Estimating Tridimensional Coordinates of Skeleton Joints in a Multicamera System

Felippe Mendonça de Queiroz*, Rodolfo Picoreti*, Clebeson Canuto dos Santos*,
Mariana Rampinelli Fernandes†, Raquel Frizera Vassallo*

* Universidade Federal do Espírito Santo
Programa de Pós-Graduação em Engenharia Elétrica
Vitória, Espírito Santo, Brasil
Email: {mendonca.felippe, rodolfo.picoreti, clebeson.canuto}@gmail.com, raquel@ele.ufes.br

† Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo – Campus Vitória
Coordenadoria de Eletrotécnica
Vitória, Espírito Santo, Brasil
Email: mariana.rampinelli@ifes.edu.br

***Resumo*—Tridimensional localization of individuals' joints or skeletons is essential for many applications, which may vary from action and gesture recognition to gate analysis and video games. Usually, RGB-D sensors are preferred due to their facility in providing 3D data. However, such sensors have limitations as small field of view, short range and restricted use for indoors. Moreover, the use of CNNs benefited new approaches for skeletons detection on conventional images, with small evaluation time. Also, it's common to find multi-camera networks in many workplaces, mainly used for surveillance. With this in mind, we present a methodology to recover the 3D position of people's joints, using just a network of conventional cameras. It involves a 2D skeleton detection followed by a grouping and matching stage for 3D localization. The method was evaluated using a well-known dataset and was already tested in our experimental infrastructure. Our recent results are presented and discussed.**

***Index Terms*—Computer Vision, Multicamera System, 3D Skeleton Detection.**

I. Introdução

DEDECTAR as juntas e esqueletos de indivíduos em imagens e vídeos tem sido o objetivo de diversas pesquisas na área de visão computacional, realidade virtual e realidade aumentada. A posição e movimentação de indivíduos na cena são muitas vezes informações essenciais para alguns trabalhos.

A detecção de esqueletos pode ser usada por exemplo para classificar ou reconhecer ações e comportamentos [1], [2]; interpretação de cena ou monitoramento do ambiente [3]; detecção da intenção de movimentos [4], utilização de gestos para comunicação e interação homem-máquina [5], [6]; análise de movimento para fisioterapia [7], assistência a saúde [8] e até mesmo biometria, quando a maneira de caminhar é usada como forma de reconhecimento de indivíduos [9]. Nos casos de realidade virtual e realidade aumentada, a inserção de objetos ou personagens virtuais e a utilização dos movimentos de uma pessoa para animação de um avatar são as aplicações mais comuns [10], [11].

Uma das principais vantagens de se usar apenas a informação das juntas, é que a quantidade de dados é bem menor do que a imagem inteira [12]. Isso torna os processos de treinamento e inferência mais rápidos, além de permitir que diferentes implementações e aplicações compartilhem os mesmos dados de entrada.

Com o surgimento dos sensores RGB-D (*Red,Green,Blue - Depth*), vários métodos de detecção de juntas e esqueletos foram desenvolvidos, unindo-se a informação visual de uma câmera convencional com as medidas feitas por um sensor de profundidade [13], [14]. Normalmente, tais abordagens são computacionalmente mais baratas e não necessitam de um sistema estéreo para obtenção de informação tridimensional.

Mesmo assim, há ainda trabalhos que adotam uma rede de sensores RGB-D para aumentar o seu pequeno campo visual e facilitar a correspondência entre os esqueletos [15], [16].

Entretanto, sensores RGB-D não funcionam bem em ambientes externos, devido à incidência de luz solar, e possuem limitação de distância de funcionamento. Em geral, são utilizados para a detecção de poucos indivíduos e em posição frontal. Além disso, quando se usa uma rede desses sensores, é necessária a instalação de um computador próximo a cada sensor, uma vez que possuem interface USB e, portanto, necessitam conectar-se diretamente a um PC.

Em contrapartida, muitos ambientes possuem redes de câmeras de monitoramento. Portanto, mesmo com custo computacional mais elevado, métodos que usam apenas informação visual ainda tem sido o tema de muitas pesquisas [17].

O uso de redes neurais convolucionais (CNNs - *Convolutional Neural Networks*) e o aprimoramento de hardware como GPUs, permitiram a implementação de métodos de detecção de esqueletos mais eficientes e precisos. O maior custo computacional concentra-se na etapa de treinamento da rede, enquanto a etapa de detecção alcança, muitas vezes, desempenho para aplicação em tempo de execução [18]–[20].

Desta forma, se métodos eficientes de correspondência e reconstrução tridimensional forem desenvolvidos, será possível aplicar detectores baseados em CNNs em sistemas

multicâmeras, obtendo-se, ao mesmo tempo, a localização 3D dos esqueletos e o desempenho em tempo de execução.

Visando tal objetivo, este trabalho propõe uma metodologia para se obter as coordenadas 3D das juntas de esqueletos a partir de imagens capturadas por uma rede de câmeras. Tal rede está instalada em um Espaço Inteligente, onde futuramente se espera usar a informação das juntas no reconhecimento de gestos dinâmicos para interação homem-robô, assim como classificação de ações e comportamento de indivíduos.

O restante deste artigo está dividido da seguinte forma: na Seção II, serão discutidos alguns trabalhos relacionados, enquanto na Seção III, serão apresentados o *dataset* usado para validação do método, bem como o detector de esqueletos em imagens. A metodologia é detalhada na Seção IV e os resultados obtidos são apresentados na Seção V. Por fim, na Seção VI, conclusões e trabalhos futuros são discutidos.

II. Trabalhos relacionados

Para comparação, aqui serão mencionados alguns trabalhos onde a detecção de esqueletos é feita usando-se uma única câmera ou uma rede delas, sejam estas câmeras convencionais ou sensores RGB-D.

Em [21], é apresentado um método conhecido como *OpenPose*, capaz de detectar vários esqueletos na mesma imagem e que, se executado em GPU, apresenta tempos de detecção em torno de dezenas de milissegundos. Tal método usa uma câmera convencional e um detector baseado em CNN, fornecendo apenas a detecção 2D de esqueletos, sem se preocupar em estimar sua localização 3D no espaço.

Já em [19], a estimativa das coordenadas tridimensionais do esqueleto é realizada, mesmo usando-se apenas uma câmera. Contudo, o método, também baseado em CNN, só é capaz de detectar as juntas de uma única pessoa na imagem.

De forma semelhante, em [20] também se estimam as posições 3D das juntas a partir de detecções 2D em uma imagem. Entretanto, os limites de ângulos entre as partes do esqueleto são utilizados para eliminar casos anatomicamente impossíveis. Assim como em [19], esse método só funciona com um indivíduo por vez.

Por sua vez, o trabalho em [15] usa uma rede de câmeras RGB-D para estimar a pose tridimensional das juntas de pessoas presentes no ambiente. A obtenção dos esqueletos 2D é feita usando-se o *OpenPose* e a informação do mapa de profundidade é usada para realizar a correspondência entre as juntas nas diversas imagens e obtenção da informação 3D.

Com uma abordagem um pouco diferente, em [17], a estimativa da localização tridimensional dos esqueletos é feita através de uma rede de câmeras convencionais, considerando um conjunto de restrições algébricas no processo de triangulação e otimização por mínimos quadrados. Apesar dos autores alegarem que o método pode ser aplicado em tempo de execução, o artigo não traz nenhuma medição de desempenho que comprove tal afirmação.

Numa abordagem híbrida, o trabalho em [16] usa uma rede de câmeras convencionais e câmeras RGB-D para gravar os movimentos de dançarinos. A fusão das informações busca selecionar a melhor correspondência entre as detecções obtidas, empregando-se uma abordagem probabilística com filtro de partículas. Todavia, este método só funciona bem com um indivíduo e requer um custo maior de instalação.

Portanto, a partir da discussão anterior, nota-se que a detecção de esqueletos em uma cena tem sido abordada das mais diversas formas. É nesse contexto que este trabalho também procura contribuir, propondo um método que aborda alguns aspectos não tratados pelos outros trabalhos da área. Resumidamente, o método proposto faz uso de uma rede de câmeras convencionais para detectar as juntas de vários indivíduos ao mesmo tempo, além de fornecer a sua localização 3D em tempo de execução.

A abordagem aqui apresentada também faz uso do *OpenPose* [21] e alcança precisão de localização semelhante a [15], mas sem fazer uso de sensores RGB-D. Além disso, possui um alcance maior, por usar câmeras RGB, além de uma implementação possível de ser executada com dezenas de milissegundos, como será apresentado na Seção V.

III. Ferramentas utilizadas

Para avaliar o método proposto, utilizou-se o *dataset*[1] gerado no estúdio *Panoptic* [22]. Esse estúdio contém 480 câmeras VGA $(640 \times 480\ pixels)$, 31 câmeras HD $(1920 \times 1080\ pixels)$ e 10 sensores *Kinect II*, além de projetores de imagens utilizados para calibração. Contudo, foram utilizadas apenas imagens de 10 câmeras HD neste trabalho, pois julgou-se como o suficiente para a validação do método, uma vez que o método proposto tem como característica utilizar poucas vistas para estimar a pose de esqueletos, diferentemente de métodos que utilizam nuvem de pontos. Os dados de calibração de todas as câmeras e *Kinects* são disponibilizados no *dataset*. Todas as imagens são usadas para gerar uma nuvem de pontos densa, que é pós-processada para se obter a posição tridimensional das juntas dos esqueletos.

O *dataset* contém 65 sequências agrupadas em 8 categorias, totalizando 5,5 horas de vídeos e 1,5 milhão de anotações de esqueletos 3D. Existem dois tipos de anotações: uma com 15 e outra com 18 juntas, sendo esta última a usada neste trabalho (vide modelo na Fig. 1a). Escolheu-se categoria *Haggling* para ser usada na avaliação, pois esta possui até 3 pessoas no mesmo instante. Assim, foi possível mostrar que o método funciona com vários indivíduos. A Fig. 1b mostra as imagens de 2 câmeras de uma das sequências da categoria *Haggling*.

Tal categoria possui três sequências, detalhadas na Tabela I e nomeadas como S1, S2 e S3[2]. Vale ainda ressaltar que, neste trabalho, foram usadas apenas as imagens de 10 câmeras com resolução HD, e não de todas as 31 disponíveis.

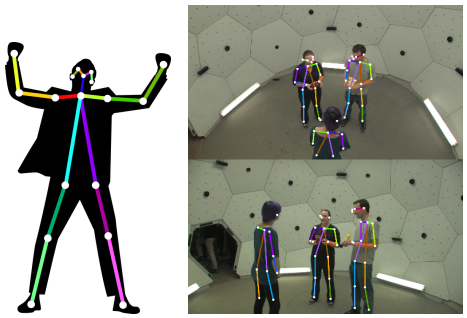
Antes da localização 3D dos esqueletos, foi preciso obter a detecção 2D das juntas nas imagens. O método aqui proposto fez uso do detector *OpenPose* [21], desenvolvido pelo mesmo grupo que gerou o *dataset* [22]. Conforme já mencionado, o *OpenPose* pode ser executado em GPU e é capaz de detectar

[1] http://domedb.perception.cs.cmu.edu/dataset.html

[2] Originalmente, tais sequências são identificadas por *160224_haggling1*, *160226_haggling1* e *160422_haggling1*.

Tabela I: Informações das sequências utilizados. O número de anotações corresponde ao total de indivíduos presentes em todas imagens, podendo ser igual a 1, 2 ou 3 em cada imagem.

Sequência	S1	S2	S3	Total
Tempo (min)	5:00	8:00	8:00	21:00
Anotações	21675	29944	32574	84193
Pessoas	12	19	18	49



(a) (b)

Figura 1: Em (a), juntas do modelo utilizado. Em (b), detecções nas imagens de 2 câmeras no mesmo instante.

vários esqueletos na mesma imagem com tempo praticamente constante.

IV. Descrição do método

O método proposto deve reconstruir tridimensionalmente as juntas dos esqueletos, detectados nas imagens capturadas por um sistema multicâmeras calibrado. Para isso, a cada instante de tempo, agrupam-se os esqueletos detectados nas imagens, associando cada um à câmera na qual foi identificado. Cada esqueleto recebe um identificador único que será usado posteriormente. A metodologia será descrita em três etapas: (1) busca de correspondências, (2) agrupamento de correspondências e (3) reconstrução tridimensional das juntas.

A. Busca de correspondências

Cada esqueleto é constituído por um conjunto de até 18 juntas devidamente identificadas e com coordenadas $(u,\ v)$ na imagem da câmera na qual foi detectada. Dado o conjunto de câmeras que compõe o sistema multicâmeras, é feita uma análise duas a duas para encontrar esqueletos correspondentes.

A correspondência entre os esqueletos em imagens de câmeras diferentes é feita através de geometria epipolar. Para melhor descrição, nas Figs. 3, 4 e 5, estão representadas as imagens de um par de câmeras, C_0 e C_1, e dois indivíduos, P_1 e P_2. Os esqueletos nas figuras são versões simplificadas com apenas 6 juntas ao invés das 18 utilizadas no trabalho. O indivíduo P_2 não aparece na imagem da câmera C_0 pois está ocluso pelo indivíduo P_1. Já na câmera C_1, os dois indivíduos estão presentes na imagem. Portanto, há três esqueletos identificados nas duas imagens e seus indivíduos correspondentes: $\{(A, P_1),\ (B, P_1),\ (C, P_2)\}$. Na Fig. 2 está apresentada uma legenda para melhor entendimento do exemplo.

Para cada esqueleto de uma imagem, busca-se o melhor correspondente nas outras imagens. Na Fig. 3, o esqueleto

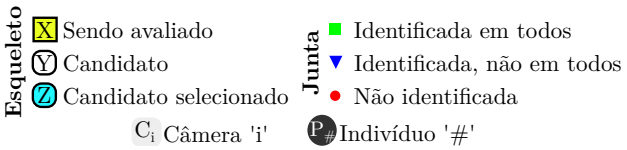


Figura 2: Legenda para as Fig.s 3, 4 e 5.

A, visto pela câmera C_0, está sendo avaliado. Na imagem da câmera C_1 há dois esqueletos, B e C. As juntas marcadas com quadrados verdes correspondem às que foram identificadas nos três esqueletos. Usa-se então a projeção dos pontos das juntas identificadas $\mathbf{m}^A_{1(0)}$, $\mathbf{m}^A_{2(0)}$ e $\mathbf{m}^A_{3(0)}$ do esqueleto A, para obter as linhas epipolares na imagem da câmera C_1. Note que na Fig. 3 a notação das variáveis foi simplificada para não poluir a imagem. O mesmo foi feito para as Figs. 4 e 5.

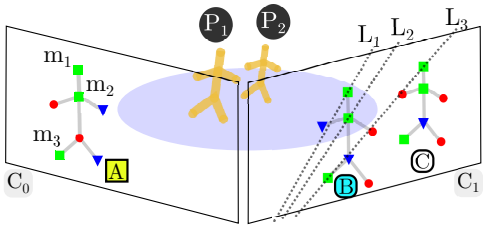


Figura 3: Exemplo do método proposto avaliando o esqueleto A.

Portanto, de modo geral, para cada ponto $\mathbf{m}^j_{k(o)}$ pertencente à câmera de origem C_o, obtém-se uma linha epipolar $L^j_{k(d)}$ na câmera de destino C_d calculada por

$$L^j_{k(d)} = \mathbf{F}^d_o\ \tilde{\mathbf{m}}^j_{k(o)} = \mathbf{F}^d_o\ \left[u^j_{k(o)},\ v^j_{k(o)},\ 1\right]^T, \qquad (1)$$

em que $u^j_{k(o)}$ e $v^j_{k(o)}$ são as coordenadas do k-ésimo ponto $\mathbf{m}^j_{k(o)}$ do j-ésimo esqueleto na imagem de câmera C_o; $L^j_{k(d)}$ é a linha epipolar na imagem da câmera de destino C_d, correspondente ao ponto $\mathbf{m}^j_{k(o)}$, sendo $\tilde{\mathbf{m}}^j_{k(o)}$ sua representação em coordenadas homogêneas; e $\mathbf{F}^d_o$ é a matriz fundamental que associa pontos na imagem da câmera C_o com suas linhas epipolares na imagem de C_d, calculada com os parâmetros de calibração intrínsecos e extrínsecos das câmeras como

$$\mathbf{F}^d_o = \mathbf{K}_d^{-T}\widehat{\mathbf{t}^d_o\mathbf{R}^d_o}\mathbf{K}_o{}^{-1}, \qquad (2)$$

onde $\mathbf{K}_o$ e $\mathbf{K}_d$ são as matrizes de parâmetros intrínsecos das câmeras C_o e C_d, respectivamente; e $\widehat{\mathbf{t}^d_o\mathbf{R}^d_o}$ corresponde ao produto vetorial entre o vetor de translação $\mathbf{t}^d_o = [t_1,\ t_2,\ t_3]^T$ e a matriz de rotação $\mathbf{R}^d_o$. O par $\mathbf{t}^d_o$ e $\mathbf{R}^d_o$ levam um ponto no referencial da câmera C_o para o referencial C_d.

Com as linhas epipolares correspondentes aos pontos $\mathbf{m}^j_{1(o)}$, $\mathbf{m}^j_{2(o)}$ e $\mathbf{m}^j_{3(o)}$ da imagem de C_o projetadas na imagem de C_d, calcula-se, para cada esqueleto candidato na imagem, a distância de cada junta à sua linha correspondente. Depois, para cada esqueleto, calcula-se a média das distâncias das juntas às linhas epipolares, da seguinte forma:

$$\bar{d}^{j,l} = \frac{\sum_k \left\langle \mathbf{m}^l_{k(o)},\ L^j_{k(d)} \right\rangle}{N}, \qquad (3)$$

em que o índice j representa o esqueleto avaliado na imagem da câmera de origem C_o; l representa o esqueleto na imagem da câmera C_d, candidato a correspondente de j; $k = 1, \cdots, N$; e N corresponde à quantidade de pontos do esqueleto j também identificados em l. Após todos os esqueletos candidatos da imagem de C_d serem avaliados, toma-se a menor distância média encontrada. Se esta for menor que um limiar, assume-se que o esqueleto associado a esta distância é o correspondente ao que está sendo avaliado na imagem de C_o. Neste trabalho foi utilizado o limiar de 50 *pixels*. O valor desse limiar deve ser proporcional à resolução das imagens utilizadas.

Tomando como exemplo a Fig. 3, se a distância média do esqueleto B for menor que a do esqueleto C e menor que 50 *pixels*, o par de esqueletos (A, B) é adicionado ao conjunto de correspondências, i.e., $\mathbf{Q} = \{(A, B)\}$.

Ainda seguindo o exemplo apresentado, uma vez que a imagem da câmera C_0 possui apenas um esqueleto, o algoritmo passa a analisar os esqueletos da imagem de C_1, como ilustrado na Fig. 4. O próximo esqueleto avaliado é, então, o esqueleto B. Nessa situação, só existe um candidato presente na imagem da câmera C_0, que é o esqueleto A. Caso seja atendida a condição de $\bar{d}^{B,A} < 50\ pixels$, adiciona-se (B, A) ao conjunto de correspondências: $\mathbf{Q} = \{(A, B),\ (B, A)\}$.

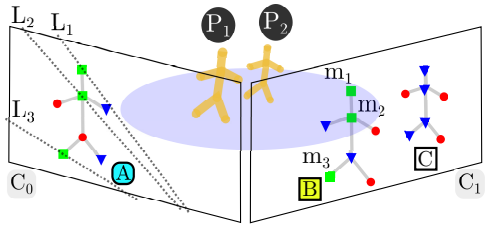


Figura 4: Exemplo do método proposto avaliando o esqueleto B.

Por fim, avalia-se o esqueleto C que possui apenas um candidato na imagem de C_0, como observado na Fig. 5. Supondo que a condição para a distância média seja atendida, mesmo que o par (C, A) seja uma falsa correspondência, este será adicionado ao conjunto: $\mathbf{Q} = \{(A, B),\ (B, A),\ (C, A)\}$.

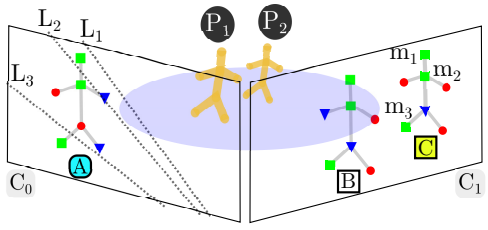


Figura 5: Exemplo do método proposto avaliando o esqueleto C.

O próximo passo, é eliminar as falsas correspondências. Para isso, é analisado se existem pares que possuam correspondentes repetidos para diferentes esqueletos na imagem de uma mesma câmera. Por exemplo, para os esqueletos da câmera C_1, há dois pares cujo o correspondente é o esqueleto A, que são (B, A) e (C, A). Como isso não é possível de ocorrer com apenas um par de câmeras, mantém-se apenas o par que possua o menor erro médio $\bar{d}$.

Assim, prováveis falsas correspondências são eliminadas, como o caso (C, A), que representava uma correspondência do indivíduo P_2, visto por C_1, com o indivíduo P_1, visto por C_0. Isso ocorre pois no campo de visão de C_0, o indivíduo P_2 está ocluso por P_1. Portanto, apenas o indivíduo P_1 será identificado nessa configuração de câmeras.

O processo continua agrupando pares de correspondências compostos pelos mesmos esqueletos. Por exemplo, os pares (A, B) e (B, A) são agrupados em um único par (A, B). Os pares restantes, não eliminados nas etapas anteriores, são adicionados a um conjunto $\bar{\mathbf{Q}}$ que agrupa todas as correspondências entre cada par de câmeras. Após todos pares de câmeras serem avaliados, realiza-se o agrupamento dessas correspondências, o qual será detalhado na próxima etapa.

B. Agrupamento de correspondências

Uma vez que as correspondências são obtidas em pares, podem haver pares com esqueletos em comum. Estes devem ser agrupados em um único grupo para realizar a etapa de reconstrução tridimensional das juntas. Este problema pode ser encarado como se fosse a busca de componentes conectados de um grafo não-dirigido, no qual cada nó representa um esqueleto e cada laço possui um anti-paralelo.

Para obter os componentes conectados, que representam os grupos de esqueletos correspondentes, basta aplicar o algoritmo DFS (do inglês, *Depth-first search*), iniciando de um nó arbitrário e percorrendo cada nó não visitado.

C. Reconstrução Tridimensional das Juntas

A partir do conjunto das correspondências agrupadas, realiza-se o processo de reconstrução 3D de cada junta. Como não há informação tridimensional a priori das juntas, só é possível a reconstrução se houver juntas detectadas em pelo menos duas câmeras. Portanto, para cada junta, verifica-se em quais câmeras ela foi detectada, e então utilizam-se suas coordenadas em cada imagem para o processo de reconstrução.

Para isso, considera-se o modelo de câmera *pinhole* conforme Equação 4, na qual λ_i corresponde a um fator de escala; $\tilde{\mathbf{m}}_i = [u_i,\ v_i,\ 1]^T$ é um ponto na imagem da câmera i; $\mathbf{K}_i$ é a matriz de parâmetros intrínsecos; Π a matriz de projeção; $[\mathbf{R}_i, \mathbf{T}_i]$ a matriz de parâmetros extrínsecos, composta, respectivamente, por uma rotação e uma translação; e, por fim, $\tilde{\mathbf{M}} = [x,\ y,\ z,\ 1]^T$ corresponde ao ponto tridimensional, no referencial global no qual as câmeras foram calibradas, que gera as projeções $\tilde{\mathbf{m}}_i$ em cada imagem. O subíndice i tem objetivo de diferenciar as câmeras, e as variáveis indicadas com $\tilde{}$ estão representadas em coordenadas homogêneas.

$$\lambda_i \tilde{\mathbf{m}}_i = \mathbf{K}_i\ \Pi\ [\mathbf{R}_i, \mathbf{T}_i]\ \tilde{\mathbf{M}} \tag{4}$$

Pode-se então para cada junta, com seus respectivos pontos $\tilde{\mathbf{m}}_i$ detectados em no mínimo duas câmeras, montar um sistema de equações para determinar o ponto $\tilde{\mathbf{M}}$, que representa a posição tridimensional da junta em questão. A Equação 4 é

reescrita na forma da Equação 5, na qual as incógnitas são λ_i e $\mathbf{M} = [x,\ y,\ z]^T$.

$$\lambda_i \tilde{\mathbf{m}}_i = \mathbf{K}_i(\mathbf{R}_i\mathbf{M} + \mathbf{T}_i) \tag{5}$$

Manipulando-se a Equação 5 para que fique na forma de um sistema de equações, tem-se

$$\lambda_i(\mathbf{K}_i\mathbf{R}_i)^{-1}\tilde{\mathbf{m}}_i - \mathbf{M} = \mathbf{R}_i^{-1}\mathbf{T}_i, \tag{6}$$

que pode ser escrita matricialmente como mostrado na Equação 7, em que $\mathbf{I}$ é uma matriz identidade 3×3

$$\begin{bmatrix}-\mathbf{I} & (\mathbf{K}_i\mathbf{R}_i)^{-1}\tilde{\mathbf{m}}_i\end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ \lambda_i \end{bmatrix} = \mathbf{R}_i^{-1}\mathbf{T}_i. \tag{7}$$

Generalizando para duas ou mais câmeras, adicionam-se fatores de escala referentes a cada uma delas como variáveis do sistema de equações, uma vez que o vetor $\mathbf{M}$ é o mesmo para todas as câmeras. Observe que cada câmera que detecta a junta fornece três equações para resolver as coordenadas do ponto tridimensional $\mathbf{M}$ e o fator de escala λ_i. Logo, tem-se a Equação 8, que representa o sistema de equações na forma matricial, onde $(\mathbf{K}_i\mathbf{R}_i)^{-1}\tilde{\mathbf{m}}_i$ é representado por $\mathbf{W}_i$, e $\mathbf{0}_{m\times n}$ é uma matriz de zeros de dimensão $m \times n$.

$$\begin{bmatrix} -\mathbf{I} & \mathbf{W}_1 & \mathbf{0}_{3\times n-1} & \\ \vdots & \vdots & \vdots & \vdots \\ -\mathbf{I} & \mathbf{0}_{3\times i-1} & \mathbf{W}_i & \mathbf{0}_{3\times n-i} \\ \vdots & \vdots & \vdots & \vdots \\ -\mathbf{I} & & \mathbf{0}_{3\times n-1} & \mathbf{W}_n \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ \lambda_1 \\ \vdots \\ \lambda_i \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1^{-1}\mathbf{T}_1 \\ \vdots \\ \mathbf{R}_i^{-1}\mathbf{T}_i \\ \vdots \\ \mathbf{R}_n^{-1}\mathbf{T}_n \end{bmatrix} \tag{8}$$

Resolvendo esse sistema de equações para cada conjunto de pontos de cada junta, obtém-se as coordenadas 3D das juntas detectadas para cada esqueleto. Esse pode ou não conter todas as 18 juntas que compõem o modelo utilizado neste trabalho.

V. Resultados e Discussões

Inicialmente, para avaliar o método proposto, foram feitos dois testes. No primeiro caso, avaliou-se apenas o erro associado às etapas de correspondência, agrupamento e reconstrução 3D. Para isso, o *ground truth* das posições 3D das juntas dado pelo *dataset* foi usado para gerar as projeções 2D nas imagens. A partir das projeções 2D, aplicaram-se as etapas do método proposto para a obtenção das posições 3D, as quais foram então comparadas aos valores de *ground truth*.

No segundo teste, foi empregada a metodologia completa, ou seja, o detector de juntas 2D foi aplicado às imagens das 10 câmeras HD e, a partir das detecções das juntas nas imagens, foi efetuada o processo de correspondência, agrupamento e reconstrução 3D. As posições tridimensionais assim obtidas foram então comparadas com o *ground truth*.

A Fig. 6 mostra o erro médio de reconstrução de cada junta bem como o erro médio geral para os dois testes realizados. Pode-se verificar que, para o primeiro teste, o erro fica em geral menor que 5 mm, enquanto para o segundo teste, no qual utilizou-se o detector, o erro atinge valores acima de 20 mm, mas com média abaixo de 15 mm. No primeiro teste, o erro está associado apenas ao erro de reprojeção e reconstrução, os quais estão relacionados ao erro de calibração das câmeras. Já para o segundo teste, além do erro inerente ao processo de reconstrução, há o erro de localização das juntas inserido pelo detector. De maneira geral, para o método proposto a precisão da localização das juntas vai depender da qualidade da calibração da rede de câmeras, da precisão do detector e da qualidade das imagens. O nível de erro aceitável vai depender da aplicação para a qual essa informação será utilizada.

Além do erro de localização, também avaliou-se a taxa de não-detecção geral para cada sequência. Os resultados estão apresentados na Tabela II, na qual pode-se verificar que, em geral, apenas cerca de 10% das anotações não foram recuperadas. Contudo, sabendo-se que diversas aplicações fazem uso da detecção de esqueletos de forma temporal, alguma técnica de rastreamento aplicada às coordenadas das juntas faria com que essa taxa se tornasse ainda menor.

Tabela II: Taxa de não-detecção (*miss-rate*) geral e para cada sequência. Os valores percentuais são em relação ao número de anotações apresentados na Tabela I.

Sequência	S1	S2	S3	Total
miss-rate (%)	4,62	21,81	4,37	10,64

A metodologia proposta também foi testada no ambiente de experimentação montado que é composto por 8 câmeras. Um teste com 4 delas foi realizado, pois estas compõem uma das áreas de sobreposição de câmeras do ambiente. Cinco indivíduos foram posicionados nesta área e uma sequência de 90 imagens por câmera foram capturadas. Na Fig. 7a, podem ser vistas as detecções das 4 câmeras, e na Fig. 7b a representação tridimensional da cena. O tempo médio para a realização do processo de correspondências e reconstrução para esta sequência foi de $22,75 \pm 6,30$ ms.

Esse tempo médio foi obtido executando-se o método em um computador com processador Intel i5-7200U @2.50GHz e 8GB de RAM. A implementação utiliza apenas uma *thread*, podendo ainda ser paralelizada uma vez que o processo de busca de correspondências 2 a 2 permite isso. Além disso, para a detecção dos esqueletos nas imagens, foram empregadas duas GPUs diferentes: uma Tesla V100 16GB e uma GeForce GTX 1080 8GB, nas quais obteve-se tempo médio de detecção de 30 e 70 ms, respectivamente, por imagem.

VI. Conclusão e Trabalhos Futuros

O trabalho apresentado propôs uma metodologia para obter as coordenadas tridimensionais das juntas dos esqueletos em um sistema multicâmera. Foi utilizado um *dataset* no qual foi possível avaliar o erro de estimativa das posições 3D reconstruídas, que na média foi de $14,2$ mm.

Além disso, a metodologia foi testada em um ambiente real em que obteve-se tempo de execução promissor para utilizar as detecções tridimensionais em aplicações que exige-se tempo real a depender da taxa desejada e do *hardware* disponível.

Como próximo passo, está a implementação modificada para o método se tornar *multi-thread*, possibilitando a execução

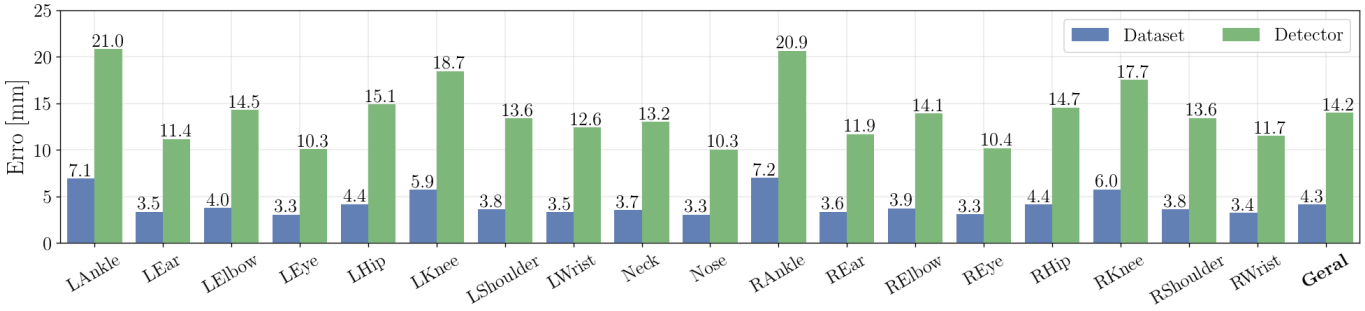
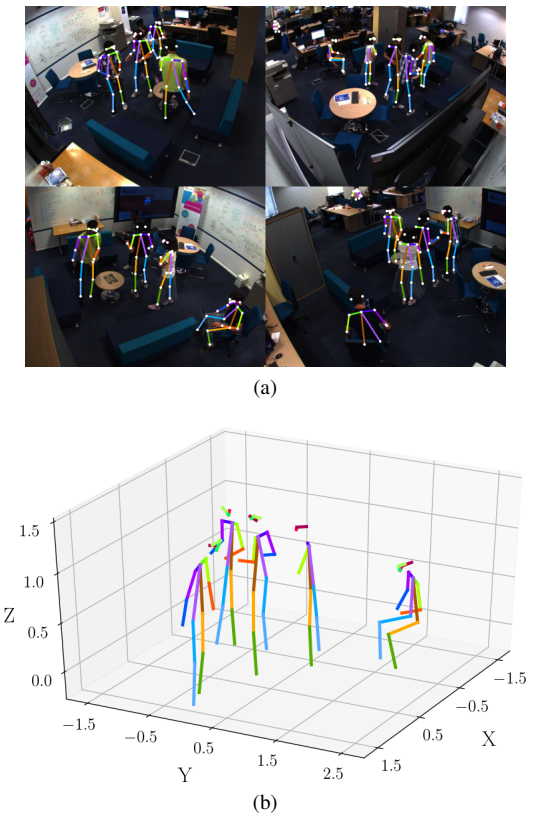


Figura 6: Erro médio de reconstrução para cada junta e geral, considerando as coordenadas dos esqueletos na imagem obtidas a partir da projeção do *ground truth* e a partir do detector utilizado neste trabalho.



(a)

(b)

Figura 7: Em (a), detecções no ambiente de experimentação, e em (b), representação 3D da cena.

em taxas de amostragem maiores. Além disso, as detecções tridimensionais das juntas, fornecidas pelo método proposto, serão futuramente utilizadas para reconhecimento de gestos e/ou ações no Espaço Inteligente onde a rede de câmeras está instalada.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro ao primeiro autor.

REFERÊNCIAS

[1] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *2013 IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 486–491.

[2] R. Lun and W. Zhao, "A Survey of Applications and Human Motion Recognition with Microsoft Kinect," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 05, p. 1555008, Aug 2015.

[3] A. Taha, H. H. Zayed, M. E. Khalifa, and E.-S. M. El-Horbaty, "Skeleton-based human activity recognition for video surveillance," *Int. Journal of Scientific Engineering Research*, vol. 6, pp. 993–1004, 2015.

[4] Z. Fang, D. Vázquez, and A. M. López, "On-board detection of pedestrian intentions," *Sensors (Basel, Switzerland)*, vol. 17(10), p. 2193, 2017.

[5] D. Casillas-Perez, J. Macias-Guarasa, M. Marron-Romera, D. Fuentes-Jimenez, and A. Fernandez-Rincon, "Full Body Gesture Recognition for Human-Machine Interaction in Intelligent Spaces," pp. 664–676, Apr 2016.

[6] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *2012 IEEE RO-MAN: The 21st IEEE Int. Symp. on Robot and Human Interactive Communication*, Sep 2012, pp. 411–417.

[7] Y. Su, Y. Wu, Y. Gao, W. Dong, Y. Sun, and Z. Du, "A upper limb rehabilitation system with motion intention detection," in *The 2nd Int. Conf. on Advanced Robotics and Mechatronics (ICARM)*, Aug 2017, pp. 510–516.

[8] E. E. Stone and M. Skubic, "Fall Detection in Homes of Older Adults Using the Microsoft Kinect," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 290–301, Jan 2015.

[9] M. Balazia and K. N. Plataniotis, "Human gait recognition from motion capture data in signature poses," *IET Biometrics*, vol. 6, no. 2, pp. 129–137, 2017.

[10] P. K. Lai and R. Laganière, "Creating immersive virtual reality scenes using a single rgb-d camera," in *Image Analysis and Recognition*. Cham: Springer International Publishing, 2017, pp. 221–230.

[11] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from rgb-d input," in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 294–310.

[12] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-Lopez, X. Baro, I. Guyon, S. Kasaei, and S. Escalera, "A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences," in *12th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2017)*, May 2017, pp. 476–483.

[13] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, June 2011, pp. 1297–1304.

[14] H. Haggag, M. Hossny, S. Nahavandi, and O. Haggag, "An adaptable system for rgb-d based human body detection and pose estimation: Incorporating attached props," in *2016 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, Oct 2016, pp. 001 544–001 549.

[15] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks," *arXiv:1710.06235*, Oct 2017.

[16] Y. Kim, "Dance motion capture and composition using multiple rgb and depth sensors," *International Journal of Distributed Sensor Networks*, vol. 13, no. 2, p. 1550147717696083, 2017.

[17] M. Lora, S. Ghidoni, M. Munaro, and E. Menegatti, "A geometric approach to multiple viewpoint human body pose estimation," in *2015 European Conference on Mobile Robots (ECMR)*, Sept 2015, pp. 1–6.

[18] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE CVPR 2017*, July 2017, pp. 1302–1310.

[19] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M.-h. Shafiei, H.-p. Seidel, D. Casas, C. Theobalt, M. Shafiei, H.-P. Seidel, and W. Xu,

"VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera," *ACM Trans. Graph. Article*, vol. 36, no. 14, 2017.

[20] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *IEEE CVPR 2015*, Jun. 2015.

[21] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *2016 IEEE CVPR*, June 2016, pp. 4724–4732.

[22] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *2015 IEEE Int. Conf. on Computer Vision (ICCV)*, Dec 2015, pp. 3334–3342.