024b). Para realização dessas tarefas, o YOLOv8 oferece diversos modelos pré-treinados para realização delas, mas é possível treinar o seu próprio modelo.

Figura 8 – Arquitetura do YOLOv8



Fonte: (WU; DONG, Y., 2023)

A arquitetura do YOLOv8 pode ser vista na Figura 8. O YOLOv8 emprega arquitetura avançadas de *backbone* e *neck* de última geração. A rede de *backbone* é responsável pela extração de recursos. Já a rede *neck* é responsável por agregar e processar os recursos extraídos pela rede *backbone*, assim, desempenha um papel crucial na integração de recursos de diferentes escalas. Dessa forma, resulta em uma melhor extração de recursos e desempenho de detecção de objetos. (ULTRALYTICS, 2024b)(WU; DONG, Y., 2023).

Uma das características do YOLOv8, é a adoção de modelo sem âncora com cabeças desacopladas para lidar com tarefas de detecção, classificação e regressão de objetos de forma independente (WU; DONG, Y., 2023). Isso contribui para uma melhor precisão e processo de detecção mais eficiente em comparação com as abordagens baseadas em âncora (ULTRALYTICS, 2024b).

3.2 TRABALHOS RELACIONADOS

A partir da pesquisa realizada, nessa seção, são apresentados alguns trabalhos relacionados aos temas abordados nesta Dissertação. Os artigos foram descritos de forma simplificada na Tabela 1 como também de forma mais extensa no decorrer do texto.

3.2.1 Trabalhos sobre Estimativa de Pose

O primeiro trabalho, (ANAND; PALANISWAMY, 2023) trata da capacidade de estimar com precisão as poses de indivíduos capturados em vídeos e gerar animações 3D realistas. A partir disso, surge duas abordagens, a primeira é para realizar a estimativa simultânea de poses de várias pessoas, enquanto a segunda, é a geração de animações 3D dinâmicas com base nas ações capturadas por câmeras diferentes simultaneamente. Para a primeira abordagem, utilizou-se o algoritmo MoveNet Lightning, um modelo de aprendizado profundo projetado para estimativa robusta de poses em cenários com muitas pessoas. Para a segunda abordagem, utilizou-se o YOLOv7, que é uma estrutura de detecção de objetos de última geração, baseada em rede neural convolucional profunda que permite a detecção eficiente e em tempo real de objetos em imagens e vídeos.

A proposta de (ZHANG, Y. *et al.*, 2024) é para a realização de uma rede leve de ponta a ponta chamada SP-YOLO para realizar estimativa de pose humana em tempo real. Essa proposta vem para suprir o problema que alguns algoritmos de aprendizado profundo para estimativa de pose humana têm baixa precisão de posicionamento, baixa confiança no reconhecimento de pessoas e operações de parâmetros grandes.

Já o autor (LI, J. *et al.*, 2024) propõe o BP-YOLO, um novo modelo de venda automática não tripulado. Para detecção de produtos, propôs otimizar o desempenho da detecção introduzindo um mecanismo de atenção 3D e um módulo de convolução deformável na rede de *backbone* e na rede de pirâmide de recursos da estrutura BP-YOLO, respectivamente. Para reconhecimento de compras, usou-se um modelo de fusão visual que combina geração de pontos de esqueleto e estimativa de pose para rastrear o comportamento de compras dos consumidores. Em seguida, aplicou-se BP-YOLO para reconhecer dinamicamente os produtos escolhidos pelos consumidores e seu comportamento de compras em tempo real.

Outra proposta de abordagem é do autor (LI, Xin *et al.*, 2023), que propõe uma abordagem leve e completa de estimativa de pose humana com base na atenção de coordenadas multiescala que não tem o problema de erro de quantização dos métodos anteriores baseados em mapa de calor. Esse método garante uma rede leve e apenas alguns parâmetros e custos de computação são adicionados, o que permite uma representação mais eficiente dos recursos de várias escalas e melhora o desempenho da rede. A rede proposta é melhorada com a rede Yolo-Pose como estrutura.

O trabalho de (DONG, C.; TANG; ZHANG, L., 2024) propõe uma metodologia chamada MDA-YOLO Person para estimativa de pose humana, que é baseada na popular arquitetura de estimativa de pose humana YOLOv7-pose com melhorias. Esse modelo emprega uma estratégia de baixo para cima baseada em regressão de âncora para executar simultaneamente a detecção de pessoas e regredir todos os pontos-chave, aliviando assim o problema de falha na estimativa de pose causada por pontos-chave perdidos.

Enquanto (MOU *et al.*, 2022) propõe um esquema de controle prático e eficaz desenvolvido para realizar a tarefa de manipulação robótica de inserção de cabos aeronáuticos e industriais em aplicações de fabricação. É um método visual robusto e preciso para resolver o problema de estimativa de pose de conectores industriais manipulados. A arquitetura aplicada é baseada em dois YOLOs, para a tarefa de detecção do conector do cabo para detectar instâncias de conectores com uma determinada característica. Inclui uma operação de localização do conector e uma operação de classificação do conector. Enquanto a estimativa de pose é baseada em visão para estimar os estados do conector.

O (SOARES, 2023) propôs um sistema capaz de detectar e classificar os integrantes nas sessões de terapia do TEA. Para isso, utilizou a tarefa de detecção de objetos do YOLOv5. Após a realização das predições de detecção e classificação, é feita a predição de detecção de interação (a partir do toque) entre os integrantes. Com essa predição, os profissionais da saúde obtém um *feedback* dos possíveis momentos que possuem essa interação, e dessa forma basta analisar manualmente se realmente houve interação apenas nesses momentos, permitindo que diminua o trabalho de análise manual do profissional de toda a sessão.

Por fim, este trabalho pretende usar as tarefas de detecção de objetos e estimativa de pose, para a realização do esqueleto das pessoas utilizando o YOLOv8, que oferece um suporte das duas técnicas. Após a realização dessas predições, irá ocorrer a análise da detecção de interação baseada na heurística proposta.

3.2.2 Trabalhos sobre Detecção de interação

A detecção de interação é uma abordagem que possui limitações por situações em que o humano tem interação com mais de um objeto, ou vários humanos possuem

interação com o mesmo objeto (WANG, T. *et al.*, 2020).

Com o objetivo de melhorar a precisão e eficiência da detecção de interação humano-objeto, (WANG, T. *et al.*, 2020) propõe o IP-NET, um método para identificar pontos de interação através da detecção e agrupamento de pontos-chave. Em seu trabalho, (ZHONG *et al.*, 2021) propõe a GGNet, um modelo que utiliza etapas de "olhar"e "fixar"para inferir pontos com reconhecimento de ação, que representam com maior precisão a área de interação. A proposta do (GKIOXARI *et al.*, 2018) é o InteractNet, modelo baseado em uma abordagem centrada no ser humano. Explora a hipótese que a pose, roupa e ações das pessoas, é um indicativo para localizar os objetos nas quais estão interagindo.

3.2.3 Trabalhos sobre TEA

Existe diferentes trabalhos que usam algum tipo de automação para verificar ações comportamentais e tentar resolver problemas associados a diagnósticos. Abaixo são apresentados dois trabalhos, porém nesta dissertação o foco não é o diagnóstico do TEA, mas auxiliar em avaliação de progresso do paciente a partir de interação baseada em toque.

O trabalho do (RAMÍREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018) tem como objetivo apresentar o projeto e implementação de um cenário de interação supervisionada criança-robô que utiliza análise automatizada de pistas não-verbais para auxiliar no diagnóstico de TEA. Ele utiliza um ambiente equipado com câmeras nas paredes, além de pessoas na sala ao lado analisando os resultados. Também conta com alguns brinquedos colados nas paredes, e o robô na mesa. A intenção foi comparar a interação das crianças com TEA e das crianças que não possuem o diagnóstico. As análises automatizada de sinais não-verbais contou com algumas etapas: Detecção e reconhecimento facial da criança; Análise facial, pontos de referência, pose da cabeça e olhar fixo; Fusão de dados e Campo de Visão e Foco Visual de Atenção.

O artigo do (KOŁAKOWSKA *et al.*, 2017) utiliza métodos de aprendizado de máquina para verificar a ideia de reconhecimento do progresso da terapia. Os métodos incluíram algoritmos de treinamento supervisionado aplicados para treinar classificadores de duas classes: progresso e sem progresso. Foram propostos cinco jogos de tablet para as análises: Caixas, Compartilhamento, Cata-vento, Criatividade e Gato e Cachorro.

Tabela 1 – Trabalhos relacionados

Trabalho	Técnica utilizada	Proposta	Aplicação
ANAND; PALANISWAMY, 2023	Estimativa de pose e reconstrução 3D	Estimar com precisão a pose de várias pessoas e gerar animações 3D dinâmicas	Reconhecimento de atividades, análise de comportamento, realidade virtual e experiências interativas.

ZHANG, Y. et al., 2024	Estimativa de pose	Propôs a rede SP-YOLO para estimativa de pose humana em tempo real.	Estimativa de pose humana em tempo real e multi-pessoa.
LI, J. et al., 2024	Detecção de objetos e estimativa de pose	Propôs o BP-YOLO, um modelo de venda automática não tripulado.	Detecção de produtos e reconhecimento de comportamento de compra em máquinas de venda automáticas inteligentes.
LI, X. et al., 2023	Estimativa de pose	Propõe uma abordagem leve e completa de estimativa de pose humana	Estimativa de pose humana em tempo real.
DONG, C.; TANG; ZHANG, L., 2024	Estimativa de pose	Propõe uma metodologia chamada MDA-YOLO Person para estimativa de pose humana.	Estimativa 2D da pose humana em tempo real, inclusive em multidões e situações com oclusão.
MOU et al., 2022	Detecção de região de interesse e estimativa de pose	Propõe um esquema para integrar pose humana para controle/interação de sistemas robóticos	Sistemas robóticos interativos.
SOARES, 2023	Detecção de objetos	Propôs um sistema capaz de detectar e classificar os integrantes presentes em sessões de terapia do TEA e detectar interações baseadas em toque (eventos).	Auxiliar o profissional na revisão e tomada de decisão, nos possíveis momentos em que ocorrem interações nas sessões
WANG, T. et al., 2020	Detecção de interação	Propõe o IP-NET para identificar pontos de interação através da detecção e agrupamento de pontos-chave	Reconhecimento doe humano, ação, objeto em imagens.
ZHONG et al., 2021	Detecção de interação	Propõe a GGNet, um modelo que utiliza etapas de "olhar"e "fixar" para inferir pontos com reconhecimento de ação, que representam com maior precisão a área de interação	Interação Humano-Objeto em uma única etapa, rápido e eficiente.
GKIOXARI et al., 2018	Detecção de interação	Propôs o InteractNet, modelo baseado em uma abordagem centrada no ser humano. Explora a hipótese que a pose, roupa e ações das pessoas, é um indicativo para localizar os objetos nas quais estão interagindo	Reconhecimento de ações e objetos associados em imagens.
RAMÍREZ-DUQUE; FRIZERA-NETO; BASTOS, 2018	Análises de foco visual e movimentos	Criar um sistema robótico para apoiar diagnóstico de autismo por meio de análise automática de pistas não verbais.	Apoio ao diagnóstico clínico de Transtorno do Espectro Autista.
KOŁAKOWSKA et al., 2017	Dados comportamentais coletado por sensores de tablet	Reconhecer automaticamente o progresso terapêutico de crianças com TEA usando parâmetros gerados enquanto jogam cinco jogos desenvolvidos para coleta de dados.	Acompanhamento automático da evolução terapêutica de crianças com TEA usando jogos em tablets.

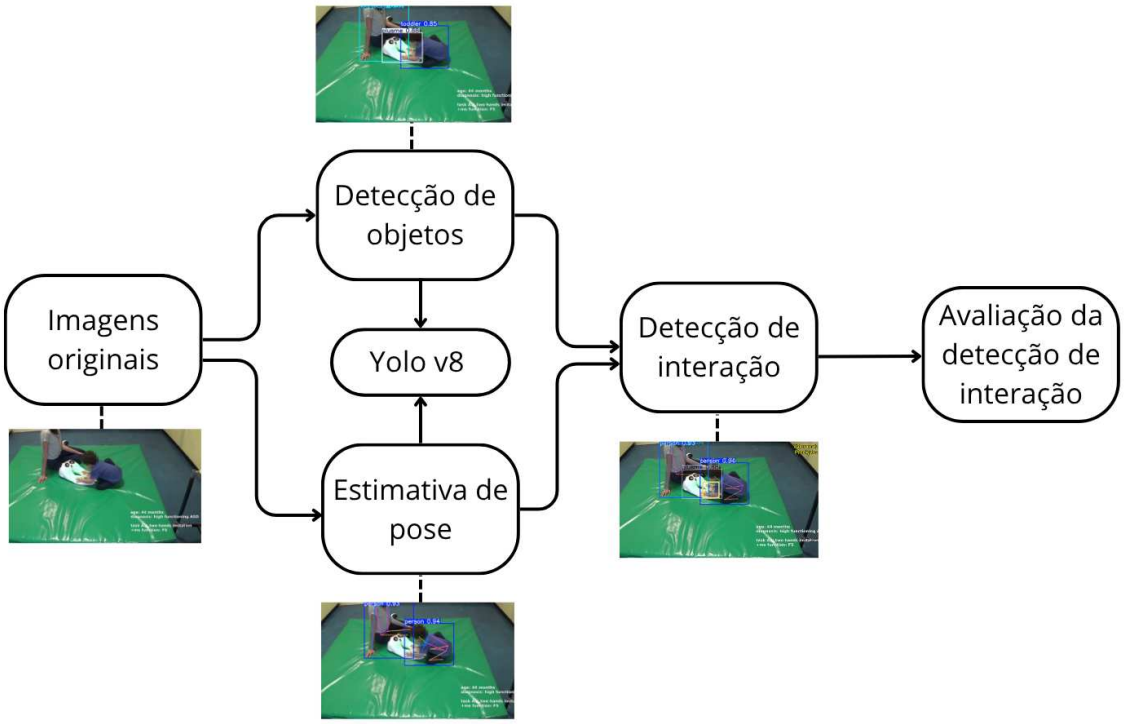
Trabalho do autor	Estimativa de pose e Detecção de objetos	Este trabalho tem a proposta de detectar interações baseadas em toque em imagens de sessões de terapia com crianças com TEA	Auxiliar o profissional na revisão e tomada de decisão, nos possíveis momentos em que ocorrem interações nas sessões, através da predição de interação em imagens utilizando a combinação de técnicas da visão computacional.

4 METODOLOGIA DO TRABALHO

Este trabalho tem como proposta realizar a detecção de interação entre os participantes das imagens das sessões de terapia do TEA. Para atingir esse objetivo, o trabalho foi desenvolvido em etapas que estão representadas na Figura 9.

A partir das imagens originais do conjunto de dados de teste, foram realizadas predições de detecção de objetos utilizando o YOLOv8 e de estimativa de pose utilizando o YOLOv8-Pose. Em seguida, com os resultados dessas predições, foi feita a detecção de interação baseada nas heurísticas propostas. Por fim, aplicou-se métricas de avaliação, para verificar se os resultados obtidos foram satisfatórios.

Figura 9 – Etapas do trabalho.



Fonte: Autor

4.1 DETECÇÃO DE OBJETOS

Para detectar os participantes nas sessões de terapia, utilizou-se a abordagem de detecção de objetos em tempo real YOLO (You Only Look Once) na versão 8 (Yolov8). Essa abordagem foi escolhida por ser um dos principais modelos de detecção de objetos atuais.

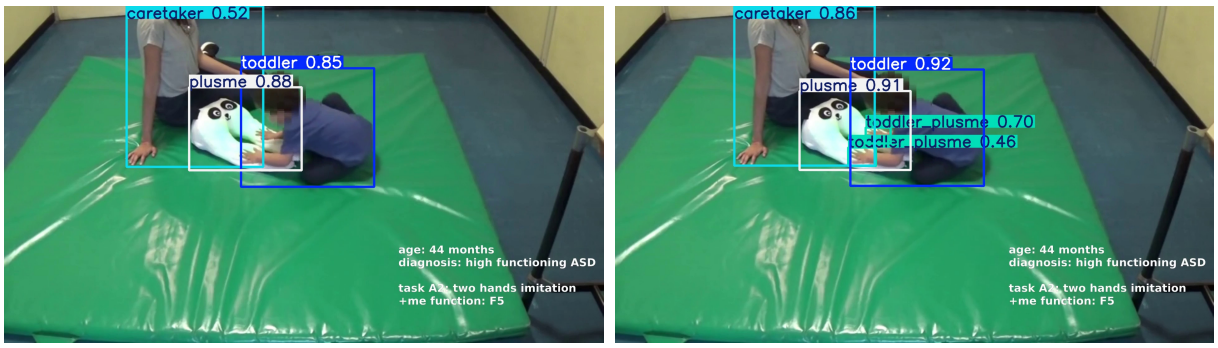
Foram realizados dois treinamentos para a tarefa de detecção de objetos. No primeiro treinamento, utilizou-se o conjunto de dados que possui apenas as três classes principais: Toddler, Caretaker e Plusme.

Para o primeiro treinamento, foi aplicado o *fine-tuning*, técnica que consiste em usar um modelo pré-treinado, e treiná-lo em um novo conjunto de dados específico da tarefa desejada (ULTRALYTICS, 2025). Escolheu-se o modelo pré-treinado *extra-large*, devido a característica de ser mais preciso que os demais.

Os parâmetros utilizados no treinamento, foram os mesmos utilizados por (SOARES, 2023) com o objetivo de obter uma melhor comparação entre os trabalhos. O treinamento foi realizado durante 200 épocas, uma imagem com tamanho 256, portanto o yolo redimensiona a imagem do tamanho original para o tamanho escolhido. Um *early stopping* de paciência igual a 100 épocas e um *batch* igual a 1. Foram avaliados outros modelos com uma imagem de tamanho padrão 640, porém os resultados visuais nas imagens do conjunto de teste não obtiveram resultados significativos. Além disso, modelos treinados sem o uso do *fine-tuning* foram avaliados com a imagem de tamanho 256 e foi observado que os dados apontaram *overfitting* nas últimas 10 épocas de treinamento.

A Figura 10-esquerda demonstra um caso da detecção de objetos aplicada a imagem do conjunto de teste, onde é detectado tanto as caixas delimitadoras dos participantes nas imagens quanto suas respectivas classes. Nessa imagem, foram detectadas a terapeuta, a criança e o Plusme.

Figura 10 – Casos de detecção de objetos aplicados em imagens do conjunto de teste.



Fonte: Autor

No segundo treinamento, utilizou-se o conjunto de dados que possui as 6 classes: Toddler, Caretaker, Plusme, Toddler_Plusme, Caretaker_Plusme e Toddler_Caretaker, o treinamento também foi realizado durante 200 épocas, aplicado ao *fine tuning*, um *early stopping* de paciência igual a 100 épocas e um *batch* igual a 1, porém com uma imagem de tamanho 480.

A Figura 10-direita demonstra um caso da detecção de objetos aplicada a imagem do conjunto de teste, onde é detectado tanto as caixas delimitadoras dos participantes nas imagens quanto as caixas delimitadoras correspondentes às interações entre a criança e o Plusme e suas respectivas classes.

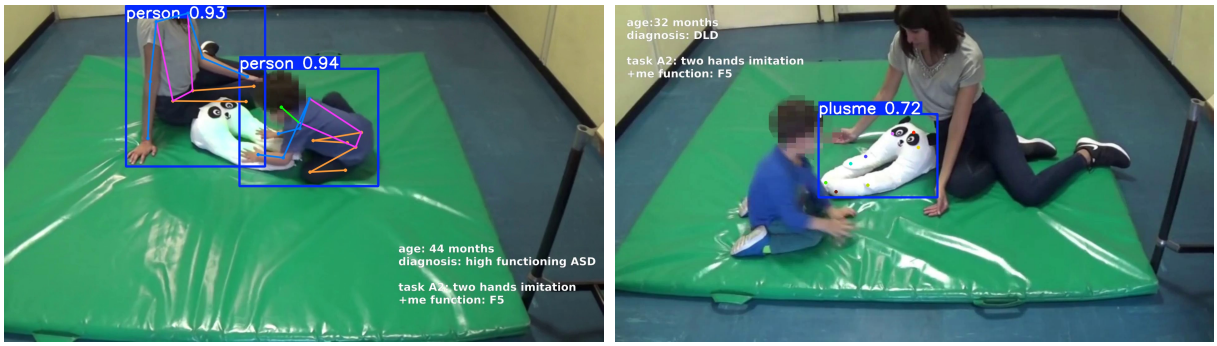
4.2 ESTIMATIVA DE POSE

Para estimar a pose dos seres humanos nas imagens do conjunto de teste conforme a Figura 4, utilizou-se o modelo *extra-large* pré-treinado do YOLOv8 - Pose (ULTRALYTICS, 2024b). Um caso da aplicação da estimativa de pose das pessoas é demonstrado na Figura 11-esquerda, onde mostra os esqueletos preditos da criança e da terapeuta.

O modelo para estimar a pose do plusme foi treinado com *fine tuning*, previsto para 200 épocas, porém com o uso do *early stopping* de paciência igual a 100 épocas, o treinamento ocorreu durante 123 épocas. Uma imagem com tamanho 256 e *batch* igual a 1.

A Figura 11-direita demonstra um exemplo de caso da predição da estimativa de pose do Plusme, onde é possível observar os pontos-chave preditos de sua pose.

Figura 11 – Casos de estimativa de pose aplicados em imagens do conjunto de teste para predizer as poses das pessoas e do Plusme.



Fonte: Autor

4.3 DETECÇÃO DE INTERAÇÃO

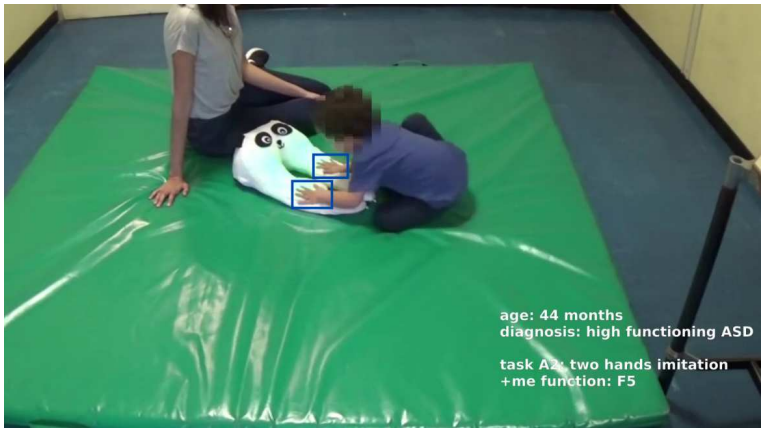
Conforme o diagrama proposto na Figura 9, para detectar a interação, antes é necessário realizar as predições de detecção de objetos e estimativa de pose. Logo após a realização desses passos, a detecção de interação entre os participantes torna-se possível de ser realizada.

A detecção de interação conta com a implementação de algoritmos na linguagem de programação Python para verificar se os participantes possuem interação entre si nas imagens.

Para realizar as análises, considerou-se como interação real as imagens que ocorrem o toque das mãos da criança ou da terapeuta no Plusme. Dessa forma, as predições de interação foram comparadas com as interações reais.

A Figura 12 mostra um exemplo de uma interação real entre a criança e o Plusme. Essa interação é delimitada pelas caixas na cor azul indicando onde ela acontece.

Figura 12 – Exemplo de interação real (toque) da mão da criança no Plusme



Fonte: Autor

4.4 HEURÍSTICAS

Para fins de comparação entre as técnicas de detecção de objetos e a estimativa de pose para identificar interação entre os participantes, estabeleceu-se heurísticas, que são detalhadas abaixo.

4.4.1 Heurística 1

A **Heurística 1** foi utilizada pelo (SOARES, 2023). Ela é relacionada apenas à detecção de objetos: **Será considerado interação entre um par de objetos quando houver sobreposição das suas caixas delimitadoras**.

Neste trabalho, utilizou-se esta heurística utilizada pelo (SOARES, 2023). Além desta, foram criadas outras sete heurísticas, sendo seis delas, combinadas com a estimativa de pose e detecção de objetos.

4.4.2 Heurística 2

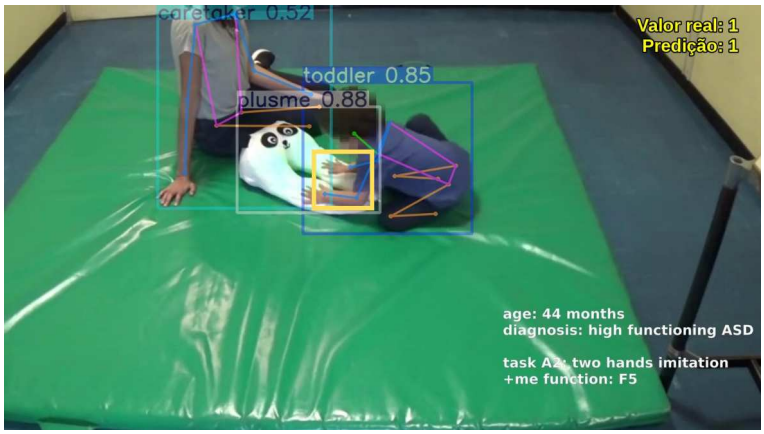
Conforme comentado anteriormente na subseção 2.2.2, o esqueleto humano é composto por pontos-chave e arestas, como na Figura 4. Esses pontos-chave representam as articulações do corpo, então cada ponto representa uma parte específica do corpo. Eles são definidos nas imagens com pontos coloridos, neste caso, os pontos de maior relevância no trabalho, são os pontos em azul, que estão próximos às mãos das pessoas.

Como pretende-se analisar toque entre as pessoas e o Plusme, utilizou-se o ponto-chave mais próximo das mãos, ou seja, o ponto-chave do pulso. Então originou-se a **Heurística 2: Será considerado interação quando houver algum ponto-chave do pulso das pessoas dentro da caixa delimitadora do Plusme**.

Na Figura 13 está ilustrado um exemplo de detecção de interação entre a criança

e Plusme. O quadrado amarelo representa a interação, onde os pontos-chave dos pulsos da criança estão dentro da caixa delimitadora do Plusme, caracterizando assim uma interação.

Figura 13 – Exemplo de interação utilizando a Heurística 2



Fonte: Autor

4.4.3 Heurística 3

Observou-se que alguns pontos dos pulsos não estavam sendo preditos como o esperado, pois como a intenção é avaliar o toque, espera-se que os pontos-chave estejam próximos das mãos, porém, alguns pontos-chave estavam longe das mãos, então criou-se um novo ponto-chave na cor vermelha para poder representar a mão.

Novas análises foram realizadas a partir da **Heurística 3: Será considerado interação quando houver algum ponto-chave da mão das pessoas dentro da caixa delimitadora do Plusme**.

Figura 14 – Detecção de interação com o ponto-chave da mão.

Fonte: Autor

A Figura 14 demonstra um exemplo dessa heurística. O quadrado amarelo representa a interação ocorrida entre a terapeuta e o Plusme com base na Heurística 3, ou seja, o ponto (vermelho) da mão da terapeuta está dentro da caixa do Plusme. O ponto-chave da mão, destacado em vermelho, foi criado para melhorar as análises

em que há interação real, pois em alguns casos, o ponto-chave do pulso está fora da caixa do Plusme, conforme é ilustrado na Figura.

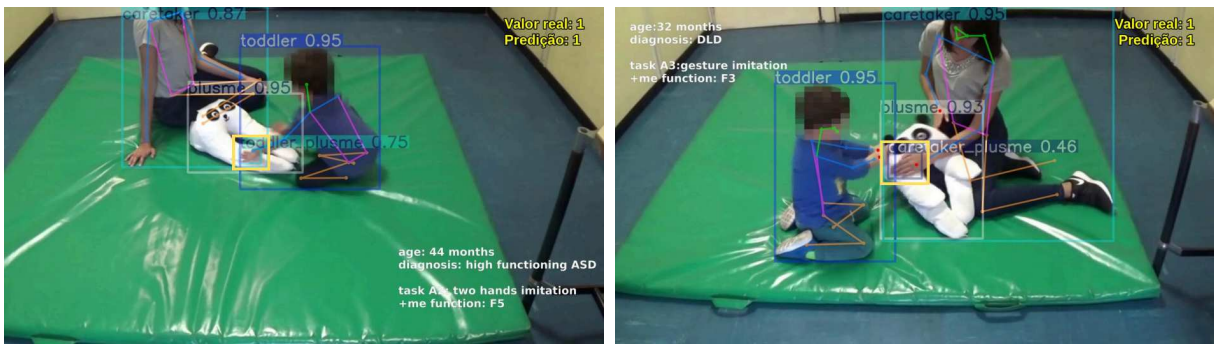
4.4.4 Heurísticas 4 e 5

Utilizando o modelo de detecção de objetos, foi predito as caixas delimitadoras de cada participante como também as caixas delimitadoras de interações entre os participantes. Para as análises a partir das caixas de interação e dos pontos-chave dos pulsos das pessoas, surgiu a **Heurística 4: Será considerado interação quando houver algum ponto-chave dos pulsos das pessoas dentro das caixas delimitadoras de interação**.

Enquanto para análises com as caixas de interação e dos pontos-chave das mãos das pessoas, criou-se a **Heurística 5: Será considerado interação quando houver algum ponto-chave das mãos das pessoas dentro das caixas delimitadoras de interação**.

A aplicação dessas 2 heurísticas podem ser visualizadas na Figura 15. O quadrado amarelo indica a área em que ocorre a interação em cada imagem. Na imagem à esquerda, o ponto-chave do pulso (ponto azul) está dentro da caixa delimitadora de interação, representada na cor ciano; na imagem à direita, o ponto-chave da mão (ponto vermelho) está dentro da caixa delimitadora de interação, representada na cor azul escuro.

Figura 15 – Detecção de interação com os pontos-chave do pulso e da mão.



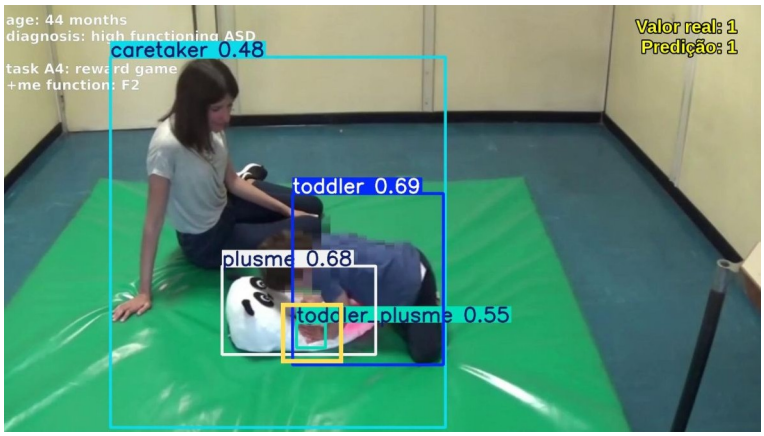
Fonte: Autor

4.4.5 Heurística 6

Considerando ainda as caixas delimitadoras de interação, foi criada a **Heurística 6** de sobreposição de caixas delimitadoras com a caixa do Plusme: **Será considerado interação quando houver sobreposição entre as caixas delimitadoras de interação e a caixa delimitadora do Plusme**.

Na Figura 16, o quadrado amarelo representa a interação detectada por sobreposição da caixa delimitadora do Plusme (caixa branca) e caixa delimitadora de interação (caixa na cor ciano).

Figura 16 – Exemplo de interação utilizando a Heurística 6



Fonte: Autor

4.4.6 Heurísticas 7 e 8

Por fim, para as análises com as predições da pose do Plusme, utilizou-se as caixas delimitadoras das pessoas, onde surgiu a **Heurística 7: Será considerado interação quando houver qualquer ponto-chave do Plusme dentro das caixas delimitadoras da criança e da terapeuta**.

Também foi analisada a pose do Plusme com as caixas delimitadoras de interação, surgindo a **Heurística 8: Será considerado interação quando houver qualquer ponto-chave do Plusme dentro das caixas delimitadoras de interação**.

As Heurísticas 7 e 8 são demonstradas na Figura 17. O quadrado amarelo representa a área em que os pontos-chave do Plusme estão dentro das caixas das pessoas (imagem à esquerda) e dentro das caixas de interação (imagem à direita).

Figura 17 – Detecção de interação com os pontos-chave do Plusme.

Fonte: Autor

4.5 DESAFIOS ENCONTRADOS NAS PREDIÇÕES DE ESTIMATIVA DE POSE E DETECÇÃO DE OBJETOS

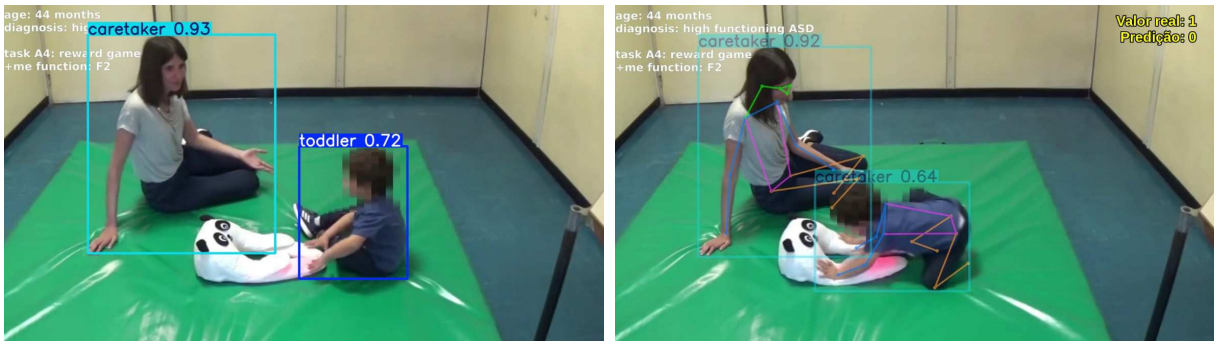
Conforme mencionado anteriormente, para predizer a interação nas imagens, antes é necessário predizer as caixas delimitadoras e o esqueleto dos participantes. Nessas predições, foram encontrados alguns desafios, que são descritos abaixo.

4.5.1 Desafios da Detecção de objetos

Na detecção de objetos, ocorreram situações em que houve dificuldades para reconhecer e detectar os participantes nas imagens, ou seja, algumas caixas delimitadoras dos participantes não foram definidas. Nos casos em que a caixa delimitadora do Plusme não é detectada, não é possível predizer interação naquela imagem quando são analisadas as heurísticas que precisam da caixa do Plusme.

A Figura 18 representa o desafio de não detectar as caixas delimitadoras do Plusme. Dessa forma, não é possível predizer interação em nenhum desses casos, pois na primeira imagem utilizaria a sobreposição da caixa Plusme e caixa das pessoas; enquanto na segunda imagem não é possível verificar se o ponto-chave do pulso está dentro da delimitação da caixa Plusme.

Figura 18 – Desafios em detectar caixas delimitadoras

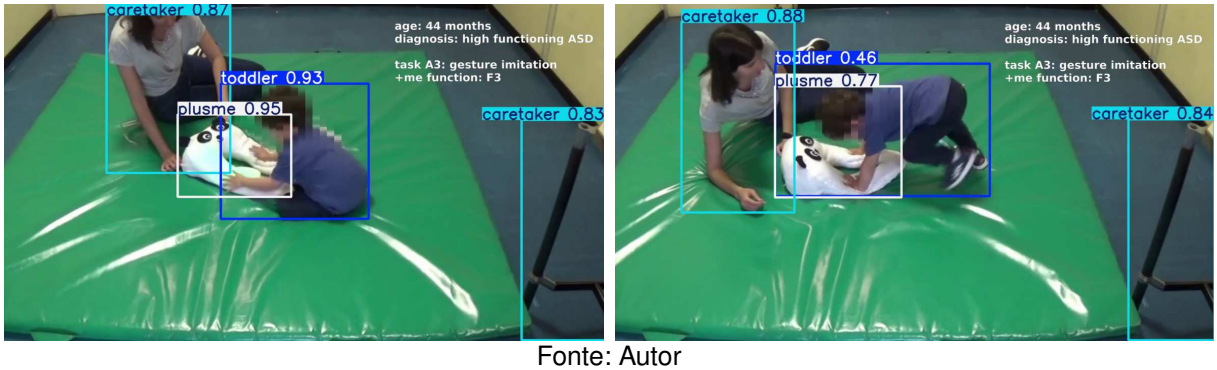


Fonte: Autor

Esse desafio citado anteriormente, também ocorreu nas imagens preditas com a detecção das caixas delimitadoras da criança, terapeuta, Plusme e suas respectivas caixas de interação. Em muitas imagens não foram preditas as caixas de interação. A Figura 19 demonstra casos em que não foram detectadas as caixas delimitadoras de interação ocorridas entre a criança e o Plusme.

Outro tipo de desafios encontrados, foi em relação ao tamanho das caixas delimitadoras do PlusMe em algumas imagens. Estas tiveram uma grande abrangência na parte superior, e resultaram em falsos positivos, pois não havia interação entre a pessoa e o PlusMe, porém o ponto-chave estava inserido dentro das delimitações da caixa do Plusme.

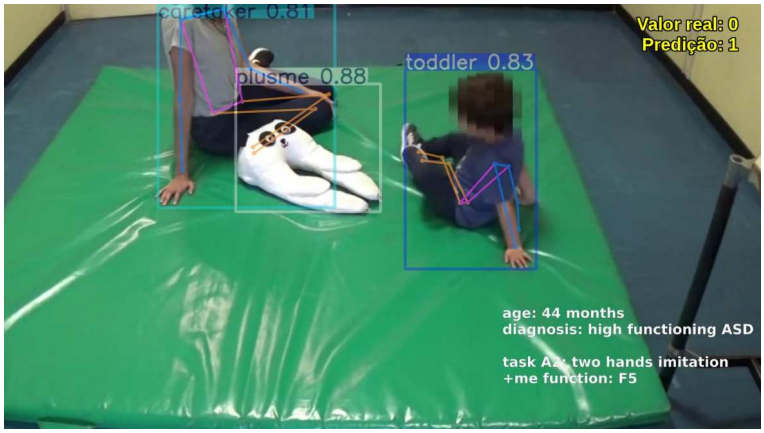
Figura 19 – Desafios em detectar caixas delimitadoras de interação



Fonte: Autor

A Figura 20 ilustra esse caso, onde a caixa delimitadora Plusme é maior na parte superior. Com isso, o ponto-chave do pulso da terapeuta está dentro da delimitação dessa caixa, predizendo interação mesmo sem possuir interação real.

Figura 20 – Desafio da grande abrangência na parte superior da caixa Plusme



Fonte: Autor

Por fim, houve situações em que as classificações não corresponderam ao objeto em questão. Porém as classificações das caixas delimitadoras de forma errada, não influenciam nos resultados.

4.5.2 Desafios da Estimativa de Pose

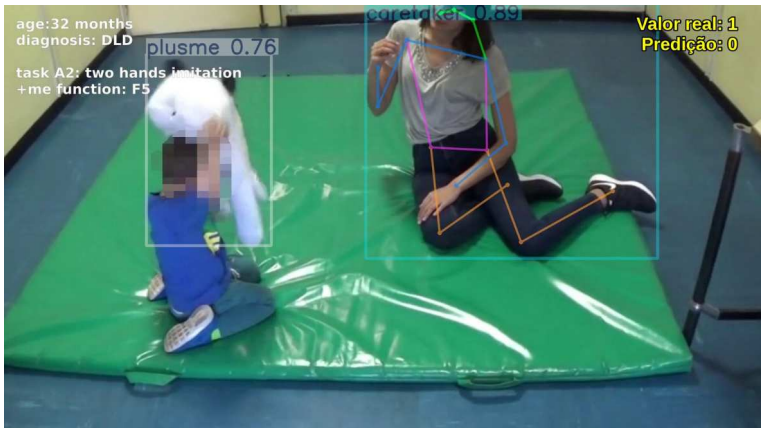
Na estimativa de pose, ocorreram situações em que o esqueleto humano da criança ou terapeuta não foram estimados. Portanto, não é possível verificar interação entre este esqueleto e o PlusMe.

Com esses desafios, caso houver interação real na imagem, vai resultar em um falso negativo, pois não será possível detectá-las.

A Figura 21 mostra que não foi predita a pose da criança. Nesse caso, não é possível verificar se possui interação entre a criança e o Plusme na heurística baseada

em estimativa de pose, pois não possui os pontos-chave para analisar dentro da caixa Plusme.

Figura 21 – Caso que mostra o desafio em não predizer o esqueleto da pessoa.



Fonte: Autor

As predições realizadas com a pose do Plusme, ocorreram situações em que os pontos-chave detectados não estão nas localidades dentro do Plusme conforme foram anotadas. Esse desafio é representado na Figura 22.

Figura 22 – Desafio nas localidades dos pontos-chave preditos da pose do Plusme

Fonte: Autor

Estes foram os desafios observados ao fazer as predições de estimativa de pose e detecção de objetos. Na seção de resultados, são propostas soluções para alguns desses desafios, com o propósito de melhorar as análises e obter resultados mais satisfatórios.

5 METODOLOGIA DOS EXPERIMENTOS

5.1 CONJUNTO DE DADOS

O conjunto de dados utilizado é o mesmo descrito em (SOARES, 2023). Foram utilizados 819 frames aleatórios do total de frames dos vídeos de sessões de terapia com crianças com TEA disponibilizados pelo (INSTITUTO DE CIÊNCIAS E TECNOLOGIAS COGNITIVAS, 2024). As sessões de terapia foram gravadas com 8 crianças, sendo 6 homens e 2 mulheres com idades entre 2 e 4 anos. Além das crianças, os frames das sessões têm a participação de uma terapeuta e do Plusme, conforme é ilustrado na Figura 23.

Figura 23 – Frames de cada trecho do vídeo disponibilizado

Fonte: Autor.

O conjunto de dados foi dividido conforme a Tabela 2. O conjunto de treinamento é composto pelos frames dos vídeos 1 a 4 (419 frames). O conjunto de validação é composto pelos frames dos vídeos 5 e 6 (200 frames) e o conjunto de teste é composto pelos frames dos vídeos 7 e 8 (200 frames).

Tabela 2 – Distribuição do conjunto de dados

Conjunto	Quantidade de frames	Trecho do vídeo
Treinamento	419	1 a 4
Validação	200	5 e 6
Teste	200	7 e 8

Para a detecção de objetos, utilizou-se o conjunto de dados já anotado com as classes: Toddler, Caretaker e Plusme, que correspondem às caixas delimitadoras da criança, da terapeuta e do Plusme. Além disso, foram utilizadas as anotações: Toddler_Plusme, Caretaker_Plusme e Toddler_Caretaker que correspondem às interações ocorridas entre esses participantes.

Para a estimativa de pose do Plusme, foi realizada a anotação do conjunto de dados com a classe Plusme.

5.2 MÉTRICAS DE AVALIAÇÃO

Para avaliar as interações detectadas nas imagens entre as pessoas e o Plusme, utilizou-se algumas métricas de avaliação para verificar o desempenho das predições no conjunto de teste. As métricas utilizadas foram a acurácia, precisão, recall, confiança F1 e a matriz de confusão.

5.2.1 Acurácia

A acurácia representa o desempenho do modelo nas classes. É calculada como a proporção das amostras classificadas corretamente em relação a contagem total de amostras (KAUR; SINGH, S., 2023). Sua equação é demonstrada abaixo.

$$P = \frac{TP + TN}{TP + TN + FP + FN}$$

5.2.2 Precisão

A precisão representa a exatidão do modelo, ou seja, quantas identificações positivas foram realmente corretas. É calculada a partir da razão entre o número de amostras positivas identificadas com precisão para a contagem total de amostras positivas (KAUR; SINGH, S., 2023)(CHERAPANAMJERI; RAO, 2022). A equação correspondente a precisão pode ser vista abaixo.

$$P = \frac{TP}{TP + FP}$$

5.2.3 Recall

O *recall* representa quantos positivos reais foram identificados corretamente. É avaliado como a proporção do número de amostras positivas corretamente identificadas para o número total de amostras positivas reais (KAUR; SINGH, S., 2023).

$$R = \frac{TP}{TP + FN}$$

5.2.4 Confiança F1

A métrica de confiança F1, representa a média harmônica entre a precisão e *recall* em diferentes limites de confiança. O valor dessa métrica será alto se tanto a precisão quanto o recall forem altos, caso contrário, o valor baixo de F1 mostra um desequilíbrio entre a precisão e *recall* (KAUR; SINGH, S., 2023)(ULTRALYTICS, 2024a). O cálculo é feito da seguinte forma:

$$F1 = 2 \times \frac{\text{precisão} \times \text{recall}}{\text{precisão} + \text{recall}}$$

5.2.5 Matriz de confusão

A matriz de confusão permite visualizar o desempenho de um algoritmo. Os valores da diagonal principal representam as previsões corretas realizadas para cada classe. Enquanto os demais valores representam as classificações erradas para cada classe (ULTRALYTICS, 2024a).

- **Verdadeiro Positivo (VP)**: Casos positivos classificados corretamente. As imagens possuem interação real e as interações foram preditas corretamente.

- **Falso Positivo (FP)**: Casos negativos classificados incorretamente como positivos. As imagens não possuem interação real, mas foram preditas interações.

- **Verdadeiro Negativo (VN)**: Casos negativos classificados corretamente. As imagens não possuem interação real e foram preditas corretamente.

- **Falso Negativo (FN)**: Casos positivos classificados incorretamente como negativos. As imagens possuem interação real, mas não foram preditas.

5.3 EXPERIMENTOS

Conforme relatado anteriormente, foram realizados dois treinamentos com o YOLOv8 para a tarefa de detecção de objetos. As predições resultantes das caixas delimitadoras da criança, do terapeuta e do Plusme foram analisadas nas Heurísticas 1, 2 e 3.

5.3.1 Experimentos do primeiro treinamento com YOLOv8

A Heurística 1 considera interação quando há sobreposição das caixas delimitadoras. Inicialmente, considerou-se a sobreposição de quaisquer caixas delimitadoras (criança, terapeuta e Plusme), que representa a interação criança-plusme, terapeuta-plusme e criança-terapeuta.

Como as análises para as Heurísticas 2 e 3 que utilizam estimativa de pose, abrangem apenas as interações das pessoas com o Plusme, ou seja, criança-plusme ou terapeuta-plusme, foi realizada uma segunda análise da Heurística 1 de sobreposição de caixas delimitadoras, considerando apenas as sobreposições existentes com a caixa delimitadora do Plusme. Essa análise foi utilizada para fins de comparação com as heurísticas que empregam estimativa de pose.

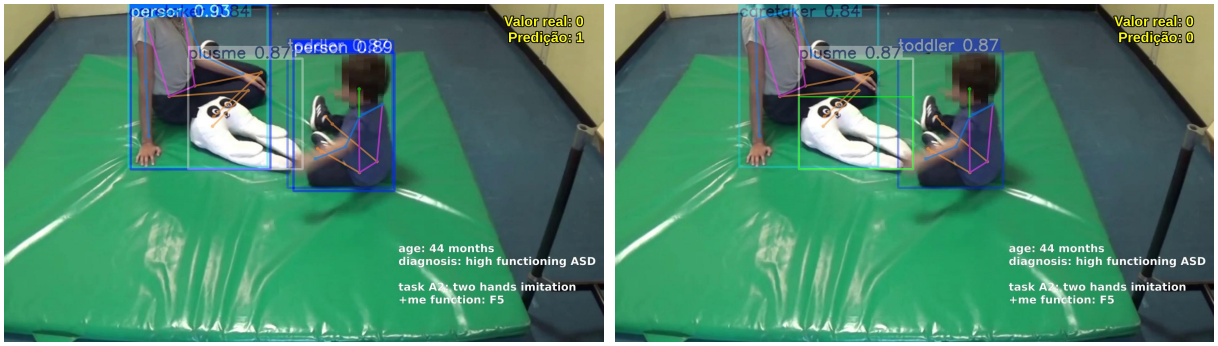
5.3.1.1 Soluções para os desafios encontrados

Para melhores resultados nas análises que utilizam a pose, buscou-se solucionar alguns desafios encontrados, que foram descritas na Subseção 4.5.1 - *Desafios encontrados nas predições de Estimativa de Pose e Detecção de Objetos*. Como houve ocorrências em que a pose das pessoas não foram detectadas, optou-se por incluir

a sobreposição de caixas entre o Plusme e as pessoas, caso esta fosse identificada. Assim, nas imagens em que tanto a pose da pessoa quanto a caixa do Plusme foram detectadas, apenas a análise dos pontos-chave dentro da caixa do Plusme foi utilizada. Porém, nos casos em que somente a caixa do Plusme foi detectada, aplicou-se a heurística de sobreposição de caixas delimitadoras, para abranger a imagem na análise de detecção de interação.

Inicialmente, muitos falsos positivos foram detectados em imagens devido à alta abrangência da parte superior da caixa do Plusme, que ocupava mais espaço que o necessário na sua delimitação. Para superar esse desafio, as imagens que resultaram em falsos positivos foram aplicadas em um código que ajusta a delimitação da caixa do Plusme a partir da análise da cor branca (cor do dispositivo Plusme). Dessa forma, foi possível criar caixas delimitadoras mais proporcionais ao seu tamanho, conforme a Figura 24, que demonstra o tamanho predito original da caixa do Plusme na imagem à esquerda, enquanto a imagem da direita mostra a delimitação da nova caixa do Plusme na cor verde.

Figura 24 – A imagem à esquerda demonstra a grande abrangência na parte superior da caixa do Plusme, enquanto a imagem à direita mostra a nova caixa delimitada na cor verde



Fonte: Autor

Posteriormente, esses novos dados foram aplicados nas análises das Heurísticas 1 e 2, que usam a estimativa de pose, ou seja, caso o ponto-chave do pulso ou da mão da pessoa estiver dentro da caixa do Plusme, o que resultou em significativa redução dos falsos positivos.

5.3.2 Experimentos do segundo treinamento com YOLOv8

O segundo treinamento de detecção de objetos com o YOLOv8, resultou em predições das caixas delimitadoras da criança, do terapeuta, do Plusme e das caixas de interação. Elas foram analisadas nas Heurísticas 4, 5, 6, 7 e 8.

Para as Heurísticas 4 e 5, inicialmente foram realizadas as análises em qualquer ponto-chave do pulso ou da mão das pessoas dentro das caixas de interação detectadas. Com os resultados, observou-se que obteve um número elevado de FN,

devido o fato que em algumas imagens, as caixas delimitadoras que correspondem as interações não foram detectadas.

Para melhorar os resultados e reduzir os FN, foi realizada uma nova análise com etapas adicionais de refinamento. Primeiramente, nas imagens em que não foram detectadas caixas de interação, mas em que a caixa do PlusMe foi identificada, verificou-se se os pontos-chave dos pulsos e das mãos das pessoas estavam contidos na caixa do dispositivo, a fim de considerar uma interação. Em seguida, nas imagens classificadas como FN, aplicou-se uma nova delimitação na parte superior da caixa do Plusme, conforme já adotado nas heurísticas anteriores.

A Heurística 6, segue o mesmo princípio da Heurística 1, de analisar interações a partir da sobreposição de caixas delimitadoras apenas com a caixa do Plusme. Porém, neste caso, a sobreposição ocorre entre a caixa do Plusme e as caixas de interação.

As Heurísticas 7 e 8, envolvem a estimativa de pose do Plusme. Para a análise de interação, foram utilizadas as caixas delimitadoras das pessoas e das interações. A avaliação consistiu em verificar se algum ponto-chave do PlusMe estava contido nessas caixas. Nenhuma etapa adicional de aprimoramento ou ajuste foi aplicada para melhorar os resultados.

5.3.3 Resultados

Para a visualização dos resultados obtidos em cada heurística, utilizou-se uma tabela. Esses resultados foram derivados da avaliação dos percentuais calculados para cada métrica, complementada por análises visuais das imagens. Não foi realizada análise estatística.

A Tabela 3 apresenta a comparação entre os resultados das heurísticas, evidenciando as melhorias obtidas. Nessa tabela são demonstrados os resultados das métricas de avaliação aplicadas em cada heurística. As linhas correspondentes às Heurísticas 2, 3, 4 e 5, após a aplicação das etapas destinadas a solucionar os desafios, referem-se aos ajustes implementados para superar essas dificuldades encontradas.

Conforme a Tabela 3, os resultados obtidos nas Heurísticas 1, 2 e 3 antes da aplicação das etapas destinadas a solucionar os desafios, foram semelhantes. Já nas Heurísticas 4, 5, 6, 7 e 8, foram obtidos altos valores de precisão ou de *recall*, acima de 90%, enquanto os demais resultados permaneceram baixos.

Os valores obtidos nas métricas não atingiram o desempenho máximo, devido aos desafios encontrados relatados na Subseção 4.5.1. Os principais fatores que impactaram esses resultados foram: a não detecção das caixas delimitadoras do PlusMe, das caixas de interação e da pose das pessoas. Como as heurísticas dependem da detecção dos participantes, sempre que houver interação real em uma imagem, mas algum desses elementos não for detectado, a interação não poderá ser corretamente

Tabela 3 – Comparação entre heurísticas

Heurísticas	**Métricas**			
	Acurácia	**Precisão**	**Recall**	**Confiança F1**
I (Quaisquer interações)	68%	71.23%	82.54%	76.47%
I (Apenas interações com Plusme)	68.5%	72.03%	81.75%	76.58%
II	67.5%	71.94%	79.36%	75.47%
III	67%	70.83%	80.95%	75.55%
II (Após a aplicação das etapas destinadas a solucionar os desafios)	**80.5%**	**87.83%**	**80.16%**	**83.81%**
III (Após a aplicação das etapas destinadas a solucionar os desafios)	**78.5%**	**83.74%**	**81.75%**	**82.73%**
IV	54%	97.22%	27.78%	43.21%
IV (Após a aplicação das etapas destinadas a solucionar os desafios)	**79.5%**	**82.94%**	**84.92%**	**83.92%**
V	59.5%	97.87%	36.51%	53.18%
V (Após a aplicação das etapas destinadas a solucionar os desafios)	**73%**	**74.32%**	**87.30%**	**80.29%**
VI	61.5%	98.04%	39.68%	56.49%
VII	63.5%	64.97%	91.27%	75.91%
VIII	52.5%	96.97%	25.39%	40.25%

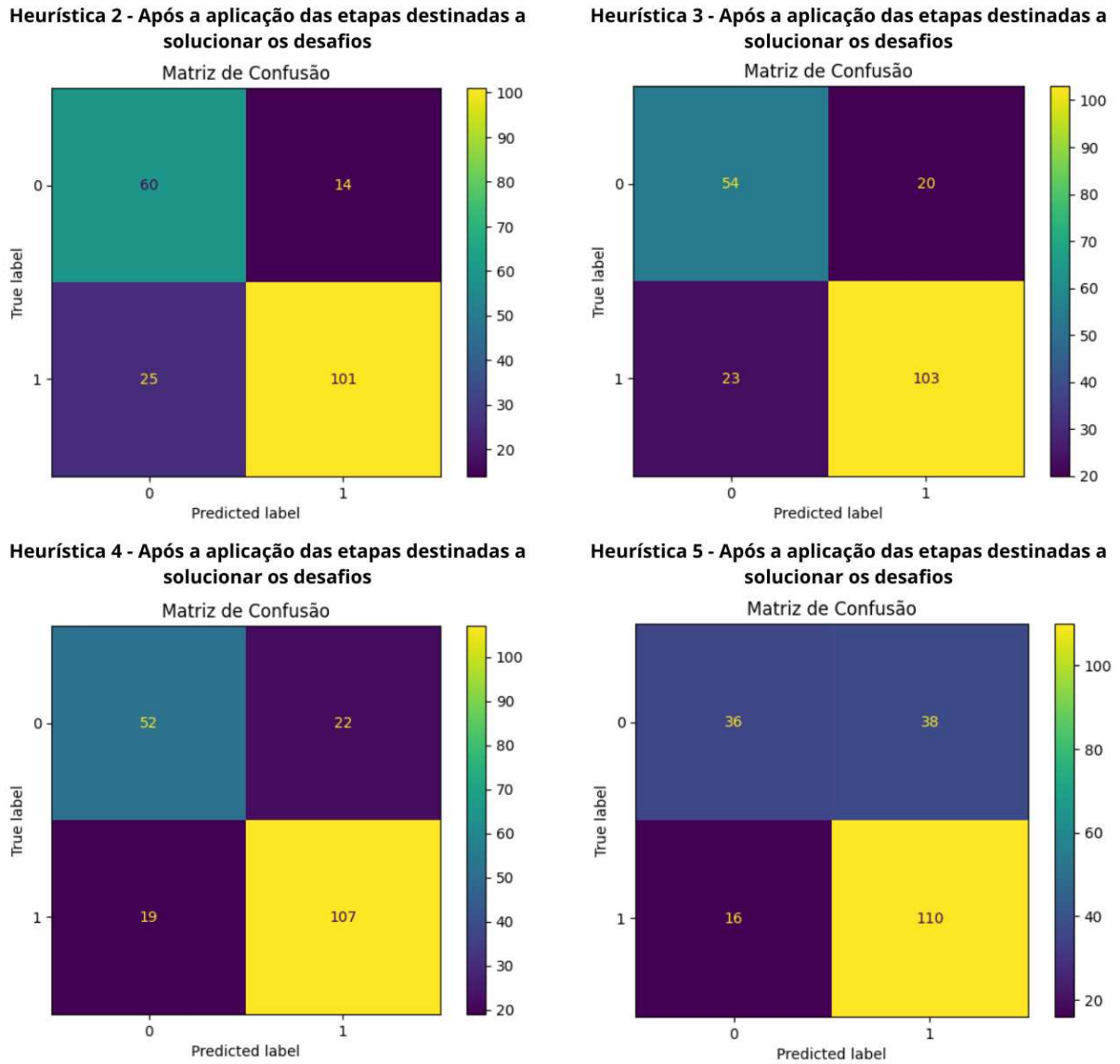
predita.

Após a implementação das melhorias propostas para superar os desafios identificados, foram obtidos resultados satisfatórios nas Heurísticas 2, 3, 4 e 5, que combinam as técnicas de estimativa de pose com detecção de objetos. Esses resultados superam os da Heurística 1, que utiliza apenas detecção de objetos e considera exclusivamente as interações com o dispositivo Plusme.

As métricas de acurácia, precisão e *recall* foram examinadas para definir a melhor heurística de pose a partir de 3 objetivos: de todas as predições verdadeiras e falsas, quantas o modelo acertou (acurácia); de todas as predições verdadeiras, ou seja, as predições que possuem interação, quantas o modelo acertou (precisão); e de todas as possibilidades verdadeiras, ou seja, de todas as imagens que possuem interação real, quantas o modelo acertou (*recall*).

As Heurísticas 2 e 4 tiveram os maiores valores de acurácia com 80.5% e 79.5%

Figura 25 – Matriz de confusão das heurísticas que utilizam a estimativa de pose das pessoas.



Fonte: Autor

respectivamente. Enquanto a precisão das Heurísticas 2 e 3 foram de 87.83% e 83.74%. Já o *recall* alcançou o valor de 87.30% na Heurística 5 e 84.92% na Heurística 4. Também foi analisada a matriz de confusão resultante de cada uma dessas heurísticas, mostrada na Figura 25.

A matriz de confusão mostra a quantidade de imagens que foram classificadas corretamente, se possuem ou não interação, e a quantidade de imagens que foram classificadas incorretamente. Destacam-se as matrizes de confusão das Heurísticas 2 e 4. Enquanto a Heurística 2 obteve mais valores VN (60) e menos valores FP (14) em relação as demais, a Heurística 4 alcançou 107 imagens classificadas como VP e 19 imagens classificadas como FN. Embora a Heurística 5 resultou em mais imagens classificadas como VP, ela também foi a que apresentou o maior número de FP.

Portanto, com base nessas observações, os melhores resultados foram obtidos

pelas Heurísticas 2 e 4 após a implementação das melhorias propostas para superar os desafios. Ambas heurísticas utilizam os pontos-chave dos pulsos das pessoas para detectar interação, desde que estejam dentro das caixas delimitadoras, sendo a Heurística 2 com a caixa do Plusme e a Heurística 4 com as caixas de interação. Elas demonstraram desempenho superior em comparação à heurística baseada apenas na sobreposição de caixas delimitadoras.

6 CONSIDERAÇÕES FINAIS

Neste trabalho, foi proposta uma comparação entre heurísticas que utilizam técnicas de detecção de objetos e estimativa de pose para predizer interações entre participantes presentes em imagens de sessões de terapia com crianças com TEA. Para alcançar esse objetivo, foram empregadas técnicas de Visão Computacional: detecção de objetos, para predizer as caixas delimitadoras dos participantes, e estimativa de pose, para predizer o esqueleto das pessoas e do dispositivo PlusMe. Em seguida, essas predições foram utilizadas para identificar interação, por meio de heurísticas desenvolvidas a partir da combinação dessas duas técnicas.

Foram utilizadas oito heurísticas, sendo que uma delas foi proposta no trabalho do (SOARES, 2023), no qual esta Dissertação se baseia. As demais heurísticas foram desenvolvidas para explorar diferentes possibilidades de experimentação. Entre elas, as que apresentaram melhores resultados foram aquelas que combinam o esqueleto das pessoas com as caixas do Plusme e com as caixas de interação.

Durante as análises, foram observados diversos desafios em algumas imagens. Muitos deles foram solucionados por meio de etapas adicionais aplicadas no processo; contudo, outros permaneceram sem solução e foram deixados como proposta para trabalhos futuros.

Com base na questão apresentada na introdução, conclui-se, a partir das análises dos resultados, que as heurísticas que utilizam tanto a estimativa de pose quanto a detecção de objetos — após a implementação das melhorias propostas para superar os desafios identificados — apresentaram desempenho superior à heurística baseada apenas na sobreposição de caixas delimitadoras.

As Heurísticas 2 e 4, após a aplicação das etapas destinadas a solucionar os desafios, obtiveram os melhores resultados, com o aumento de até 12% na acurácia, 15.8% na precisão, 3.17% no *recall* e 7.34% no *F1-score*, em comparação aos resultados da Heurística 1, baseada exclusivamente na sobreposição das caixas delimitadoras do Plusme.

Parte deste trabalho foi publicado em um artigo no site da IEEE (MENDES; TYSKA; GRELLERT, 2025). O artigo foi aceito no 38º Simpósio Internacional IEEE sobre Sistemas Médicos Baseados em Computador (CBMS), realizado entre 18 e 20 de junho de 2025, em Madri, na Espanha, onde foi apresentado. Nele, também são discutidas heurísticas para detecção de interação, utilizando as mesmas técnicas exploradas nesta Dissertação.

6.1 PERSPECTIVAS FUTURAS

Para buscar resolução para os desafios de não detecção das caixas delimitadoras e esqueleto, que não foram solucionados neste trabalho, recomenda-se a

realização de novos treinamentos, utilizando versões mais recentes do YOLO e explorando diferentes configurações de parâmetros. Além disso, destaca-se a importância de ampliar o conjunto de dados, permitindo que o modelo seja treinado com mais exemplos, a fim de aumentar as chances de detecção das caixas e do esqueleto.

Para possibilitar comparações mais robustas, recomenda-se o desenvolvimento de novas heurísticas e utilização de outras técnicas da Visão Computacional, como a segmentação de objetos, com o objetivo de delimitar melhor a área ocupada pelo Plusme e obter resultados mais precisos no que diz respeito à interação por toque.

REFERÊNCIAS

AMRUTHA, K.; PRABU, P.; PAULOSE, Joy. Human Body Pose Estimation and Applications. *In*: 2021 Innovations in Power and Advanced Computing Technologies (i-PACT). [*S.l.*: *s.n.*], 2021. p. 1–6.

ANAND, RN; PALANISWAMY, Suja. Multi Person Pose Estimation and 3D Pose Detection Animation. *In*: IEEE. 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). [*S.l.*: *s.n.*], 2023. p. 1–5.

ANTOUN, Maya; ASMAR, Daniel. Human object interaction detection: Design and survey. **Image and Vision Computing**, Elsevier, v. 130, p. 104617, 2023.

BARBOSA, Lucas Amparo; JESUS, Leone da Silva de. Como as Máquinas Enxergam? Um apanhado teórico e prático sobre Visão Computacional aplicada a problemas cotidianos utilizando Aprendizado de Máquina e Inteligência Artificial. **Sociedade Brasileira de Computação**, 2020.

BOSA, Cleonice Alves. Autismo: intervenções psicoeducacionais. **Brazilian Journal of Psychiatry**, Associação Brasileira de Psiquiatria, v. 28, s47–s53, mai. 2006. ISSN 1516-4446.

CAMPISI, Lisa; IMRAN, Nazish; NAZEER, Ahsan; SKOKAUSKAS, Norbert; AZEEM, Muhammad Waqar. Autism spectrum disorder. **British Medical Bulletin**, v. 127, n. 1, p. 91–100, ago. 2018. ISSN 0007-1420. eprint: `https://academic.oup.com/bmb/article-pdf/127/1/91/28035519/ldy026.pdf`.

CAZZATO, Dario; CIMARELLI, Claudio; SANCHEZ-LOPEZ, Jose Luis; VOOS, Holger; LEO, Marco. A survey of computer vision methods for 2d object detection from unmanned aerial vehicles. **Journal of Imaging**, MDPI, v. 6, n. 8, p. 78, 2020.

CHERAPANAMJERI, Jyothsna; RAO, B Narendra Kumar. Neural networks based object detection techniques in computer vision. *In*: IEEE. 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA). [*S.l.*: *s.n.*], 2022. p. 1092–1099.

DONG, Chengang; TANG, Yuhao; ZHANG, Liyan. MDA-YOLO Person: a 2D human pose estimation model based on YOLO detection framework. **Cluster Computing**, Springer, p. 1–18, 2024.

GKIOXARI, Georgia; GIRSHICK, Ross; DOLLÁR, Piotr; HE, Kaiming. Detecting and Recognizing Human-Object Interactions. *In*: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [*S.l.*: *s.n.*], 2018. p. 8359–8367.

INSTITUTO DE CIÊNCIAS E TECNOLOGIAS COGNITIVAS, ISTC-CNR. **Project overview**. 2024. Disponível em: `https://www.plusme-h2020.eu/overview/`. Acesso em: 18 jul. 2024.

JANIESCH, Christian; ZSCHECH, Patrick; HEINRICH, Kai. Machine learning and deep learning. **Electronic Markets**, Springer, v. 31, n. 3, p. 685–695, 2021.

KAUR, Ravpreet; SINGH, Sarbjeet. A comprehensive review of object detection with deep learning. **Digital Signal Processing**, v. 132, p. 103812, 2023. ISSN 1051-2004.

KOŁAKOWSKA, Agata; LANDOWSKA, Agnieszka; ANZULEWICZ, Anna; SOBOTA, Krzysztof. Automatic recognition of therapy progress among children with autism. **Scientific Reports**, Springer Science e Business Media LLC, v. 7, n. 1, 2017.

LI, Jingxiang; TANG, Fuquan; ZHU, Chao; HE, Shiwei; ZHANG, Shujin; SU, Yu. BP-YOLO: A Real-Time Product Detection and Shopping Behaviors Recognition Model for Intelligent Unmanned Vending Machine. **IEEE Access**, IEEE, 2024.

LI, X.; ZENG, L.; ZHENG, L. Improvement of the key point detection algorithm based on yolov8. *In*: PROCEEDINGS of the International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023). [*S.l.*: *s.n.*], 2023. P. 25.

LI, Xin; GUO, Yuxin; PAN, Weiguo; LIU, Hongzhe; XU, Bingxin. Human pose estimation based on lightweight multi-scale coordinate attention. **Applied Sciences**, MDPI, v. 13, n. 6, p. 3614, 2023.

LORD, Catherine; COOK, Edwin H.; LEVENTHAL, Bennett L.; AMARAL, David G. Autism Spectrum Disorders. **Cell Press**, v. 28, p. 355–363, 2000.

LOU, Haitong; DUAN, Xuehu; GUO, Junmei; LIU, Haiying; GU, Jason; BI, Lingyun; CHEN, Haonan. DC-YOLOv8: small-size object detection algorithm based on camera sensor. **Electronics**, MDPI, v. 12, n. 10, p. 2323, 2023.

MAENNER, M. J.; SHAW, K. A.; BAKIAN, A. V.; ET AL. Prevalence and characteristics of autism spectrum disorder among children aged 8 years — Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2018-. **MMWR Surveill Summ**, v. 70, SS-11, p. 1–16, 2021.

MENDES, Brenda Caroline Santos; TYSKA, Jônata; GRELLERT, Mateus. Interaction Detection in Images of Therapy Sessions with Children with Autism Spectrum Disorder. *In*: 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS). [*S.l.*: *s.n.*], 2025. p. 585–590.

MOU, Fangli; REN, Hao; WANG, Bin; WU, Dan. Pose estimation and robotic insertion tasks based on YOLO and layout features. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 114, p. 105164, 2022.

NIGAM, Swati; SINGH, Rajiv; MISRA, AK. A review of computational approaches for human behavior detection. **Archives of Computational Methods in Engineering**, Springer, v. 26, n. 4, p. 831–863, 2019.

NOGAY, Hidir Selcuk; ADELI, Hojjat. **Reviews in the Neurosciences**, v. 31, n. 8, p. 825–841, 2020.

OZCAN, Beste; SPERATI, Valerio; GIOCONDO, Flora; SCHEMBRI, Massimiliano; BALDASSARRE, Gianluca. Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. *In*: PROCEEDINGS of the 21st Annual ACM Interaction Design and Children Conference. [*S.l.*: *s.n.*], 2022. p. 625–628.

RAMÍREZ-DUQUE, Andrés A.; FRIZERA-NETO, Anselmo; BASTOS, Teodiano Freire. Robot-Assisted Diagnosis for Children with Autism Spectrum Disorder Based on Automated Analysis of Nonverbal Cues. *In*: 7TH IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob). Enschede, Holanda: [*s.n.*], 2018. p. 456–461.

REDMON, Joseph; DIVVALA, Santosh; GIRSHICK, Ross; FARHADI, Ali. You only look once: Unified, real-time object detection. *In*: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [*S.l.*: *s.n.*], 2016. p. 779–788.

SHARMA, Samata R.; GONDA, Xenia; TARAZI, Frank I. Autism Spectrum Disorder: Classification, diagnosis and therapy. **Pharmacology Therapeutics**, v. 190, p. 91–104, 2018. ISSN 0163-7258.

SHUJAH ISLAM, M. Computer vision-based approach for skeleton-based action recognition, SAHC. **Signal, Image and Video Processing**, Springer, v. 18, n. 2, p. 1343–1354, 2024.

SOARES, Alexandre Soli. **Uma Ferramenta para auxílio a terapias do Transtorno do Espectro Autista usando Aprendizado de Máquina**. 2023. F. 65. Monografia (monografia) – Faculdade em Engenharia Eletrônica, UNIVERSIDADE FEDERAL DE SANTA CATARINA, Florianópolis.

ULTRALYTICS. **F1 confidence, Precision-Recall curve, Precision-Confidence curve, Recall Confidence curve and Confusion matrix**. 2024. Disponível em: `https://github.com/ultralytics/ultralytics/issues/7307`. Acesso em: 15 jul. 2024.

ULTRALYTICS. **Fine-tuning**. 2025. Disponível em: https://www.ultralytics.com/glossary/fine-tuning. Acesso em: 19 fev. 2025.

ULTRALYTICS. **Ultralytics YOLO Docs**. 2024. Disponível em: https://docs.ultralytics.com/pt. Acesso em: 28 jun. 2024.

WANG, Hansen; LI, Dongju; ISSHIKI, Tsuyoshi. Energy-Efficient Implementation of YOLOv8, Instance Segmentation, and Pose Detection on RISC-V SoC. **IEEE Access**, v. 12, p. 64050–64068, 2024.

WANG, Jia; SHUAI, Hong-Han; LI, Yung-Hui; CHENG, Wen-Huang. Human-Object Interaction Detection: An Overview. **IEEE Consumer Electronics Magazine**, p. 1–14, 2023.

WANG, Tiancai; YANG, Tong; DANELLJAN, Martin; KHAN, Fahad Shahbaz; ZHANG, Xiangyu; SUN, Jian. Learning Human-Object Interaction Detection Using Interaction Points. *In*: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [*S.l.*: *s.n.*], jun. 2020.

WU, Tianyong; DONG, Youkou. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. **Applied Sciences**, MDPI, v. 13, n. 24, p. 12977, 2023.

XU, Shuyuan; WANG, Jun; SHOU, Wenchi; NGO, Tuan; SADICK, Abdul-Manan; WANG, Xiangyu. Computer vision techniques in construction: a critical review. **Archives of Computational Methods in Engineering**, Springer, v. 28, p. 3383–3397, 2021.

ZHANG, Yuting; WANG, Zongyan; LI, Menglong; GAO, Pei. SP-YOLO: an end-to-end lightweight network for real-time human pose estimation. **Signal, Image and Video Processing**, Springer, v. 18, n. 1, p. 863–876, 2024.

ZHONG, Xubin; QU, Xian; DING, Changxing; TAO, Dacheng. Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection. *In*: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [*S.l.*: *s.n.*], 2021. p. 13229–13238.