

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO  
CENTRO TECNOLÓGICO  
DEPARTAMENTO DE ENGENHARIA ELÉTRICA  
PROJETO DE GRADUAÇÃO



LUCAS FEHLBERG FOLLADOR

ESTUDO E IMPLEMENTAÇÃO DE MÉTODO DE  
ESTIMAÇÃO E RASTREAMENTO 3D DE  
MÚLTIPLAS PESSOAS EM UM ESPAÇO  
INTELIGENTE

VITÓRIA-ES

DEZEMBRO/2023

Lucas Fehlberg Follador

# ESTUDO E IMPLEMENTAÇÃO DE MÉTODO DE ESTIMAÇÃO E RASTREAMENTO 3D DE MÚLTIPLAS PESSOAS EM UM ESPAÇO INTELIGENTE

Parte manuscrita do Projeto de Graduação do aluno Lucas Fehlberg Follador, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Vitória-ES

Dezembro/2023

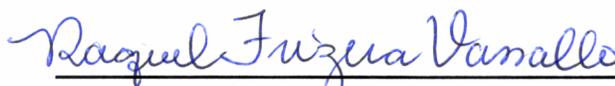
Lucas Fehlberg Follador

# ESTUDO E IMPLEMENTAÇÃO DE MÉTODO DE ESTIMAÇÃO E RASTREAMENTO 3D DE MÚLTIPLAS PESSOAS EM UM ESPAÇO INTELIGENTE

Parte manuscrita do Projeto de Graduação do aluno Lucas Fehlberg Follador, apresentado ao Departamento de Engenharia Elétrica do Centro Tecnológico da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

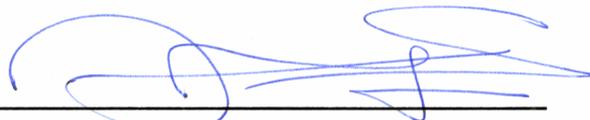
Aprovado em 19 de dezembro de 2023.

COMISSÃO EXAMINADORA:



---

**Profa. Dra. Raquel Frizera Vassallo**  
Universidade Federal do Espírito Santo  
Orientador



---

**Prof. Dr. Jorge Leonid Aching  
Samatelo**  
Universidade Federal do Espírito Santo  
Examinador



---

**Msc. Felipe Mendonça de Queiroz**  
Universidade Federal do Espírito Santo  
Examinador

Vitória-ES

Dezembro/2023

*Aos meus pais, por todo o apoio, incentivo e carinho.*

## AGRADECIMENTOS

Gostaria de agradecer aos meus pais, Claudécir e Tereza, por todo o apoio prestado durante todos esses anos, incentivando-me sempre a estudar e dar o melhor de mim.

Aos meus amigos de curso: Danilo, Eric, Felipe, Luiz, Pedro, Witalo e muitos outros que tive o privilégio de conhecer durante essa jornada, que me acompanharam nos momentos de estudo e descontração.

À minha orientadora, Raquel Frizera Vassallo, por toda a ajuda e orientação não só durante o desenvolvimento deste trabalho, mas desde quando entrei em seu laboratório em busca de uma iniciação científica.

Aos membros do Laboratório VISIO por todo o apoio, principalmente na realização dos experimentos deste trabalho.

À banca examinadora pela aceitação do convite e pelo tempo investido para leitura e avaliação deste trabalho.

Finalmente, agradeço à Universidade Federal do Espírito Santo pela minha formação.

## RESUMO

Com os avanços nas áreas de eletrônica e computação, objetos comuns estão se tornando mais inteligentes, assim como ambientes que, equipados com sensores e atuadores, formam os chamados espaços inteligentes. No Laboratório de Visão Computacional e Robótica (Lab VISIO) da Universidade Federal do Espírito Santo (UFES), foi desenvolvido um Espaço Inteligente Programável (PIS) que oferece diversos serviços aos usuários, além da capacidade de criar novas funcionalidades e aplicações. Essas aplicações visam interagir com o usuário, possibilitando uma interação humano-robô ou humano-ambiente, como, por exemplo, o controle de dispositivos, como acender ou apagar uma luz através de gestos, ou pedir ajuda. Muitas dessas aplicações requerem um serviço de estimação de pose humana 3D e rastreamento das poses geradas. Desta forma, com o intuito de aprimorar ainda mais a interação com as pessoas no ambiente, este trabalho examinou métodos presentes na literatura, escolhendo um deles para a implementação da estimação e rastreamento 3D de poses humanas em um sistema de múltiplas câmeras. A fim de assegurar que a solução escolhida fosse adequada ao PIS, foram conduzidos dois experimentos. O primeiro teve como objetivo avaliar o desempenho em tempo real, enquanto o segundo visou validar a reconstrução e rastreamento 3D, proporcionando um exemplo prático de sua aplicação.

**Palavras-chave:** Espaço Inteligente; Pose Humana; Rastreamento; Reconstrução 3D; Tempo Real.

## ABSTRACT

Due to the advances in the fields of electronics and computing, everyday objects are becoming smarter, as are environments equipped with sensors and actuators, forming what are known as smart spaces. At the Laboratório de Visão Computacional e Robótica at the Universidade Federal do Espírito Santo (UFES), a Programmable Intelligent Space (PIS) has been developed that offers various services to users, along with the ability to create new functionalities and applications. These applications aim to interact with the user, enabling human-robot or human-environment interaction, such as controlling devices by gestures, as turning a light on or off, or asking for assistance. Many of these applications require a 3D human pose estimation service and tracking of the generated poses. Therefore, with the goal of further enhancing interaction with people in the environment, this work examined methods present in the literature, selecting one of them to be implemented for 3D human pose estimation and tracking in a multi-camera system. In order to ensure that the chosen solution was suitable for the PIS, two experiments were conducted. The first aimed to evaluate real-time performance, while the second aimed to validate 3D reconstruction and tracking, providing a practical example of its application.

**Keywords:** 3D Reconstruction; Human Pose; Intelligent Space; Real-time; Tracking.

## LISTA DE FIGURAS

Figura 1 – Modelos de representação do corpo humano. . . . .	15
Figura 2 – Métodos de Estimção de Pose Humana 2D para uma única pessoa. . .	16
Figura 3 – Abordagens de Estimção de Pose Humana 2D para Múltiplas Pessoas.	17
Figura 4 – Comparação entre tempos de inferência de alguns estimadores de pose humana 2D. . . . .	17
Figura 5 – Exemplo de um dos quadros disponíveis no conjunto de dados <i>Shelf</i> e seu <i>ground-truth</i> 3D. . . . .	21
Figura 6 – Visão geral do processo de reconstrução e rastreamento do algoritmo selecionado. . . . .	23
Figura 7 – Fluxograma do processo de reconstrução 3D. . . . .	26
Figura 8 – Visão geral do processo do serviço de reconstrução e rastreamento da pose 3D. . . . .	27
Figura 9 – Exemplo de um dos quadros disponíveis no conjunto de dados <i>Campus</i> e seu <i>ground-truth</i> 3D. . . . .	29
Figura 10 – Exemplo de imagens utilizada para obter os parâmetros do algoritmo de reconstrução e rastreamento. . . . .	30
Figura 11 – Exemplo de uma reconstrução 3D realizada por meio de um serviço PIS. Na representação gráfica, a pessoa identificada como ID 2 acabou sobrepondo a representação da pessoa identificada como ID 1, no entanto, foi adequadamente reconstruída e rastreada. . . . .	32
Figura 12 – Exemplo de uma reconstrução 3D realizada por meio de um serviço PIS.	32
Figura 13 – Tempo de execução total por número de pessoas presentes no ambiente.	36
Figura 14 – Pessoa 0 e Pessoa 1 com pose neutra. . . . .	37
Figura 15 – Pessoa 0 com braço esquerdo levantado e Pessoa 1 com braço direito levantado. . . . .	37
Figura 16 – Pessoa 0 com braço direito levantado e Pessoa 1 com ambos os braços levantados. . . . .	38
Figura 17 – Pose identificada ao longo dos quadros. A cor verde representa a inferência de gesto simples para a pessoa 0, enquanto a cor roxa indica a inferência de gesto simples para a pessoa 1. . . . .	38
Figura 18 – Comparação entre os gestos rotulados manualmente (em azul) e os inferidos através da reconstrução 3D das poses humanas (em vermelho) para a Pessoa 0. . . . .	39
Figura 19 – Comparação entre os gestos rotulados manualmente (em azul) e os inferidos através da reconstrução 3D das poses humanas (em vermelho) para a Pessoa 1. . . . .	39

Figura 20 – Falha na reconstrução da pose da Pessoa 1, gerando uma inferência errada para a detecção de gestos. . . . .	39
Figura 21 – Visão geral do sistema de reconstrução e rastreamento da pose 3D proposto como trabalho futuro. . . . .	42

## LISTA DE TABELAS

Tabela 1 – Comparação quantitativa no Dataset Shelf dos métodos avaliados usando a métrica PCP3D. . . . .	22
Tabela 2 – Comparação dos resultados e tempo de processamento de cada conjunto de detector de pessoas e poses 2D. . . . .	29
Tabela 3 – Descrição dos campos de <code>ObjectAnnotations</code> . . . . .	31
Tabela 4 – Desempenho no <i>Dataset Campus</i> . . . . .	34
Tabela 5 – Desempenho no <i>Dataset Shelf</i> . . . . .	34
Tabela 6 – Tempo de Execução em Função do Número de Pessoas . . . . .	35

## LISTA DE ABREVIATURAS E SIGLAS

DLT	<i>Direct Linear Transformation</i>
PCP3D	<i>Percentage of Correct Parts 3D</i>
PIS	<i>Programmable Intelligent Space</i>
UFES	Universidade Federal do Espírito Santo

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>12</b>
<b>1.1</b>	<b>Apresentação</b>	<b>12</b>
<b>1.2</b>	<b>Objetivos</b>	<b>13</b>
1.2.1	Objetivo Geral	13
1.2.2	Objetivos Específicos	13
<b>2</b>	<b>REFERENCIAL TEÓRICO</b>	<b>14</b>
<b>2.1</b>	<b>Espaço Inteligente Programável</b>	<b>14</b>
<b>2.2</b>	<b>Pose Humana</b>	<b>14</b>
2.2.1	Pose Humana 2D	15
2.2.2	Pose Humana 3D	18
<b>2.3</b>	<b>Rastreamento</b>	<b>19</b>
<b>3</b>	<b>PROPOSTA E METODOLOGIA</b>	<b>20</b>
<b>3.1</b>	<b>Seleção do Algoritmo</b>	<b>20</b>
<b>3.2</b>	<b>Algoritmo Selecionado</b>	<b>22</b>
3.2.1	Estimação de Pose Humana 2D	23
3.2.2	Associação 2D-3D	23
3.2.3	Reconstrução de Pose 3D	24
<b>3.3</b>	<b>Serviço do PIS</b>	<b>26</b>
3.3.1	Obtenção das Imagens das Câmeras	27
3.3.2	Detector de Pose Humana 2D	28
3.3.3	Parâmetros do Algoritmo de Reconstrução e Rastreamento	30
3.3.4	Disponibilização das Poses 3D rastreadas	31
<b>4</b>	<b>EXPERIMENTOS E RESULTADOS</b>	<b>33</b>
<b>4.1</b>	<b>Recursos Computacionais</b>	<b>33</b>
<b>4.2</b>	<b>Avaliação da Reconstrução da Pose 3D</b>	<b>33</b>
<b>4.3</b>	<b>Experimento I - Tempo de Execução</b>	<b>35</b>
<b>4.4</b>	<b>Experimento II - Detecção de Gestos Simples</b>	<b>36</b>
<b>5</b>	<b>CONCLUSÃO E TRABALHOS FUTUROS</b>	<b>40</b>
<b>5.1</b>	<b>Conclusão</b>	<b>40</b>
<b>5.2</b>	<b>Trabalhos Futuros</b>	<b>41</b>
	<b>REFERÊNCIAS</b>	<b>43</b>

# 1 INTRODUÇÃO

## 1.1 Apresentação

Com o crescente avanço de áreas como eletrônica e computação, coisas comuns do dia a dia, como telefones, relógios e televisões, estão ficando cada vez mais inteligentes, agregando novas funcionalidades para auxiliar seus usuários. Essa tendência se fez presente também em ambientes que, quando equipados com uma rede de sensores e atuadores, são capazes de tomar decisões baseadas nas informações coletadas, interagindo com as pessoas do ambiente. Tais ambientes são comumente chamados de Espaços Inteligentes (LEE; HASHIMOTO, 2002; CARMO, 2021).

Um espaço inteligente pode ter diversas aplicações e, para torná-las mais versáteis e abrangentes, é interessante compreender e acompanhar as pessoas que interagem com esses ambientes. Sendo assim, conseguir estimar e rastrear a pose 3D das pessoas torna-se uma ferramenta muito útil.

Há como estimar a pose 3D de múltiplas pessoas utilizando uma câmera ou várias delas. Métodos que utilizam apenas uma câmera apresentam diversos problemas de ambiguidade, já que diferentes poses 3D podem corresponder a imagens 2D muito semelhantes, além de serem mais suscetíveis a falhas devido a oclusões. Em contrapartida, métodos multicâmera normalmente requerem um trabalho maior, como a sincronização das requisições das imagens e/ou um sistema de câmeras calibradas.

O rastreamento de cada pose humana 3D detectada é outra técnica muito importante. Com ela é possível atribuir um identificador único a cada pessoa, permitindo analisar melhor suas ações, já que sua identificação é mantida durante o tempo. Em alguns casos, também permite um melhor desempenho das estimações de pose 3D.

Outro aspecto importante, visto que a intenção dos serviços presentes em um espaço inteligente é interagir com os usuários, é o tempo de processamento. Altos tempos de resposta podem gerar uma má experiência para as pessoas que estão utilizando o espaço inteligente ou suas aplicações. Logo, obter um tempo de resposta que esteja dentro dos requisitos de tempo do sistema é indispensável.

No Laboratório de Visão Computacional e Robótica (Lab VISIO), presente na Universidade Federal do Espírito Santo (UFES), foi desenvolvido um Espaço Inteligente Programável (PIS, do inglês *Programmable Intelligent Space*), proposto por Carmo (2021). Nele, várias aplicações se beneficiam da estimação de pose humana 3D, como a localização temporal

de gestos (QUEIROZ, 2019) e a construção de um mapa de ocupação (CUSTODIO et al., 2020), ambas executadas em tempo real.

Entretanto, o estimador de pose humana 3D atualmente implementado no Lab VISIO (QUEIROZ, 2019) não realiza o rastreamento das poses geradas, o que limita algumas aplicações. Além disso, o sistema de localização temporal de gestos proposto por Queiroz (2019), por exemplo, só é capaz de realizar a tarefa para um único indivíduo.

Nesse sentido, este trabalho propõe estudar métodos, atuais e existentes na literatura, de estimação e rastreamento de pose humana 3D em um sistema multicâmeras para implementá-lo na forma de um serviço do PIS. Para validar a proposta, dois experimentos serão executados e avaliados.

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

O objetivo deste trabalho é desenvolver um novo serviço de estimativa de pose 3D e rastreamento de múltiplas pessoas para um Espaço Inteligente Programável, usando um sistema de múltiplas câmeras.

### 1.2.2 Objetivos Específicos

Com o intuito de atingir o objetivo geral, estabelecem-se os seguintes objetivos específicos:

- Estudar os métodos existentes de estimação e rastreamento de pose humana 3D para múltiplas pessoas em sistemas de múltiplas câmeras;
- Escolher um método adequado para os requisitos de funcionamento do PIS;
- Construir um serviço para o PIS que permita a reconstrução e o rastreamento da pose 3D de múltiplas pessoas utilizando as poses 2D obtidas do sistema de múltiplas câmeras;
- Validar a aplicação por meio de experimentos no PIS do laboratório.

## 2 REFERENCIAL TEÓRICO

### 2.1 Espaço Inteligente Programável

Segundo Lee e Hashimoto (2002), um Espaço Inteligente é um ambiente físico equipado com sensores que possibilitam a captação e compreensão dos eventos que ocorrem nele, permitindo assim a oferta de serviços úteis aos usuários. Essa ideia de integrar recursos computacionais ao ambiente deriva do conceito de computação ubíqua de Weiser (1991), que propôs que o avanço tecnológico naturalmente levará à invisibilidade da computação, pois se tornará parte integrante da rotina das pessoas.

O espaço inteligente utilizado neste trabalho é baseado em visão computacional, utilizando um sistema de câmeras calibradas como principal meio de sensoriamento. Além disso, é fundamentado na classe de espaço inteligente proposta por Carmo (2021), chamado *Programmable Intelligent Space* (PIS), que abstrai a infraestrutura e os dispositivos do ambiente, tornando sua utilização mais fácil.

O PIS adota uma arquitetura baseada em microsserviços, o que proporciona, entre outros benefícios, alta modularidade (LEWIS; FOWLER, 2014). Essa abordagem facilita a introdução de novas funcionalidades e permite a reutilização de serviços já existentes, permitindo que os desenvolvedores aproveitem soluções originalmente projetadas para outros contextos específicos (CARMO et al., 2020).

A comunicação do PIS ocorre por meio de um *broker*, que processa as mensagens publicadas nos tópicos e as entrega aos consumidores de interesse, seguindo o padrão *Publisher-Subscriber*. Adicionalmente, é possível realizar requisições por meio de um RPC (Remote Procedure Call), no formato *Request/Reply*. As mensagens são padronizadas pela biblioteca *is-msgs*, desenvolvida pelo Lab VISIO utilizando o formato padrão *Protocol Buffers* para garantir uniformidade.

### 2.2 Pose Humana

A pose humana refere-se à maneira em que o corpo humano se encontra disposto, ou seja, a sua postura. Detectar a pose humana em vídeos e fotos é um desafio que vem sendo estudado há anos em visão computacional, sendo útil nas mais diversas áreas, como esporte

(WANG et al., 2019), realidade virtual (ZHANG et al., 2021), entretenimento (WILLETT et al., 2020) e reconhecimento de gestos (ANGELINI et al., 2018).

Para poder estimar a pose humana, foram desenvolvidos alguns modelos para representar os dados extraídos de uma imagem e conseguir representar o corpo humano e suas diversas partes e articulações. Os três modelos mais comuns são o modelo cinemático, o modelo planar e o modelo volumétrico, os quais estão representados na Figura 1.

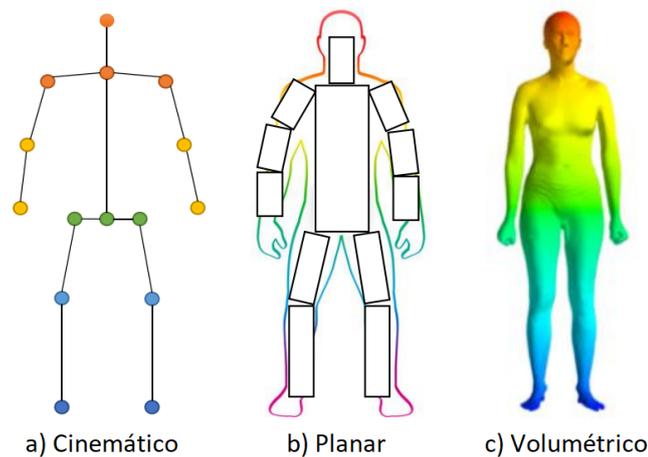


Figura 1 – Modelos de representação do corpo humano.  
Fonte: Adaptado de Zheng et al. (2020).

O modelo cinemático é amplamente utilizado tanto na estimação de pose humana 2D quanto na 3D, e utiliza um conjunto de pontos que, quando conectados, representam os membros e as articulações do corpo humano. O modelo planar busca representar não apenas a estrutura do corpo humano, mas também sua forma e textura, utilizando formas geométricas, normalmente retângulos, para representá-las. Por outro lado, o modelo volumétrico é empregado em algumas estimativas de pose humana 3D e retorna uma malha com estrutura e formato semelhantes aos de um corpo humano real.

### 2.2.1 Pose Humana 2D

Os métodos de estimação de pose humana 2D buscam obter a localização em 2D dos membros e articulações do corpo humano em uma imagem ou vídeo, e têm se beneficiado significativamente com o avanço das redes de aprendizado profundo e da disponibilidade de conjunto de dados como COCO, *MPII Human Pose Dataset* e LSP (MUNEA et al., 2020). Esses métodos podem ser classificados em estimação de pose para uma única pessoa ou para múltiplas pessoas. Enquanto os primeiros são capazes de obter a pose de apenas um

único indivíduo na cena, os últimos calculam a estimativa de pose para todas as pessoas presentes na imagem.

Para os métodos de estimação de pose para uma única pessoa, são comumente utilizadas abordagens de regressão ou de mapas de calor. Na abordagem de regressão, os pontos desejados são obtidos diretamente da imagem de entrada. O DeepPose (TOSHEV; SZEGEDY, 2014) é um dos modelos mais conhecidos nesse tipo de abordagem. Já na abordagem de mapas de calor, são gerados mapas de calor para cada um dos pontos a serem estimados e, em seguida, esses mapas são associados para formar a pose humana 2D. O OpenPose (CAO et al., 2021), um dos modelos mais conhecidos de estimação de pose 2D, é um exemplo de modelo que utiliza a abordagem por mapas de calor. A Figura 2 ilustra o funcionamento de ambas as abordagens.

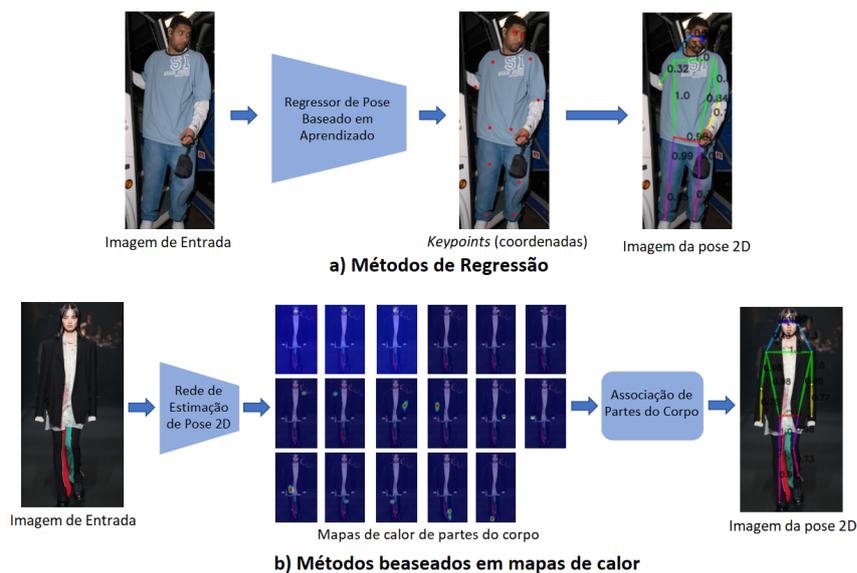


Figura 2 – Métodos de Estimação de Pose Humana 2D para uma única pessoa.

Fonte: Adaptado de Zheng et al. (2020).

Trazendo um maior nível de complexidade, os métodos capazes de lidar com múltiplas pessoas podem ser divididos em dois tipos de abordagem: *top-down* e *bottom-up*. Na abordagem *top-down*, é utilizado um detector de pessoas para obter a imagem de cada pessoa presente na cena separadamente e, em seguida, as imagens são passadas por um estimador de pose 2D. Um exemplo de modelo que utiliza essa abordagem é o AlphaPose (FANG et al., 2017).

Já na abordagem *bottom-up*, primeiro são localizados todos os pontos de todas as pessoas presentes na imagem, que posteriormente são conectados para gerar a pose de múltiplas pessoas. O OpenPose é um modelo que utiliza essa abordagem. A Figura 3 ilustra o funcionamento de ambas as abordagens.

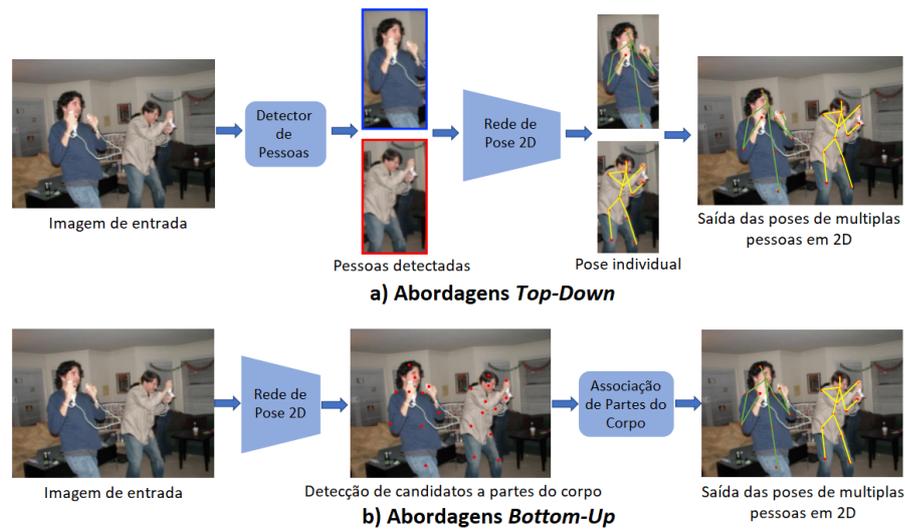


Figura 3 – Abordagens de Estimação de Pose Humana 2D para Múltiplas Pessoas.  
Fonte: Adaptado de Zheng et al. (2020).

Um ponto importante a ser comparado nessas duas abordagens é o custo computacional envolvido. Enquanto os métodos *bottom-up* calculam os mapas de calor apenas uma vez e, em seguida, os conectam para cada pessoa, os métodos *top-down* instanciam um estimador de pose 2D para cada pessoa detectada, tornando seu tempo de execução diretamente proporcional à quantidade de pessoas na cena. Isso geralmente torna os métodos *bottom-up* mais rápidos em termos de tempo de processamento (ZHENG et al., 2020). A Figura 4 apresenta uma comparação entre o tempo de inferência do OpenPose (*bottom-up*), AlphaPose (*top-down*) e Mask R-CNN (*top-down*), evidenciando as diferenças de desempenho entre essas abordagens.

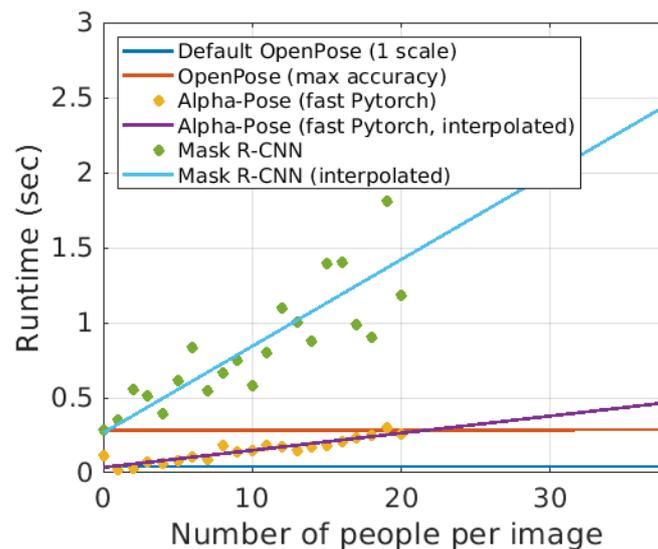


Figura 4 – Comparação entre tempos de inferência de alguns estimadores de pose humana 2D.  
Fonte: Cao et al. (2021).

### 2.2.2 Pose Humana 3D

Possibilitando uma representação mais completa do que a 2D, os métodos de estimação de pose humana 3D vêm sendo cada vez mais estudados. No entanto, eles enfrentam diversos desafios, como oclusões, ambiguidades de profundidade e falta de dados de treinamento adequados. Essa escassez de dados ocorre devido à dificuldade de anotação, que geralmente depende de um sistema de captura de movimento. Além disso, existe uma falta de generalização, uma vez que a maioria dos dados disponíveis são de ambientes internos específicos (ZHENG et al., 2020).

Existem várias abordagens que podem ser utilizadas na estimativa de pose humana 3D. É possível obter a pose utilizando uma única câmera ou um sistema de múltiplas câmeras, obter o resultado de maneira direta ou utilizando a pose humana 2D e localizar uma ou várias pessoas na cena. Neste trabalho será dada uma ênfase maior nas abordagens com múltiplas câmeras envolvendo várias pessoas, visto que é o que mais se adéqua ao tema proposto.

Sistemas de múltiplas câmeras para estimar a pose humana 3D de várias pessoas têm a capacidade de reduzir ambiguidades e oclusões parciais, porém enfrentam um desafio adicional: encontrar a correspondência de cada pessoa nas diferentes vistas. Uma vez que as correspondências são estabelecidas, é possível realizar a fusão das informações provenientes de todas as câmeras para cada pessoa na cena e, assim, estimar sua pose 3D.

Tanke e Gall (2019) propõem um estimador de pose humana 3D para múltiplas pessoas em um ambiente com um conjunto de câmeras calibradas. No trabalho, o problema de encontrar correspondências entre as vistas é formulado como um grafo  $k$ -partido entre as poses 2D, tendo como peso a distância média entre as linhas epipolares. Esse problema é resolvido através de um algoritmo guloso e a reconstrução 3D é realizada por meio da triangulação.

Já no trabalho de Dong et al. (2019), para resolver o problema das correspondências, é proposto um método de agrupamento das detecções de pose 2D da mesma pessoa entre as diferentes vistas. Esse método leva em consideração tanto aspectos geométricos quanto características relacionadas à aparência da pessoa. Em relação à reconstrução 3D, os autores propõem uma abordagem utilizando uma estrutura pictórica 3D, que apresenta melhores resultados ao custo de um maior tempo de processamento, mas também realizam testes utilizando a triangulação.

### 2.3 Rastreamento

Segundo Ning, Pei e Huang (2020), o rastreamento de pose consiste em estimar a pose humana de várias pessoas em uma cena e atribuir a elas uma identificação única, que será mantida nos quadros subsequentes, sendo muito útil em aplicações como reconhecimento de gestos humanos, compreensão de interações humanas, captura de movimento e animação. Essa técnica permite acompanhar a mudança da estimação da pose humana ao longo do tempo, podendo ser usada para suavizar a captura de movimentos, lidar com ambiguidades de pose e até mesmo diminuir o tempo de processamento (CHU et al., 2021).

No trabalho proposto por Tanke e Gall (2019), o rastreamento de poses 3D é realizado através da solução de um problema de correspondência de grafo bipartido, onde o custo para a associação entre duas detecções é a média da distância euclidiana entre os membros que estão presentes nas duas poses. Em casos em que não há nenhum membro em comum, é utilizada a distância euclidiana entre os centroides da pose humana projetados no plano  $xy$ .

Os autores Dong et al. (2022) apresentam um rastreador que leva em consideração tanto a distância entre as poses 3D quanto as características visuais das pessoas na cena. Além disso, propõem um método para assegurar a aquisição das características visuais da cena de maneira a agregar informações mais significativas sobre a aparência. Este método visa selecionar as características visuais obtidas da vista frontal, pois, de acordo com o autor, as vistas de perfil tendem a ser menos distintas e podem introduzir características visuais ruidosas.

## 3 PROPOSTA E METODOLOGIA

Entre os objetivos atuais dos projetos de pesquisa desenvolvidos no Lab VISIO está o reconhecimento de gestos dinâmicos em espaços inteligentes. Para isso, uma das abordagens possíveis faz uso da detecção e rastreamento 3D de juntas de esqueletos das pessoas presentes no ambiente. Desta forma, diante da necessidade de um serviço voltado ao PIS capaz de mapear de forma mais precisa os indivíduos em cena, foi concebida uma proposta para implementar um sistema de reconstrução e rastreamento 3D das poses humanas em tempo real. O algoritmo selecionado será aquele que atenda aos requisitos propostos, oferecendo um bom compromisso entre precisão e tempo de resposta. A precisão será medida pela métrica de Porcentagem de Partes Corretas 3D (PCP3D, do inglês *Percentage of Correct Parts 3D*), e será adaptado para se adequar às demandas do PIS. Esta seção detalha o processo de desenvolvimento de todas essas etapas, visando alcançar os objetivos estabelecidos para este trabalho.

### 3.1 Seleção do Algoritmo

Três autores propuseram artigos na literatura que se destacaram como promissores para o serviço do PIS de reconstrução e rastreamento tridimensional de poses humanas: Zhang et al. (2020), Tanke e Gall (2019) e Chu et al. (2021). Vale ressaltar que todos esses estudos disponibilizaram seus códigos-fonte em um repositório online, facilitando a compreensão e a aplicação dos métodos propostos.

O estudo conduzido por Zhang et al. (2020) integra a análise por visualização, a correspondência entre visualizações e o rastreamento temporal em um único *framework* de otimização em 4D. Este trabalho aborda igualmente e simultaneamente cada dimensão envolvida, ou seja, o espaço, o ponto de vista e o tempo. A implementação do algoritmo foi realizada em linguagem de programação C++.

Tanke e Gall (2019) apresenta um método que aborda a associação de poses 2D entre cenas por meio da formulação dessa associação como uma partição de grafos k-partidos. Esta abordagem emprega um algoritmo “guloso” para lidar de forma eficiente com a associação. Após a correlação de todas as poses, a reconstrução é realizada por meio de um processo de triangulação. O rastreamento das poses 3D é então concretizado ao resolver um problema de correspondência em um grafo bipartido, onde o custo associado à união

de duas detecções é calculado como a média da distância euclidiana entre os membros presentes em ambas as poses. O código-fonte disponível está na linguagem Python.

O método proposto por Chu et al. (2021) apresenta uma abordagem que aproveita a consistência temporal para associar as poses 2D estimadas na cena de cada câmera com os esqueletos 3D previamente construídos, possibilitando a realização simultânea das associações entre as cenas e as associações temporais. Para lidar com estimativas imprecisas ou ruidosas, o autor introduziu o chamado *part-aware measurement*, que é uma medida entre partes específicas em vez de considerar o corpo como um todo, avaliando apenas as partes que tem afinidade positiva. Adicionalmente, um filtro capaz de lidar com *outliers* em estimações 2D durante o processo de reconstrução também foi proposto. A implementação está disponível em Python.

Os três métodos foram aplicados ao *Dataset Shelf*, apresentado por Belagiannis et al. (2014), com o propósito de avaliar a detecção de pose 3D em múltiplas pessoas. Este conjunto de dados é composto por quatro indivíduos interagindo enquanto desmontam uma prateleira, sendo capturados por cinco câmeras diferentes. As marcações das articulações foram realizadas manualmente para as câmeras 2, 3 e 4. Para a obtenção do *ground-truth* 3D, foi realizada a triangulação entre os pontos dessas três vistas. Além disso, os pontos 3D foram empregados para projetar as coordenadas 2D das articulações nas câmeras não anotadas. O ator 4, devido à oclusão na maior parte das vistas das câmeras, não é usado para avaliação. A Figura 5 mostra as anotações e o *ground-truth* 3D de um dos quadros do conjunto de dados.

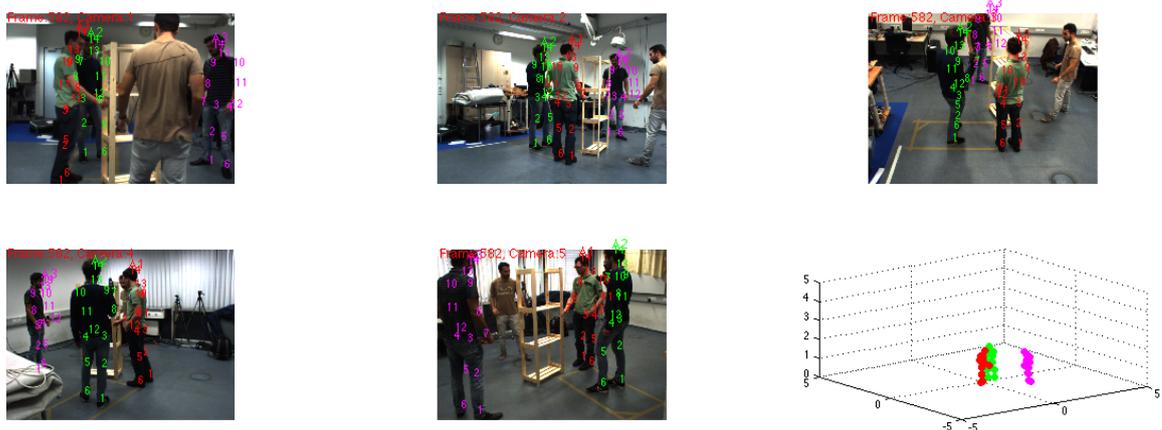


Figura 5 – Exemplo de um dos quadros disponíveis no conjunto de dados *Shelf* e seu *ground-truth* 3D. Fonte: Belagiannis et al. (2016).

Como meio de comparação, a métrica PCP3D foi empregada para avaliar o desempenho de cada método, sendo essencialmente uma adaptação da métrica PCP para avaliação de reconstruções de pose em três dimensões. A métrica PCP é definida como a porcentagem

de membros que foram corretamente estimados (WANG et al., 2021), e seu cálculo é realizado por meio da seguinte equação:

$$\frac{\|s_n - \hat{s}_n\| + \|e_n - \hat{e}_n\|}{2} \leq \alpha \|s_n - e_n\|$$

onde  $s_n$  e  $e_n$  representam as posições de início e término, respectivamente, do *ground-truth* do membro  $n$ , enquanto  $\hat{s}_n$  e  $\hat{e}_n$  indicam as posições correspondentes estimadas. O parâmetro  $\alpha$  é utilizado como um limiar (*threshold*). Comumente, adota-se  $\alpha = 0.5$ , o que sugere que um membro é considerado corretamente estimado se o erro na sua posição for, no máximo, a metade do comprimento total do membro. É importante notar que essa abordagem tem a desvantagem de penalizar membros mais curtos (ZHENG et al., 2020).

A Tabela 1 exibe a comparação quantitativa entre os métodos propostos pelos três autores, utilizando a métrica PCP3D com  $\alpha = 0.5$  para avaliar os resultados dos algoritmos no *Dataset Shelf*.

Tabela 1 – Comparação quantitativa no Dataset Shelf dos métodos avaliados usando a métrica PCP3D.

Método	PCP3D (%)		
	Zhang et al. (2020)	Tanke e Gall (2019)	Chu et al. (2021)
Ator 1	99.0	99.8	99.1
Ator 2	96.2	90.0	95.4
Ator 3	97.6	98.0	97.6
Média	97.6	95.9	97.4

Fonte: Produção do próprio autor.

Embora o método de Zhang et al. (2020) tenha demonstrado um desempenho superior, foi decidido priorizar a utilização do método proposto por Chu et al. (2021) devido à maior proficiência dos membros do Lab VISIO na linguagem Python. Isso foi feito visando facilitar futuras tarefas de manutenção e aprimoramento do sistema. Assim, optou-se pelo método de Chu et al. (2021), o qual também obteve um resultado satisfatório e bastante semelhante ao método de Zhang et al. (2020).

### 3.2 Algoritmo Selecionado

Conforme mencionado na Seção 3.1, o método escolhido para a implementação do novo serviço de estimativa de pose 3D e rastreamento de múltiplas pessoas para o PIS do Lab VISIO foi o proposto por Chu et al. (2021). Agora, será explorado em maior profundidade o funcionamento desse método. A Figura 6 dá uma visão geral do processo.

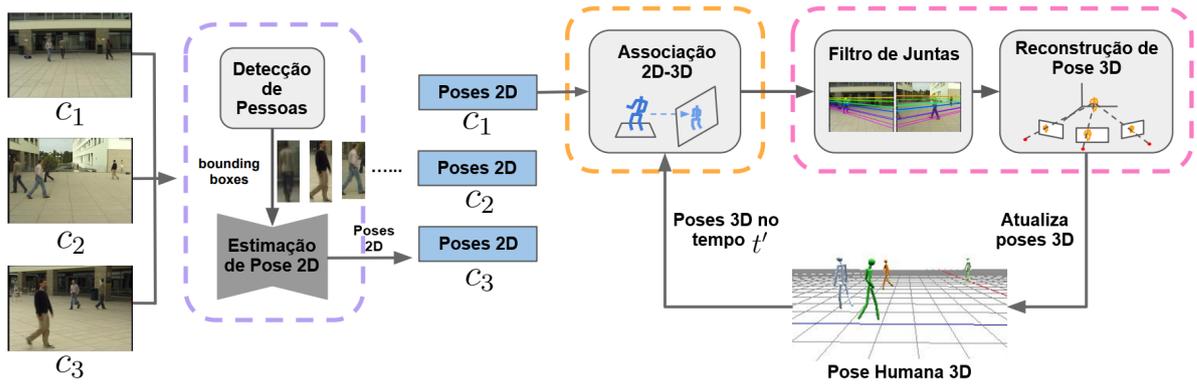


Figura 6 – Visão geral do processo de reconstrução e rastreamento do algoritmo selecionado.  
Fonte: Adaptado de Chu et al. (2021).

### 3.2.1 Estimção de Pose Humana 2D

O método emprega uma abordagem indireta na estimativa da pose humana 3D em um sistema multicâmeras. Essa estratégia baseou-se na obtenção das poses 2D das pessoas registradas por cada câmera, posteriormente reconstruindo-as em 3D. Além disso, a estimção da pose humana 2D adotou a abordagem *top-down*, utilizando o YOLOv3 (REDMON; FARHADI, 2018) como detector de pessoas e o HRNet (SUN et al., 2019) como estimador de pose humana 2D. A precisão nessa etapa é crucial, já que exerce influência significativa nos estágios subsequentes, como a associação entre poses de diferentes perspectivas e na subsequente reconstrução 3D (CHU et al., 2021).

### 3.2.2 Associação 2D-3D

Sendo  $\mathbf{p}$  as poses humanas 2D e  $\mathbf{P}$  o conjunto de poses 3D, Chu et al. (2021) propõe a associação das poses humanas entre as cenas e a associação temporal (rastreamento) de forma simultânea, através de um cálculo de afinidade entre as poses 2D  $\mathbf{p}$  e as poses 3D reconstruídas no tempo anterior  $\mathbf{X}_{t'} \in \mathbf{P}$ , que são reprojetaadas em 2D para cada câmera. Através das restrições geométricas de reprojeção entre as poses 2D, a afinidade geométrica  $\mathcal{G}(\mathbf{x}_{t,c}, \mathbf{X}_{t'})$  é calculada para cada junta  $n$ , de cada pose 2D  $\mathbf{x}_{t,c} \in \mathbf{p}$ , pela seguinte equação:

$$\mathcal{G}^n(\mathbf{x}_{t,c}^n, \mathbf{X}_{t'}^n) = \left(1 - \frac{\|\mathbf{x}_{t,c}^n - \tilde{\mathbf{x}}_{t,c}^n\|}{\alpha_{2D}(t - t')}\right) \cdot e^{-\lambda_\alpha(t-t')} \quad (3.1)$$

onde  $\tilde{\mathbf{x}}_{t,c}^n$  é a reprojeção 2D de  $\mathbf{X}_{t'}^n$  na câmera  $c$ ,  $\alpha_{2D}$  o limite de velocidade 2D e  $\lambda_\alpha$  a taxa de penalidade do intervalo de tempo.

Após o cálculo de todas as afinidades, a abordagem de Chu et al. (2021) seleciona exclusivamente as juntas com afinidade positiva para o cálculo da média, seguindo a estratégia de partes específicas *part-aware*, que visa minimizar os impactos de estimativas ruidosas. Adicionalmente, foi estabelecido um limiar  $\varepsilon$ , no qual valores de afinidade inferiores a esse limiar são considerados como zero. Por fim, a afinidade geométrica  $\mathcal{G}(\mathbf{x}_{t,c}, \mathbf{X}_{t'})$  é determinada pela média entre os valores das juntas com afinidade positiva.

O último passo é associar as poses 2D às poses 3D reconstruídas. Para isso é feita uma matriz de afinidades  $\mathbf{A} \in \mathbb{R}^{|\mathbb{P}| \times |\mathbb{P}|}$  que as associa. Como se trata de um problema de atribuição, o Algoritmo Húngaro (KUHN, 1955) é utilizado, o que resolve não só a associação temporal, mas também a associação das poses 2D entre câmeras, já que se trata das poses 2D relacionadas a uma mesma pose 3D.

### 3.2.3 Reconstrução de Pose 3D

Após todas as poses 2D serem associadas a uma pose 3D rastreada  $\mathbf{X}_{t'}$ , sua pose pode ser atualizada através da reconstrução dos pontos das novas poses 2D. Entretanto, para casos onde apenas uma única pose 2D foi associada a uma pose 3D, Chu et al. (2021) propôs reunir um conjunto  $\mathbb{P} = \{\mathbf{x}_{t'_c,c} | c \in \mathbf{C}; 0 \leq t - t'_c \leq \tau\}$ , onde  $t'_c$  é o tempo da última pose 2D corretamente associada da câmera  $c$  e  $t'_c \leq t$ , para recuperar as últimas poses 2D associadas em cada câmera, dentro de um intervalo  $\tau$ , para realizar a triangulação.

Outro procedimento que é sempre realizado é a aplicação do filtro de juntas. Esse método emprega a restrição epipolar para eliminar valores discrepantes, realizando o cálculo da distância entre a linha epipolar e o ponto correspondente, lidando com cada junta de maneira independente. A matriz de afinidade epipolar  $\mathbf{E} \in \mathbb{R}^{|\mathbb{P}| \times |\mathbb{P}|}$  é calculada para cada junta  $n$  através da seguinte equação:

$$\mathbf{E}_{i,j}(\mathbf{x}_i^n, \mathbf{x}_j^n) = 1 - \frac{d(\mathbf{x}_i^n, L_j^n) + d(\mathbf{x}_j^n, L_i^n)}{2\alpha_{epi}} \quad (3.2)$$

onde  $i$  e  $j$  representam o par de poses 2D associadas comparadas,  $\mathbf{x}_i^n$  e  $\mathbf{x}_j^n$  os pontos 2D correspondentes à junta  $n$  para a imagem de cada pose 2D,  $L_i^n$  e  $L_j^n$  são as linhas epipolares referentes à junta  $n$  e projetada de uma câmera para outra,  $d$  a função de distância ponto-linha e  $\alpha_{epi}$  o limite de erro de distância aceitável. Caso  $\mathbf{E}_{i,j}$  seja um valor negativo, indica que há pelo menos uma junta discrepante em  $\mathbf{x}_i^n$  e  $\mathbf{x}_j^n$ , que é removida através de um método guloso que os compara com  $\hat{\mathbf{X}}_t^n$ , que é  $\mathbf{X}_t^n$  estimado por um modelo

de movimento. Para isso, o raio de retroprojeção da pose 2D (*back-project-ray*) é calculado pela equação:

$$\mathbf{r}_{i,c}^n = (\mathbf{KR})_c^{-1} \cdot \mathbf{x}_{i,c}^n + \mathbf{o}_c \quad (3.3)$$

onde  $\mathbf{K}, \mathbf{R} \in \mathbb{R}^{3 \times 3}$  são, respectivamente, a matriz intrínseca e a matriz de rotação, e  $\mathbf{o}_c$  é o centro da câmera  $c$  em relação à coordenada global do sistema. Por fim, a distância entre o ponto 3D  $\hat{\mathbf{X}}_t^n$  e a linha 3D retroprojetada  $\mathbf{r}_i^n$  é calculada, sendo a junta com maior distância removida para que se tenha uma reconstrução mais robusta (CHU et al., 2021).

Para a reconstrução 3D, Chu et al. (2021) utiliza a Transformação Linear Direta (DLT, do inglês *Direct Linear Transformation*) para realizar a triangulação, ou seja, precisa de pelo menos dois pontos para realizar o procedimento. Para casos onde apenas uma pose 2D foi associada a um esqueleto 3D, o conjunto  $\mathbb{P}$  de poses 2D coletadas no pequeno intervalo de tempo  $\tau$  consegue contornar esse problema, acrescentando apenas uma penalização por intervalo de tempo. Já para casos onde todas as juntas são estimadas incorretamente, normalmente ocorrendo com antebraços e pernas (CHU et al., 2021), elas são consideradas como juntas faltantes e substituídas por  $\hat{\mathbf{X}}_t^n$ . Ademais, a informação temporal também é utilizada para suavizar a trajetória da pose 3D com filtro gaussiano, removendo possíveis estimações ruidosas que sobraram, mesmo após o filtro de juntas.

Após o processo de associação 2D-3D descrito em 3.2.2, algumas poses 2D podem não corresponder a nenhuma pose 3D já reconstruída, indicando que possivelmente se trata de uma nova pessoa na cena e, portanto, uma nova pose 3D deve ser reconstruída. Para associar as poses 2D sem correspondências entre as cenas, Chu et al. (2021) faz um conjunto  $\mathbb{U}_c$  de poses 2D sem correspondências na câmera  $c$ , e usa as restrições epipolares para calcular a correlação entre elas e as presentes nas outras cenas usando a Equação 3.2. Com isso, é gerada a matriz de afinidade  $\mathbf{E} = \sum_{n=1}^N \mathbf{E}^n$ , que é resolvida através do Algoritmo Húngaro (KUHN, 1955). O processo é realizado para cada câmera, sendo que a cada iteração, ainda pode haver poses sem correspondência  $\mathbb{U}'_c$ . Isso significa que  $\mathbb{U}'_c$  foi capturada por outras câmeras, mas não pela câmera  $c$ , portanto,  $\mathbb{U}'_c$  é incluído em  $\mathbb{U}_c$ .

Após concluir a associação para todas as cenas, o método de filtro de juntas também é aplicado, fornecendo uma reconstrução mais robusta para a inicialização da primeira pose 3D. Mas, diferente de quando há uma pose 3D reconstruída para servir de referência, neste caso é utilizado a matriz de  $n$  juntas  $\mathbf{E}^n$ , onde uma junta é considerada discrepante quando o valor de  $\mathbf{E}_{i,j}^n$  é negativo. Sendo assim, método de Chu et al. (2021) computa a  $\sum_{j=0} \mathbf{E}_{i,j}^n$  de  $\mathbf{x}_i^n$  e  $\mathbf{x}_j^n$  e remove a junta com a menor soma, ou seja, com a menor afinidade

em relação às outras. Em seguida, a pose 3D é reconstruída. Todo o processo pode ser visualizado no fluxograma presente na Figura 7.

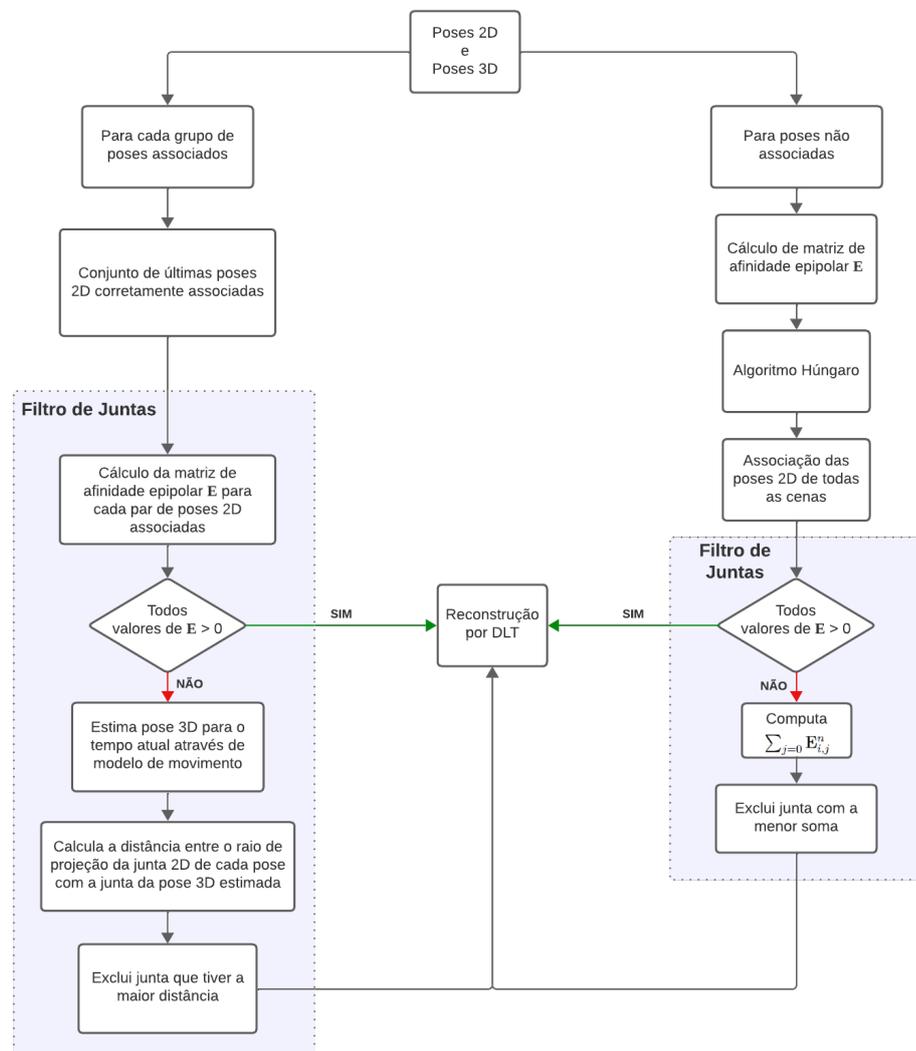


Figura 7 – Fluxograma do processo de reconstrução 3D.

Fonte: Produção do próprio autor.

### 3.3 Serviço do PIS

O PIS proposto por Carmo (2021) destaca-se pela capacidade de incorporar novas funcionalidades e possibilitar a reutilização de serviços já existentes. Para facilitar a integração, a comunicação entre os diversos serviços existentes é realizada por meio do padrão *publisher-subscriber*. Nesse contexto, um serviço publica uma mensagem em um formato específico em um tópico, e os programas inscritos nesse tópico recebem a mensagem correspondente. A Figura 8 representa o *pipeline* proposto para esse trabalho.

O serviço de reconstrução e rastreamento da pose 3D inicia obtendo imagens das câmeras

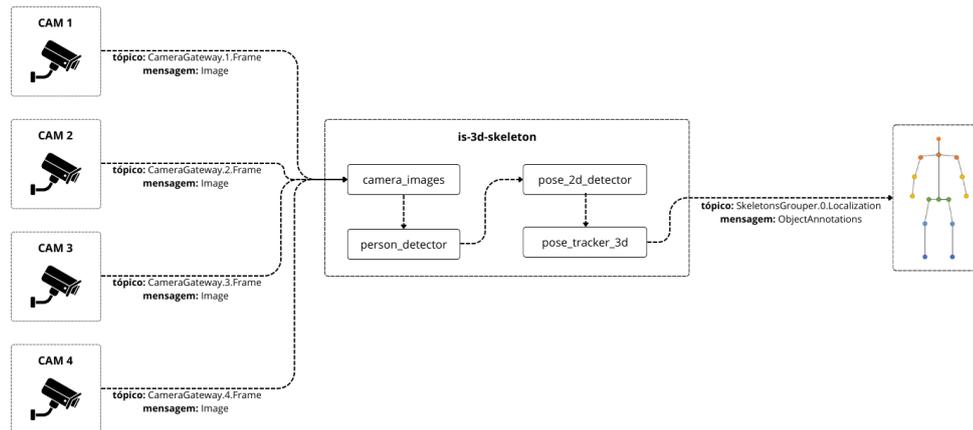


Figura 8 – Visão geral do processo do serviço de reconstrução e rastreamento da pose 3D.  
Fonte: Produção do próprio autor.

disponíveis no laboratório, as quais são então encaminhadas para o módulo de detecção de pessoas. Após a identificação de todas as pessoas nas imagens provenientes de cada câmera, o resultado, apresentado como uma lista de *bounding boxes*, é encaminhado para o módulo de detecção de pose humana 2D. Nesse módulo, a imagem associada a cada *bounding box* é isolada e utilizada como entrada para um detector de pose humana 2D, que fornece como saída a estimativa de localização de cada junta da pessoa.

Ao concluir esta etapa, as informações sobre a localização das juntas estimadas para todas as pessoas identificadas nas várias imagens são enviadas ao módulo de rastreamento e reconstrução. Este último retorna os pontos 3D correspondentes e os identificadores únicos de cada pessoa, publicando-os em um tópico do PIS, consolidando assim o processo de reconstrução e rastreamento da pose 3D. No entanto, para realizar corretamente a reconstrução e o rastreamento, é essencial configurar seus parâmetros de maneira adequada. As seções a seguir explicam com mais detalhes cada etapa da elaboração do serviço de reconstrução e rastreamento da pose 3D.

### 3.3.1 Obtenção das Imagens das Câmeras

O PIS do Lab VISIO dispõe de quatro câmeras, as quais, quando ativadas, publicam cada frame em um tópico específico no espaço inteligente. Para garantir uma boa reconstrução, a sincronização entre elas deve ser a mais precisa possível. Parte do serviço é responsável por consumir a imagem de todas as câmeras no mesmo instante, aguardando um breve intervalo de tempo para a chegada de novas mensagens nos tópicos de cada câmera. Após esse intervalo, cada mensagem do tipo *Image*, proveniente da biblioteca *is-msgs*, é

processada e convertida em uma `array` Numpy <sup>1</sup>. Quando todas as matrizes estão prontas, são retornadas na forma de um dicionário, no qual a chave corresponde ao ID da câmera e o valor à `array` Numpy associada.

### 3.3.2 Detector de Pose Humana 2D

O sistema proposto por Chu et al. (2021) incorpora em seu *pipeline* um detector *top-down*, o que implica a necessidade de um detector de pessoas acoplado a um detector de pose humana 2D. Essa escolha foi mantida para o serviço proposto, uma vez que o OpenPose (CAO et al., 2021), principal detector de pose humana 2D *bottom-up*, não apresenta bom desempenho em placas de vídeo mais recentes, não sendo compatível com Cudnn8 e versões de CUDA 11. Além disso, abordagens *top-down* geralmente são mais precisas. Como o laboratório onde o PIS está instalado é pequeno e não é comum haver muitas pessoas presentes simultaneamente, o fato de o tempo de inferência desse tipo de abordagem aumentar linearmente com o número de pessoas não será um problema.

No que diz respeito ao detector de pessoas, foram avaliados três modelos: YOLOv3 (REDMON; FARHADI, 2018), YOLOv8<sub>n</sub> (JOCHER; CHAURASIA; QIU, 2023) e RTMDet<sub>n</sub> (LYU et al., 2022), sendo este último convertido para TensorRT, que é um kit de desenvolvimento de software (SDK) para inferência de aprendizado profundo de alto desempenho. Quanto aos detectores de pose humana 2D, foram analisados HRPose (SUN et al., 2019) e RTMPose<sub>s</sub>(JIANG et al., 2023), sendo este último também convertido para TensorRT. Para a escolha do melhor conjunto de modelos foi realizado um teste no conjunto de dados *Campus*.

O *Dataset Campus*, assim como o *Shelf*, foi inicialmente apresentado por Belagiannis et al. (2014). Este conjunto de dados consiste em registros de interações envolvendo três indivíduos em um ambiente de *campus*, capturados por meio de três câmeras distintas. As juntas dos três atores foram manualmente anotadas para as duas primeiras câmeras. A reconstrução 3D foi realizada empregando a triangulação entre os pontos anotados, e o resultado dessa reconstrução foi projetado para a terceira câmera. A Figura 9 ilustra as anotações e o *ground-truth* 3D referentes a um dos quadros do conjunto de dados.

A comparação entre o conjunto de detectores de pessoas e de pose humana 2D foi realizada em uma máquina equipada com uma placa gráfica NVIDIA GeForce GTX 1650, contendo 4 GB de memória. A análise considerou tanto a qualidade da reconstrução, avaliada pela

<sup>1</sup> Numpy é uma biblioteca de computação numérica em Python, sendo `array` um dos seus tipos fundamentais de estrutura de dados.

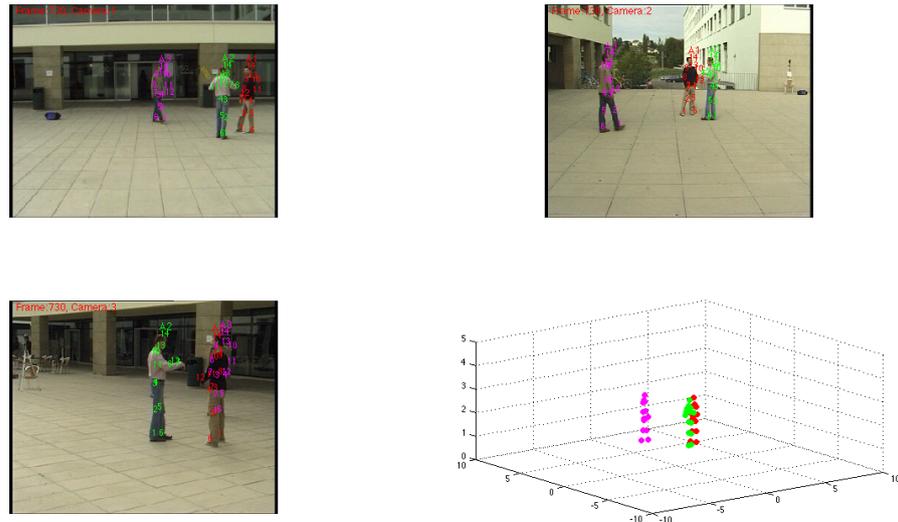


Figura 9 – Exemplo de um dos quadros disponíveis no conjunto de dados *Campus* e seu *ground-truth* 3D. Fonte: Belagiannis et al. (2016).

métrica PCP3D, quanto o tempo de processamento dos modelos de detecção. A Tabela 2 apresenta os resultados obtidos para cada conjunto.

Tabela 2 – Comparação dos resultados e tempo de processamento de cada conjunto de detector de pessoas e poses 2D.

Modelos Detector de Pessoas + Detector de Pose	PCP3D Médio (%)	Tempo Médio de Processamento Detector de Pessoas (ms)	Tempo Médio de Processamento Detector de Pose (ms)
RTMDet + HRPose	96.59	7.51	313.59
YOLOv3 + HRPose	96.79	105.19	313.97
YOLOv8 <sub>n</sub> + HRPose	96.78	38.50	323.54
RTMDet <sub>n</sub> + RTMPose <sub>s</sub>	95.51	8.77	26.13
YOLOv3 + RTMPose <sub>s</sub>	96.05	111.38	23.84
YOLOv8 <sub>n</sub> + RTMPose <sub>s</sub>	95.72	45.24	28.92

Fonte: Produção do próprio autor.

Analisando a Tabela 2, observa-se que o conjunto de modelos que proporciona a melhor reconstrução é composto pelo detector de pessoas YOLOv3 (REDMON; FARHADI, 2018) e o detector de pose humana 2D RTMPose<sub>s</sub> (JIANG et al., 2023). No entanto, ao analisar o tempo médio de processamento do YOLOv3 (REDMON; FARHADI, 2018), percebe-se que este está significativamente acima do YOLOv8<sub>n</sub> (JOCHER; CHAURASIA; QIU, 2023), que alcançou um PCP3D médio praticamente idêntico. O modelo RTMDet<sub>n</sub> (LYU et al., 2022) também parecia interessante, devido ao seu baixo tempo de processamento, mas, ao testá-lo com imagens do laboratório, seu resultado foi bem inferior aos demais, provocando muitas ausências de detecções. Portanto, optou-se por selecionar os modelos YOLOv8<sub>n</sub>

(JOCHER; CHAURASIA; QIU, 2023) em conjunto com o RTMPose<sub>s</sub> (JIANG et al., 2023) para compor o detector de pose humana 2D *top-down* do serviço.

### 3.3.3 Parâmetros do Algoritmo de Reconstrução e Rastreamento

O método proposto por Chu et al. (2021) apresenta diversos parâmetros, tais como: limiar de velocidade 2D ( $\alpha_{2D}$ ), limiar de erro de distância epipolar ( $\alpha_{epi}$ ), limiar de intervalo de tempo ( $\tau$ ) e limiar de número de afinidades positivas ( $\varepsilon$ ), além da taxa de penalidade de intervalo de tempo ( $\lambda_\alpha$ ). Cada um desses parâmetros foi selecionado empiricamente para cada conjunto de dados utilizado em seu estudo. Nota-se que  $\alpha_{2D}$ ,  $\alpha_{epi}$  e  $\varepsilon$  são influenciados pelo tamanho da imagem e pela distância entre as pessoas e a câmera. Por outro lado,  $\tau$  e  $\lambda_\alpha$  dependem da quantidade de quadros por segundo da aquisição das imagens.



Figura 10 – Exemplo de imagens utilizada para obter os parâmetros do algoritmo de reconstrução e rastreamento.

Fonte: Produção do próprio autor.

Para a seleção dos parâmetros no Lab VISIO, inicialmente, foram capturadas uma série de imagens com as quatro câmeras sincronizadas, conforme exemplificado na Figura 10. As gravações ocorreram a uma taxa de 10 quadros por segundo, que representa o mínimo desejado para o sistema. Em seguida, os parâmetros foram ajustados e, como o conjunto de dados não foi rotulado, as escolhas basearam-se na percepção subjetiva. Pôde-se observar que, para baixos valores de  $\alpha_{epi}$  perde-se algumas reconstruções. Além disso, valores reduzidos de  $\alpha_{2D}$  provoca oscilações na reconstrução, resultando na perda das poses reconstruídas em algumas iterações.

### 3.3.4 Disponibilização das Poses 3D rastreadas

Atualmente, o serviço abrange integralmente o ciclo operacional, desde a aquisição das imagens provenientes das câmeras até a obtenção das poses em 2D e a subsequente reconstrução e rastreamento em 3D. Ao término desse procedimento, uma mensagem do tipo `ObjectAnnotations`, pertencente à biblioteca `is-msgs`, é publicada em um tópico específico do PIS, como, por exemplo, `SkeletonsGrouper.0.Localization`. A estrutura da mensagem está presente na Tabela 3. O principal campo de mensagem usado é o `ObjectAnnotation`, que tem o `label` `repeated` por se tratar de uma lista. Cada `ObjectAnnotation` refere-se a uma pose, guardando as coordenadas tridimensionais de cada junta reconstruída, juntamente com o identificador único de rastreamento.

Tabela 3 – Descrição dos campos de `ObjectAnnotations`

Campo	Tipo	Label	Descrição
<code>objects</code>	<code>ObjectAnnotation</code>	<code>repeated</code>	Lista de objetos e suas respectivas anotações.
<code>resolution</code>	<code>Resolution</code>		Resolução original da imagem ao anotar uma imagem.
<code>frame_id</code>	<code>int64</code>		ID do quadro de referência usado para localizar os vértices ao anotar objetos no espaço.

Fonte: LabVISIO (2023).

Com isso, todos os demais serviços do PIS que desejem utilizar as poses 3D reconstruídas pelo serviço proposto devem se inscrever em seu tópico, tornando possível o consumo das mensagens publicadas. Um exemplo desses serviços é o de representar graficamente as poses reconstruídas, incluindo seus identificadores correspondentes, conforme ilustrado na Figura 11. Na representação gráfica, a pessoa identificada como ID 2 acabou sobrepondo a representação da pessoa identificada como ID 1, no entanto, foi adequadamente reconstruída e rastreada. A Figura 12 apresenta outro exemplo de reconstrução utilizando o conjunto de dados montado, desta vez com elevação e ângulo de azimute diferentes.

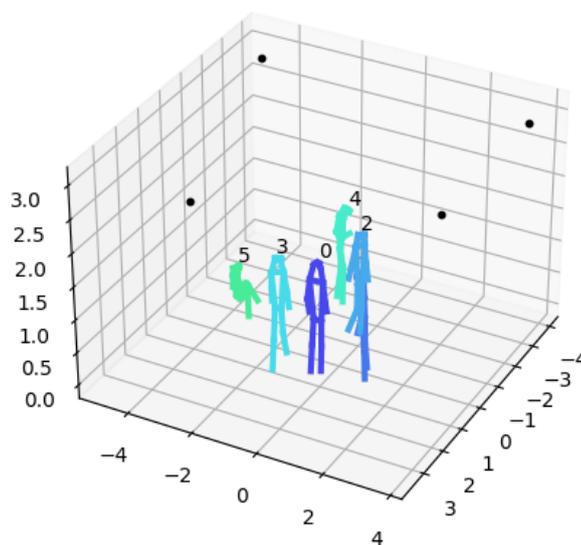
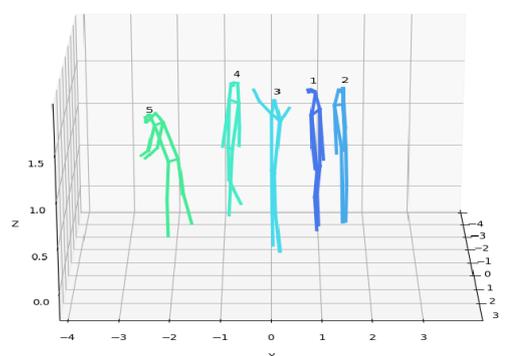


Figura 11 – Exemplo de uma reconstrução 3D realizada por meio de um serviço PIS. Na representação gráfica, a pessoa identificada como ID 2 acabou sobrepondo a representação da pessoa identificada como ID 1, no entanto, foi adequadamente reconstruída e rastreada.

Fonte: Produção do próprio autor.



(a) Poses 2D obtidas das imagens das quatro câmeras do laboratório.



(b) Reconstrução e rastreamento 3D resultante das poses 2D de (a).

Figura 12 – Exemplo de uma reconstrução 3D realizada por meio de um serviço PIS.

Fonte: Produção do próprio autor.

## 4 EXPERIMENTOS E RESULTADOS

Neste capítulo, são apresentados os experimentos realizados e os resultados obtidos. Adicionalmente, são detalhados os recursos computacionais, tanto de *hardware* quanto de *software*, empregados na execução deste trabalho.

### 4.1 Recursos Computacionais

#### Recursos de *Software*

A linguagem de programação Python foi empregada no desenvolvimento do serviço destinado ao espaço inteligente, bem como nos demais códigos necessários para sua validação. Para a implementação dos detectores de pessoas e da pose humana 2D, foram utilizados os *frameworks* de aprendizado profundo PyTorch. Por último, o software Docker foi empregado para empacotar a aplicação em containers.

#### Recursos de *Hardware*

O hardware utilizado pode ser descrito como quatro câmeras RGB do modelo Blackfly GigE BFLY-PGE-09S2C-CS do Espaço Inteligente Programável. Além disso, para realizar os experimentos, foi utilizada uma máquina com as seguintes configurações: (*i*) sistema operacional Linux, distribuição Ubuntu Server 22.04; (*ii*) processador Intel Xeon E5-2660, 2.0GHz com 14 núcleos físicos; (*iii*) 64GB de memória, DDR4; (*iv*) conexão ethernet de 1Gbps; (*v*) SSD de 480GB; (*vi*) 1 NVIDIA GeForce RTX 3090.

### 4.2 Avaliação da Reconstrução da Pose 3D

No estudo conduzido por Queiroz (2019), a métrica utilizada para avaliar a reconstrução 3D consiste no erro médio da localização das articulações, expresso em milímetros, calculada sobre o conjunto de dados *CMU Panoptic* (JOO et al., 2015). Esta métrica difere da abordagem adotada neste trabalho, que emprega a métrica PCP3D avaliada nos conjuntos de dados *Campus* e *Shelf*. Além disso, o método proposto por Queiroz (2019) realiza a reconstrução quadro a quadro, sem incorporar qualquer processo de rastreamento, enquanto o método empregado neste serviço não apenas inclui o rastreamento, mas também o utiliza

para aprimorar sua reconstrução. Em razão dessas diferenças, não foi realizada uma análise comparativa entre ambos os métodos, dado que seria injusta e demandaria o cálculo da métrica do erro médio da localização das articulações no conjunto de dados *CMU Panoptic*.

Desta forma, para mostrar a eficiência da reconstrução com o serviço proposto com mais detalhes, foi utilizado novamente o *Dataset Campus* para realizar a reconstrução 3D de seus atores, utilizando os modelos de detecção de pessoas e pose humana escolhidos na Seção 3.3.2. Os dados estão disponíveis na Tabela 4, a qual apresenta o PCP3D obtido com a reconstrução dos grupos de membros de cada ator. Aqui é possível notar que a métrica tem como desvantagem penalizar os membros mais curtos, principalmente os antebraços.

Tabela 4 – Desempenho no *Dataset Campus*

Grupo de Membros	Ator 0 (%)	Ator 1 (%)	Ator 2 (%)	Média (%)
Cabeça	100.00	100.00	100.00	100.00
Torso	100.00	100.00	100.00	100.00
Braços	98.98	100.00	98.19	99.06
Antebraços	84.69	65.87	88.41	79.66
Parte Superior das Pernas	100.00	100.00	100.00	100.00
Parte Inferior das Pernas	100.00	100.00	99.64	99.88
Total	96.73	93.17	97.25	95.72

Fonte: Produção do próprio autor.

Para avaliar a reconstrução da pose 3D em outra situação, o *Dataset Shelf* também foi utilizado. Diferente do *Dataset Campus*, que os atores se apresentam bem distantes das câmeras (Figura 9) aqui os atores estão mais próximos (Figura 5). Os resultados obtidos estão presentes na Tabela 5.

Tabela 5 – Desempenho no *Dataset Shelf*

Grupo de Membros	Ator 0 (%)	Ator 1 (%)	Ator 2 (%)	Média (%)
Cabeça	95.70	97.30	92.55	95.18
Torso	100.00	100.00	100.00	100.00
Braços	100.00	93.24	96.27	96.51
Antebraços	96.59	68.92	96.27	87.26
Parte Superior das Pernas	100.00	100.00	100.00	100.00
Parte Inferior das Pernas	100.00	100.00	100.00	100.00
Total	98.89	92.16	97.76	96.27

Fonte: Produção do próprio autor.

Esta avaliação, conduzida em dois conjuntos de dados distintos, é relevante não apenas para mostrar a correta funcionalidade do método de reconstrução 3D escolhido para

integrar o serviço PIS, mas também para demonstrar sua eficácia em diversas perspectivas e distâncias de câmera. Esses resultados evidenciam a robustez do serviço proposto neste trabalho, confirmando sua capacidade de atender aos requisitos essenciais.

Como última observação, é importante lembrar que a métrica PCP é calculada com base no número de membros corretamente estimados sobre o número total de membros, ou seja, o PCP3D total para cada ator, presente nas Tabelas 4 e 5, não é calculado através da média entre cada grupo de membros.

### 4.3 Experimento I - Tempo de Execução

Este experimento visa analisar o tempo de execução do serviço em relação ao número de pessoas presentes no Espaço Inteligente Programável do Lab VISIO. Para atingir esse objetivo, o serviço foi executado, e os tempos de inferência para as poses humanas 2D das imagens de todas as câmeras foram registrados, juntamente com os tempos de reconstrução e rastreamento 3D. Os dados foram coletados em cenários que envolviam diferentes números de pessoas. As informações detalhadas estão disponíveis na Tabela 6.

Tabela 6 – Tempo de Execução em Função do Número de Pessoas

Número de Pessoas	Tempo Médio Poses 2D (ms)	Tempo Médio Poses 3D e Rastreamento (ms)	Tempo Total (ms)
1	36.38	4.96	41.34
2	45.68	8.78	54.46
3	51.78	12.27	64.05
4	56.39	15.44	71.83
5	62.64	18.45	81.09
6	69.25	20.78	90.03

Fonte: Produção do próprio autor.

Observando a Tabela 6 é possível notar que tanto o tempo de inferência das poses 2D quanto o tempo de reconstrução e rastreamento 3D aumentam de forma praticamente linear. A Figura 13 mostra o gráfico do tempo de execução total em função do número de pessoas no ambiente.

A Figura 13 revela que o sistema é capaz de operar a 10 quadros por segundo, contanto que o ambiente contenha, no máximo, 6 pessoas. Se o número de pessoas ultrapassar esse limite, isso não implica necessariamente em falha no serviço. No entanto, devido aos parâmetros

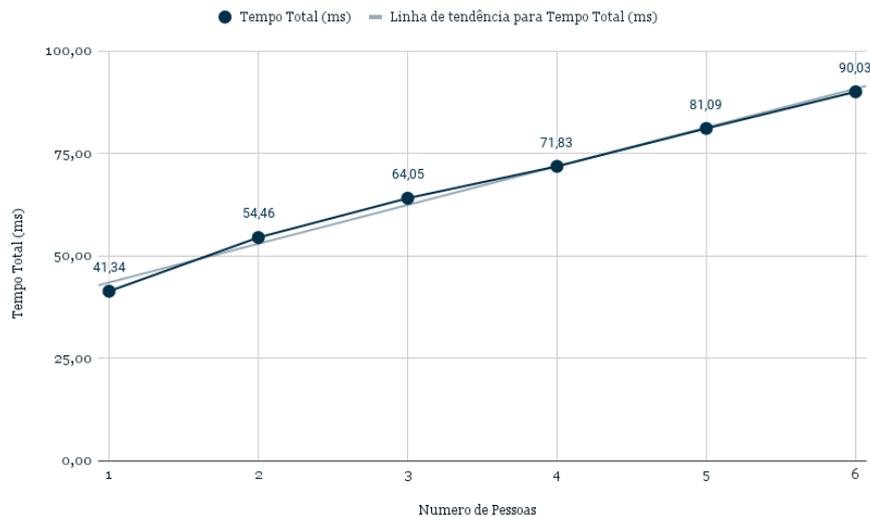


Figura 13 – Tempo de execução total por número de pessoas presentes no ambiente.  
Fonte: Produção do próprio autor.

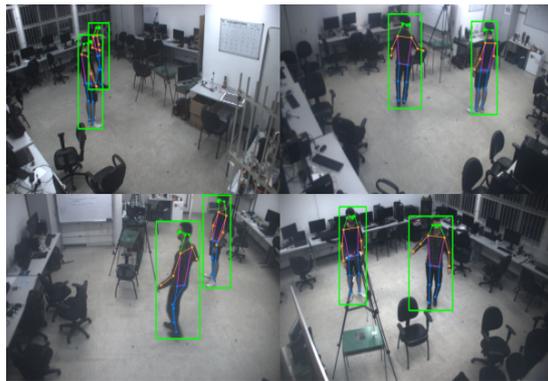
do sistema terem sido definidos considerando a operação a 10 quadros por segundo, existe a possibilidade de um aumento no índice de falhas na reconstrução e rastreamento.

#### 4.4 Experimento II - Detecção de Gestos Simples

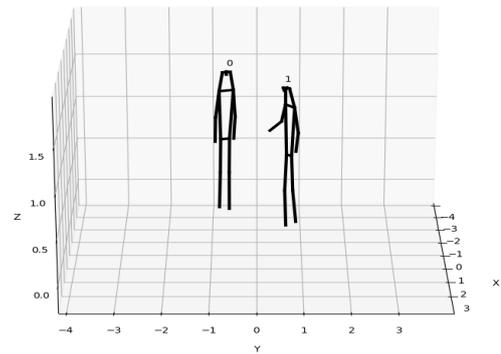
Este experimento teve como objetivo investigar uma das potenciais aplicações da reconstrução e rastreamento do serviço proposto. Para realizar essa análise, inicialmente, foi adquirido um conjunto de imagens sincronizadas de cada câmera do PIS, nas quais duas pessoas percorriam o ambiente executando quatro gestos simples: pose neutra (PN), mão esquerda levantada (MEL), mão direita levantada (MDL) e ambas as mãos levantadas (AML).

Desenvolveu-se um código em Python para processar as mensagens do tipo `ObjectAnnotations` e analisar a pose tridimensional de cada pessoa, atribuindo rótulos aos gestos realizados, rotulando uma mão como levanta se a junta do pulso estiver a uma altura superior a junta do nariz. Posteriormente, as poses foram representadas em gráficos, utilizando cores específicas para identificar cada gesto: preto para PN, vermelho para MEL, verde para MDL e azul para AML. As Figuras 14, 15 e 16 mostram as imagens das câmeras com as poses 2D das pessoas identificadas e a respectiva reconstrução 3D.

Após o processamento de todas as imagens no conjunto de dados, foi gerado um gráfico que fornece informações sobre o rótulo do gesto simples realizado por cada pessoa em cada quadro, atribuindo valores numéricos específicos: 0 para PN, 1 para MEL, 2 para MDL e



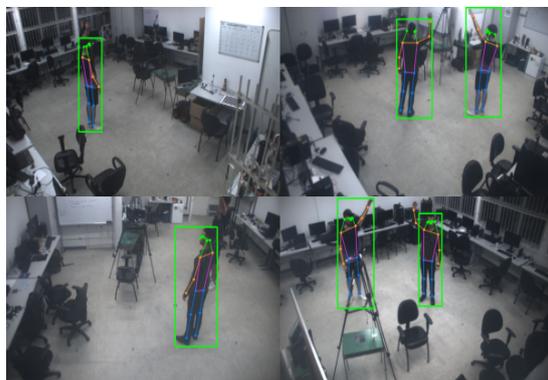
(a) Poses 2D obtidas das imagens das quatro câmeras do laboratório.



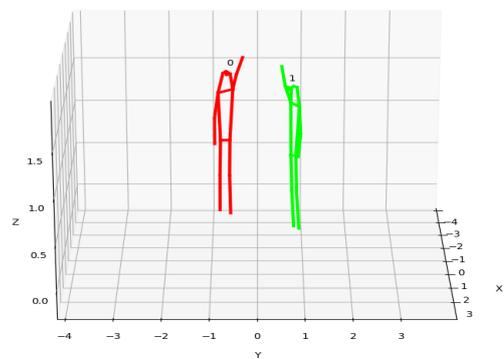
(b) Reconstrução e rastreamento 3D resultante das poses 2D de (a).

Figura 14 – Pessoa 0 e Pessoa 1 com pose neutra.

Fonte: Produção do próprio autor.



(a) Poses 2D obtidas das imagens das quatro câmeras do laboratório.



(b) Reconstrução e rastreamento 3D resultante das poses 2D de (a).

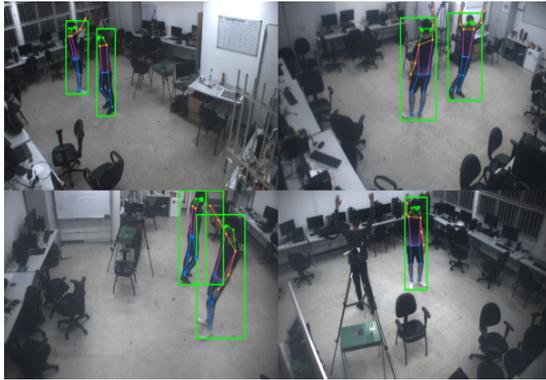
Figura 15 – Pessoa 0 com braço esquerdo levantado e Pessoa 1 com braço direito levantado.

Fonte: Produção do próprio autor.

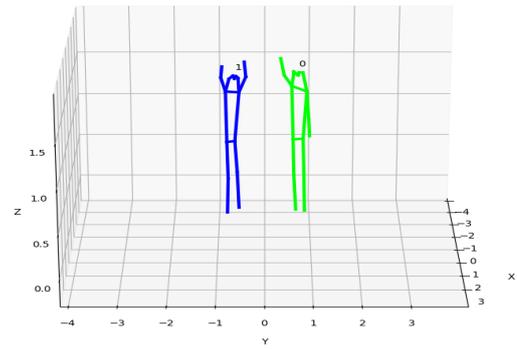
3 para AML. A Figura 17 apresenta o referido gráfico.

Para avaliar o quão bem os gestos simples foram identificados, utilizou-se o índice de Jaccard, muito usado para mensurar a similaridade entre um conjunto de amostras, podendo ser definido como a interseção sobre a união dos conjuntos. Para calculá-lo, todos os quadros do conjunto de dados foram rotulados manualmente, resultando no *ground-truth* dos gestos simples para ambas as pessoas presentes na cena. As Figuras 18 e 19 ilustram a comparação entre os gestos simples inferidos e o *ground-truth* para ambas as pessoas.

No cálculo do índice de Jaccard para as pessoas, os dois primeiros quadros foram desconsiderados devido à ausência de instâncias das poses 3D. Ao calcular o índice para a Pessoa 0, obteve-se um valor de 0,9823, indicando que 98,23% dos gestos simples inferidos foram corretamente identificados. Para a Pessoa 1, o índice de Jaccard foi de 0,9225, evidenciando uma taxa de precisão de 92,25% na estimativa dos gestos simples.



(a) Poses 2D obtidas das imagens das quatro câmeras do laboratório.



(b) Reconstrução e rastreamento 3D resultante das poses 2D de (a).

Figura 16 – Pessoa 0 com braço direito levantado e Pessoa 1 com ambos os braços levantados.  
 Fonte: Produção do próprio autor.

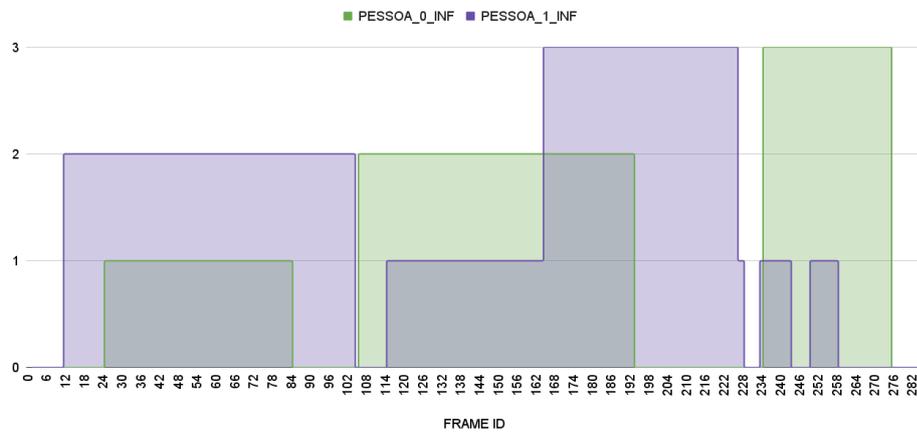


Figura 17 – Pose identificada ao longo dos quadros. A cor verde representa a inferência de gesto simples para a pessoa 0, enquanto a cor roxa indica a inferência de gesto simples para a pessoa 1.  
 Fonte: Produção do próprio autor.

Observando a Figura 19, é possível identificar uma falha na inferência dos gestos da Pessoa 1 entre os quadros 230 e 250. Isso ocorreu devido à incapacidade em detectar a Pessoa 1 de forma precisa, seja não a identificando ou realizando uma identificação incompleta, resultando na exclusão da mão levantada da *bounding box* (Figura 20a). Uma vez que as imagens das *bounding boxes* são encaminhadas para o detector de pose 2D, essa falha teve um impacto direto na reconstrução da pose 3D (Figura 20b).

Através deste experimento, embora aparentemente simples, conseguiu-se demonstrar a viabilidade do rastreamento de poses ao longo do tempo, conforme representado pelo ID do quadro. Este fato não apenas valida a eficácia do método, mas também abre portas para aplicações mais amplas, como o potencial uso do localizador de gestos proposto por Queiroz (2019) para monitorar simultaneamente múltiplos indivíduos, assim como a utilização do serviço de reconstrução e rastreamento 3D de poses para o uso em reconhecimento de gestos dinâmicos.

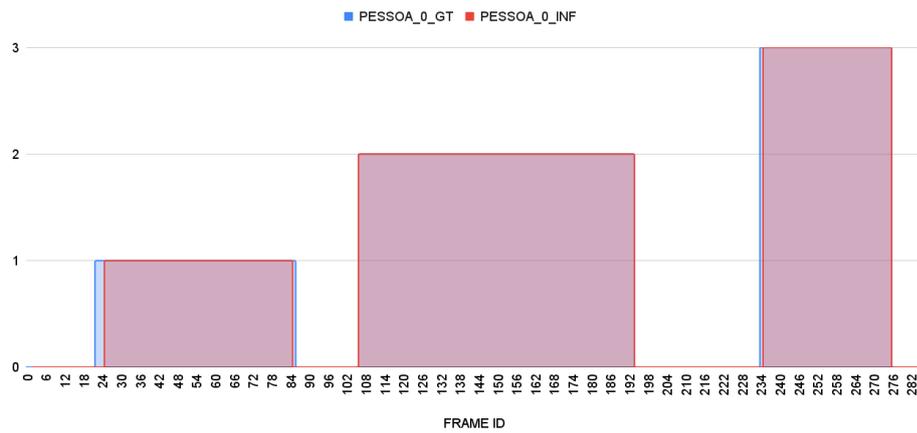


Figura 18 – Comparação entre os gestos rotulados manualmente (em azul) e os inferidos através da reconstrução 3D das poses humanas (em vermelho) para a Pessoa 0.

Fonte: Produção do próprio autor.

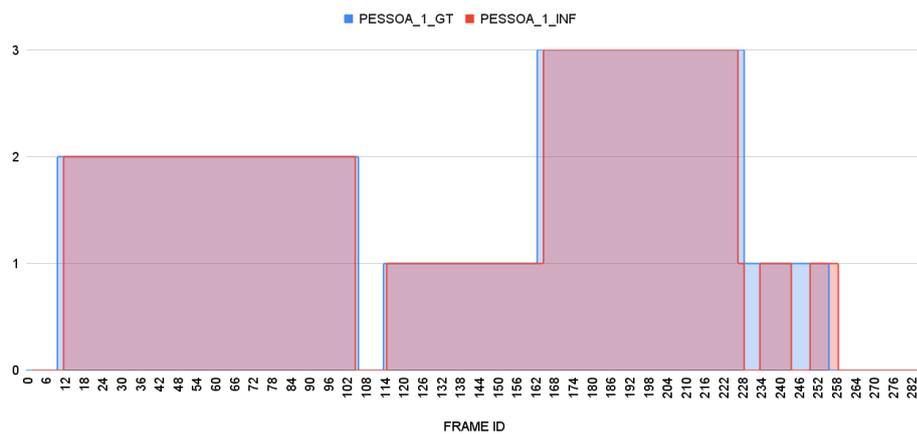
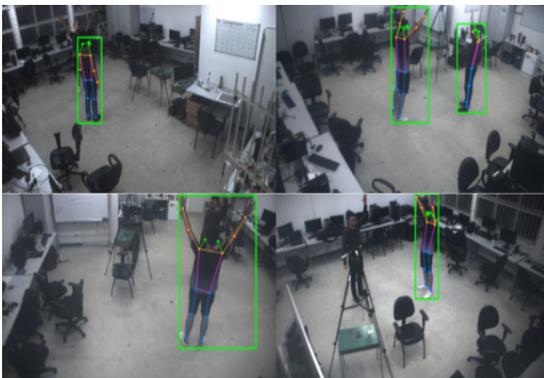
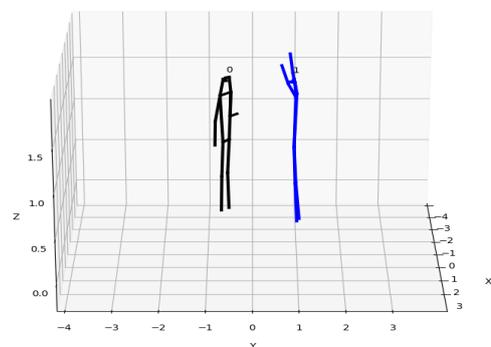


Figura 19 – Comparação entre os gestos rotulados manualmente (em azul) e os inferidos através da reconstrução 3D das poses humanas (em vermelho) para a Pessoa 1.

Fonte: Produção do próprio autor.



(a) Poses 2D obtidas das imagens das quatro câmeras do laboratório.



(b) Reconstrução e rastreamento 3D resultante das poses 2D de (a).

Figura 20 – Falha na reconstrução da pose da Pessoa 1, gerando uma inferência errada para a detecção de gestos.

Fonte: Produção do próprio autor.

## 5 CONCLUSÃO E TRABALHOS FUTUROS

### 5.1 Conclusão

O objetivo principal deste trabalho foi propor um novo serviço destinado à estimativa de pose 3D e rastreamento de múltiplas pessoas para o PIS, empregando um sistema de múltiplas câmeras. O serviço desenvolvido demonstrou eficácia na reconstrução e rastreamento das pessoas presentes no ambiente, atendendo aos requisitos de tempo real.

Inicialmente, realizou-se uma busca na literatura para a seleção do método mais apropriado para o contexto do laboratório. Com a metodologia escolhida, promoveram-se diversas adaptações, incluindo a substituição dos detectores de pessoas e de pose 2D anteriormente utilizados. Adicionalmente, os parâmetros de reconstrução e rastreamento foram ajustados às condições específicas do Lab VISIO. Para garantir a pertinência da solução proposta ao PIS, conduziu-se uma avaliação e dois experimentos.

A avaliação teve como objetivo validar o método escolhido para a reconstrução com rastreamento. Devido à inviabilidade de comparação, tanto quantitativa quanto qualitativa, entre o método atual do laboratório e o proposto para ser adotado neste serviço, foram empregados os conjuntos de dados *Campus* e *Shelf*. O resultado foi satisfatório, evidenciando que o serviço é capaz de operar em diversas perspectivas e distâncias de câmeras, revelando-se robusto e em conformidade com os requisitos necessários.

O primeiro experimento teve como objetivo avaliar o tempo de execução de cada reconstrução em relação ao número de pessoas presentes no PIS. Para isso, o serviço foi executado e, para diferentes quantidades de pessoas presentes, foram registrados os tempos necessários para obter as inferências de pose 2D e realizar o rastreamento e reconstrução 3D. Observou-se um claro aumento linear no tempo de execução, indicando que o sistema é capaz de operar a 10 quadros por segundo até um número máximo de 6 pessoas em cena.

O segundo experimento evidenciou uma das potenciais aplicações da reconstrução com o rastreamento proporcionado pelo serviço proposto, que é a identificação de gestos. Utilizando um conjunto de dados provenientes das câmeras do PIS, o serviço foi implementado e as poses 3D de cada pessoa em cena, referenciadas como Pessoa 0 e Pessoa 1, foram categorizadas em quatro gestos simples. Ao comparar os rótulos inferidos com o *ground-truth*, obteve-se um índice de Jaccard de 0,9823 para a Pessoa 0 e 0,9225 para a Pessoa 1. O

resultado inferior desta última deveu-se à sua não identificação por parte do detector de pessoas em alguns quadros. Este experimento demonstrou a viabilidade do rastreamento de pose ao longo do tempo, abrindo possibilidades para novas aplicações que se utilizem dessa valiosa informação

Por fim, almeja-se que este trabalho possa contribuir para o desenvolvimento do PIS, proporcionando um serviço capaz de realizar a reconstrução e rastreamento 3D em tempo real dos indivíduos presentes no ambiente. Ao viabilizar o rastreamento de poses, espera-se a criação de novos serviços, bem como a abertura de novas perspectivas para serviços já existentes.

## 5.2 Trabalhos Futuros

O serviço proposto neste trabalho segue o processo mostrado na Figura 8, sendo encarregado de obter imagens das câmeras do PIS de maneira sincronizada, gerar poses 2D para cada pessoa em cada imagem e, por fim, reconstruir e rastrear essas poses em 3D. Para facilitar a orquestração e proporcionar maior modularidade, uma sugestão para trabalhos futuros é dividir o serviço em dois componentes distintos: um encarregado de obter as poses 2D e outro responsável pela reconstrução e rastreamento, assim como mostrado na Figura 21. Uma vez que a parte do código destinada à reconstrução e rastreamento já está configurada para receber mensagens do tipo `ObjectAnnotations`, será necessária apenas uma modificação para que o serviço possa receber eficientemente esse tipo de informação das quatro câmeras. Isso abre a possibilidade de, por exemplo, executar simultaneamente diferentes instâncias de detectores de pose 2D. Dessa forma, é possível, contanto que haja poder de processamento disponível, reduzir o tempo de processamento mediante a paralelização dessa etapa.

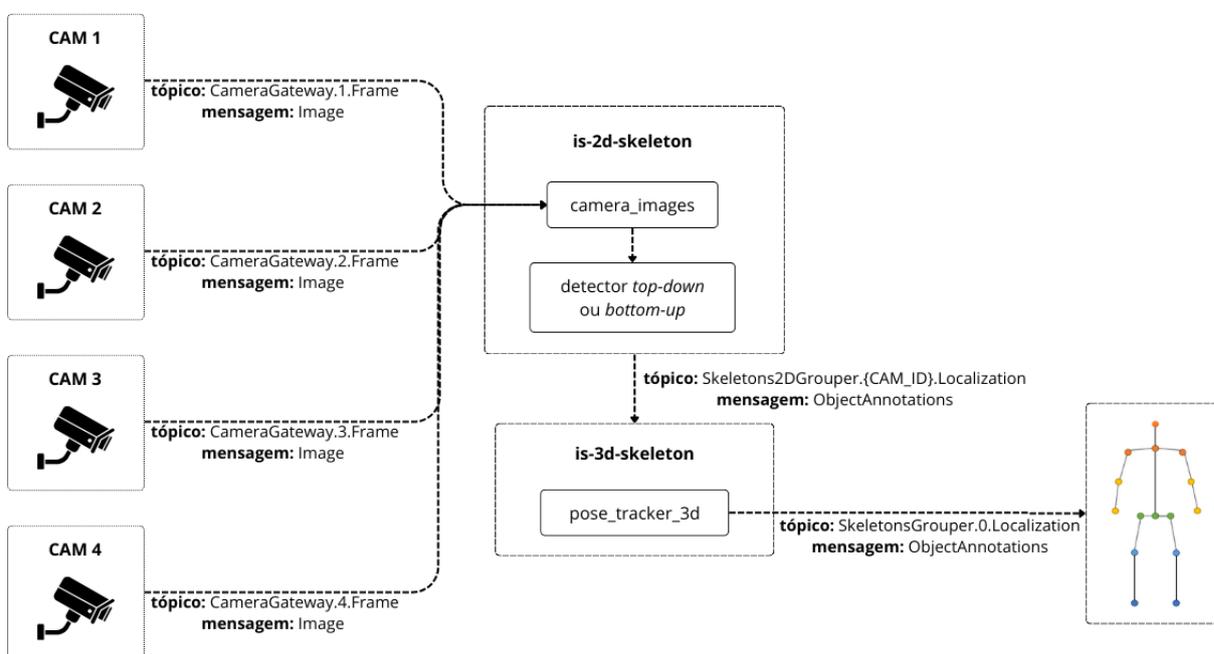


Figura 21 – Visão geral do sistema de reconstrução e rastreamento da pose 3D proposto como trabalho futuro.

Fonte: Produção do próprio autor.

## REFERÊNCIAS

- ANGELINI, F.; FU, Z.; LONG, Y.; SHAO, L.; NAQVI, S. M. ActionXPose: A Novel 2D Multi-view Pose-based Algorithm for Real-time Human Action Recognition. 2018. Disponível em: <<https://arxiv.org/abs/1810.12126>>. Citado na página 15.
- BELAGIANNIS, V.; AMIN, S.; ANDRILUKA, M.; SCHIELE, B.; NAVAB, N.; ILIC, S. 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 21 e 28.
- BELAGIANNIS, V.; AMIN, S.; ANDRILUKA, M.; SCHIELE, B.; NAVAB, N.; ILIC, S. Multiple Human Pose Estimation (3D Pictorial Structures). 2016. Accessed: 2023-12-5. Disponível em: <<https://campar.in.tum.de/Chair/MultiHumanPose>>. Citado 2 vezes nas páginas 21 e 29.
- CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 43, n. 1, p. 172–186, 2021. Citado 3 vezes nas páginas 16, 17 e 28.
- CARMO, A. do; QUEIROZ, F. de; SANTOS, C. dos; SILVA, L.; VASSALLO, R. Uso de um espaço inteligente baseado em visão computacional para o controle de formação de robôs móveis. In: Anais do XII Simpósio Brasileiro de Computação Ubíqua e Pervasiva. Porto Alegre, RS, Brasil: SBC, 2020. p. 171–180. ISSN 2595-6183. Disponível em: <<https://sol.sbc.org.br/index.php/sbcup/article/view/11223>>. Citado na página 14.
- CARMO, A. P. do. Uma Arquitetura de Microsserviços centrada na Observabilidade Multinível para Espaços Inteligentes baseados em Visão Computacional. Tese (Doutorado) — UFES, 2021. Disponível em: <<https://engenhariaeletrica.ufes.br/pt-br/pos-graduacao/PPGEE/detalhes-da-tese?id=14968>>. Citado 3 vezes nas páginas 12, 14 e 26.
- CHU, H.; LEE, J.-H.; LEE, Y.-C.; HSU, C.-H.; LI, J.-D.; CHEN, C.-S. Part-Aware Measurement for Robust Multi-View Multi-Human 3D Pose Estimation and Tracking. 2021. Disponível em: <<https://arxiv.org/abs/2106.11589>>. Citado 9 vezes nas páginas 19, 20, 21, 22, 23, 24, 25, 28 e 30.
- CUSTODIO, P.; QUEIROZ, F. de; ALMONFREY, D.; COTTA, W.; CARMO, A. do; VASSALLO, R. Proposta de um ambiente interacional baseado em um espaço inteligente programável. In: Anais do XII Simpósio Brasileiro de Computação Ubíqua e Pervasiva. Porto Alegre, RS, Brasil: SBC, 2020. p. 181–190. ISSN 2595-6183. Disponível em: <<https://sol.sbc.org.br/index.php/sbcup/article/view/11224>>. Citado na página 13.
- DONG, J.; FANG, Q.; JIANG, W.; YANG, Y.; HUANG, Q.; BAO, H.; ZHOU, X. Fast and robust multi-person 3d pose estimation and tracking from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 44, n. 10, p. 6981–6992, 2022. Citado na página 19.

- DONG, J.; JIANG, W.; HUANG, Q.; BAO, H.; ZHOU, X. Fast and robust multi-person 3d pose estimation from multiple views. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2019. p. 7784–7793. Citado na página 18.
- FANG, H.-S.; XIE, S.; TAI, Y.-W.; LU, C. Rmpe: Regional multi-person pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). [S.l.: s.n.], 2017. p. 2353–2362. Citado na página 16.
- JIANG, T.; LU, P.; ZHANG, L.; MA, N.; HAN, R.; LYU, C.; LI, Y.; CHEN, K. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. arXiv, 2023. Disponível em: <<https://arxiv.org/abs/2303.07399>>. Citado 3 vezes nas páginas 28, 29 e 30.
- JOCHER, G.; CHAURASIA, A.; QIU, J. YOLO by Ultralytics. 2023. Disponível em: <<https://github.com/ultralytics/ultralytics>>. Citado 3 vezes nas páginas 28, 29 e 30.
- JOO, H.; LIU, H.; TAN, L.; GUI, L.; NABBE, B.; MATTHEWS, I.; KANADE, T.; NOBUHARA, S.; SHEIKH, Y. Panoptic studio: A massively multiview system for social motion capture. In: 2015 IEEE International Conference on Computer Vision (ICCV). [S.l.: s.n.], 2015. p. 3334–3342. Citado na página 33.
- KUHN, H. W. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, v. 2, p. 83–97, 1955. Citado 2 vezes nas páginas 24 e 25.
- LABVISIO. IS Messages - Protocol Documentation. 2023. Accessed: 2023-12-6. Disponível em: <<https://github.com/labvisio/is-msgs/blob/master/docs/README.md#is.vision.ObjectAnnotations>>. Citado na página 31.
- LEE, J.-H.; HASHIMOTO, H. Intelligent space — concept and contents. Advanced Robotics, Taylor Francis, v. 16, n. 3, p. 265–280, 2002. Citado 2 vezes nas páginas 12 e 14.
- LEWIS, J.; FOWLER, M. Microservices: a definition of this new architectural term. 2014. Disponível em: <<https://martinfowler.com/articles/microservices.html>>. Citado na página 14.
- LYU, C.; ZHANG, W.; HUANG, H.; ZHOU, Y.; WANG, Y.; LIU, Y.; ZHANG, S.; CHEN, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. arXiv, 2022. Disponível em: <<https://arxiv.org/abs/2212.07784>>. Citado 2 vezes nas páginas 28 e 29.
- MUNEA, T. L.; JEMBRE, Y. Z.; WELDEGEBRIEL, H. T.; CHEN, L.; HUANG, C.; YANG, C. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. IEEE Access, v. 8, p. 133330–133348, 2020. Citado na página 15.
- NING, G.; PEI, J.; HUANG, H. Lighttrack: A generic framework for online top-down human pose tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). [S.l.: s.n.], 2020. p. 4456–4465. Citado na página 19.
- QUEIROZ, F. M. de. Sistema de Localização de Gestos Utilizando um Sistema Multicâmeras. Dissertação (Mestrado) — UFES, 2019. Disponível em: <<https://engenhariaeletrica.ufes.br/pt-br/pos-graduacao/PPGEE/detalhes-da-tese?id=13945>>. Citado 3 vezes nas páginas 13, 33 e 38.

- REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. ArXiv, abs/1804.02767, 2018. Disponível em: <<https://arxiv.org/abs/1804.02767>>. Citado 3 vezes nas páginas 23, 28 e 29.
- SUN, K.; XIAO, B.; LIU, D.; WANG, J. Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2019. p. 5686–5696. Citado 2 vezes nas páginas 23 e 28.
- TANKE, J.; GALL, J. Iterative greedy matching for 3d human pose tracking from multiple views. In: FINK, G. A.; FRINTROP, S.; JIANG, X. (Ed.). Pattern Recognition. Cham: Springer International Publishing, 2019. p. 537–550. ISBN 978-3-030-33676-9. Citado 4 vezes nas páginas 18, 19, 20 e 22.
- TOSHEV, A.; SZEGEDY, C. DeepPose: Human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014. Disponível em: <<https://doi.org/10.1109%2Fcvpr.2014.214>>. Citado na página 16.
- WANG, J.; QIU, K.; PENG, H.; FU, J.; ZHU, J. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In: Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2019. (MM '19), p. 374–382. ISBN 9781450368896. Disponível em: <<https://doi.org/10.1145/3343031.3350910>>. Citado na página 15.
- WANG, J.; TAN, S.; ZHEN, X.; XU, S.; ZHENG, F.; HE, Z.; SHAO, L. Deep 3d human pose estimation: A review. Computer Vision and Image Understanding, v. 210, p. 103225, 2021. ISSN 1077-3142. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1077314221000692>>. Citado na página 22.
- WEISER, M. The computer for the 21 st century. Scientific american, JSTOR, v. 265, n. 3, p. 94–105, 1991. Citado na página 14.
- WILLETT, N. S.; SHIN, H. V.; JIN, Z.; LI, W.; FINKELSTEIN, A. Pose2pose: Pose selection and transfer for 2d character animation. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. New York, NY, USA: Association for Computing Machinery, 2020. (IUI '20), p. 88–99. ISBN 9781450371186. Disponível em: <<https://doi.org/10.1145/3377325.3377505>>. Citado na página 15.
- ZHANG, H.; SCIUTTO, C.; AGRAWALA, M.; FATAHALIAN, K. Vid2player: Controllable video sprites that behave and appear like professional tennis players. ACM Transactions on Graphics (TOG), ACM New York, NY, v. 40, n. 3, p. 1–16, 2021. Citado na página 15.
- ZHANG, Y.; AN, L.; YU, T.; LI, X.; LI, K.; LIU, Y. 4d association graph for realtime multi-person motion capture using multiple video cameras. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2020. p. 1321–1330. Citado 2 vezes nas páginas 20 e 22.
- ZHENG, C.; WU, W.; YANG, T.; ZHU, S.; CHEN, C.; LIU, R.; SHEN, J.; KEHTARNAVAZ, N.; SHAH, M. Deep learning-based human pose estimation: A survey. CoRR, abs/2012.13392, 2020. Disponível em: <<https://arxiv.org/abs/2012.13392>>. Citado 5 vezes nas páginas 15, 16, 17, 18 e 22.