UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PEDRO LUIZ CASON CALDATO

RECONHECIMENTO DE MARCHA HUMANA UTILIZANDO POSES 3D ESTIMADAS DE MÚLTIPLAS POSES 2D

BAURU
Novembro/2023

PEDRO LUIZ CASON CALDATO

RECONHECIMENTO DE MARCHA HUMANA UTILIZANDO POSES 3D ESTIMADAS DE MÚLTIPLAS POSES 2D

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Faculdade de Ciências, da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus Bauru.
Orientador: Prof. Assoc. Aparecido Nilceu Marana

BAURU
Novembro/2023

C145r

Caldato, Pedro Luiz Cason
Reconhecimento de Marcha Humana Utilizando Poses 3D Estimadas de Múltiplas Poses 2D / Pedro Luiz Cason Caldato. -- Bauru, 2023
36 p.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: Aparecido Nilceu Marana

1. Redes Neurais. 2. Biometria. 3. Reconhecimento de marcha. 4. Estimação de poses 2D. 5. Estimação de poses 3D. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Pedro Luiz Cason Caldato

RECONHECIMENTO DE MARCHA HUMANA UTILIZANDO POSES 3D ESTIMADAS DE MÚLTIPLAS POSES 2D

Banca Examinadora

Prof. Assoc. Aparecido Nilceu Marana
Orientador
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Prof. Dra. Simone Das Graças Domingues Prado
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Prof. Dr. Kelton Augusto Pontara da Costa
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Faculdade de Ciências
Departamento de Computação

Bauru, 14 de Novembro de 2023.

Resumo

O reconhecimento biométrico de indivíduos é um campo de estudo amplamente explorado nos dias atuais, impulsionado pelos avanços recentes na computação e pela necessidade premente de reforçar a segurança em diversas configurações, incluindo vigilância e processos de autenticação em aplicativos como bancos e gerenciamento de documentos. Dentre as diversas abordagens biométricas desenvolvidas, destacam-se o reconhecimento facial, análise de impressões digitais, leitura da íris e reconhecimento de voz. No entanto, esses métodos geralmente exigem interação direta com os indivíduos, equipamentos especializados (no caso do reconhecimento de impressões digitais) ou imagens de alta resolução (no caso do reconhecimento facial ou de íris). Em situações que demandam reconhecimento biométrico indireto, a análise da marcha se mostra valiosa. Essa abordagem envolve a avaliação do padrão de caminhada de um indivíduo, bem como a análise de medidas relacionadas ao comprimento dos membros, como braços e pernas, a fim de identificar a pessoa comparando-a com um banco de dados existente. Este projeto tem como objetivo estabelecer uma abordagem de reconhecimento biométrico com base na marcha, utilizando representações tridimensionais de poses humanas extraídas de múltiplas imagens bidimensionais. Uma das vantagens notáveis dessa abordagem é sua maior resistência a oclusões, juntamente com um aumento na precisão do reconhecimento da marcha, alcançado por meio da incorporação de um conjunto mais abrangente de dados correlacionados provenientes de diferentes fontes de câmeras. O método proposto faz uso de uma Rede Neural Convolucional (CNN) para calcular um vetor descritivo por meio da agregação temporal das poses tridimensionais estimadas a partir de múltiplas imagens bidimensionais. A eficácia desse método foi avaliada usando as bases de dados CASIA GAIT-A e CASIA GAIT-B, resultando em taxas de precisão de 90,00% e 86,10%, respectivamente.

Palavras-chave: Redes Neurais, Biometria, Reconhecimento de marcha, Estimação de poses 2D, Estimação de poses 3D.

Abstract

Biometric recognition of individuals is a widely explored field in the present day, driven by recent advancements in computing and the urgent need to enhance security in various settings, including surveillance and authentication processes in applications such as banking and document management. Among the various biometric approaches developed, notable ones include facial recognition, fingerprint analysis, iris scanning, and voice recognition. However, these methods typically require direct interaction with individuals, specialized equipment (in the case of fingerprint recognition), or high-resolution images (in the case of facial or iris recognition). In situations that require indirect biometric recognition, gait analysis proves valuable. This approach involves evaluating an individual's walking pattern and analyzing measurements related to limb length, such as arms and legs, in order to identify the person by comparing them to an existing database. This project aims to establish a novel approach to biometric recognition based on gait, using three-dimensional representations of human poses extracted from multiple two-dimensional images. One notable advantage of this approach is its increased resistance to occlusions, along with an improvement in the accuracy of gait recognition achieved by incorporating a more comprehensive set of correlated data from different camera sources. The proposed method utilizes a Convolutional Neural Network (CNN) to calculate a descriptor vector through the temporal aggregation of three-dimensional poses estimated from multiple two-dimensional images. The effectiveness of this method was evaluated using the CASIA GAIT-A and CASIA GAIT-B databases, resulting in accuracy rates of 90.00% and 86.10%, respectively.

Keywords: Neural Networks, Biometry, Gait Recognition, 2D poses estimation, 3D poses estimation.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

CNN	*Convolutional Neural Network*
PCA	*Principal Component Analysis* (Análise de Componentes Principais, em português)
RNC	Redes Neurais de Convolução
RNA	Redes Neurais Artificiais
RNG	*Random Number Generator* (Gerador de Número Aleatório, em português)

Lista de símbolos

ϕ	Função de ativação.
$\mathcal{R}$	Conjunto dos números reais.
$\mathcal{Z}$	Conjunto dos números inteiros.
$\mathbf{A}$	Matriz $n \times m$.
$\mathbf{v}$	Vetor coluna.
$\hat{y}$	Vetor de comparação.
$\tilde{y}$	Vetor estimado.
$(.)$	Multiplicação matricial.
$\odot$	Produto escalar entre vetores.
$f(t)$	Função contínua em $f : \mathcal{R} \rightarrow \mathcal{R}$.
$f[n]$	Função discreta em $f : \mathcal{R} \rightarrow \mathcal{R}$ e $n \in \mathcal{Z}$.
$(f \star g)(t)$	Correlação cruzada entre $f(t)$ e $g(t)$.
$*$	Multiplicação por escalar.
$\|\|.\|\|$	Norma euclidiana.
E[x]	Esperança, ou média, de x.
Var[x]	Variância de x.

Sumário

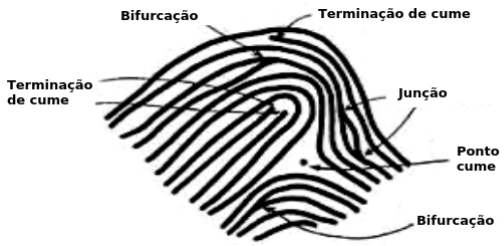
1 Introdução

Nas últimas decadas, com o avanço da tecnologia, o reconhecimento biométrico de pessoas tornou-se fundamental em aplicações que envolvem segurança de dados. Pode-se citar os aplicativos de bancos que, em geral, utilizam de impressão digital para liberar o acesso ao aplicativo, o reconhecimento facial em portarias eletrônicas e pagamentos autorizados por meio de íris (Thales Group, 2023).

O método através de reconhecimento facial baseia-se em utilizar uma imagem com o rosto de uma pessoa, extrair características únicas que fazem com que o algoritmo utilizado seja capaz de diferenciá-la entre as demais presente em um banco de dados. Os algoritmos empregados neste método é o PCA, em Kaur e Himanshi (2015), e RNC, em Alansari et al. (2023).

O uso de impressão digital utiliza, em geral, um sensor biométrico que aplica um pré-processamento na imagem antes de realizar de fato o reconhecimento. Após a coleta da imagem, é aplicado uma correção na orientação da imagem e, então, aplica-se um extrator de características. Com as características extraídas, como ilustra a Figura 1, é feito a verificação desta impressão digital com as presentes no banco de dados para achar a mais semelhante (ALI et al., 2016).

Figura 1 – Exemplo de características extraídas de impressão digital.



Fonte: Adaptado de Ali et al. (2016).

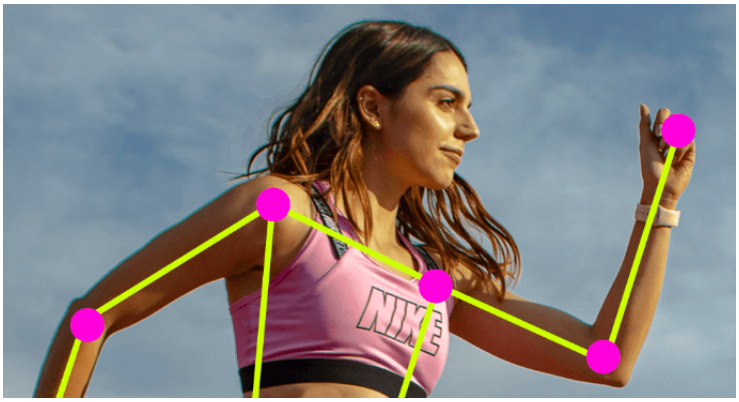
Em Rashad et al. (2011), o método proposto para reconhecimento biométrico baseado em íris baseia-se em, primeiramente, detectar o centro geométrico da pupila para auxiliar na segmentação de toda a irís visível na imagem. Em seguida, utiliza-se da imagem segmentada da íris e cria-se um histograma com a mesma. E, por fim, é feito a comparação desse histograma com os histogramas presentes no banco de dados.

Os métodos supracitados são métodos que pertencem ao estado-da-arte quando refere-se ao reconhecimento biométrico de pessoas devido a sua performance. Entretanto, a qualidade da imagem é crucial para um bom resultado (ISLAM, 2023). Em aplicações com videos de baixa-resolução, o reconhecimento biométrico através de marcha torna-se viável (JANGUA;

MARANA, 2020). O reconhecimento biométrico através de marcha possui, também, a vantagem de preservar a identidade facial da pessoa. E, além disso, também é capaz de trabalhar com imagens de baixa resolução, tornando-se, assim, uma opção viável caso não seja possível substituir os equipamentos já instalados em um ambiente onde deseja-se aplicar reconhecimento biométrico. Outra notável vantagem é que o reconhecimento biométrico por marcha pode operar de forma suscinta, ou seja, se um indivíduo tentar acobertar sua face, o reconhecimento biométrico por marcha mantém sua acuracidade, já que não utiliza de características faciais.

Em Jangua e Marana (2020), os autores propuseram o reconhecimento biométrico de pessoas por marcha humana utilizando poses 2D. A pose 2D de um corpo humano refere-se ao conjunto de juntas de partes do corpo. Estas juntas são representadas em coordenadas de pixel, no plano da imagem de uma câmera.

Figura 2 – Exemplo de pose 2D em corpo humano.

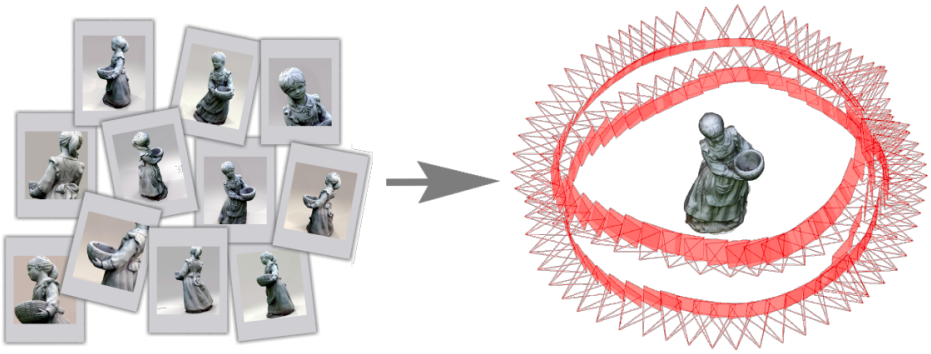


Fonte: Elaborado pelo autor.

Na Figura 2, cada círculo representa uma junta. Geralmente, os modelos de detecção de pose 2D, tais como Kreiss, Bertoni e Alahi (2021) e Cao et al. (2019), detectam a coordenada de pixel do ombro esquerdo, ombro direito, cotovelo esquerdo, cotovelo direito, pulso esquerdo, pulso direito e etc... Existe, também, uma relação entre as juntas, denotada pelas linhas em amarelo na Figura 2. Quando traçadas estas linhas, cria-se, então, o que chama-se de esqueleto do corpo humano (KREISS; BERTONI; ALAHI, 2021).

É possível representar a mesma pose 2D em três dimensões (3D) utilizando técnicas de fotogrametria. Fotogrametria é a ciência cujo objetivo é obter medidas de objetos e cenários reais a partir de imagens (HARTLEY; ZISSERMAN, 2004). Através de imagens de múltiplas câmeras, em ângulos e posições diferentes, é possível estimar a estrutura observada, exemplo na Figura 3.

Figura 3 – Exemplo de estimação de estruturas através de múltiplas vistas.



Fonte: Bianco, Ciocca e Marelli (2018)

Com os algoritmos de estimação de estruturas, é possível estimar a pose 3D a partir de múltiplas poses 2D. Um exemplo de pose 3D é representado na Figura 4.

Figura 4 – Exemplo de pose 3D.

Fonte: Elaborado pelo autor.

A proposta deste trabalho de pesquisa é a elaboração de um método para reconhecimento biométrico de pessoas através de reconhecimento de marcha humana utilizando poses 3D estimadas a partir de múltiplas poses 2D. Em Jangua e Marana (2020), os autores propuseram o reconhecimento biométrico de pessoas por marcha humana utilizando poses 2D. No relatório de acuracidade do artigo, os autores mostram os pontos em que o algoritmo desenvolvido atinge uma performance inferior. Tal momento ocorre quando há oclusão dos membros, em vistas laterais. Contudo, quando utiliza-se pose 3D, é esperado que não haja oclusão, dado que existem múltiplas observações do mesmo indivíduo de diferentes perspectivas. Sendo assim, este método tende a ser mais robusto quando há oclusões parciais.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver um método baseado em rede neural artificial (RNA) para reconhecimento biométrico de pessoas através de reconhecimento de marcha humana utilizando poses 3D obtidas a partir de múltiplas poses 2D.

1.1.2 Objetivos Específicos

Considerando o desenvolvimento do trabalho e o objetivo geral apresentado, destacam-se os seguintes objetivos específicos:

- Pesquisar sobre reconhecimento biométrico de pessoas;
- Realizar levantamento bibliográfico de trabalho correlatos;
- Extração de poses 3D a partir de múltiplas poses 2D;
- Modelagem de RNA;
- Realização de testes com a RNA proposta;
- Analisar os resultados obtidos.

1.2 Organização

O presente trabalho está organizado da seguinte forma:

Capítulo 2: Apresenta os conceitos relacionados a RNA, bem como suas aplicações em reconhecimento de pessoas. Apresenta também conceitos de pose 2D e 3D, câmeras e reconhecimento biométrico;

Capítulo 3: Apresenta o método proposto por este trabalho de pesquisa;

Capítulo 4: Apresenta a discussão dos resultados;

Capítulo 5: Apresenta a conclusão deste trabalho e propostas para trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo são apresentados os conceitos de RNA, suas aplicações no reconhecimento biométrico de pessoas, o conceito de reconhecimento de marcha humana, sua importância e a relevância de utilizar poses em 3D.

2.1 Redes Neurais Artificiais

De acordo com Haykin (2009), uma rede neural artificial (RNA) é um processador distribuído massivamente paralelo composto de unidades de processamento simples, que tem uma propensão natural para extrair conhecimento experiencial e disponibilizá-lo para uso. Uma RNA assemelha-se ao cérebro humano em dois aspectos:

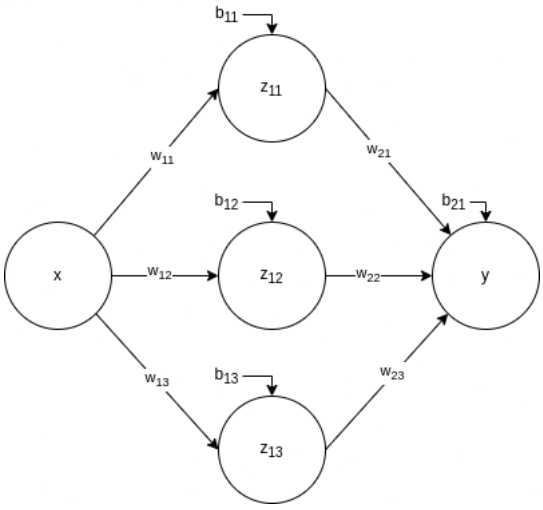
1. O conhecimento é adquirido pela RNA através de um processo de aprendizado;

2. As conexões entre os neurônios, conhecidos como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido.

2.1.1 Perceptron

Uma das arquiteturas de redes neurais mais simples que foram desenvolvidas no início é o perceptron. Um exemplo de perceptron é apresentado na Figura 5.

Figura 5 – Arquitetura do Perceptron.



Fonte: Elaborado pelo autor.

Da Figura 5, os termos são os seguintes:

- x: entrada da rede neural.

- w_{ij}: peso j da camada i.
- b_{ij}: viés j da camada i.
- ϕ: função de ativação da rede neural.
- y: saída da rede neural.

A entrada x segue a seguinte regra $x \in \mathcal{R}$. Os pesos w_{ij} são valores pertencentes à matriz W_i. A variável b_{ij} representa o elemento j da camada i que pertence ao vetor de viés. E, por fim, a função de ativação, denominada por ϕ, pode ser qualquer função $f : \mathcal{R} \to \mathcal{R}$, em geral, não-linear.

O cálculo da saída y é feito, primeiramente, calculando os valores z_{ij} conforme Equação 2.1:

$$z_{ij} = w_{ij} \cdot x + b_{ij}. \tag{2.1}$$

Então, a saída y segue como a equação 2.2:

$$y = w_{ij} \cdot z_{ij} + b_{ij}. \tag{2.2}$$

2.2 Reconhecimento Biométrico

Nesta seção é apresentado um breve histórico do reconhecimento biométrico e, em seguida, sobre o reconhecimento de marcha humana e trabalhos correlatos a este projeto de pesquisa.

2.2.1 Histórico

O reconhecimento de pessoas utilizando inteligência artificial remonta às décadas de 1960 e 1970 (Speech Ocean, 2022), quando os primeiros esforços foram feitos para automatizar a identificação de indivíduos por meio de suas características faciais. Nessa época, também, começaram a ser desenvolvidos os primeiros sensores ópticos, que proporcionaram o avanço neste segmento. No entanto, foi apenas nas últimas duas décadas que os avanços significativos começaram a ocorrer. Um marco importante foi a competição *Face Recognition Grand Challenge* (FRGC) realizada em 2005 pelo Instituto Nacional de Padrões e Tecnologia (NIST) dos Estados Unidos, que impulsionou o desenvolvimento de algoritmos de reconhecimento facial mais precisos. Desde então, tem-se observado um rápido crescimento, com aplicativos que vão desde segurança e vigilância até autenticação biométrica em dispositivos móveis e redes sociais (NIST, 2021).

A primeira rede neural convolucional (RNC), denominada *LeNet*, serviu como base para reconhecimento de caracteres, sendo este o início das redes neurais convolucionais (LECUN LéON BOTTOU; HAFFNER, 1998). Desde então, elas foram aprimoradas significativamente, aumentando sua capacidade de identificar faces em imagens e vídeos, e permitindo aplicações em tempo real e em larga escala. Além disso, os avanços no aprendizado profundo e no treinamento de grandes conjuntos de dados de faces humanas contribuíram para a precisão e robustez dos sistemas de reconhecimento biométrico de pessoas.

Taigman et al. (2014) atingiram o estado-da-arte, em métricas de acurácia, no conjunto de dados proposto por Huang et al. (2007), denominado *Labeled Faces in the Wild*, estabelecendo um novo marco da tecnologia de reconhecimento facial através de imagens e redes neurais. Desde então, diversos algoritmos e melhorias foram propostas como consta em Masi et al. (2018).

2.2.2 Reconhecimento de Marcha Humana

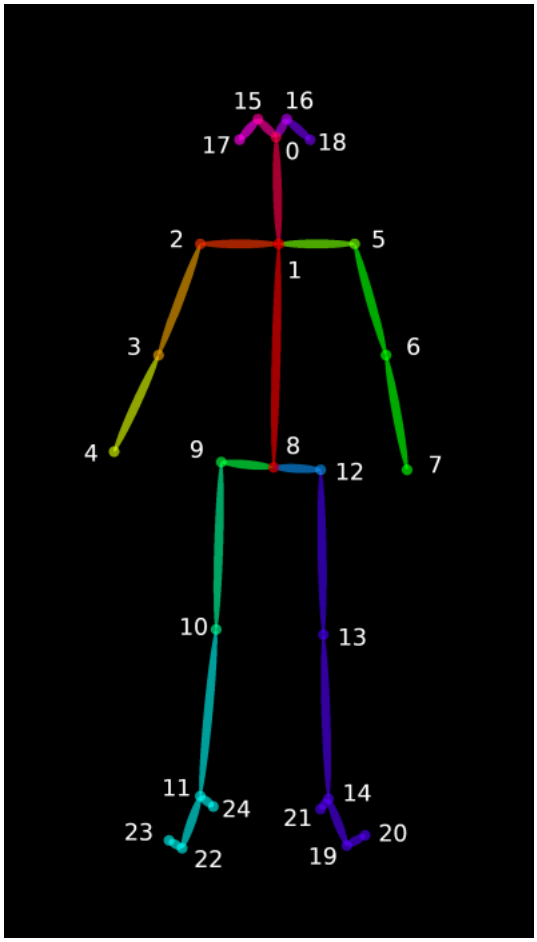
Como citado anteriormente, o reconhecimento biométrico por face, íris e impressão digital dependem da qualidade da imagem para garantir a qualidade do método. Então, em ambientes onde câmeras possuem baixa resolução, ou estão distantes, o reconhecimento biométrico de marcha pode ser empregado. Além disso, também destaca-se a necessidade, eventual, de aplicar o reconhecimento biométrico de forma anônima. Nesse contexto, as câmeras, até então com o intuito de monitoramento, podem ser utilizadas para reconhecimento biométrico de pessoas também.

Jangua e Marana (2020) utilizam do reconhecimento de marcha humana através de poses 2D. A pose contempla um conjunto de pontos, em coordenadas de pixel na imagem, que formam uma representação, simplificada, de um esqueleto humano, como ilustra a Figura 6.

Para a extração de poses é utilizada uma RNC denominada *Openpose* (Cao et al., 2019). O objetivo desta RNC é estimar as coordenadas dos pixels de 25 pontos, denominados juntas, que formam o esqueleto humano, ilustrado na Figura 7.

Após realizar a detecção das 25 juntas do esqueleto humano, os autores computam os ângulos e as distâncias das juntas para um conjunto de imagens sequenciais. Em seguida, cria-se um histograma com estes dados para cada junta. Este histograma serve como o vetor de entrada para um classificador baseado em k-NN. O classificador, por fim, que irá classificar cada indivíduo. O diagrama de blocos do método encontra-se na Figura 8.

Figura 6 – Esqueleto com a pose 2D estimado pela RNC *OpenPose*.

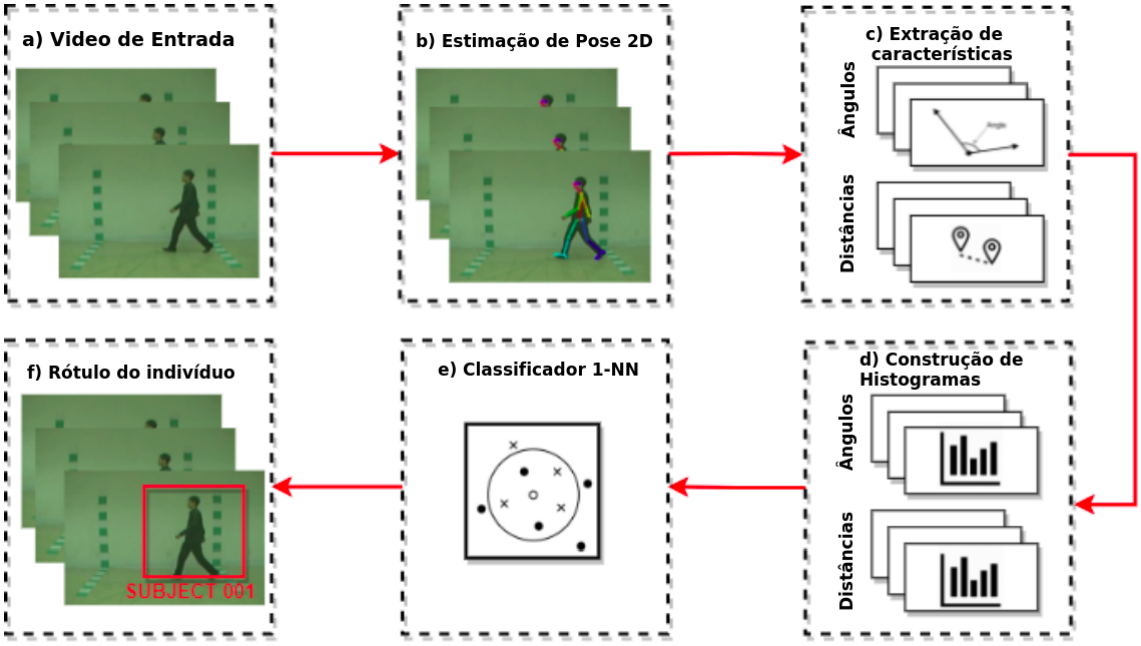


Fonte: Jangua e Marana (2020).

Figura 7 – Exemplo de pose 2D detectado pela RNC *OpenPose*.

Fonte: Cao et al. (2019).

Figura 8 – Diagrama de blocos do método proposto por Jangua e Marana (2020).



Fonte: Adaptado de Jangua e Marana (2020).

3 Material e Métodos

O objetivo deste capítulo é apresentar o material utilizado neste trabalho e, principalmente, o método proposto para o reconhecimento biométrico com base na marcha, utilizando representações tridimensionais de poses humanas extraídas de múltiplas imagens bidimensionais.

3.1 Conjuntos de Dados

Neste trabalho, utilizou-se dois conjuntos de dados CASIA Gait Dataset-A e CASIA Gait Dataset-B propostos em (Center for Biometrics and Security Research, 2010). CASIA Gait Dataset-A possui 20 indivíduos distintos, para cada pessoa existem 12 sequências de imagens, 4 sequências para cada uma das três direções, isto é, 0 graus, 45 graus e 90 graus, com relação ao plano de imagem. A Figura 9 apresenta três exemplos de imagens contidas neste conjunto de dados.

Figura 9 – Exemplo de imagens no conjunto de dados CASIA Dataset-A.

Fonte: Center for Biometrics and Security Research (2010).

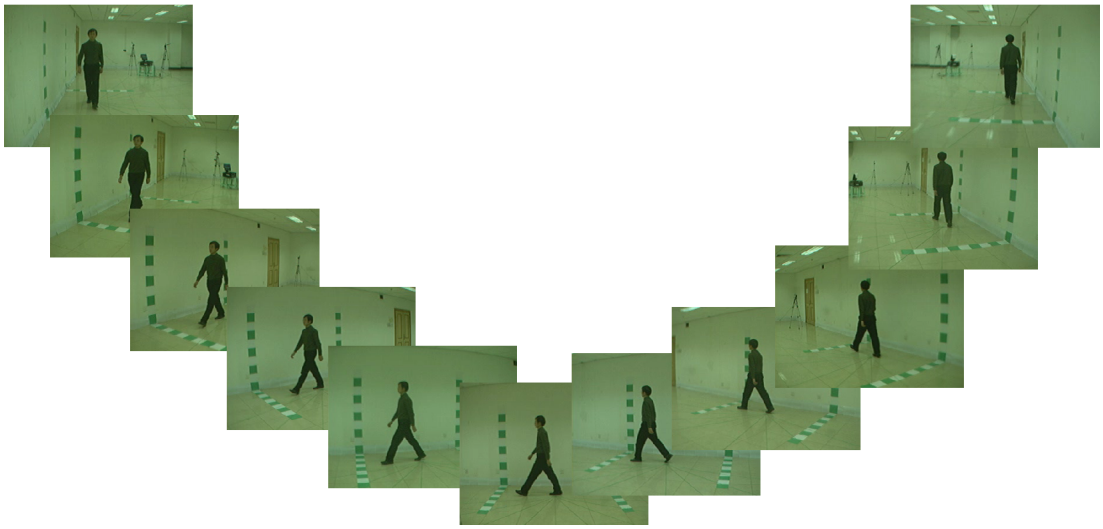
Já o conjunto de dados CASIA Gait Dataset-B possui 124 indivíduos distintos, cada um destes indivíduos foram filmados através de 11 câmeras, posicionadas em ângulos e posições diferentes. A Figura 10 apresenta exemplos de imagens contidas neste conjunto de dados.

3.2 Método Proposto

O fluxo de dados e os procedimentos essenciais para a estimativa da pose 3D de um indivíduo e seu subsequente reconhecimento são ilustrados na Figura 11.

Na abordagem apresentada, as imagens provenientes de diferentes câmeras, todas sincronizadas e pertencentes ao mesmo indivíduo, constituem os dados de entrada. Inicialmente, a pose 2D é extraída para cada imagem utilizando a rede neural convolucional OpenPifPaf, conforme detalhado na Subseção 3.2.1. Posteriormente, algoritmos de fotogrametria são

Figura 10 – Exemplo de imagens no conjunto de dados CASIA Dataset-B.



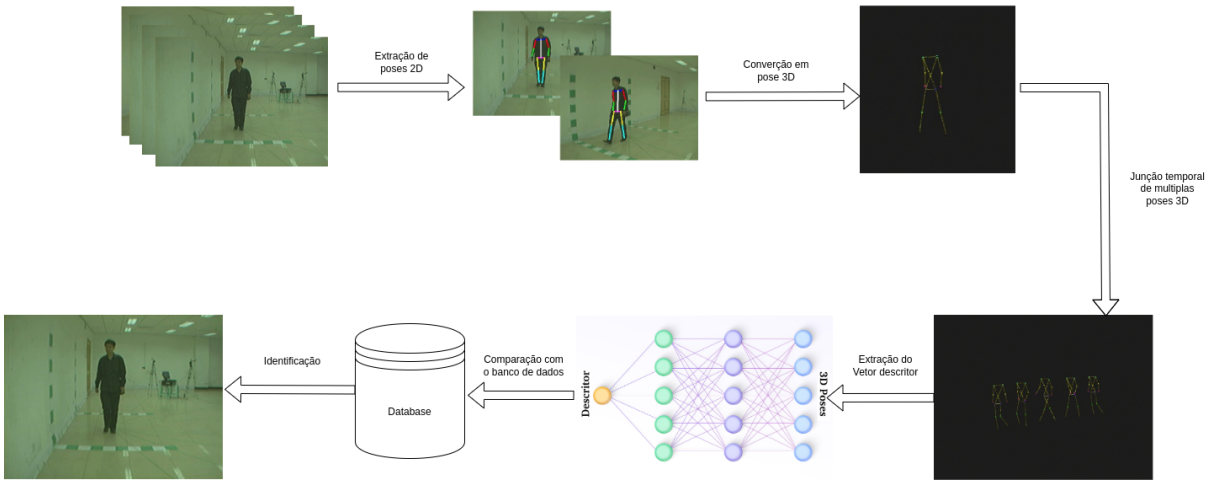
Fonte: Center for Biometrics and Security Research (2010).

empregados para converter as poses 2D em poses 3D, conforme detalhado nas Subseções 3.2.2 e 3.2.3.

As quarenta poses 3D temporais do mesmo indivíduo são agrupadas e encaminhadas para a rede neural proposta, detalhada na Subseção 3.2.4. Essa rede tem como objetivo gerar um vetor descritor que represente o padrão de caminhada derivado das poses 3D fornecidas. A função de custo utilizada por esta rede neural está descita na Subseção 3.2.5.

Por fim, realiza-se uma comparação entre o vetor descritor gerado e o banco de dados. Ao calcular a menor distância cossenoidal entre o vetor descritor atual e os registros do banco de dados, é possível identificar de maneira única o indivíduo em questão.

Figura 11 – *Diagrama de bloco do método proposto*.



Fonte: Elaborado pelo autor.

3.2.1 Extração de Poses 2D

Para realizar a extração de poses 2D, utilizou-se do modelo de RNC *OpenPifPaf* (KREISS; BERTONI; ALAHI, 2021). A escolha se deu ao fato do modelo possuir código de licença aberta para pesquisa e facilidade em executá-lo. As Figuras 12 e **??** representam a detecção da pose 2D, do mesmo indivíduo, em videos diferentes.

Figura 12 – Exemplo de poses 2D detectada pela RNC *OpenPifPaf*.

Fonte: Elaborado pelo autor.

A *OpenPifPaf*, diferentemente do método OpenPose proposto por Cao et al. (2019), utilizado por Jangua e Marana (2020), estima 17 juntas, ao invés de 25. A pose 2D produzida pelo método *OpenPifPaf* é ilustrada na Figura 13.

3.2.2 Estimação de pose 3D a partir de múltiplas poses 2D

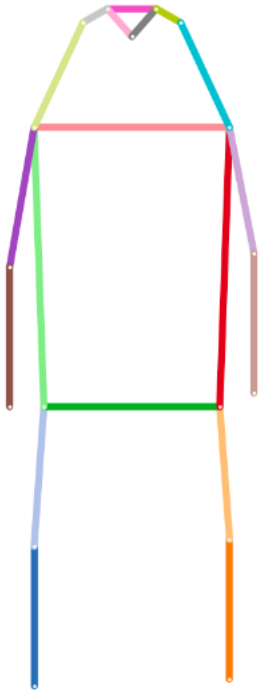
Para ser possível estimar as poses 3D a partir de múltiplas poses 2D, é necessário realizar o processo de triangulação (HARTLEY; ZISSERMAN, 2004). A triangulação consiste em estimar a posição 3D de um ponto dado a observação do mesmo em duas, ou mais, imagens.

A Figura 14 ilustra o processo de triangulação, onde o ponto P é observado nas imagens O_1 e O_2 como sendo o ponto p_1 e p_2, respectivamente.

Para obter o ponto P, primeiro precisa-se definir como ele é projetado no plano da imagem de uma câmera.

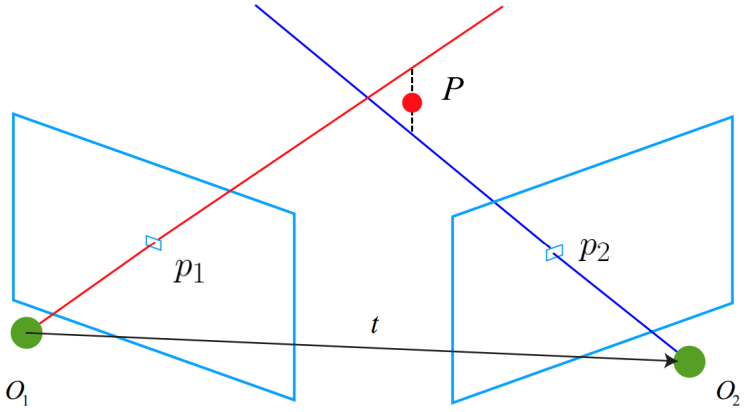
O modelo de projeção de câmera mais comum é o *pinhole* (GAO; ZHANG, 2021). Neste modelo, tem-se uma matriz $\mathbf{K}$ com os parâmetros de distância focal horizontal e vertical, denominados f_x e f_y, respectivamente, e os seus centros ópticos, denominados c_x e c_y, respectivamente. A Equação 3.1 demonstra como esses parâmetros estão organizados. A Equação 3.1 é denominada matriz de parâmetros intrínsecos.

Figura 13 – Pose humana 2D estimada pelo *OpenPifPaf*.



Fonte: Kreiss, Bertoni e Alahi (2021).

Figura 14 – Processo de triângulação para obtenção de ponto 3D.



Fonte: Kreiss, Bertoni e Alahi (2021).

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3.1}$$

Dado um ponto $P = (x, y, z) \in \mathcal{R}$, ele é projetado no plano da imagem, como sendo coordenadas de pixel $p = (u, v)$ através de:

$$\begin{bmatrix} u \\ v \\ z \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.2}$$

Da Equação 3.2, tem-se:

$$\begin{cases} u &= f_x x + c_x z \\ v &= f_y y + c_y z \\ z &= z \end{cases} . \tag{3.3}$$

Agora, da Equação 3.3, dividi-se todos os termos por z, este processo é chamado de normalização (GAO; ZHANG, 2021). Então, tem-se:

$$\begin{cases} u &= \frac{f_x x}{z} + c_x \\ v &= \frac{f_y y}{z} + c_y \end{cases} . \tag{3.4}$$

Logo, da Equação 3.4, o par (u, v) é a coordenada de pixel no plano da imagem do ponto 3D P.

Agora, ao realizar o processo contrário, ou seja, tentar obter o ponto P a partir do ponto p, percebe-se que é não será possível recuperar a escala, ou a profundidade do ponto. Isto se deve ao processo natural de projeção de um ponto 3D em uma imagem. Entretanto, é possível recuperar este ponto caso exista mais observadores, ou seja, mais imagens de um mesmo ponto em posições e orientações diferentes (HARTLEY; ZISSERMAN, 2004). Para esta finalidade, é feito, enfim, o processo de triangulação, que consiste em escrever como o ponto P, da Figura 14, é projetado na imagem O_1 e O_2.

Para a imagem O_1:

$$z p_1 = \mathbf{KP}, \tag{3.5}$$

e para O_2

$$z p_2 = \mathbf{K}(\mathbf{RP} + \mathbf{t}). \tag{3.6}$$

Nas Equações 3.5 e 3.6, z é o mesmo ponto z conforme descrito na Equação 3.3. Na Equação 3.6, observa-se que $\mathbf{R}$ e $\mathbf{t}$ são a matriz de rotação e o vetor de translação do observador O_2, respectivamente. Esses dois parâmetros mostram que houve um deslocamento e uma rotação entre o observador O_2 e o observador O_1.

Das Equações 3.5 e 3.6, é possível obter a profundidade z se for conhecido a matriz de parâmetros intrínsecos K e a rotação e deslocamento do observador O_2, em relação ao observador O_1 isolando as variáveis.

A forma de obtenção de tais valores é descrito na subseção 3.2.3.

3.2.3 COLMAP

O COLMAP, disponível em COLMAP (2023), é um software de código aberto cujo objetivo é realizar a reconstrução de modelos 3D a partir de múltiplas imagens 2D. Ele utiliza a técnica de Estrutura a partir do Movimento, *Structure-from-Motion* (SfM) em inglês, para estimar a matriz de projeção entre múltiplas câmeras em um sistema de visão estéreo. A matriz de projeção de uma câmera é geralmente representada como $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}]$, onde $\mathbf{K}$ é a matriz de calibração intrínseca da câmera, $\mathbf{R}$ é a matriz de rotação que descreve a orientação da câmera e $\mathbf{t}$ é o vetor de translação que descreve a posição da câmera. O COLMAP estima essas matrizes usando correspondências de pontos 2D-3D, que relacionam os pontos do mundo 3D aos pontos de imagem 2D observados nas imagens. O algoritmo de SfM minimiza o erro de reprojeção, onde para cada correspondência de ponto, a diferença entre a projeção 3D do ponto no espaço e sua projeção 2D observada nas imagens é minimizada. Isto é feito resolvendo um sistema de equações não-linear, minimizando a função de custo de reprojeção, conforme mostra a Equação 3.7. O resultado é a estimativa das matrizes $\mathbf{R}$ e $\mathbf{t}$ que descrevem a transformação entre as câmeras estéreo.

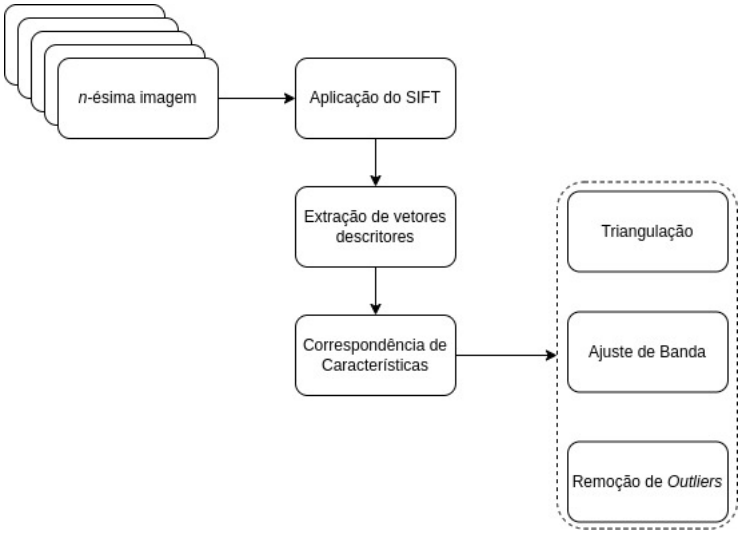
$$\min_{R,t} \sum_i \|x_i - P_1 X_i\|^2 , \tag{3.7}$$

onde x_i são as coordenadas 2D dos pontos correspondentes nas imagens e X_i são as coordenadas 3D dos pontos correspondentes no espaço.

A Figura 15 mostra as etapas do processo de estimação da matriz de projeção através do COLMAP.

Do diagrama apresentado na Figura 15, a primeira etapa é aplicar as imagens coletadas no algoritmo *SIFT*. O SIFT é um extrator de características, que detecta pontos distintos na imagem de entrada e, também, cria um vetor descritor de dimensão 128 para representá-lo vetorialmente (LOWE, 2004). Com as características detectadas pelo *SIFT*, é feita a correspondência entre imagens. Este processo serve para determinar se as imagens estão observando pontos em comum.

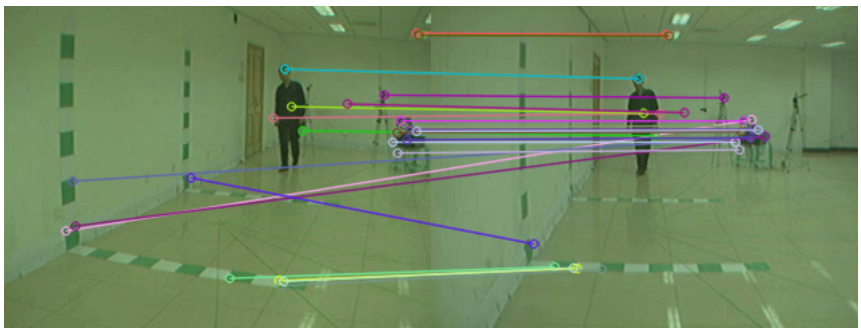
Figura 15 – Etapas do processo de estimação da matriz de projeção através do COLMAP.



Fonte: Elaborado pelo autor.

Na Figura 16, as linhas desenhadas representam onde o algoritmo de correspondência encontrou pontos semelhantes nas imagens. O algoritmo de correspondência utilizado pelo COLMAP é o *brute-force matching*, correspondência por força-bruta em tradução livre. Este algoritmo irá analisar a combinação de correspondência entre todos os vetores de características detectados. Este processo, em geral, é lento devido a busca exaustiva.

Figura 16 – Correspondência de características usando SIFT como descritor.



Fonte: Elaborado pelo autor.

A última etapa consiste em um problema de otimização denominado *Bundle Adjustment*, ou Ajuste de Banda. Este problema de otimização pode ser formulado pela Equação 3.8:

$$\mathcal{C} = \text{argmin} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \|\pi(\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i, \mathbf{P}_j) - \mathbf{p}_{i,j}\|. \tag{3.8}$$

A Equação (3.8), obtida de Tang e Tan (2019), tem o intuito de encontrar os parâmetros de projeção $\mathbf{K}_i$, $\mathbf{R}_i$, $\mathbf{t}_i$ e $\mathbf{P}_j$ que minimizem a distância euclidiana do ponto 2D detectado $\mathbf{p}_{i,j}$. Na Equação 3.8, o termo $\mathbf{K}_i$ representa a matriz de parâmetros intrínsecos i-ésima imagem.

Contudo, assume-se, para simplificar o problema, que todas as imagens foram capturadas pela mesma câmera, com foco fixo. A função $\pi(.)$ é uma outra forma de representar a projeção de um ponto 3D no plano de imagem como foi descrito nas Equações 3.2 e 3.3.

Uma vez que é atingido o platô da otimização, isto é, conforme mais iterações são alcançadas, a função de custo $\mathcal{C}$ não diminui, obtem-se a matriz de parâmetros intrínsecos $\mathbf{K}$, e as matrizes de rotação $\mathbf{R}$ e o vetor de translação $\mathbf{t}$ para cada uma das imagens inclusas no *software* COLMAP.

Para o conjunto de dados CASIA GAIT-A, selecionou-se três imagens, pois só existem três câmeras. As câmeras foram gravadas de forma síncrona e com a mesma taxa de aquisição, logo todos os vídeos possuem a mesma duração.

Para o conjunto de dados CASIA GAIT-B, selecionou-se 11 imagens, pois no conjunto de dados existem 11 câmeras posicionadas em ângulos e posições distintas.

Com os resultados, estimou-se o esqueleto em três dimensões resolvendo o sistema de equações das Equações 3.5 e 3.6. Para resolver este sistema específico, utilizou-se a biblioteca *OpenCV* Bradski (2000), com a função *cv2.triangulatePoints*, em *Python*. O resultado final são os pontos 3D obtidos pelas matrizes de projeção, *i.e.*, o conjunto $\mathbf{K}_i, \mathbf{R}_i$ e $\mathbf{t}_i$. conforme ilustra a Figura 17.

Para realizar a estimação, conforme as Equações 3.5 e 3.6, necessita apenas de dois observadores, *i.e.*, duas câmeras. Dado que existem diversas observações em ambos os conjuntos de dados, escolheu-se as vistas frontal e oblíqua para o conjunto GAIT-A e as vistas de 0 graus e 53 graus para o GAIT-B. O motivo dessa escolha ocorreu pois é possível observar todos as partes dos membros, mesmo em momentos de oclusão parcial. Garantindo, assim, precisão na reconstrução 3D.

3.2.4 Modelo de RNA

O modelo de RNA proposto para este trabalho baseou-se em dois principais artigos. Em (JANGUA; MARANA, 2020), os autores condensaram as informação de ângulos e distâncias das juntas através de um histograma. Este histograma serviu como o extrator de características para realizar o reconhecimento biométrico através de marcha.

Para criar esse vetor de caracterísitcas, foi selecionado uma RNC similar a Arandjelović et al. (2016). O objetivo da RNC de Arandjelović et al. (2016) é condensar as características de uma imagem em um vetor de características, também chamado de vetor descritor. Este vetor descritor é utilizado para fazer busca por imagem em um banco de dados. Neste banco de dados, cada item é composto por um vetor descritor também. Escolheu-se RNC devido a sua capacidade de aprender características relevantes e redução de dimensionalidade (LECUN LéON BOTTOU; HAFFNER, 1998).

Figura 17 – Pose 3D obtida a partir de duas poses 2D.

Fonte: Elaborado pelo autor.

A RNC de Arandjelović et al. (2016) tem como objetivo criar uma representação vetorial, em um espaço vetorial de menor dimensão, de uma imagem. Este problema comumente é chamado de redução de dimensionalidade (KAUR; HIMANSHI, 2015).

A Tabela 1 ilustra a RNC projetada para este trabalho. Ela se assemelha a RNC de Arandjelović et al. (2016), entretanto, optou-se por não utilizar o bloco *VLAD*, para simplificar o treinamento da RNC.

A entrada x é composta pela agregação temporal de 40 amostras, consecutivas, dos pontos das juntas do esqueleto em três dimensões.

As poses 3D agregadas temporalmente, em 40 amostras consecutivas (escolheu-se 40 empiricamente), são fornecidas para o modelo de RNC, cujo resultado é um vetor que condensa a informação e a representa em um espaço vetorial de menor dimensão. Pode-se escrever que a RNC, denotada por $\mathbf{h}(\mathbf{x})$, é uma função que satisfaz $\mathbf{h} : \mathcal{R}^{40\times17\times3} \rightarrow \mathcal{R}^{256}$.

A última camada, escrita como *L2-NORM* é a normalização do vetor descritor final. É computado a distância euclidiana do vetor e, então, divide-se ela por todos os elementos do

Tabela 1 – Modelo de RNA proposto.

Camada	Canais de Entrada	Canais de Saída
$\mathbf{x}$	$40 \times 17 \times 3$	-
CONV 2D	$40 \times 17 \times 3$	$64 \times 17 \times 3$
BATCH NORM	$64 \times 17 \times 3$	$64 \times 17 \times 3$
RELU	$64 \times 17 \times 3$	$64 \times 17 \times 3$
CONV 2D	$64 \times 17 \times 3$	$128 \times 17 \times 3$
BATCH NORM	$128 \times 17 \times 3$	$128 \times 17 \times 3$
RELU	$128 \times 17 \times 3$	$128 \times 17 \times 3$
CONV 2D	$128 \times 17 \times 3$	$256 \times 17 \times 3$
BATCH NORM	$256 \times 17 \times 3$	$256 \times 17 \times 3$
RELU	$256 \times 17 \times 3$	$256 \times 17 \times 3$
FLATTEN	$256 \times 17 \times 3$	13056×1
LINEAR	13056×1	512×1
BATCH NORM	512×1	512×1
RELU	512×1	512×1
LINEAR	512×1	256×1
L2-NORM	256×1	256×1

Fonte: Elaborada pelo autor.

vetor descritor. Transformando, assim, o vetor descritor com norma 1.

3.2.5 Função de Custo

A função de custo empregada para o treinamento da RNA proposta foi a *Triplet Loss*, modelada através da equação 3.9:

$$l(a, p, n) = max(d(a_i, p_i) - d(a_i, n_i) + 1, 0), \tag{3.9}$$

onde $d(x, y)$ é o cosseno entre os vetores x e y calculado através do produto escalar, assim como na equação 3.10:

$$\cos\theta = \frac{x \bigodot y}{||x|| \cdot ||y||}. \tag{3.10}$$

Onde, da equação 3.9:

- a_i Exemplo âncora.
- p_i Exemplo positivo.
- n_i Exemplo negativo.

Um exemplo âncora e um exemplo positivo referem-se a exemplos pertencentes a uma mesma instância, uma mesma pessoa para o nosso caso, porém com uma modificação, de

forma que a entrada não seja exatamente a mesma. Aqui, para criar um exemplo positivo, aplicou-se uma técnica de aumento de dados.

Técnicas de aumento de dados são empregadas quando há poucas amostras em um conjunto de treinamento. Dado que a entrada de nossa rede neural é o conjunto histórico das poses em três dimensões, aplicou-se as operações de rotação e translação nos pontos, conforme ilustra a equação 3.11:

$$\mathbf{P}' = \mathbf{R} \cdot \mathbf{P} + \mathbf{t}, \tag{3.11}$$

onde $\mathbf{R}$ e $\mathbf{t}$ são a matriz de rotação, pertencente ao grupo $SO(3)$, e o vetor de translação pertencente ao conjunto $\mathcal{R}$, respectivamente (GAO; ZHANG, 2021). Estes valores foram gerados utilizando RNG da biblioteca *NumPy*.

O emprego da *Triplet Loss* se deve ao fato dela fornecer um aprendizado similar ao de uma função de custo comumente usada para classificação. A grande vantagem é que aprender um vetor descritor e realizar uma comparação por métrica não exige realizar o treinamento do modelo de RNC, ou RNA, novamente se houver alguma classe nova (HOFFER; AILON, 2018). Comumente, nos modelos de classificação, possuem um número exato de classes que eles serão treinados para realizar a classificação. Entretanto, ao utilizar um modelo de RNC, ou RNA, que aprende a condensar a informação, em outras palavras: redução de dimensionalidade, é vantajoso pois não há necessidade de realizar o treinamento tendo que saber, exatamente, quantas classes de objetos devem ser classificados. A classificação é feita por comparação, através de uma função de distância. Para este trabalho utilizou-se da distância cossenoidal, expressa na Equação 3.10.

Como a função $\cos(x)$ é limitada no intervalo real $[-1, 1]$, os vetores descritores $\mathbf{x}$ e $\mathbf{y}$ serão similares, *i.e.* pertencem ao mesmo indivíduo, se a distância cossenoidal for próxima de 1.

3.3 Hardware utilizado

O hardware utilizado para fins de treinamento, e avaliação, do modelo de rede neural apresentado foram:

- GPU GeForce RTX 3060 com 6Gb de VRAM;
- Processador Intel i7-11800H de 16 threads e *clock* de 2.30 GHz;
- Memória RAM 16Gb com DDR5;
- Sistema Operacional Ubuntu 20.04 LTS.

4 Resultados e Discussão

Neste capítulo é descrito como foi feita a avaliação do modelo de RNC proposto e comparam-se seus resultados com trabalhos correlatos.

4.1 Métrica de Avaliação

A métrica utilizada para realizar a identificação do indivíduo foi a distância cossenoidal, descrita na Subseção 3.2.5.

Para um mesmo indivíduo, é computado seu descritor médio, levando em consideração todo o conjunto sequential de frames, tomados 40 à 40. Ou seja, se existem 400 imagens no vídeo, foram computados 10 vetores descritores para cada intervalo de 40 imagens e, ao final, é feito a média aritmética deles como na equação 4.1:

$$\hat{\mathbf{x}} = \frac{1}{K} \sum_{k=0}^{K} \mathbf{x}_k, \tag{4.1}$$

onde $K = \textit{floor}\left[\frac{N}{40}\right]$, em que N é o número de imagens presentes no vídeo e *floor* é a função que arredonda a divisão para o inteiro inferior.

Estes vetores descritores $\hat{\mathbf{x}}$ são armazenados em um arquivo para ser utilizado no processo de checagem de acurácia.

Para realizar a classificação, é computado o vetor descritor da sequência em análise, de 40 imagens sequenciais, e calcula-se a distância cossenoidal entre este vetor e todos os presentes do banco de dados. Seleciona-se, então, o resultado cujo valor foi mais próximo de 1. Este processo também é denominado na literatura como *Top-1* (WOJKE; BEWLEY, 2018).

4.2 Resultados Experimentais

A avaliação da RNC proposta foi feita de duas formas. A primeira consiste em avaliar a acurácia da RNC no mesmo *dataset* em que foi treinada. A segunda consiste em avaliar a acurácia da RNC no *dataset* oposto ao seu treinamento.

Os resultados deste trabalho e de trabalhos de base, de Jangua e Marana (2020) e de Lima e Schwartz (2019) são apresentados nas Tabelas 2 e 3.

Tabela 2 – Resultado e comparativo no conjunto de dados CASIA GAIT-A em *Top-1*.

Método	**Acurácia**
Wang et al. (2003)	88.75%
Liu et al. (2016)	89.17%
Lima e Schwartz (2019)	95.42%
Jangua e Marana (2020) com distância Euclidiana	87.92%
Jangua e Marana (2020) com distância Chi-quadrado	91.67%
Método Proposto Treinado no Conjunto GAIT-A	90.00%
Método Proposto Treinado no Conjunto GAIT-B	85.00%

Fonte: Elaborada pelo autor.

Tabela 3 – Resultado e comparativo no conjunto de dados CASIA GAIT-B em *Top-1*.

Método	**Acurácia**
Yu et al. (2007)	83.50%
Chen et al. (2009)	91.10%
Lima e Schwartz (2019)	98.00%
Jangua e Marana (2020) com distância Euclidiana	91.26%
Jangua e Marana (2020) com distância Chi-quadrado	94.22%
Método Proposto Treinado no Conjunto GAIT-A	56.45%
Método Proposto Treinado no Conjunto GAIT-B	86.10%

Fonte: Elaborada pelo autor.

Na Tabela 2, o método proposto neste trabalho de pesquisa apresentou uma métrica de acurácia competitiva, quando comparada a outros métodos. Alguns dados da Tabela 2 foram obtidos de Jangua e Marana (2020), foram utilizados os valores de acurácia média, uma vez que os autores calcularam a acurácia para as três vistas individualmente.

É interessante ressaltar que o método proposto também obteve resultados competitivos no conjunto de dados GAIT-B, exceto quando a RNC foi treinada no GAIT-A e avaliada no GAIT-B. Acredita-se que isso se deve ao fato de haver poucos indivíduos no conjunto GAIT-A, que são 20, enquanto que no conjunto GAIT-B são 124. Este argumento é reforçado dado que, na Tabela 3, o método proposto, quando treinado no GAIT-B e avaliado em GAIT-A obteve resultados similares e, também, competitivos.

No conjunto de dados GAIT-B também está disponível uma versão dos dados em que, os mesmos indivíduos, estão utilizando casacos, de forma a cobrirem parte do corpo. A Tabela 4 mostra os resultados de Lima e Schwartz (2019), de Jangua e Marana (2020) e do método proposto nesse trabalho. Percebe-se que a diferença de acurácia é baixa em relação ao conjunto de dados sem casaco, quando comparado com os métodos de Lima e Schwartz (2019) e Jangua e Marana (2020).

Tabela 4 – Resultado e comparativo no conjunto de dados CASIA GAIT-B com Casaco *Top-1*.

Método	**Acurácia**
Lima e Schwartz (2019)	95.16%
Jangua e Marana (2020) com distância Euclidiana	86.29%
Jangua e Marana (2020) com distância Chi-quadrado	89.72%
Método Proposto Treinado no Conjunto GAIT-A	56.15%
Método Proposto Treinado no Conjunto GAIT-B	85.80%

Fonte: Elaborada pelo autor.

5 Conclusão

Foi apresentado o diferencial deste trabalho, que é a construção do conjunto de dados de poses em três dimensões através de algoritmos de fotogrametria e estimação de estruturas, com a finalidade de reconhecer pessoas pelo padrão de caminhada. Este trabalho também contribuiu com uma expansão do conjunto de dados CASIA Gait B, onde outros pesquisadores poderão utilizá-lo e, também, propor novas metodologias para melhorar a acurácia e eficiência de algoritmos de reconhecimento através de poses humanas e seu padrão de caminhada.

Após o desenvolvimento e análise dos resultados apresentados, constatou-se que o método proposto não demonstrou uma vantagem significativa durante os testes com os conjuntos de dados CASIA Gait A e CASIA Gait B. No entanto, é importante destacar que os métodos anteriores, utilizados para a comparação dos resultados, baseiam-se exclusivamente em poses 2D, ao contrário do enfoque adotado neste trabalho, que se concentra em poses 3D.

A distinção entre esses métodos é crucial, uma vez que as poses 2D exigem dados de maior qualidade, nos quais a oclusão do indivíduo, por exemplo, não é tolerada. Assim, é fundamental ressaltar que, ao empregar um conjunto de dados que contemple oclusões, simulando, por exemplo, um ambiente de vigilância com câmeras, espera-se uma melhoria na acurácia e robustez ao utilizar o método proposto neste trabalho, quando comparado com métodos baseados em poses 2D. Isso se deve à utilização de dados provenientes de diversas câmeras, mitigando o problema de oclusões inerente ao uso de uma única câmera.

Outro aspecto no qual o presente método se destaca em relação aos trabalhos anteriores (não abordado apenas pela análise de acurácia) é a invariância em relação ao ângulo da câmera e à orientação do indivíduo a ser reconhecido. Ao empregar poses 2D, essa representação ocorre no quadro de referência da imagem. Dessa forma, se o banco de dados apresentar apenas imagens do indivíduo lateralmente, torna-se mais desafiador reconhecê-lo quando se utiliza uma imagem frontal, o que pode resultar em uma diminuição na acurácia nesse tipo de abordagem. No entanto, ao adotar poses 3D representadas no quadro de referência do mundo, esse problema é mitigado.

Por fim, como melhorias para trabalhos futuros, pode-se citar a inclusão de novas técnicas de aumento de dados, a ampliação do conjunto de dados e teste de outras arquiteturas de RNA. Além disso, também vale salientar o teste de outros algoritmos utilizando o conjunto de poses em três dimensões, dado que os trabalhos de comparação utilizaram poses em duas dimensões.

Não foi explorado, mas também é possível fazer a análise da acurácia do modelo levando em consideração o segundo melhor candidato, terceiro melhor candidato e assim por diante. Como foi descrito, utilizou-se apenas o de menor distância (*rank-1*).

Referências

ALANSARI, M.; HAY, O. A.; JAVED, S.; SHOUFAN, A.; ZWEIRI, Y.; WERGHI, N. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, v. 11, p. 35429–35446, 2023.

ALI, M. M.; MAHALE, V. H.; YANNAWAR, P.; GAIKWAD, A. T. Overview of fingerprint recognition system. In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. [S.l.: s.n.], 2016. p. 1334–1338.

ARANDJELOVIć, R.; GRONAT, P.; TORII, A.; PAJDLA, T.; SIVIC, J. *NetVLAD: CNN architecture for weakly supervised place recognition*. 2016.

BIANCO, S.; CIOCCA, G.; MARELLI, D. Evaluating the performance of structure from motion pipelines. *Journal of Imaging*, v. 4, n. 8, 2018. ISSN 2313-433X. Disponível em: <https://www.mdpi.com/2313-433X/4/8/98>.

BRADSKI, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y. A. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Center for Biometrics and Security Research. *Chinese Academy of Sciences (CASIA)*. 2010. Acesso em: 10 de Outubro, 2023. Disponível em: <http://www.cbsr.ia.ac.cn/english/index.asp>.

CHEN, C.; LIANG, J.; ZHAO, H.; HU, H.; TIAN, J. Frame difference energy image for gait recognition with incomplete silhouettes. In: *Pattern Recognition Letters*. [S.l.: s.n.], 2009.

COLMAP. *Tutorial - COLMAP*. 2023. Acesso em: 3 de Outubro, 2023. Disponível em: <https://colmap.github.io/tutorial.html>.

GAO, X.; ZHANG, T. *Introduction to Visual SLAM*. 1st. ed. Singapura: Springer, 2021. ISBN 9811649383.

HARTLEY, R.; ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. 2nd. ed. England: Cambridge, 2004. ISBN 052154018.

HAYKIN, S. *Neural Networks and Learning Machines*. 3rd. ed. Hamilton, Ontario, Canada: Pearson, 2009. ISBN 978-0-13-147139-9.

HOFFER, E.; AILON, N. *Deep metric learning using Triplet network*. 2018.

HUANG, G. B.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. [S.l.], 2007.

ISLAM, K. *Deep Learning for Video-based Person Re-Identification: A Survey*. 2023.

JANGUA, D.; MARANA, A. A new method for gait recognition using 2d poses. In: *Anais do XVI Workshop de Visão Computacional*. Porto Alegre, RS, Brasil: SBC, 2020. p. 69–74. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/wvc/article/view/13483>.

KAUR, R.; HIMANSHI, E. Face recognition using principal component analysis. In: *2015 IEEE International Advance Computing Conference (IACC)*. [S.l.: s.n.], 2015. p. 585–589.

KREISS, S.; BERTONI, L.; ALAHI, A. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, p. 1–14, March 2021.

LECUN LéON BOTTOU, Y. B. Y.; HAFFNER, P. Gradient-based learning applied to document recognition. In: IEEE. [S.l.], 1998.

LIMA, V. de; SCHWARTZ, R. Gait recognition using pose estimation and signal processing. In: *Iberoamerican on Pattern Ecognition*. [S.l.]: CIARP, 2019. ISSN 0000-0000.

LIU, D.; YE, M.; X., Z.; LIN, L. Memory-based gait recognition. *British Machine Vision Conference (BMVC 2016)*, v. 12, n. 9, p. 82.1–82.12, 2016.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. In: *Internal Journal of Computer Vision*. [S.l.]: University of British Columbia, 2004. p. 91–110.

MASI, I.; WU, Y.; HASSNER, T.; NATARAJAN, P. Deep face recognition: A survey. In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2018. p. 471–478.

NIST. *Face Recognition Grande Challenge (FRGC)*. 2021. Acesso em: 3 de Outubro, 2023. Disponível em: <https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc>.

RASHAD; SHAMS; NOMIR; AWADY, E. IRIS recognition based on LBP and combined LVQ classifier. *International Journal of Computer Science and Information Technology*, Academy and Industry Research Collaboration Center (AIRCC), v. 3, n. 5, p. 67–78, oct 2011. Disponível em: <https://doi.org/10.5121%2Fijcsit.2011.3506>.

Speech Ocean. *The history of face recognition and the technical process*. 2022. Acesso em: 3 de Outubro, 2023. Disponível em: <https://en.speechocean.com/Cy/487.html>.

TAIGMAN, Y.; YANG, M.; RANZATO, M.; WOLF, L. Deepface: Closing the gap to human-level performance in face verification. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1701–1708.

TANG, C.; TAN, P. *BA-Net: Dense Bundle Adjustment Network*. 2019.

Thales Group. *Biometrics: definition, use cases, latest news*. 2023. Acesso em: 10 de Outubro, 2023. Disponível em: <https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/biometrics>.

WANG, L.; TAN, T.; HU, W.; NING, H. Automatic gait recognition based on statistical shape analysis. *IEEE Transactions on Image Processing*, v. 12, n. 9, p. 1120–1131, 2003.

WOJKE, N.; BEWLEY, A. Deep cosine metric learning for person re-identification. In: . [S.l.: s.n.], 2018.

YU, S.; TAN, D.; HUANG, K.; TAN, T. Reducing the effect of noise on human contour in gait recognition. In: *International Conference on Biometrics*. [S.l.: s.n.], 2007.