Body Orientation Estimation Through Graph Representation: Expanding Accuracy with Data Augmentation and Gradient Boosting

Pedro Victor Vieira de Paiva, Murillo Rehder Batista, and Josué Junior Guimarães Ramos

***Abstract*—Body Orientation Estimation (BOE) is important for a wide array of applications, including robotics, surveillance and consumer analysis. Although multi-sensor approaches are effective, they are not a viable option for in the wild scenarios; the usual approach in such cases is to use single camera images, with imprecise results. Some applications that deal with people benefit from obtaining 2D human skeletons for gesture recognition, and these skeletons bring valuable information about the person's pose. It is proposed to build a 2D skeleton via OpenPose and using its data as training data on XGBoost to detect BOE. To evaluate predictions considering real situations based on a single camera, the TUD Multiview Pedestrian dataset is used and extended considering that a single person is originally considered in images where more people were often identified. It is compared the proposed approach against various state-of-the-art methods and our results indicate better performance. Finally, it is proved that our method is viable for BOE in real-time scenarios by presenting case studies on simulated scenes.**

***Index Terms*—Body orientation estimation, Pose estimation, XGBoost, Computer Vision**

I. Introdução

A utilização da pose de uma pessoa detectada (posição e orientação) é um recurso valioso para muitas aplicações, como robôs sociais e vigilância. Define-se a orientação corporal como o ângulo entre o vetor definido no plano da imagem contendo um corpo humano em relação ao ponto de vista de captura da imagem. A obtenção da orientação de uma pessoa sem recorrer a sensores no corpo ou nas roupas da pessoa requer estratégias de estimação baseadas em imagens, como as abordagens utilizando Redes Neurais Convolucionais (CNNs) vistas na literatura [1].

Outra estrutura que pode ser útil é a de um esqueleto 2D dessa pessoa construído a partir de articulações-chave do corpo humano, doravante chamadas de juntas por conveniência. Esse esqueleto pode ser usado para estimar gestos, inferir se uma pessoa está ociosa, caminhando ou correndo, entre outros usos. Uma biblioteca comum para estimar o esqueleto 2D de pessoas é a OpenPose [2]. No entanto, combinar OpenPose e outra abordagem de aprendizado profundo para estimativa de orientação pode ser muito intensivo em termos de computação devido à execução de inferências em redes profundas em paralelo. Uma característica particular dos esqueletos humanos 2D obtidos a partir de uma imagem é que algumas articulações podem não ser identificadas dependendo da pose da pessoa. Tratar tal incompletude é uma obrigação para aplicações no mundo real.

Este trabalho tem como objetivo contribuir na estimação da orientação de pessoas a partir de esqueletos incompletos associados à imagens. Ele apresenta uma estratégia baseada no treinamento de um sistema Extreme Gradient Boosting (XGBoost) com os ângulos e distâncias articulares de um esqueleto para obter a orientação de uma pessoa. O algoritmo XGBoost foi escolhido devido à sua capacidade de tratar dados incompletos, como na ausência de determinadas partes do corpo por oclusão, eficiência e velocidade.

As contribuições deste trabalho são as seguintes:

- Uma estratégia de estimativa de orientação de pessoas baseada em esqueleto 2D, robusta à ausência de partes deste esqueleto;
- A aplicação bem-sucedida do algoritmo XGBoost para estimar a orientação de uma pessoa, apresentando resultados melhores do que os métodos de classificação comparados;
- Um método de expansão de *datasets* clássicos de orientação de pessoas por imagens utilizando estimação de esqueletos;
- A aplicação da técnica proposta em cenários simulados.

Este artigo está organizado da seguinte forma. Na Seção II, são apresentados artigos que propõem soluções para o problema de estimação da orientação corporal. Na Seção III, a base para a construção de nossa solução é apresentada. Na Seção IV, nosso método proposto é detalhado. Na Seção V, experimentos e resultados são mostrados e discutidos. A Seção VI apresenta estudos de caso em cenários simulados. Este artigo é concluído na Seção VII.

II. Trabalhos Relacionados

Nesta Seção, são relacionados trabalhos que tratam da estimação da orientação de pessoas através de imagens RGB. Um *dataset* relevante para este problema é o TUD (2010) [3], que apresenta imagens de pedestres em oito faixas de orientação distintas, separadas igualmente em 45 graus. O *dataset* apresenta imagens de pessoas andando em ambientes urbanos abertos, sempre em deslocamento, e várias imagens pertencem à mesma sequência temporal. A partir do TUD, foi treinado um algoritmo SVM para a identificação da orientação.

Abordagens a partir de redes profundas também são encontradas na literatura devido ao sucesso das mesmas em problemas que envolvem visão computacional. Raza (2018) incorpora vários *datasets* para treinar uma abordagem baseada em aprendizagem profunda para obter a orientação da cabeça e corpo de pedestres [4]. Para testar a técnica, foram utilizados dois *datasets*, sendo um destes o TUD, e sequências de vídeo. Considerando uma resolução de 45 graus, foi obtida uma acurácia média de 91% para estimação da orientação da cabeça e de 92% para a orientação do corpo.

Pedro Paiva, Murillo Batista and Josué Ramos are with Cyber-physical Systems Division, Renato Archer IT Center, Campinas, Brazil. e-mail: {pvpaiva, mbatista, jgramos}@cti.gov.br

Trabalhos como o de Lee (2019) têm abordagens que permitem tratar a estimação de orientação de pessoas como um problema de regressão, o que permitiu o trabalho em diferentes *datasets* [5]. Utilizando uma floresta convolucional de projeção randômica, foi obtida uma classificação de acurácia de 81.3% e uma regressão com acurácia de 1.12 graus.

Wu (2020) propõe o dataset MEBOW (Monocular Estimation of Body Orientation in the Wild) [6]. Os rótulos de orientação consistem em 130 mil pessoas e propiciam uma acurácia mais precisa do que a apresentada no TUD, de 5 graus. O dataset foi validado realizando um treinamento prévio na base de dados COCO 2D, realizando a validação através do TUD. Foram obtidos resultados compatíveis com a resolução dos *datasets*. Apesar da robustez da base MEBOW, pouco trabalhos citam esse *dataset* inviabilizando assim a comparação em experimentos.

Lewandowski *et al.* (2019) estimou o problema de orientação usando uma CNN [1] rápida. Foi relatado que as dificuldades na localização da pessoa foram a principal fonte de erros de inferência. Outro trabalho que usa a mesma estratégia de aprendizado profundo é apresentado por Muller *et al.*(2020) [7]. De forma semelhante a esse trabalho, o OpenPose é usado para localizar pessoas no ambiente, enquanto que para rastrear pessoas previamente identificadas é aplicada uma rede convolucional (CNN) para reconhecimento de faces.

Sebti e Hassanpour (2017) [8] propuseram um conjunto de classificadores de regressão logística para estimar a orientação de uma pessoa. A estimativa se reduz ao problema de determinar o ângulo aproximado considerando oito ângulos diferentes igualmente separados em 45 graus. Cada classificador é uma abordagem *logitboost* de aumento de regressão logística, e é responsável por uma região retangular do corpo. As características foram obtidas a partir de um Histograma de Gradientes da caixa delimitadora da pessoa. Embora o método proposto tenha obtido um resultado melhor do que algumas abordagens SVM, a estimativa alcançou uma precisão de 59,7%.

Para obter a orientação de uma pessoa, Wengefeld et al. (2019) [9] usou nuvens de pontos de um sensor Kinect e agrupou pessoas a partir desses dados. Em seguida, os atributos foram retirados da nuvem de pontos da pessoa e treinados por duas técnicas de aumento de árvore, AdaBoost e XGBoost. Os resultados da estimativa de orientação foram bons quando comparados com estratégias de extração de esqueleto 3D.

Outro trabalho que usa um esqueleto 2D do OpenPose para inferir a orientação de uma pessoa é apresentado por Islam et al. (2020) [10]; seis pontos do corpo de uma pessoa são usados para extrair ângulos, que são então entregues a um classificador SVM. No entanto, não foram dados detalhes sobre as comparações com outros métodos ou sobre a eficácia do método. Além disso, não há discussão no caso de oclusões de um subconjunto de pontos.

Não foram encontrados outros trabalhos que utilizam a mesma metodologia e técnica proposta por este trabalho.Neste sentido, este trabalho preenche as lacunas relacionadas à estimação de orientação e compara os resultados obtidos com demais trabalhos que usam o mesmo *dataset* TUD como referência.

III. Fundamentação Teórica

Nesta Seção, são descritos os conceitos e fundamentos relativos à técnica de extração de esqueletos a partir de imagens proposta por [2] e o princípio de funcionamento do classificador XGBoost [11]. Além disso, apresenta-se uma definição matemática inicial das principais etapas de cada técnica.

A. Extração de Esqueletos a partir de Imagens - OpenPose

Localizar pontos-chave anatômicos em imagens, também conhecida como estimativa de pose humana, é um problema desafiador para a visão computacional. No entanto, avanços recentes em GPUs e redes neurais convolucionais ascendentes tornaram possível inferir esqueletos de várias pessoas em tempo real. Uma arquitetura CNN amplamente usada para estimativa de pose é a OpenPose, uma técnica que supera a maioria das abordagens de última geração. Proposto por Cao et al. [2], a OpenPose usa representações não paramétricas para aprender como associar partes do corpo com indivíduos nas imagens. As informações do corpo são obtidas por meio de três etapas principais: mapas de confiança, campos de afinidade e correspondência bipartida.

Os mapas de confiança são representações bidimensionais da possibilidade de uma articulação do esqueleto ocorrer em um determinado local. Cao et al. [2] propõem uma distribuição gaussiana para gerar mapas de confiança individuais para cada ponto ponderado por um valor de referência. Formalmente, dado x como uma localização conhecida para uma parte do corpo e σ como o coeficiente de propagação, o mapa de confiança S na posição p é definido como na Equação 1.

$$S = \exp(-\frac{\|p - x\|_2^2}{\sigma^2}) \tag{1}$$

As partes do corpo por si mesmas não são muito descritivas; para aumentar as informações do corpo, cada parte detectada deve ser associada aos pares, e isso é feito usando campos de afinidade. Para indicar a direção para a qual uma parte do corpo está apontando, vetores 2D são usados para codificar campos de afinidade unindo suas duas partes associadas. Resumindo o método, o alinhamento entre os pontos E é medido calculando a integral de linha sobre o vetor de afinidade como na Equação 2 para dois locais d_i, d_j e a linha segmento L.

$$E = \int_{u=0}^{u=1} L \cdot \frac{d_i - d_j}{\|d_i - d_j\|} \, du, \tag{2}$$

Finalmente, o método húngaro [12] é usado para obter a combinação ótima entre as partes do corpo e pessoas encontradas na imagem. Para aumentar o desempenho, o primeiro e o segundo estágios do método são calculados simultaneamente, dividindo a rede em dois ramos.

B. Otimização de Árvores de Decisão - XGBoost

XGBoost é uma estrutura para implementar o aumento de gradiente, que é uma técnica onde vários preditores fracos são construídos consecutivamente não apenas a partir das amostras de treinamento, mas também dos erros de modelos anteriores. Esses modelos são, geralmente, árvores de decisão [11].

O modelo construído pelo XGBoost é treinado por uma abordagem aditiva, adicionando uma árvore de regressão que minimiza a Equação 3, onde l é uma função de perda diferenciável que mede a diferença entre a previsão $\hat{y}_i$ e o alvo y_i e Ω é uma medida de complexidade de uma árvore; T é o número de folhas de uma árvore, w são os pesos das folhas e λ e γ são constantes [11]. É importante minimizar Ω para permitir a generalização do modelo.

$$\mathfrak{L}^{(\phi)} = \sum_i l(\hat{y}_i, y_i)) + \sum_k \Omega(f_k)$$
$$\Omega(f) = \gamma T + 0.5\lambda||w||^2 \quad (3)$$

Uma característica fundamental para a escolha do XGBoost neste trabalho é sua capacidade de lidar com dados esparsos, pois em muitas situações não é possível identificar todo o esqueleto de uma pessoa. Ele faz isso abordando uma direção padrão em cada divisão. Durante a construção de uma árvore, o ganho obtido de cada lado da divisão é calculado, e o lado com o maior ganho é escolhido como direção padrão.

IV. Metodologia

Um objeto em um espaço tridimensional possui três ângulos de rotação: arfagem, guinada e rolagem (do inglês *pitch, yaw and roll*). Neste trabalho, como de costume no problema de estimativa da orientação corporal, apenas um dos três ângulos de rotação, a guinada, é considerado, ou a rotação do corpo humano ao longo do eixo vertical. A abordagem é baseada na hipótese de que a saída do esqueleto do OpenPose é um conjunto de dados informativos ricos. Características simples, como distâncias e ângulos das juntas do corpo, devem ser suficientes para discriminar a orientação. Outra suposição é que a orientação quantizada é uma boa prática. Na abordagem utilizada, semelhante a Baltieri *et al.* [13], reduz-se os ângulos individuais de 360° a oito classes separadas por 45° esperando uma classificação de alta precisão. Outra razão para fazer isso é comparar com outros trabalhos que o fazem adequadamente [8]. Neste trabalho, denomina-se classificação exata quando é estimada na classe correta, e não exata quando é classificada como uma das classes vizinhas.

Um classificador Extreme Gradient Boosting é treinado com base em dois atributos do esqueleto: triangulação vetorial 2D e distância de Bray-Curtis [14]. Essas métricas são calculadas em pares (para distância) e em grupos de três (para ângulo) entre todas as partes do corpo disponíveis. A abordagem utilizada está resumida na Figura 1.

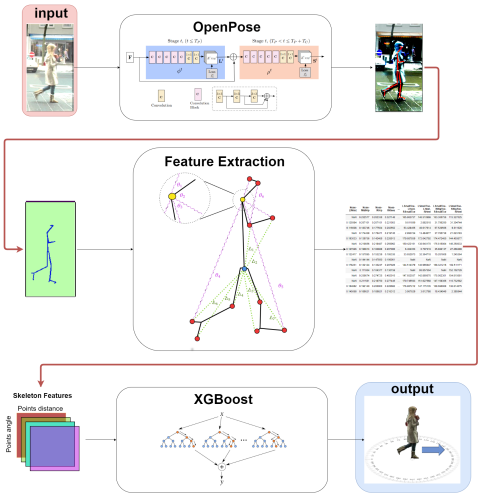


Fig. 1. Diagrama geral da abordagem proposta.

Basicamente, pode-se descrever o método em 3 etapas: (a) obtenção de esqueletos 2D a partir de imagens usando a arquitetura OpenPose; (b) calcular características de ângulo e distância entre as partes do corpo encontradas; e (c) treinar um classificador Extreme Gradient Boosting para inferir a orientação da pose a partir dos atributos selecionados.

A. Obtenção de Esqueletos Didimensionais

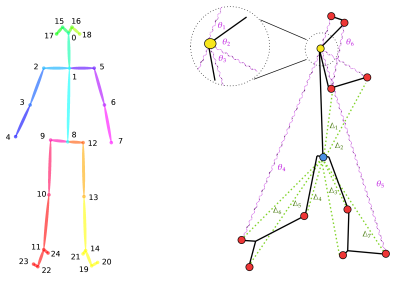
Existem duas alternativas de construção de esqueleto como saída no OpenPose. O formato de saída da pose escolhido é BODY_25, ilustrado na Figura 2(a). Esta alternativa de esqueleto é formada por 25 pontos corporais, conforme apresentado.

Uma matriz contendo as localizações das partes do corpo e a confiança de detecção é o formato de saída bruto do OpenPose. Esta matriz é composta pelas coordenadas x e y, localizações dos *pixels* e um valor que varia entre [0 - 1) representando a confiança. A estrutura de dados usada pelo OpenPose tem duas limitações: a representação é bidimensional e, quando ocorre oclusão, os pontos não são estimados. Para a identificação da orientação, são extraídas informações associadas às distâncias desses pontos e os ângulos entre esses segmentos, como mostrado a seguir.

B. Atributos do Esqueleto: Distâncias e Ângulos

Para detectar a orientação, certos ângulos de pontos e distâncias são mais informativos do que outros. No entanto, assumir um determinado conjunto de pontos corporais pode levar a uma abordagem tendenciosa ou excluir comparações expressivas. Para cobrir todas as correlações possíveis, todas as associações de pontos (pares e grupos de três) foram estimadas, o que garante uma compreensão completa das características do esqueleto. A Figura 2 ilustra as combinações de pontos. Todos os ângulos possíveis para o ponto amarelo são destacados $(\theta_1, \ldots, \theta_6)$ e todas as distâncias possíveis para o ponto azul também $(\Delta_1, \ldots, \Delta_7)$.

O número de combinações de pontos é dado por uma equação da análise combinatória, relacionada ao numero de combinações [15], que nos dá 276 combinações no esqueleto, considerando conjuntos de dois pontos, e 2034 com três amostras A Figura 2(b) ilustra algumas das amostras geradas por tais combinações.



(a) (b)

Fig. 2. (a) Modelo de esqueleto usado como referência e saída no OpenPose; (b) Representação gráfica de pares e em grupos de três combinações nos dados do esqueleto BODY_25.

Em imagens 2D, a posição é um vetor com dois valores positivos que se referem a um *pixel* em uma grade. O espaço da "imagem" se beneficia de métricas de distância que maximizam a dissimilaridade em uma distribuição regular. A distância de Bray-Curtis [14] é um sistema de coordenadas normalizado para valores positivos, variando entre 0 e 1. A distância d para dois pontos i, j no conjunto x é formulada como na Equação 4:

$$d_{i,j} = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{\sum_{k=1}^{n} (x_{ik} + x_{jk})} \quad (4)$$

Uma das premissas da medida de Bray-Curtis é que as amostras são retiradas do mesmo tamanho físico, seja área ou volume. A inferência de tamanho regular é uma propriedade

que ajuda a medida de Bray-Curtis a superar outras métricas de distância, como a distância euclidiana, no problema de orientação do esqueleto.

Os ângulos articulares são outros atributos comuns para serem utilizados quando a orientação deve ser estimada. O corpo humano tem ângulos criados naturalmente pela articulação. Projetar os ângulos naturais para um espaço R^2, como em imagens RGB, torna a triangulação vetorial 2D uma escolha adequada para calcular interseções. Para três vetores $\vec{A}, \vec{B}$ e $\vec{C} \in R^2$ dados pelo OpenPose como coordenadas de x e y em *pixel*, o ângulo interno θ entre esses vetores é dado pela Equação 5.

$$\theta = \arccos\left(\frac{\vec{BA} \cdot \vec{BC}}{\|\vec{BA}\|\|\vec{BC}\|}\right) \tag{5}$$

Quando todas as distâncias e ângulos possíveis são calculados, seus valores são usados como atributos para um algoritmo de aprendizado supervisionado. O classificador deve ser tolerante a valores ausentes.

C. Extreme Gradient Boosting

O algoritmo Extreme Gradient Boosting (XGBoost) foi usado para a classificação. Ele foi implementado usando as bibliotecas XGBoost em Python [11]. Nove parâmetros de quinze foram personalizados. Alguns dos principais ajustes e a explicação das mudanças são apresentados a seguir:

- Taxa de aprendizado (padrão 0,3) (usado 0,1): valores menores evitam que o modelo convirja muito rapidamente para uma solução subótima;
- Peso da balança (padrão 0) (usado 1): configurado para informar ao classificador sobre classes não balanceadas;
- Subamostra (padrão 1) (usado 0,8): valores mais baixos tornam o algoritmo robusto a *overfitting*;
- Gamma (default 0) (usado 0.8): valor para dividir a árvore quando ocorrer uma redução na função de perda;
- Número de árvores (padrão 100) (usado 1000): garante estimadores suficientes para o problema;
- Profundidade máxima (padrão 6) (usado 5): para reduzir a complexidade do modelo.

outros parâmetros alterados foram: função objetivo, definida como logística binária; regularização L1, reduzida para 0.3; e taxa de colapso das árvores, definido em 0.8 .

Um elemento-chave para a adoção do XGBoost é sua capacidade de lidar com valores ausentes. No mundo real, existem casos em que partes do corpo estão ausentes por vários motivos, como oclusões parciais, erros de detecção do OpenPose ou pessoas distantes do sensor. O método de aprendizado usado pelo XGBoost [11] se baseia em modelos de árvores de decisão com K funções aditivas para cada vetor de características x_i oriundos de uma dada amostra i no formato:

$$\hat{y_i} = \phi\left(x_i\right) = \sum_{k=1}^{K} f_k\left(x_i\right), \quad f_k \in \mathcal{F} \tag{6}$$

tal que $\mathcal{F}$ seja o conjunto de árvores possíveis. O simples mapeamento da função f_k para a saída $\hat{y_i}$ por si só não traria ganhos ao método; o grande ganho de eficiência está na função de regularização do XGBoost, que minimiza a função a seguir:

$$\mathcal{L}(\phi) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right) \tag{7}$$

onde

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2. \tag{8}$$

A Equação 7, l é uma função convexa de perda que estima a diferença entre os valores preditos $\hat{y_i}$ e os reais y_i. O termo expandido na Equação 8 restringe a complexidade do modelo quanto ao número de ramos por árvore T e o número de vetores de *score* na folhas ω. A implementação do método proposto está disponível no github.com.

V. Experimentos e Resultados

Nesta Seção, pretende-se provar que a abordagem proposta supera os métodos mais modernos para a estimativa da orientação do corpo humano. Compara-se o método com outras abordagens existentes usando o mesmo conjunto de dados, divisão de treinamento / teste e métricas de avaliação. Testou-se o método no conjunto de dados TUD Multiview Pedestrians [3], que contém 4730 imagens para treinamento, 248 para validação e 248 para teste. As amostras não são distribuídas igualmente no conjunto de dados, conforme apresentado na Tabela I, evidenciando este como um conjunto desequilibrado.

TABELA I
DISTRIBUIÇÃO DAS AMOSTRAS NO *dataset*
TUD Multiview Pedestrians [3].

Ângulo	90°	135°	180°	225°	270°	315°	0°	45°
Amostras	339	754	480	766	392	617	735	589

O desempenho do classificador é quantificado usando métricas de classificação extraídas da matriz de confusão, combinando valores verdadeiro-positivo (TP), verdadeiro-negativo (TN), falso-positivo (FP) e falso-negativo (FN).

A Tabela II apresenta os valores de desempenho e a Figura 3 apresenta os resultados da classificação em matriz de confusão para o método proposto. Todas as oito classes no conjunto de dados foram usadas para a previsão exata; nenhuma normalização de classe foi conduzida.

TABELA II
DESEMPENHO DE ESTIMATIVA DE ORIENTAÇÃO PARA 8 CLASSES NOS CASOS DE TESTE DO CONJUNTO DE DADOS TUD MULTIVIEW PEDESTRIANS.

$\sim$ **Ângulo**	**Precision**	**Recall**	**F1-Score**	**Acurácia**	**Amostras**
90°	92.86	50.00	65.00	53.80	26
135°	60.71	89.47	72.64	89.57	16
180°	88.10	90.24	89.16	87.84	41
225°	80.49	86.84	83.54	86.86	38
270°	93.33	60.87	73.68	65.23	23
315°	87.18	91.89	89.47	94.63	37
0°	85.37	89.74	87.50	87.21	39
45°	78.57	88.00	83.02	88.00	25

A. Discussão

Comparando os resultados obtidos com outros resultados da literatura usando o mesmo conjunto de dados e divisão de treinamento / teste (apresentado na Tabela III), pode-se ver que a abordagem proposta teve melhor desempenho na previsão exata, ou seja, sem considerar o intervalo não exato. É importante destacar que a previsão exata é a mais utilizável para muitas aplicações de orientação. O ganho de 15,2 %

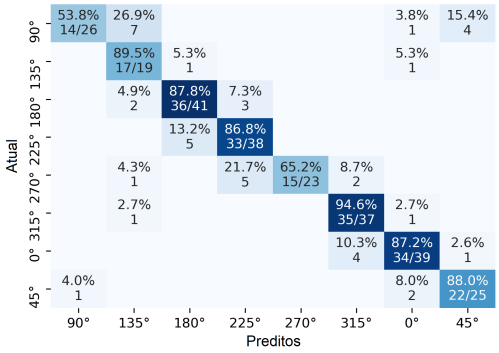


Fig. 3. A matriz de confusão para a abordagem proposta no conjunto de dados TUD Multiview Pedestrians.

na previsão exata, comparando o método apresentado com o segundo melhor [16], corrobora a contribuição do trabalho.

Outro fator que indica a assertividade na abordagem proposta é o conjunto de valores próximo à diagonal da matriz de confusão na Figura 3. Mesmo quando o classificador perde a classe de amostra, o rótulo previsto geralmente está próximo ao correto. Este comportamento fica claro nos falsos negativos das classes 90º e 270º, à direita e à esquerda, que foram as que apresentaram menor tamanho amostral (ver Tabela I). A classificação "correta" não exata é visível nos valores do F1-Score (Tabela II) de fato.

TABELA III
COMPARAÇÃO DE PRECISÃO MÉDIA ENTRE A ABORDAGEM PROPOSTA E DEMAIS MÉTODOS.

Método	**Acurácia**	
	0° (Exato)	±45° (não-exato)
Proposta sugerida	**82.6%**	**83.9%**
MoAWG [13]	67.4%	*
PLS-RF [17]	66.3%	*
HOG+LRC [16]	57.9%	83.7%
HOG+SVM+PCA [18]	53.2%	78.8%
ERT+MoAWG [13]	53.0%	81.5%

*valores não disponíveis na obra referenciada.

Deve-se enfatizar que era necessário poder comparar os resultados deste trabalho com outros trabalhos. Portanto, usou-se a mesma abordagem para testar a solução proposta, ou seja, uma definição prévia dos conjuntos de treinamento e teste, que não é uma prática usual como as técnicas de avaliação *k-fold*. Uma observação importante é que o conjunto de dados foi previamente dividido em pastas de treinamento e teste, o que não é a política de avaliação ideal.

O conjunto de dados usado neste trabalho tem um número total de 4978 imagens. A quantidade de amostras de treinamento parece ser menor do que a desejável; melhores resultados poderiam ser alcançados se mais dados estivessem disponíveis. Para provar isso, avaliamos a curva de aprendizado do classificador e a seleção de atributos. A curva de aprendizado é calculada aumentando o número de amostras iterativamente, enquanto uma validação cruzada de 5 vezes é conduzida usando uma divisão de treinamento / teste 80 / 20. Podemos ver claramente na Figura 4 que a pontuação de treinamento ainda está em torno do máximo e a pontuação de validação poderia ser aumentada com mais amostras de treinamento.

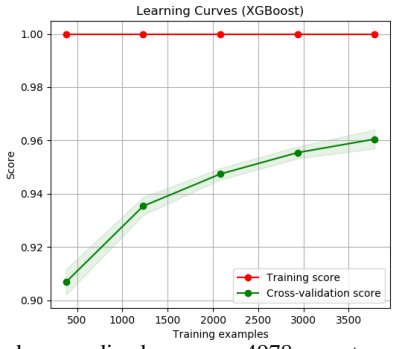


Fig. 4. Curva de aprendizado para as 4978 amostras de treinamento no conjunto de dados TUD Multiview Pedestrians.

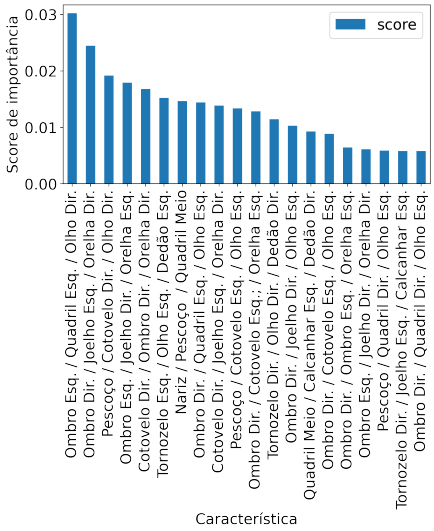


Fig. 5. Atributos selecionados para avaliar o tamanho da amostra. Cada atributo é composto por combinações de juntas de esqueleto: duas para distâncias e três para ângulos.

Outra evidência que sugere que o número de amostras era muito pequeno para treinamento é a seleção de atributos. Selecionando apenas os vinte atributos de maior pontuação (ver Figura 5) para treinar o classificador, que corresponde a cerca de 9 % do total de atributos, uma precisão média de 79,84 % é obtida. Isso enfatiza que, neste caso, para aumentar a precisão, as amostras são mais necessárias do que os atributos.

B. Ampliando o Número de Amostras via Extração de Esqueletos

É possível aumentar o número de amostras no conjunto de dados TUD Multiview Pedestrians pela otimização da rotulagem de orientação das imagens. Conforme ilustrado na Figura 6, apenas uma orientação é fornecida (destacada em verde), enquanto outros esqueletos (previstos pelo OpenPose) são ignorados. Comparando os esqueletos encontrados e as amostras rotuladas no conjunto de dados, fica evidente a subutilização das informações.

Com o intuito de melhor aproveitar o *dataset*, propôs-se um processo de reamostragem de seus dados. Ao serem inferidos os esqueletos de todos os humanos presentes em cada uma das imagens do *dataset*, é atribuída a mesma classe de rotação original a todos os humanos da imagem. Dada a característica de orientação dos pedestres presentes nas imagens, a maior

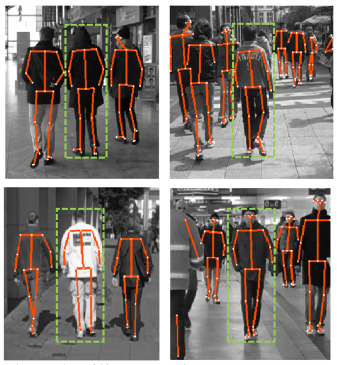


Fig. 6. Exemplo de subutilização de amostra no conjunto de dados TUD Multiview Pedestrians.

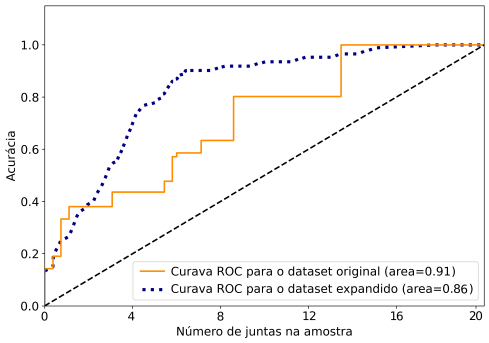


Fig. 8. Curva de ROC correlacionando número médio de juntas nas amostras e taxa predição correta.

parte das novas amostras possuem a mesma orientação do pedestre central. A fim de manter-se a corretude dos dados, uma verificação manual dos rótulos também foi feita. Ao se reamostrar a base de dados TUD Multiview Pedestrians com a extração de todos os esqueletos presentes nas cenas, o número total de amostras aumentou para 14.534 esqueletos, gerando um incremento de 292%. Ao se treinar o método proposto com as novas amostras presentes no conjunto de dados, foi alcançada uma acurácia média de 91.52%, utilizando validação cruzada 10 k-fold.

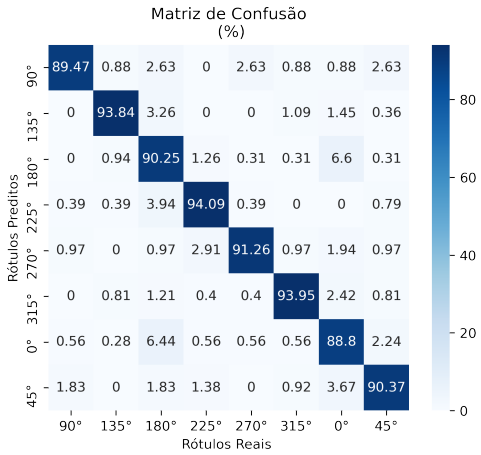


Fig. 7. Matriz de confusão com as médias por classes após o incremento da base (validação cruzada 10 k-fold).

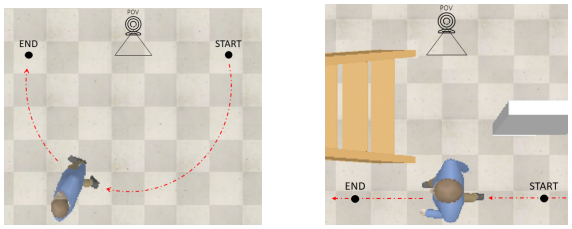
C. Robustez à Oclusão Parcial do Corpo

Para avaliar a robustez da abordagem proposta em relação a oclusão de partes do corpo na detecção correta da orientação, foi analisada a relação entre número de juntas (do esqueleto) presentes na amostra e o nível médio de acurácia do estimador. A Figura 8 sumariza essa análise.

As curvas dos conjuntos (original e expandido) provam que o aumento de amostra expande a capacidade do método estimar corretamente orientações com um número menor de juntas visíveis. No ponto de 7 juntas visíveis o modelo alcança mais de 80% de acurácia quando novos dados são inseridos, o que só ocorre no *dataset* original após as 15 juntas.

VI. Estudo de Caso

Para demonstrar como a abordagem proposta é aplicada em um robô, foi feito um cenário de aproximação humana simulada que usa nossa estratégia de estimativa de orientação. O simulador *CoppeliaSim* [19] foi utilizado nesses experimentos.



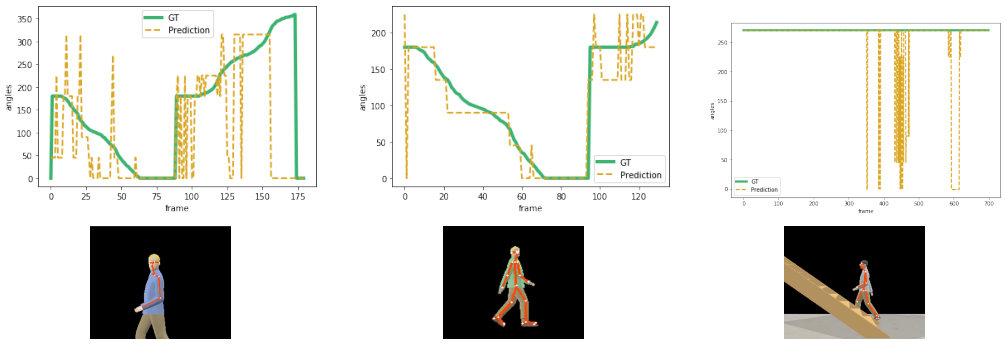
(a) (b)

Fig. 9. Rotas simuladas: (a) trajeto em U e (b) caminho em linha reta com obstáculos entre a pessoa e o robô.

Duas cenas com uma pessoa caminhando foram construídas considerando situações que um robô móvel social enfrentaria. O primeiro mostra uma pessoa fazendo trajetória em forma de U (ilustrado na Figura 9-a). O segundo apresenta uma câmera em movimento seguindo uma pessoa com obstáculos entre a câmera e a pessoa, que ocluem partes da pessoa (ilustrada na Figura 9-b).

No experimento, três cenários são apresentados: (i) parte inferior do corpo ocluída, (ii) visão de corpo inteiro e (iii) corpo parcialmente / totalmente ocluído. São apresentados gráficos para os ângulos verdadeiros e preditos e o erro quadrático médio (RMSE) correspondente entre as curvas obtidas (vídeo disponível em: vimeo.com).

O caso (i) reflete situações em que o robô está próximo ao humano, o que geralmente compromete a percepção visual das partes inferiores do corpo. Neste cenário, apresentado na Figura 10-a, o método sofre com a falta de informações do esqueleto. No entanto, mesmo com a indisponibilidade de dados, a abordagem pode prever a tendência de variação, conforme visto no gráfico do gráfico 10-a, RMSE = $67,4(+/-50,9)$. Contrastando com o Caso (i), o Caso (ii) apresenta um cenário ótimo, em que o corpo inteiro é visível. Neste caso, o método proposto apresenta um bom desempenho como visto em 10-b e um RMSE inferior = $30,3(+/-25,8)$. Por fim, o Caso (iii) simula uma caminhada lado a lado com oclusão ocasional. A abordagem proposta foi capaz de lidar com obstáculos desafiadores, conforme mostrado no vídeo do link. Nesse caso, 10-c prova que, na maioria das vezes, os valores de previsão e reais são os mesmos. O RMSE obtido é igual a $30,9(+/-29,8)$.



(a) (b) (c)

Fig. 10. Gráfico de desempenho comparando o valor real com o valor estimado e o esqueleto detectado na visão lateral para os cenários: (a) parte inferior do corpo ocluída, (b) corpo inteiro visível e (c) parcialmente ocluído.

VII. Conclusão

Neste artigo, propõe-se o uso de uma representação em grafo para prever a orientação do corpo humano. Essa abordagem é baseada nas partes do corpo e no alinhamento detectado pelo OpenPose, distâncias de Bray-Curtis e ângulos extraídos dos dados do esqueleto e aprendizado supervisionado usando Extreme Gradient Boosting. O método foi aplicado com sucesso em um conjunto estendido de dados TUD Multiview Pedestrians. Experimentos demonstraram a validade da técnica e seu melhor desempenho em comparação aos métodos de estimativa de orientação do estado da arte. Também demonstrou-se que a abordagem foi capaz de atingir maior precisão com a expansão da base de dados.

Agradecimentos

Os autores agradecem ao Programa PCI do CTI Renato Archer. Agradecimento especial ao Dr. Reid Simmons. Este trabalho recebe apoio do Projeto FAPESP 2020/07074-3.

Referências

[1] B. Lewandowski, D. Seichter, T. Wengefeld, L. Pfennig, H. Drumm, and H.-M. Gross, "Deep orientation: Fast and robust upper body orientation estimation for mobile robotic applications," in *will be published on International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[3] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 623–630, IEEE, 2010.

[4] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao, "Appearance based pedestrians' head pose and body orientation estimation using deep learning," *Neurocomputing*, vol. 272, pp. 647–659, 2018.

[5] D. Lee, M.-H. Yang, and S. Oh, "Head and body orientation estimation using convolutional random projection forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 107–120, 2017.

[6] C. Wu, Y. Chen, J. Luo, C.-C. Su, A. Dawane, B. Hanzra, Z. Deng, B. Liu, J. Z. Wang, and C.-h. Kuo, "Mebow: Monocular estimation of body orientation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3451–3461, 2020.

[7] S. Müller, T. Wengefeld, T. Q. Trinh, D. Aganian, M. Eisenbach, and H.-M. Gross, "A multi-modal person perception framework for socially interactive mobile service robots," *Sensors*, vol. 20, no. 3, p. 722, 2020.

[8] A. Sebti and H. Hassanpour, "Body orientation estimation with the ensemble of logistic regression classifiers," *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 23589–23605, 2017.

[9] T. Wengefeld, B. Lewandowski, D. Seichter, L. Pfennig, and H.-M. Gross, "Real-time person orientation estimation using colored pointclouds," in *2019 European Conference on Mobile Robots (ECMR)*, pp. 1–7, 2019.

[10] M. M. Islam, A. Lam, H. Fukuda, Y. Kobayashi, and Y. Kuno, "A person-following shopping support robot based on human pose skeleton data and lidar sensor," in *International Conference on Intelligent Computing*, pp. 9–19, 2019.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[12] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[13] D. Baltieri, R. Vezzani, and R. Cucchiara, "People orientation recognition by mixtures of wrapped distributions on random trees," in *European conference on computer vision*, pp. 270–283, Springer, 2012.

[14] C. J. Bray Jr, "An ordination of the upland forest communities of southern wisconsin. ecol monogr. 27: 325–349," 1957.

[15] D. Zwillinger, *CRC standard mathematical tables and formulae*. CRC press, 2002.

[16] A. Sebti and H. Hassanpour, "Body orientation estimation with the ensemble of logistic regression classifiers," *Multimedia Tools and Applications*, vol. 76, no. 22, pp. 23589–23605, 2017.

[17] I. Ardiyanto and J. Miura, "Partial least squares-based human upper body orientation estimation with combined detection and tracking," *Image and Vision Computing*, vol. 32, no. 11, pp. 904–915, 2014.

[18] M. Enzweiler and D. M. Gavrila, "Integrated pedestrian classification and orientation estimation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 982–989, IEEE, 2010.

[19] E. Rohmer, S. P. Singh, and M. Freese, "Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework," in *Proceedings of The International Conference on Intelligent Robots and Systems (IROS),(Tokyo). Available online at: www. coppeliarobotics. com*, 2013.

MSc. Pedro V. V. Paiva received the B.Sc. in Computer Science from Universidade Federal de Alagoas (2017), the M.Sc. in Computer Vision from Universidade Estadual de Campinas (2019). He is current fellow researcher at CTI Renato Archer. His research interests are within the fields of intelligent systems, image understanding and Human-Robot Interaction.

Dr. Murillo R. Batista received the B.Sc., M.Sc. and PhD in Computer Science from Universidade de São Paulo. He is currently a fellow researcher at CTI Renato Archer. His research interests are Human-Robot Interaction, Social Navigation and Multi-Robot Systems.

Dr Josue J. G. Ramos holds a degree in Electrical Engineering from the Federal University of Santa Catarina - UFSC (1979). The Master's Degree in Electrical Engineering from the State University of Campinas (1986) and the Ph.D. in Electrical Engineering from UFSC (2002) had an emphasis on robotic systems. In 2004 and 2013 he was a visiting researcher at the Robotics Institute at Carnegie Mellon University, USA. Since 1983, he has been working in the area of Robotics at the Renato Archer Information Technology Center, and since 2013 the emphasis is on Human Robot Interaction.