



Maria Luisa Leandro de Lima

**UM ESTUDO DE RECONHECIMENTO DE EMOÇÕES BASEADO
EM LINGUAGEM CORPORAL E MARCHA**

Trabalho de Graduação



Universidade Federal de Pernambuco

secgrad@cin.ufpe.br

www.portal.cin.ufpe.br/graduacao

RECIFE

2023



Universidade Federal de Pernambuco
Centro de Informática
Graduação em Engenharia da Computação

Maria Luisa Leandro de Lima

**UM ESTUDO DE RECONHECIMENTO DE EMOÇÕES BASEADO
EM LINGUAGEM CORPORAL E MARCHA**

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientadora: Veronica Teichrieb

Co-orientador: Willams de Lima Costa

RECIFE

2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Lima, Maria Luisa Leandro de.

Um estudo de reconhecimento de emoções baseado em linguagem corporal e marcha / Maria Luisa Leandro de Lima. - Recife, 2023.

36 p. : il., tab.

Orientador(a): Veronica Teichrieb

Cooorientador(a): Willams de Lima Costa

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Engenharia da Computação - Bacharelado, 2023.

1. reconhecimento de emoções. 2. marcha. 3. linguagem corporal. I. Teichrieb, Veronica. (Orientação). II. Costa, Willams de Lima. (Cooorientação). IV. Título.

600 CDD (22.ed.)

Trabalho de graduação apresentado por **Maria Luisa Leandro de Lima** ao programa de Graduação em Engenharia da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título **Um estudo de reconhecimento de emoções baseado em linguagem corporal e marcha**, orientada pela **Profa. Veronica Teichrieb**, co-orientada pelo **MsC.Willams de Lima Costa** e aprovada pela banca examinadora formada pelos professores:

Prof. Cleber Zanchettin
Centro de Informática/UFPE

Profa. Veronica Teichrieb
Centro de Informática/UFPE

RECIFE

2023

*Dedico este trabalho aos meus pais, minhas amigas e a
todos que me incentivaram a chegar aqui.*

Agradecimentos

Agradeço aos meus pais, Paulo e Joseane, por todo o seu esforço e dedicação à minha educação, além de todo o apoio que me dão.

Agradeço também às minhas amigas Albertyna, Amanda, Eneri, Lara, Maria Eduarda e Rafaella. Vocês são a minha rocha.

Agradeço ao Voxar Labs por toda a oportunidade de crescimento profissional e acadêmico, que começou pelo estágio e que se concretiza ainda mais aqui, neste Trabalho de Graduação. Em especial, agradeço Will, que sempre acreditou no meu trabalho e no meu potencial, e Rafaela, minha gerente e amiga que me orienta nos caminhos profissionais. Agradeço também a Maria e VT, pois sem elas não seria possível ter chegado até aqui.

Por fim, agradeço a Bubu, minha companheira de todas as horas, minha cachorrinha que sempre me ajudou quando o estresse ameaçava me tomar.

Peço perdão caso tenha esquecido algum nome, mas suas contribuições estão no meu coração.

*Emotion pulls the levers of our lives, whether it be by the song in our heart,
or the curiosity that drives our scientific inquiry.*

—ROSALIND PICARD

Resumo

O reconhecimento de emoções é a tarefa de extrair características relacionadas a emoções, permitindo que sistemas inteligentes entendam seu usuário de forma mais profunda e que mudem sua interação com o mesmo de acordo. Dentre as técnicas de reconhecimento de emoções mais comuns está a combinação de reconhecimento de expressão facial e contexto. No entanto, como sugerido pela literatura de psicologia de comportamento, a linguagem corporal também é uma importante fonte de características afetivas e poderia ser utilizada para o reconhecimento de emoções. Neste trabalho, propomos um *framework* profundo para reconhecimento de emoções baseado em marcha, que é a forma de andar de uma pessoa e que é bastante modificada de acordo com as emoções da mesma. Realizamos experimentos e avaliações no conhecido *dataset* E-Gait, passando o estado da arte com um ganho de $\approx 5\%$ em acurácia. Também mostramos que nosso modelo converge mais rápido, reduzindo os custos computacionais necessários para realizar possíveis refinamentos desta técnica para resolver problemas mais específicos em outros contextos.

Palavras-chave: reconhecimento de emoções, marcha, linguagem corporal

Abstract

Emotion recognition is the task of extracting features related to emotion, allowing intelligent systems to understand their user on a deeper level and changing their interaction accordingly. Among the more common emotion recognition approaches is the combination of facial expression and context recognition. However, as suggested by the behavioral psychology literature, body language is also a significant source of affective features and could be used for emotion recognition. In this work, we propose a deep framework for emotion recognition based on gait, which is a person's manner of walking and which is severely modified according to their emotions. We experimented and evaluated our technique in the well-known E-Gait dataset, surpassing the current state-of-the-art with a gain of $\approx 5\%$ in accuracy. We also show that our model is able to converge faster, therefore lowering the computational resources required to fine-tune this approach to solve more specific problems in different settings.

Keywords: emotion recognition, gait, body language

Lista de Figuras

2.1	Arquitetura do STEP. Fonte: BHATTACHARYA et al. (2020)	19
3.1	Arquitetura proposta para uma ferramenta de inferência de emoções que utiliza o Spatial Temporal Emotion Perception with ST-GCN++ (STEP++) a partir de uma janela de <i>frames</i> de vídeos. A ferramenta extrai os esqueletos, realizando a conversão de formato para o adequado ao Spatial Temporal Emotion Perception (STEP). Os esqueletos são então processados pelo STEP++, que realiza a classificação da emoção presente na janela de <i>frames</i> enviada para o mesmo. . .	21
3.2	Arquitetura proposta para o STEP++, composta de 3 blocos Spatio-Temporal Graph Convolutional Networks++ (ST-GCN++) com saídas de tamanho 32, 64 e 64, seguidos de um <i>Average Pooling 2D</i> e de uma <i>Convolução 2D pointwise</i> , que reduz a dimensionalidade de 64 para 4. Finalmente, tem-se a função <i>Softmax</i> como saída da saída da rede, a fim de realizar a classificação das 4 emoções descritas neste trabalho.	24
4.1	Juntas e composição de esqueletos do Emotion-Gait (E-Gait)	27
4.2	Exemplos de esqueletos do dataset E-Gait de cada uma das emoções presentes no mesmo: Angry, Happy, Neutral e Sad	27
5.1	Matrizes de confusão da inferência no conjunto de teste do STEP e STEP++ . .	30
5.2	Exemplos de inferências do STEP e STEP++ no conjunto de teste e os valores verdadeiros das amostras.	31

Lista de Tabelas

4.1	Espaço de busca definido para a busca <i>bayesiana</i>	28
5.1	Parâmetros de treinamento do STEP.	29
5.2	Parâmetros de treinamento do STEP++.	30
5.3	Valores de acurácia do STEP e STEP++	30

Lista de Acrônimos

CNN	Convolutional Neural Network	18
COCO	Common Objects in COntext	22
E-Gait	Emotion-Gait	26
ELMD	Edinburgh Locomotion Mocap Database	26
GCN	Graph Convolutional Network	18
HRNet	High Resolution Network	22
LSTM	Long Short-Term Memory	20
STEP	Spatial Temporal Emotion Perception	19
STEP++	Spatial Temporal Emotion Perception with ST-GCN++	21
ST-GCN	Spatio-Temporal Graph Convolutional Networks	18
ST-GCN++	Spatio-Temporal Graph Convolutional Networks++	19
PySKL	Skeleton data with PyTorch	22
TCN	Temporal Convolutional Network	23
WandB	Weights & Biases	25

Sumário

1	Introdução	13
2	Revisão da Literatura	15
2.1	Aspectos Psicológicos	15
2.2	Técnicas	18
2.2.1	Análise espaçotemporal de juntas do corpo	18
2.2.2	Classificação de emoções através de marcha	19
3	Método	21
3.1	Extração de pose	22
3.2	Análise no aspecto espaçotemporal	22
3.3	Classificação de emoções através de marcha	24
3.4	Busca <i>bayesiana</i>	25
4	Experimentos	26
5	Resultados	29
5.1	Discussão	32
6	Conclusão	34
	Referências	35

1

Introdução

Dentre o vasto potencial de aplicação de técnicas de reconhecimento de comportamento humano, o reconhecimento de emoções é uma das tarefas de grande interesse hoje por cientistas e pesquisadores. Pesquisadores da área de psicologia do comportamento vêm tentando conectar o comportamento visível com o pensamento e emoção que é inerente à pessoa desde Darwin ([DARWIN, 1872](#)) através de comunicação não verbal.

Reconhecer emoções permite que sistemas inteligentes possam, por exemplo, entender seu usuário e adaptar sua forma de comunicação e interação a ele, ao levar em consideração o fator humano. Sendo, desta forma, criados sistemas interativos intuitivos para as pessoas, que devem ser capazes de inteagir com a tecnologia da forma mais natural possível ([VALLI, 2005](#)), permitindo o desenvolvimento de várias tecnologias no campo de cidades inteligentes, saúde e bem estar, segurança, dentre outros.

A computação afetiva traz a inferência de características afetivas, utilizando as chamadas características afetivas como expressão facial, olhar, gestos, movimentos corporais, expressão vocal e indicadores fisiológicos como frequência cardíaca ([KLEINSMITH; BIANCHI-BERTHOUE, 2012](#); [TAO; TAN, 2005](#)). As características são extraídas de diversas formas mas um problema comum é que os usuários geralmente estão presos na frente do computador, e neste tipo de restrição, fica-se limitado a um usuário interagindo com um totem, ou com um computador no espaço, por exemplo.

Dado que a marcha também contribui com informações sobre o estado afetivo de um indivíduo e como podemos extrair essa informação sem adicionar limites à movimentação do usuário, permitindo, por exemplo, cenários de *smart cities* onde os cidadãos se movimentam livremente pelas cidades e podem ser monitorados, para fins de segurança pública ou coleta de informações relevantes sobre o bem estar do cidadão para fins de urbanismo, por exemplo, sem interferência em seu deslocamento. Ainda, pode-se utilizar ou em cenários de acompanhamento

clínico em hospitais onde é necessário realizar o acompanhamento psicológico de pacientes internados, sem aumentar seu estresse.

Neste trabalho, propomos uma abordagem para a tarefa de reconhecimento de emoções através de marcha. Através do uso de blocos de processamento espaçotemporal, superamos limitações presentes no atual estado da arte que possui uma capacidade limitada de processamento neste aspecto. Mostramos também que ao superar esta limitação, nosso resultado na métrica quantitativa proposta também foi melhor do que o atual estado da arte. Desta forma, nossas contribuições neste trabalho foram:

- Um método de reconhecimento de emoções através de marcha e linguagem corporal, compatível com as descobertas da área de psicologia do comportamento, com um aumento de $\approx 5\%$ de acurácia em relação ao atual estado da arte.
- Um modelo com convergência mais rápida no treinamento, gerando economia de tempo e recursos computacionais, que permite implantação e validação de experimentos com maior velocidade.

2

Revisão da Literatura

Este capítulo realiza uma revisão da literatura sobre o reconhecimento de emoções através de linguagem corporal, e é estruturado da seguinte forma: primeiro, discutimos a revisão da literatura a partir de uma perspectiva da psicologia do comportamento, onde analisamos o impacto bidirecional entre emoções e linguagem corporal. Após essa análise, examinaremos as técnicas de ponta na área, destacando suas limitações e comparando-as com nossa proposta.

2.1 Aspectos Psicológicos

É fato que as pessoas são intuitivamente perceptivas, observando comportamentos umas das outras a fim de identificar emoções como um fator a mais no processo de realização de decisões do dia a dia. Esta habilidade é fundamental para o desenvolvimento de relações sociais. Esta perceptibilidade chamou a atenção de pesquisadores, que passaram a estudar os meios pelos quais as pessoas conseguiam realizar estas codificações, uma vez que não poderia ser apenas algum *sexto sentido*, mas algo que poderia ser investigado cientificamente (PICARD, 1997; COSTA et al., 2022).

Foi-se descobrindo que existem expressões de emoções que, à despeito da vontade e esforço do indivíduo, são expressadas através de seu corpo. Logo, a partir de elementos como tom de voz, expressão facial, posicionamento do corpo e jeito de andar, entre outros, as pessoas codificam as emoções de outrem para entendê-lo. Um exemplo de expressão não voluntária de emoção é uma pessoa que apesar de apresentar uma face neutra, se encontra tremendo e com as mãos frias, expressando seu nervosismo involuntariamente. É aceito, portanto, pelos pesquisadores, um componente fisiológico nas emoções, de forma que o corpo passa a agir como um mensageiro do estado emocional do sujeito. A detecção de estados emocionais, portanto, se torna possível através das medidas de características observáveis, os comportamentos corporais, destes estados.

Tais medidas podem, de fato, levar ao reconhecimento das expressões tanto voluntárias como involuntárias de emoções (PICARD, 1997).

Vários estudos apontam a importância da expressão corporal das emoções em seu reconhecimento por pessoas, tornando-se, em alguns casos, melhor do que a expressão facial, por exemplo. Mostra-se, inclusive, que as mudanças emocionais que são refletidas na postura corporal também auxiliam o indivíduo a lidar com as emoções sentidas. Ao observar a postura corporal é possível, inclusive, inferir a ação que será tomada por uma pessoa com medo, por exemplo, ao contrário do que seria possível com o uso das expressões faciais, que apenas indicariam que há uma ameaça (KLEINSMITH; BIANCHI-BERTHOUE, 2012).

Um trabalho de altíssimo impacto nesta investigação é o de Darwin, que investiga a relação de movimentos e expressões faciais na comunicação de emoção. Darwin determinou que existem comportamentos, que são espontâneos e involuntários, associados à expressão de emoções específicas, como sequências de posturas (WALLBOTT, 1998; COSTA et al., 2022). Alguns exemplos são a associação de movimentos *sem propósito*, pulos, pisadas fortes, corpo ereto e cabeça elevada para a alegria; Movimentação passiva e lenta com a cabeça baixa para a tristeza; Corpo trêmulo, movimentos *sem propósito* e frenéticos, cabeça elevada, peito estufado, entre outros, para a raiva. Darwin também determinou que os movimentos corporais revelavam não apenas a quantidade (intensidade) de emoções mas também sua qualidade (tipo) (WALLBOTT, 1998). Isto foi confirmado pelo estudo de WALLBOTT (1998).

Um estudo de 1993 realizado por CAMRAS; SULLIVAN; MICHEL (1993) encontrou que a expressão corporal realizada por pessoas em desconforto ou com raiva era considerada mais frenética e ativa quando comparada com tristeza. Também há o relato de que algumas emoções podem gerar tensões musculares que acabam por sua vez influenciando comportamentos motores e posturas. Inclusive, postura e atividade corporal têm sido estudadas também em pacientes depressivos, revelando baixa movimentação corporal e postura não ereta. Outro estudo, realizado por BULL (1978) em 1978 encontrou associação entre poses corporais e emoções, como a baixa da cabeça durante a tristeza (WALLBOTT, 1998).

WALLBOTT (1998) encontrou que 66% das categorias definidas em seu estudo, de movimento e postura corporais, permitiam distinguir entre emoções. Também encontrou que, de fato,

a movimentação e postura do indivíduo está ligada às emoções estudadas.

[TRACY; MATSUMOTO \(2008\)](#) encontrou que mesmo indivíduos nascidos com cegueira congênita são capazes de expressar, posturalmente, orgulho e vergonha de forma coerente com indivíduos sem cegueira, indicando que tais expressões não verbais de emoções são inerentes às emoções em si, mesmo em indivíduos que nunca puderam aprender observando as expressões de outros indivíduos ao seu redor. Já [PICARD \(1997\)](#) traz que emoções como culpa e vergonha geram expressões corporais diferentes entre si, na postura e no modo de andar.

Através de gravações em vídeo, sem áudio, de cenas com atores representando emoções, [WALLBOTT; SCHERER \(1986\)](#) encontrou uma alta acurácia (63%) na classificação de emoções por pessoas colaboradoras do estudo, demonstrando a importância da linguagem corporal na classificação de emoções, sendo os principais fatores não verbais encontrados os gestos realizados com as mãos e orientação da cabeça. Foi encontrado, inclusive, que raiva e tristeza são as emoções mais diferentes entre si (dentre alegria, tristeza, raiva e surpresa). A tristeza, por exemplo, mostrou-se caracterizada por cabeça baixa e gestos com as mãos relativamente frequentes, com movimentos pouco expansivos e ativos. Além disso, foi encontrado que características como movimentos energéticos, ativos, rápidos e expansivos eram característicos de alegria. Demonstrando a diferenciação inerente das emoções através de dicas como linguagem corporal e marcha ([WALLBOTT; SCHERER, 1986](#)).

Ainda, os movimentos do corpo na a marcha são menos suscetíveis a serem controlados pelo sujeito, em situações onde podem estar sendo observados e desejam alterar seu comportamento, do que expressões faciais, além de poderem ser usados para detectar gênero, emoções, sentimentos e identidade. Além disso, na marcha, não só a postura é importante como a evolução de cada junta do corpo ao longo do tempo. Informações importantes como o balançar de mãos, movimentações da cabeça e dos pés traz informações que ajudam a identificar a emoção de um sujeito ([RANDHAVANE et al., 2019](#)). Desta forma, percebe-se que a marcha, foco deste estudo, pode ser utilizada por modelos para a classificação de emoções.

O uso de esqueletos, utilizado na análise de marcha, é muito realizado na literatura tanto por seu formato compacto, quando comparado com imagens, por exemplo, quanto por trazer informações robustas de variações de cena que não sofrem com mudanças de iluminação ([DUAN](#)

et al., 2022).

2.2 Técnicas

2.2.1 Análise espaçotemporal de juntas do corpo

O Graph Convolutional Network (GCN) foi desenvolvido para aplicar filtros convolucionais a dados em forma de grafos, no aspecto espacial apenas. Desenvolvido por [YAN; XIONG; LIN \(2018\)](#), o Spatio-Temporal Graph Convolutional Networks (ST-GCN) estende os GCNs, que focam nas arestas conectando os vértices representativos de pontos do corpo, abrangendo também o aspecto espaço temporal, conectando as juntas do corpo entre si através de arestas temporais, permitindo uma análise do movimento corporal e seus padrões ao longo do tempo. Desta forma, é permitida a compreensão do aspecto espacial em análises envolvendo esqueletos, uma vez que métodos anteriores utilizavam vetores do esqueleto com as coordenadas das juntas a cada ponto no tempo para realizar análises no tempo, ou criavam relações manualmente para realizar as análises no aspecto temporal ([YAN; XIONG; LIN, 2018](#)).

Com o ST-GCN, não apenas ocorre esta análise no espaço, como o seu comportamento ao longo do tempo está também influenciando nos resultados da rede, de forma automática. Isso é fundamental para aplicações que envolvem esqueletos, uma vez que um esqueleto é de fato um grafo, mas que apenas a pose de corpo do esqueleto não traz informações o suficiente. Desta forma, a análise temporal das mudanças ocorridas no esqueleto em um intervalo de tempo t se mostra benéfico ([BHATTACHARYA et al., 2020](#); [YAN; XIONG; LIN, 2018](#)).

[YAN; XIONG; LIN \(2018\)](#) considera no ST-GCN que não apenas as arestas espaciais indicam vizinhança entre os vértices do grafo, como também as arestas temporais. O *kernel* temporal indica a quantidade de vizinhos a serem considerados na operação de convolução. O ST-GCN também se tornou bastante popular pois, ao contrário de técnicas de análise de atividades baseadas em esqueleto que utilizavam Convolutional Neural Networks (CNNs), as GCNs são mais leves, devido à natureza dos dados, que são esparsos.

[DUAN et al. \(2022\)](#) percebeu que as técnicas que se baseavam em GCNs utilizavam vários métodos de pre-processamento diferentes e procurou desenvolver as melhores práticas para a

obtenção de melhor performance em um *benchmark*. Notou, então, que tais melhores práticas, como pré-processamento dos dados, *augmentation* espacial e temporal, e configuração de hiperparâmetros, influenciavam mais na performance do que uma arquitetura complicada. Neste processo, desenvolveu o Spatio-Temporal Graph Convolutional Networks++ (ST-GCN++), uma simplificação do ST-GCN, que conseguiu atingir uma performance comparável ao estado da arte.

2.2.2 Classificação de emoções através de marcha

Sendo a marcha uma sequência temporal de um esqueleto em movimento, BHATTACHARYA et al. (2020) trouxe uma técnica que identifica emoções baseada em marcha. Baseando-se no fato de que as dicas emocionais podem ser identificadas principalmente através de certos movimentos como a velocidade do balançar dos braços, do tamanho e velocidade da passada, entre outros, e no fato de que a marcha é um movimento periódico de uma sequência de movimentos de esqueleto em 3 dimensões, BHATTACHARYA et al. (2020) decidiu utilizar GCNs para a classificação de emoções. Estes movimentos descritos compõem o que é tido como marcha. A arquitetura do Spatial Temporal Emotion Perception (STEP) segundo o autor pode ser vista na figura 2.1.

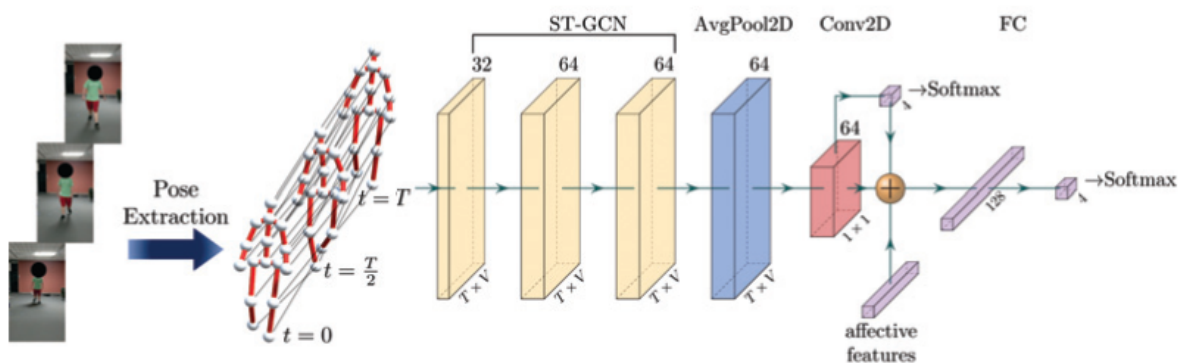


Figura 2.1: Arquitetura do STEP. Fonte: BHATTACHARYA et al. (2020)

Dessa forma, o STEP é definido por uma arquitetura de um classificador de emoções baseado em marcha composto de 3 blocos ST-GCN de 32, 64 e 64 *kernels* cada. Em seguida, a última camada do ST-GCN é passada por um *average pooling* e por uma convolução 2D de *kernel* 1×1 para finalmente ser seguido por um *Softmax* (BRIDLE, 1989) para as 4 categorias de emoções.

RANDHAVANE et al. (2019) chegou a se inspirar no STEP na utilização de marcha para a

detecção de comportamento enganoso, utilizando de uma rede profunda baseada em Long Short-Term Memory (LSTM), e a usar a rede STEP para a inferência em seu dataset DeceptiveWalk, obtendo uma acurácia de 83,68%, a fim de validação e comparação de sua técnica, uma vez que [RANDHAVANE et al. \(2019\)](#) foi o primeiro trabalho a tentar detectar comportamento enganoso a partir de marcha. [RANDHAVANE et al. \(2019\)](#) obteve uma acurácia de 72,56% em seu dataset com seu modelo proposto utilizando apenas a marcha.

3

Método

Dado um vídeo $V \in \mathbb{R}^{N \times H \times W \times 3}$ com N frames, altura H e largura W e um conjunto K de emoções, desejamos inferir, a partir da marcha de uma pessoa presente no vídeo, a sua emoção perceptível. Para isto, propomos um modelo que realiza análises espaçotemporais em cima de uma representação do esqueleto, um conjunto de juntas que é extraído através de técnicas de estimação de pose, entendendo sua movimentação ao longo do tempo. Dessa forma, as dicas não verbais como passada, balançar dos braços, angulação da cabeça com os ombros, entre outras, podem ser extraídas de forma automática e sem a intervenção do usuário. Nominalmente, nossa lista de K emoções é composta por *Happy*, *Neutral*, *Sad* e *Angry*, um subconjunto das emoções básicas de Ekman (EKMAN, 1992). Uma visão geral da nossa arquitetura, pode ser vista na Figura 3.1. Neste capítulo, iremos abordar o método proposto, com detalhes sobre sua implementação.

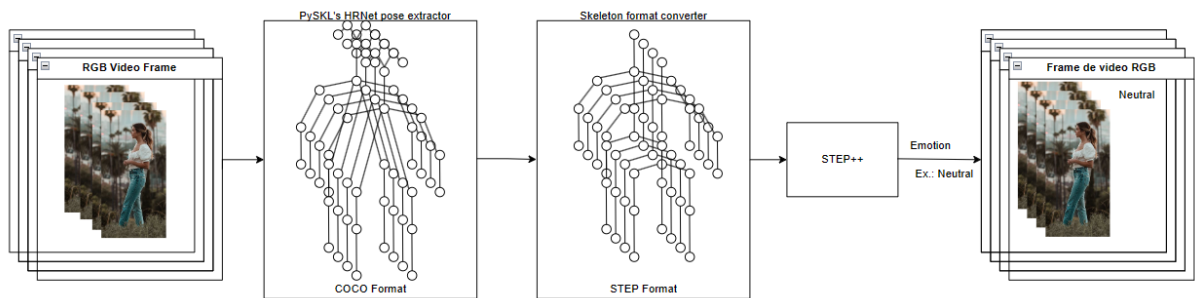


Figura 3.1: Arquitetura proposta para uma ferramenta de inferência de emoções que utiliza o Spatial Temporal Emotion Perception with ST-GCN++ (STEP++) a partir de uma janela de frames de vídeos. A ferramenta extrai os esqueletos, realizando a conversão de formato para o adequado ao STEP. Os esqueletos são então processados pelo STEP++, que realiza a classificação da emoção presente na janela de frames enviada para o mesmo.

3.1 Extração de pose

Vamos descrever nesta seção a metodologia utilizada para extrair os esqueletos de pose a partir de vídeos e utilizá-los na inferência, a fim de utilizar o STEP++, uma vez treinado. Como foi proposto o uso de uma extensão das GCNs para a análise espaçotemporal das informações do corpo da pessoa analisada, primeiro deve-se extrair a pose da pessoa centrada no vídeo, a fim de extrair padrões de movimento e comportamento que podem ser associados a emoções.

Dada uma sequência de quadros Q de um vídeo V , utilizamos a técnica HRNet (SUN et al., 2019) para extrair o conjunto de juntas que melhor descreve o corpo da pessoa visível na cena.

Buscando manter o máximo de praticidade e compatibilidade com a biblioteca Skeleton data with PyTorch (PySKL) disponibilizada por DUAN et al. (2022), foi escolhido o detector de pose High Resolution Network (HRNet), que é o detector *default* da ferramenta. As juntas que são reconhecidas pela HRNet seguem o formato de anotação do *dataset* Common Objects in Context (COCO)¹, que é um *dataset* em larga escala criado para detecção e segmentação de objetos, e que possui um formato próprio para anotação de keypoints do corpo humano. Porém, é necessário realizar uma adaptação deste conjunto de juntas para convertê-lo a um novo formato a fim de manter compatibilidade com o formato utilizado no treinamento do STEP, que, apesar de não haver uma discussão formal sobre o formato do esqueleto por BHATTACHARYA et al. (2020), entende-se que as juntas foram selecionadas de forma empírica a fim de priorizar informações mais comuns nas dicas não verbais.

3.2 Análise no aspecto espaçotemporal

Dado que a representação de uma pose de corpo pode haver um conjunto de juntas e um conjunto de segmentos que as conectam, nota-se a similaridade desta representação com a de um grafo. Com isto, o possível uso com GCNs se torna claro. Logo, definindo que um grafo pode ser representado por $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, sendo $\mathbb{V}$ o conjunto de vértices e $\mathbb{E}$ o conjunto de arestas que conectam os vértices e $\mathbb{N} = |\mathbb{V}|$ a quantidade de nós.

Com isto, uma GCN pode propagar os dados extraídos das juntas do corpo no formato

¹<https://cocodataset.org/>

$X \in \mathbb{R}^{N \times F}$ onde cada linha $x_i \in \mathbb{R}^F$ representa uma *feature* do vértice i . Assim, a propagação é realizada de forma que $\mathbb{Z}^{(l+1)} = \sigma(\mathbb{A}\mathbb{Z}^{(l)}\mathbb{W}^{(l)})$, onde $\mathbb{Z}^l$ e $\mathbb{Z}^{l+1}$ são entradas para as camadas l e $l+1$ da rede, sendo $\mathbb{Z}^0 = \mathbb{X}$. $\mathbb{W}^l$ e $\mathbb{W}^{l+1}$ são as matrizes de peso entre as camadas l e $l+1$, $\mathbb{A}$ é a matriz de adjacência de $\mathbb{G}$ e $\sigma(\cdot)$ é uma função de ativação não linear (BHATTACHARYA et al., 2020). Cada matriz de peso $\mathbb{W}$ da rede representa um *kernel* convolucional e, assim como em CNNs, pode-se aplicar múltiplos *kernels* para se obter *features*. Por exemplo, com a aplicação de k *kernels* de dimensões $\mathbb{F} \times \mathbb{D}$ em uma entrada $\mathbb{X}$, a saída corresponde a um vetor de *features* de dimensão $N \times \mathbb{D} \times k$ (BHATTACHARYA et al., 2020). Tem-se também que $A_i^t \subseteq \mathbb{V}$ é o conjunto dos nós adjacentes a v_i^t no tempo t (BHATTACHARYA et al., 2020).

Este processo é realizado de forma automática por frameworks de desenvolvimento. Nesta proposta, utilizamos a arquitetura ST-GCN++ implementada utilizando a biblioteca PySKL, que é baseada em PyTorch, para realizar esta análise.

Topologias pré-definidas. Utilizamos, de acordo com a proposta do ST-GCN++, uma topologia de juntas que é pré-definida para inicializar as matrizes de coeficientes, que são atualizadas iterativamente com *gradient descent* durante o treinamento, sendo adicionada também uma conexão residual no módulo espacial. Isso é importante pois o ST-GCN trazia pesos aprendíveis para pesar os elementos das matrizes de coeficientes, aumentando o custo computacional do modelo.

Rede de convolução temporal. O ST-GCN++ traz uma Temporal Convolutional Network (TCN) composta de 6 ramos, sendo uma convolução 1×1 , um *Max Pooling* e 4 ramos de convoluções temporais $1D$ com um *kernel* de tamanho 3 e dilatações de 1 a 4. Isso é importante pois traz flexibilidade e paralelismo à computação das *features*, tornando o modelo mais leve.

Nesta proposta, utilizamos estes blocos para propor a seguinte arquitetura, baseada no STEP: Com a entrada de um conjunto de juntas de um esqueleto, aplicamos três blocos ST-GCN++, com saídas 32, 64 e 64. Logo após, aplicamos uma camada de *pooling*, para reduzir a espacialidade, adicionando invariância à translação e escolhendo características de maior importância. Finalmente, uma convolução *pointwise*, uma convolução com um *kernel* de 1×1 , sendo aplicada a cada elemento dos dados, é aplicada, seguido de um *Softmax* para alcançar o resultado final da classificação de emoções. Uma visão geral de um bloco do STEP++ proposto neste trabalho

pode ser visto na Figura 3.2.

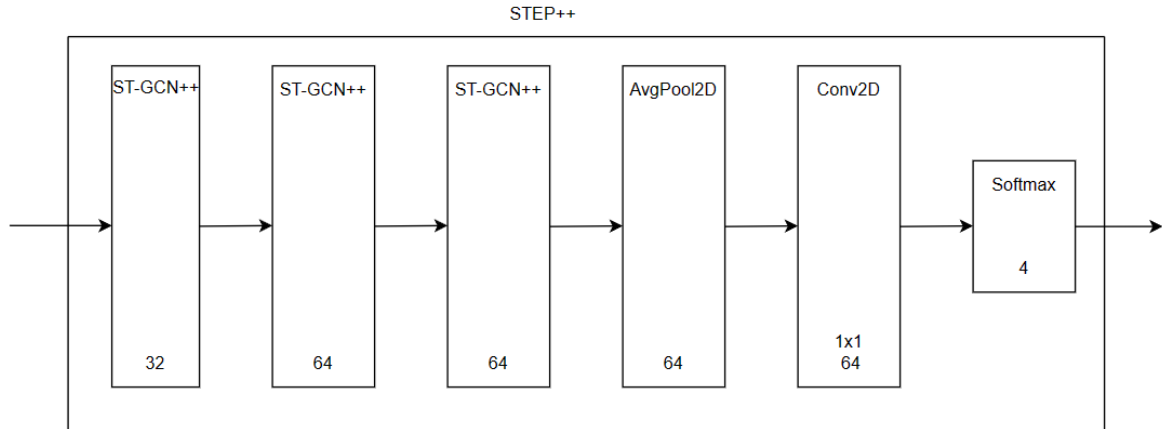


Figura 3.2: Arquitetura proposta para o STEP++, composta de 3 blocos ST-GCN++ com saídas de tamanho 32, 64 e 64, seguidos de um *Average Pooling 2D* e de uma *Convolução 2D pointwise*, que reduz a dimensionalidade de 64 para 4. Finalmente, tem-se a função *Softmax* como saída da rede, a fim de realizar a classificação das 4 emoções descritas neste trabalho.

3.3 Classificação de emoções através de marcha

Diferentes marchas $\mathbb{G}_x$ e $\mathbb{G}_y$ têm a mesma matriz de adjacência, dados T e V fixos, uma vez que podem ter conjuntos de vértices $\mathbb{V}_x$ e $\mathbb{V}_y$ diferentes mas não de arestas diferentes. Desta forma, pode-se manter uma noção de adjacência para as marchas do *dataset* e permite-se utilizar tanto do ST-GCN como do ST-GCN++ para processamento das marchas.

Além disso, seguindo a metodologia proposta por BHATTACHARYA et al. (2020), são calculadas as *affective features* para os dados utilizados. Estas *features* incluem informações sobre postura, como ângulo e distância entre juntas, área de diferentes áreas do corpo, entre outras características, além de informações de movimento, como velocidade e aceleração das juntas. Estas características são calculadas no carregamento dos dados, sendo agregadas aos mesmos antes da inserção na rede, conforme método disponibilizado por BHATTACHARYA et al. (2020) em seu repositório.

A partir do STEP, a arquitetura do STEP++ traz uma composição de 3 blocos ST-GCN++ de 32, 64 e 64 *kernels* cada. E, igualmente ao STEP, a última camada do ST-GCN++ é passada por um *average pooling* e por uma *convolução 2D* de *kernel 1x1* para finalmente ser seguido por um *Softmax* para as 4 categorias de emoções.

3.4 Busca bayesiana

Há casos onde há uma variada possibilidade de combinações de hiperparâmetros a serem utilizados para treinar um modelo. A busca *bayesiana* é utilizada, então, a fim de investigar a melhor combinação de tais hiperparâmetros, de forma automática, a fim de otimizar alguma métrica de performance. Este algoritmo funciona baseado no teorema de *bayes* de probabilidades condicionais, que diz que a probabilidade de se observar um evento Y dado que a evidência X aconteceu é igual à probabilidade da evidência X acontecer dado que o evento Y aconteceu multiplicada pela probabilidade do evento Y acontecer, divididos pela probabilidade da evidência X acontecer. Ou, $P(Y|X) = (P(X|Y) \times P(Y)) \div P(X)$, onde $P(.)$ representa a função de probabilidade.

No contexto da otimização de hiperparâmetros mencionada, X seria a métrica a ser otimizada, no caso deste trabalho, a acurácia do conjunto de validação, e Y seria a combinação de hiperâmetros sendo testada. A busca *bayesiana* cria um modelo probabilístico da função objetivo do treinamento do modelo, testando as diversas combinações definidas para o espaço de busca, buscando o maior valor para a métrica definida. O algoritmo utilizado foi o disponibilizado automaticamente na plataforma Weights & Biases (WandB) ² ao se configurar um *sweep*, ou busca.

²<https://wandb.ai/site/articles/bayesian-hyperparameter-optimization-a-primer>

4

Experimentos

Dataset. Utilizamos o *dataset* Emotion-Gait (E-Gait) ([BHATTACHARYA et al., 2020](#)) para realizar nossos experimentos. Este dataset é ideal pois contém dados reais de marcha que foram utilizados para treinar o modelo STEP, com o qual queremos comparar.¹

Mediante uma análise visual e algumas tentativas de treinamento utilizando todos os dados do E-Gait, entendemos que os dados sintéticos criados por [BHATTACHARYA et al. \(2020\)](#) não apresentam uma qualidade boa o suficiente para o treinamento, ainda mais por estarem em tão alta quantidade em relação aos dados reais, quando comparado com o *dataset* reportado no artigo. Por este motivo, decidimos utilizar apenas os dados reais neste momento.

Segundo [WALLBOTT; SCHERER \(1986\)](#), dados simulados de emoções apesar de não serem naturais ou espontâneos, evitando o dilema ético de infligir fortes emoções nos voluntários dos estudos, têm um grande valor na representação das emoções. Além disso, em cenários onde emoções ocorrem naturalmente, é extremamente difícil realizar a captura destes momentos chamados cândidos, uma vez que tais cenários não são controlados.

Dos dados reais, 342 amostras foram coletadas por [BHATTACHARYA et al. \(2020\)](#) e 1.835 foram tomadas do banco de dados Edinburgh Locomotion Mocap Database (ELMD) ([HABIBIE et al., 2017](#)). É importante notar que os dados do E-Gait consistem em esqueletos já extraídos de vídeos, que não foram disponibilizados, e já se encontram normalizados através do método [UMEYAMA \(1991\)](#).

Cada amostra possui o formato $T \times V$, onde T é o tempo que dura a amostra da marcha, e V é o número de coordenadas, que compõe 16 juntas, com 3 dimensões cada, sendo $V = 48$. Na figura 4.1 pode-se verificar tais juntas e a quais partes do esqueleto elas correspondem. Na figura

¹Segundo os autores do dataset, em uma *issue* em seu repositório, a versão do E-Gait disponível para o público hoje foi modificado após a publicação do artigo, e a versão original não está mais disponível. Anteriormente, o dataset consistia de 4.227 dados reais e 1.000 dados sintéticos, sendo modificado para 2.177 dados reais e 4.000 dados sintéticos. O *dataset* pode ser acessado clicando aqui.

ID	Junta	ID	Junta
0	Raiz	8	Cotovelo Direito
1	Coluna	9	Mão Direita
2	Pescoço	10	Quadril Esquerdo
3	Cabeça	11	Joelho Esquerdo
4	Ombro Esquerdo	12	Pé esquerdo
5	Cotovelo Esquerdo	13	Quadril Direito
6	Mão Esquerda	14	Joelho Direito
7	Ombro Direito	15	Pé Direito

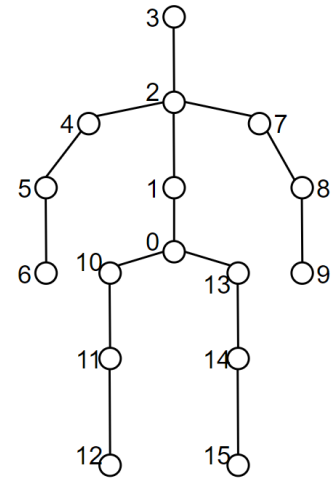


Figura 4.1: Juntas e composição de esqueletos do E-Gait

4.2 pode-se observar alguns exemplos de esqueletos do E-Gait e as emoções associadas a eles.

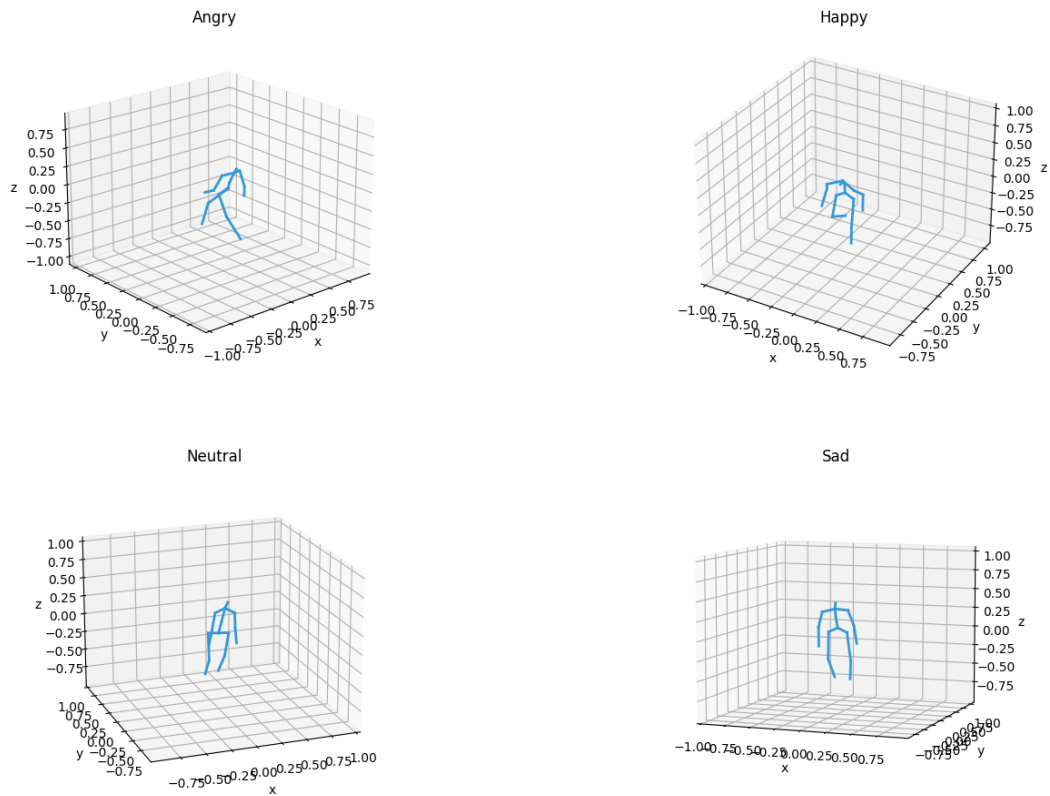


Figura 4.2: Exemplos de esqueletos do dataset E-Gait de cada uma das emoções presentes no mesmo: Angry, Happy, Neutral e Sad

Implementação. Implementamos o STEP++ utilizando o PyTorch na versão 1.7.0. Avaliemos nossa técnica na métrica acurácia no *dataset* E-Gait, sendo ela definida por $acuracia = (TP + TN) / TD$, onde TP , TN e TD são respectivamente os valores de verdadeiros positivos,

verdadeiros negativos e quantidade total de dados. A métrica é amplamente aceita e utilizada na literatura. Ela foi selecionada pois foi utilizada por BHATTACHARYA et al. (2020) em seu artigo e permite uma comparação direta entre seu método e o proposto nesse trabalho.

A fim de obter-se um novo *baseline*, uma vez que os resultados relatados por BHATTACHARYA et al. (2020) foram obtidos em uma versão modificada do *dataset* E-Gait, a técnica do STEP foi executada, seguindo os parâmetros descritos no artigo pelo autor, com os mesmos dados reais do E-Gait utilizados com a técnica STEP++ proposta neste trabalho.

Em seguida, a técnica STEP++ foi implementada a partir de uma adaptação do código disponibilizado do STEP por BHATTACHARYA et al. (2020). Foi realizada uma busca *bayesiana* à procura dos melhores parâmetros para esta implementação utilizando a ferramenta WandB e em seguida a a técnica foi treinada a partir dos dados reais do E-Gait. O espaço de busca definido para a busca *bayesiana* e os parâmetros escolhidos para o treinamento se encontram na tabela 4.1.

Tabela 4.1: Espaço de busca definido para a busca *bayesiana*

Parâmetro	Espaço de Busca	Parâmetro escolhido
Taxa de Aprendizagem Básica	0,01; 0,1; 0,3	0,01
Inicializador GCN	Importance; Offset	Offset
Otimizador	Adam; RMSProp; SGD	RMSProp
Weight Decay	0,0003; 0.0001	0,0003

Inferência. O modelo do STEP++ treinado foi utilizado para realizar algumas inferências em vídeos gratuitos *stock footage* da internet, a fim de validar o seu funcionamento. O *pipeline* seguido para a inferência foi o descrito na figura 3.1.

5

Resultados

O modelo STEP foi treinado de acordo com os parâmetros descritos pelo autor, que podem ser encontrados na tabela 5.1. O modelo STEP++ foi treinado de acordo com os melhores parâmetros encontrados na busca bayesiana. Os parâmetros para o STEP++ podem ser encontrados na tabela 5.2. Os resultados estão descritos na tabela 5.3 .

Tabela 5.1: Parâmetros de treinamento do STEP.

Parâmetro	Valores
Proporção treino/validação/teste	7:2:1
Número de Epochs	500
Tamanho do Batch	8
Otimizador	Adam
Taxa de Aprendizagem Básica	0,1
Decaimento da Taxa de Aprendizagem Básica	1/10 (após epochs 250, 375 e 438)
Momento	0,9
Weight Decay	5×10^{-4}

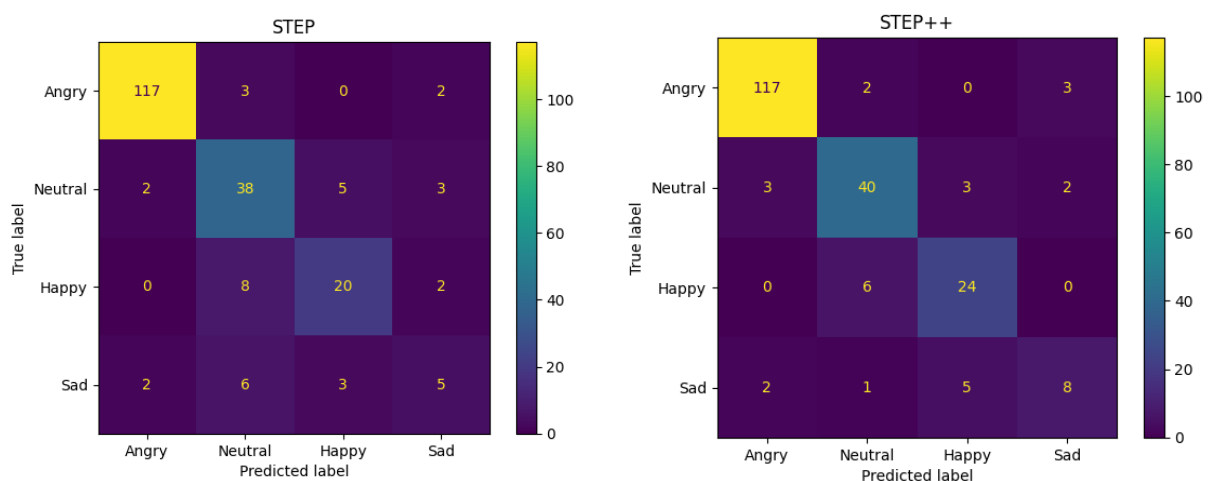
Tabela 5.2: Parâmetros de treinamento do STEP++.

Parâmetro	Valores
Proporção treino/validação/teste	7:2:1
Número de Epochs	200
Tamanho do Batch	8
Otimizador	RMSProp
Taxa de Aprendizagem Básica	0,01
Decaimento da Taxa de Aprendizagem Básica	1/10 (após epochs 250, 375 e 438)
Momento	0,9
Weight Decay	3×10^{-4}
Inicializador GCN	Offset

Tabela 5.3: Valores de acurácia do STEP e STEP++

Modelo	Acurácia de Teste	Acurácia de Validação	Acurácia de Treino	Epoch
STEP	83,33%	78,94%	82,96%	462
STEP++	87,5%	78,70%	91,84%	127

Pode-se visualizar na imagem 5.1 as matrizes de confusão dos modelos no conjunto de teste e alguns exemplos de inferências em esqueletos e seus resultados, ambos no STEP e STEP++ na imagem 5.2.

**Figura 5.1:** Matrizes de confusão da inferência no conjunto de teste do STEP e STEP++

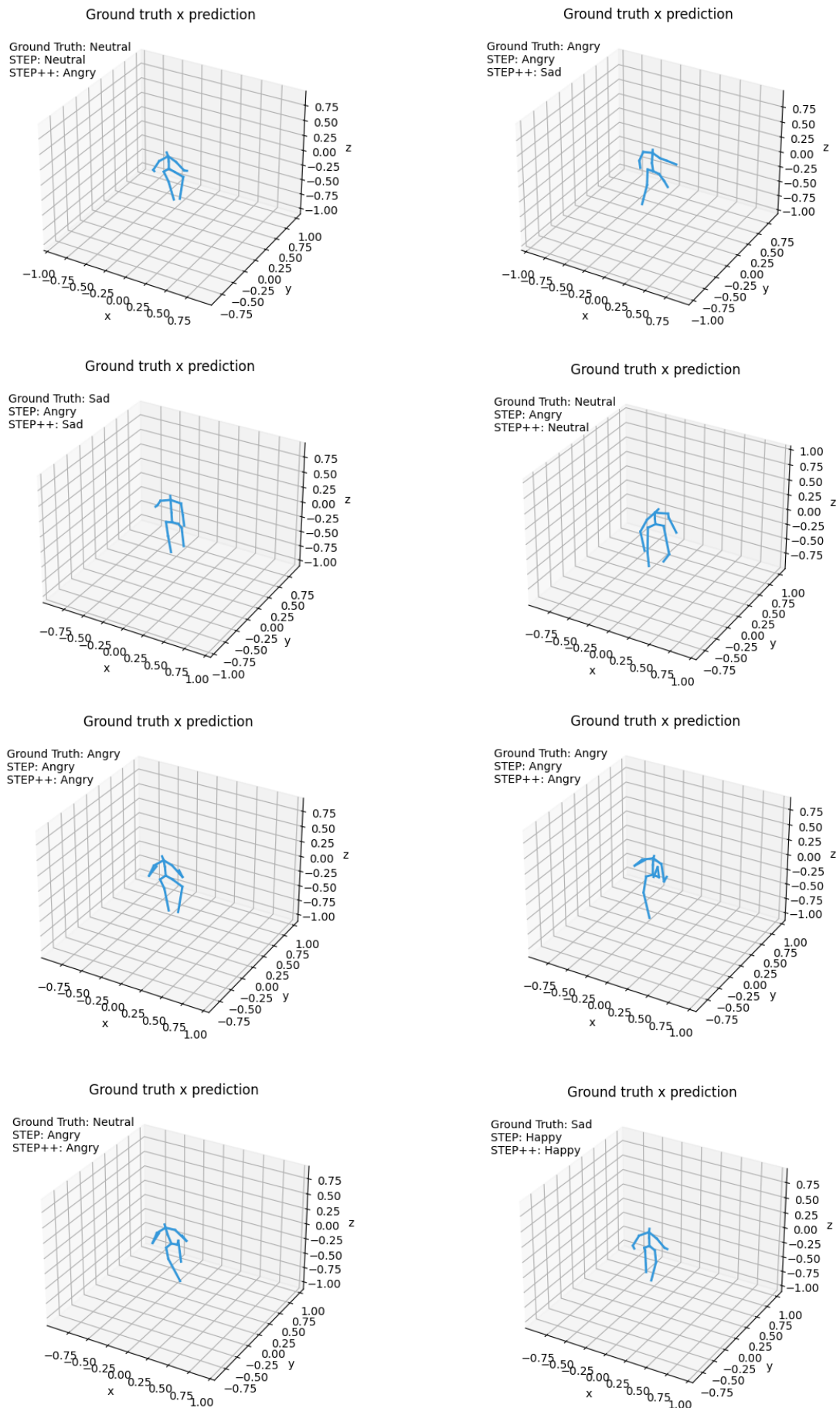


Figura 5.2: Exemplos de inferências do STEP e STEP++ no conjunto de teste e os valores verdadeiros das amostras.

5.1 Discussão

Performance, *overfitting* e ambiguidade. Objetivamente, foi obtida uma melhor performance, medida pela acurácia de teste, pelo modelo STEP++. Pode-se observar, portanto, que o modelo consegue realizar a inferência nos dados de teste, não vistos no treinamento, sendo capaz de discernir dentre as emoções.

No entanto, a partir dos valores de acurácia de treino e validação obtidos, também pode-se notar um maior *overfitting* do modelo STEP++, se comparado ao modelo STEP. Este maior *overfitting* pode se dar devido a quantidade de dados utilizados no treino, que pode se mostrar pequena para que o STEP++ aprenda sem *decorar* os dados. Ainda assim, a performance do STEP++ se mostra vantajosa.

Pode-se observar na imagem 5.1 que o modelo STEP++ possui uma diagonal mais pronunciada, mostrando sua melhor performance, confirmando o resultado da acurácia, como esperado. Observa-se também que, comparado com o STEP, o STEP++ apresenta uma melhor performance com emoções ambíguas, como *Happy* e *Neutral*. Há um maior acerto em ambas as emoções, mas também há um menor erro nas confusões entre elas, tanto entre verdadeiro *Happy* e inferido *Neutral* quanto com verdadeiro *Neutral* e inferido *Happy*. Nota-se, portanto, uma melhor performance no quesito de emoções ambíguas.

Convergência rápida. É notável a convergência mais rápida do STEP++ em comparação ao STEP, mesmo com uma taxa de aprendizagem menor, o que levaria o modelo a aprender mais lentamente. Esta convergência mais rápida é devida à simplificação da arquitetura do ST-GCN++, que apresenta uma topologia pré-definida na inicialização da matriz de coeficientes, além da paralelização do processamento dos dados através do TCN de vários ramos. Tal melhora no tempo de treinamento é propícia por dois motivos: Primeiramente, permite-se um treinamento que utiliza menos energia e requer menos recursos computacionais, criando mais oportunidade para pesquisadores com menos recursos conseguirem realizar suas pesquisas utilizando o STEP++. Também é criada a oportunidade de mais testes serem realizados com o STEP++ em menos tempo, permitindo que futuras pesquisas sejam realizadas de forma mais aprofundada e robusta.

Um ponto importante a se comentar é que, porque o STEP++ foi treinado apenas por 200

epochs, não é sabido, neste trabalho, se seria atingida uma maior acurácia posteriormente aos 200 *epochs*. No entanto, para este trabalho, pelo valor de acurácia de teste ser superior ao obtido com o STEP, este número de *epochs* será considerado o suficiente.

Dataset. Através da observação dos resultados e da visualização dos esqueletos com suas respectivas *labels*, é possível observar algumas poses de esqueleto que, para um observador, podem trazer dúvida quanto à qualidade da anotação dos dados quanto às emoções descritas. E, uma vez que não estão disponíveis no *dataset* E-Gait os vídeos que deram origem às poses, é impossível investigar a qualidade destas anotações. É possível que estas poses mais ambíguas possam ter interferido nos resultados, tanto do STEP quanto do STEP++.

Outra questão do *dataset* é o contexto cultural. Como exposto por [KLEINSMITH; BIANCHI-BERTHOUBE \(2012\)](#), a cultura apresenta uma forte influência na codificação e entendimento das dicas afetivas, afetando a comunicação das emoções. Dessa forma, é natural que um *dataset* criado em um contexto diferente do brasileiro possa apresentar uma pior performance quando inserido no mesmo.

Inferência em vídeos de *stock footage*. Devido ao escopo deste trabalho, não foi realizada uma análise qualitativa através ferramenta de inferência com o STEP++, mas foram realizados alguns testes de inferência em *stock footage*, a fim de validação do modelo e os resultados são diversos, indicando a necessidade de uma maior investigação. Um grande ponto é a extração de pose, realizada através do HRNet, que extrai esqueletos 2D, que são adaptados para esqueletos 3D para a inferência com o STEP++.

6

Conclusão

Foi obtida grande e visível melhora tanto na performance quanto no tempo de treinamento do modelo, com a modificação proposta neste trabalho. O modelo STEP++ se mostrou melhor na tarefa de detecção de emoções a partir de marcha do que o STEP, como esperado. Além disso, espera-se que a ferramenta aqui proposta, que utiliza do STEP++ possa ser utilizada para aprofundar mais a pesquisa na área.

No entanto, como discutido anteriormente, há diversos pontos a serem investigados, nominalmente:

- Realizar uma análise qualitativa profunda do modelo STEP++, a fim de indentificar os pontos de falha e de melhoria;
- Investigar a utilização de outro detector de pose, mais adequado ao formato de esqueletos 3D;
- Investigar formas de mitigar mais o *overfitting*, como através do uso de blocos de *dropout* [SRIVASTAVA et al. \(2014\)](#);
- Investigar outras técnicas que possam melhorar a acurácia do STEP++;
- Investigar o treinamento por mais *epochs* a fim de verificar uma possível maior acurácia;
- Desenvolver um *dataset* que abrange mais emoções, no contexto brasileiro, com disponibilização não apenas dos esqueletos mas também dos vídeos para uma melhor utilização por outros pesquisadores, principalmente buscando aproximar mais o desenvolvimento de tecnologias de reconhecimento de emoções da realidade brasileira.

Referências

- BHATTACHARYA, U. et al. Step: spatial temporal graph convolutional networks for emotion perception from gaits. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. **Proceedings...** [S.l.: s.n.], 2020. v.34, n.02, p.1342–1350.
- BRIDLE, J. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. **Advances in neural information processing systems**, [S.l.], v.2, 1989.
- BULL, P. The interpretation of posture through an alternative methodology to role play. **British Journal of Social and Clinical Psychology**, [S.l.], v.17, n.1, p.1–6, 1978.
- CAMRAS, L. A.; SULLIVAN, J.; MICHEL, G. Do infants express discrete emotions? Adult judgments of facial, vocal, and body actions. **Journal of Nonverbal Behavior**, [S.l.], v.17, p.171–186, 1993.
- COSTA, W. et al. A fast multiple cue fusing approach for human emotion recognition. **Available at SSRN 4255748**, [S.l.], 2022.
- DARWIN, C. **The expression of the emotions in man and animals**. [S.l.]: John Murray, 1872.
- DUAN, H. et al. Pyskl: towards good practices for skeleton action recognition. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 30. **Proceedings...** [S.l.: s.n.], 2022. p.7351–7354.
- EKMAN, P. An argument for basic emotions. **Cognition and Emotion**, [S.l.], v.6, n.3-4, p.169–200, 1992.
- HABIBIE, I. et al. A recurrent variational autoencoder for human motion synthesis. In: BRITISH MACHINE VISION CONFERENCE (BMVC). **Proceedings...** [S.l.: s.n.], 2017.
- KLEINSMITH, A.; BIANCHI-BERTHOUBE, N. Affective body expression perception and recognition: a survey. **IEEE Transactions on Affective Computing**, [S.l.], v.4, n.1, p.15–33, 2012.
- PICARD, R. W. Affective Computing. In: OF THE . **Anais...** [S.l.: s.n.], 1997.
- RANDHAVANE, T. et al. The liar's walk: detecting deception with gait and gesture. **arXiv preprint arXiv:1912.06874**, [S.l.], 2019.
- SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, [S.l.], v.15, n.56, p.1929–1958, 2014.
- SUN, K. et al. Deep high-resolution representation learning for human pose estimation. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** [S.l.: s.n.], 2019. p.5693–5703.

TAO, J.; TAN, T. Affective computing: a review. In: INTERNATIONAL CONFERENCE ON AFFECTIVE COMPUTING AND INTELLIGENT INTERACTION. **Anais...** [S.l.: s.n.], 2005. p.981–995.

TRACY, J. L.; MATSUMOTO, D. The spontaneous expression of pride and shame: evidence for biologically innate nonverbal displays. **Proceedings of the National Academy of Sciences**, [S.l.], v.105, n.33, p.11655–11660, 2008.

UMEYAMA, S. Least-squares estimation of transformation parameters between two point patterns. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, [S.l.], v.13, n.04, p.376–380, 1991.

VALLI, A. Notes on natural interaction. , [S.l.], v.5, n.2012, p.80, 2005.

WALLBOTT, H. G. Bodily expression of emotion. **European journal of social psychology**, [S.l.], v.28, n.6, p.879–896, 1998.

WALLBOTT, H. G.; SCHERER, K. R. Cues and channels in emotion recognition. **Journal of personality and social psychology**, [S.l.], v.51, n.4, p.690, 1986.

YAN, S.; XIONG, Y.; LIN, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. **Proceedings of the AAAI Conference on Artificial Intelligence**, [S.l.], v.32, n.1, Apr. 2018.