CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

REIDENTIFICAÇÃO DE PESSOAS EM LONGO PRAZO UTILIZANDO CARACTERÍSTICAS MULTIMODAIS

Jefferson Colares de Paula

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Diego Barreto Haddad, D.Sc.
Coorientador(a): Douglas O. Cardoso, D.Sc.

Rio de Janeiro,
Dezembro de 2020

REIDENTIFICAÇÃO DE PESSOAS EM LONGO PRAZO UTILIZANDO CARACTERÍSTICAS MULTIMODAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Jefferson Colares de Paula

Banca Examinadora:

Presidente, Diego Barreto Haddad, D.Sc. (CEFET/RJ) (Orientador(a))

Douglas de Oliveira Cardoso, D.Sc. (CEFET/RJ) (Coorientador(a))

Eduardo Bezerra da Silva, D.Sc. (CEFET/RJ)

Fernanda Duarte Vilela Reis de Oliveira, D.Sc. (UFRJ)

Gabriel Matos Araujo, D.Sc. (CEFET/RJ)

Rio de Janeiro,

Dezembro de 2020

DEDICATÓRIA

Dedico este trabalho a Lia Mara, cujo apoio e incentivo foram fundamentais durante esta pós-graduação e a meus mestres no CEFET, que me iniciaram nos misteriosos caminhos da Visão Computacional, em especial meu orientador Diego Haddad.

RESUMO

Reidentificação de pessoas em longo prazo utilizando características multimodais

A reidentificação de pessoas (ReID) consiste em comparar imagens contendo pessoas, obtidas por múltiplas câmeras com campos de visão não sobrepostos e inferir se as pessoas nessas imagens são as mesmas ou não. Trata-se de um problema mais complexo do que aparenta, pois as imagens comparadas costumam apresentar grandes diferenças na iluminação, no ângulo de captura, nas características óticas das lentes utilizadas, além de oclusão parcial, auto-oclusão, planos de fundos confusos e outros complicadores. A reidentificação de pessoas em longo prazo, que é o tema desse trabalho, é caracterizada pela ocorrência de um intervalo entre as capturas das imagens. Este intervalo não tem duração específica, mas em geral costuma ser superior a um dia, no qual a pessoa observada pode ter trocado de roupas ou sofrido pequenas mudanças na aparência. As mudanças na aparência ocorridas entre as capturas de imagens representam um desafio adicional, pois as cores e texturas das roupas, que são as características mais comumente utilizadas para reidentificação, não podem ser utilizadas como elementos discriminantes.

Este trabalho investiga uma solução para o problema da reidentificação de pessoas em longo prazo por meio da utilização de características da face e da forma de caminhar das pessoas como entradas para um modelo de aprendizado de máquina baseado em redes neurais. A hipótese avaliada é que a combinação destas características permitam que o modelo despreze ou minimize o efeito das trocas de roupas e ao mesmo tempo valorize as características motoras.

Os resultados obtidos mostram que a combinação das características melhoram o desempenho da reidentificação em curto prazo e também podem ser utilizadas para o reconhecimento de pessoas em longo-prazo.

Palavras-chave: Gait energy image; GEI; Reconhecimento facial; Transfer learning; Multitask learning; Redes neurais convolucionais; reidentificação de pessoas; longo prazo

ABSTRACT

Long Term Person Re-identification Using Multimodal Features

Person re-identification (ReID) consists in comparing images containing people, acquired by multiple cameras with non-overlapping views, and inferring if the people in the images are the same. The problem is more complex than it seems, because the images are prone to great differences in illumination, viewpoint, camera optics, and also partial occlusion, self-occlusion, confusing backgrounds and other challenges. Long-term re-identification, which is the subject of this work, is characterized by the occurrence of an interval between the image captures. This interval has no specific duration, but, in general, it's greater than 24 hs. In such period, the person being reidentified might have changed clothes or have some minor changes in appearance. This long interval adds complexity to the original ReID challenge because the colors and textures of the clothes, which are the most used features for person reidentification, can't be used as discriminative elements.

This work investigates a solution to the problem of long-term person re-identification by using gait and face characteristics of the person as input features for a machine learning model based on neural networks. The hypothesis to be evaluated is that the combination of gait and face characteristics can have the machine learning model to ignore or minimize the effect of changing clothes and, at the same time, value the motion characteristics.

Our results show that the combination of multi modal input characteristics improves the performance of short-term person reidentification and can also be used for long-term person reidentification.

Keywords: Gait energy image; GEI; facial recognition; Transfer learning; convolutional neural networks; person reidentification; long term

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de como as silhuetas são combinadas para formar a *gait energy image* (GEI), na última coluna [Ju Han et al., 2006]. 12

Figura 2 – Variância de GEIs de um mesmo indivíduo em múltiplos ângulos de visão: Na primeira linha, caminhando normalmente, na segunda, usando um casaco, na terceira, carregando uma bolsa [Wu et al., 2017]. 14

Figura 3 – Exemplo de delimitação da área da imagem contendo a face utilizando *bounding box*. 16

Figura 4 – Exemplo de detecção facial, adaptado de [Guo et al., 2019]. As faces detectadas são mostradas com um quadrado vermelho. O algoritmo falha em detectar as faces em ângulos oblíquos, como as duas faces no canto inferior direito. 19

Figura 5 – Exemplos de segmentação. Adaptado de [Zhao et al., 2019]. 19

Figura 6 – Amostra do dataset TUM GAID. Cada linha apresenta imagens de um mesmo indivíduo, usando diferentes roupas e calçados [Hofmann et al., 2014] 21

Figura 7 – Comparação das arquiteturas de soluções para reconhecimento de pessoas através de *gait*. Na primeira linha, encontra-se o processo manual de extração de *features*; na segunda linha, temos a utilização de rede neural convolucional e na terceira, o modelo integrado GaitNet [Song et al., 2019]. 27

Figura 8 – Visão geral do experimento desenvolvido para testar a hipótese proposta. 29

Figura 9 – Exemplos de imagens do *dataset* VGGFace2. Cada linha contém fotos de um mesmo indivíduo. 32

Figura 10 – Detalhamento das etapas do processamento do modelo MTCNN, em imagem extraída de [Zhang et al., 2016]. 35
Figura 11 – Exemplo de entrada e saída do módulo de detecção facial. 36
Figura 12 – Dados do treinamento da rede FaceNet, para reconhecimento facial, no *dataset* TUM GAID. No eixo horizontal, as épocas. No eixo vertical, a acurácia (rank-1) 37
Figura 13 – Esquema do módulo de extração de features. A entrada é uma imagem da face, recortada do frame original. A saída, um vetor de 512 posições, contendo uma representação da imagem de entrada. 38
Figura 14 – Exemplo de entrada e saída do módulo de segmentação de silhuetas. 39
Figura 15 – O módulo de geração de GEIs recebe uma sequência de silhuetas como entrada e fornece uma GEI como saída. 40
Figura 16 – Exemplos de frames descartados por não enquadrarem a silhueta completa. 40
Figura 17 – Exemplo de entrada e saída do módulo de geração de GEIs. 41
Figura 18 – O módulo de extração de *features* de GEIs recebe GEIs como entrada e fornece um vetor de características de 305 posições como saída. 42
Figura 19 – Arquitetura de rede proposta em [Min et al., 2019a] 42
Figura 20 – Dados do treinamento da rede de reconhecimento de GEI,com *dataset* TUM GAID. No eixo horizontal, as épocas. No eixo vertical, a acurácia (Rank-1) 43
Figura 21 – Exemplo de entrada (a) e saída (b) do módulo de extração de *features* de GEIs. 44
Figura 22 – No modulo de combinação de *features*, os vetores com características de face são combinados com os de *gait*, gerando novos vetores. 45
Figura 23 – Modelo de rede neural para re-identificação de pessoas. 46
Figura 24 – Dados do treinamento da rede de ReID de pessoas. Foram utilizados 18.782 lotes de 16 imagens a cada época de treinamento. No eixo horizontal, as épocas. No eixo vertical, a acurácia (Rank-1) 46

Figura 25 – Exemplos de variações na aparência de um mesmo indivíduo causada por roupas e acessórios, com suas respectivas silhuetas. 50

LISTA DE TABELAS

Tabela 1 – Quantidade anual de artigos relacionados a reidentificação de pessoas apresentados em três das principais conferências da área: CVPR, ICCV e ECCV [Wu et al., 2019]. 2

Tabela 2 – Diferentes modalidades de transferência de aprendizado, onde "Reg." denota regressão, "Clas." denota classificação, "Agr." significa agrupamento (*clusterização*) e "RD" denota redução de dimensionalidade. Adaptado de [Pan and Yang, 2010]. 10

Tabela 3 – Principais *datasets* para reconhecimento de GEI, com correspondentes quantidades de identidades e horas de vídeo disponíveis. 15

Tabela 4 – Principais *datasets* para ReID de pessoas em domínio público, com correspondentes quantidades de identidades e sequências, número de câmeras e atributos (MR = múltiplas roupas, MO = múltiplas ocasiões). 20

Tabela 5 – Comparação do desempenho dos modelos de reconhecimento facial, de reconhecimento por GEI e de re-identificação (combinação das duas *features*) no cenário de curto prazo. 48

Tabela 6 – Comparação do desempenho dos modelos de reconhecimento facial, de reconhecimento por GEI e de re-identificação (combinação das duas *features*) *em longo prazo*. 49

Tabela 7 – Comparação do desempenho dos modelos treinados apenas com sequências em condição normal e com todas as condições *em longo prazo*. 51

LISTA DE ABREVIATURAS E SIGLAS

GEI	Gait Energy Image
REID	Reidentificação De Pessoas
REID-LP	Reidentificação De Pessoas Em Longo Prazo

SUMÁRIO

1	**Introdução**	**1**
1.1	Publicações	3
1.2	Estrutura da Dissertação	3
2	**Referencial Teórico**	**5**
2.1	Redes neurais	5
2.1.1	Transfer Learning	7
2.2	Gait Energy Image	11
2.2.1	Extração de GEIs	12
2.3	Reconhecimento de Faces	15
2.3.1	Combinação de Reconhecimento facial e *Gait Energy Image*	18
2.3.2	Datasets para ReID	20
3	**Trabalhos Relacionados**	**22**
3.1	Reidentificação de Curto Prazo	22
3.2	Reidentificação de longo prazo	23
3.3	Combinação do Modo de caminhada e reconhecimento de face	25
3.4	Comentários Finais	27
4	**Metodologia de Pesquisa**	**28**
4.1	Visão Geral do Experimento	28
4.2	Datasets	30
4.2.1	Dataset TUM GAID	30
4.2.2	Dataset VGGFace2	31
4.3	Particionamento de Dados	31
4.4	Reconhecimento Facial	33
4.4.1	Detecção de Faces	33
4.4.2	Extração de características de faces	36

4.5 Reconhecimento de *Gait* 38
4.5.1 Segmentação de figuras humanas 38
4.5.2 Segmentação de silhuetas 38
4.5.3 Geração de GEIs 39
4.5.4 Extração de *features* de GEIs 41
4.6 ReID de pessoas 43
4.6.1 Combinação dos vetores de características 44
4.6.2 Reidentificação em longo prazo 45

5 Resultados 48
5.1 ReID em curto prazo 48
5.2 ReID em longo prazo 49
5.3 Robustez em condições diversas 49
5.4 Comentários Finais 51

6 Conclusão 52

Referências 53

1- Introdução

Esta dissertação propõe uma nova abordagem para o problema da reidentificação de pessoas em longo prazo (ReID-LP). Na abordagem proposta, as técnicas de reconhecimento facial e reconhecimento de Gait Energy Image (GEI) (do inglês *Gait Energy Image*) são utilizadas para obter vetores de características que, combinados, serão utilizados com o propósito de reconhecer indivíduos em longo-prazo.

A reidentificação de pessoas (ReID) consiste em processar imagens de pessoas, obtidas por múltiplas câmeras com campos de visão não sobrepostos, como as utilizadas em circuitos-internos de TV, e inferir se as pessoas nessas imagens são as mesmas ou não [Baabou et al., 2018]. Trata-se de um problema mais complexo do que aparenta, pois as imagens comparadas por diferentes câmeras costumam apresentar grandes diferenças na iluminação, no ângulo de captura, nas características óticas das lentes utilizadas, além de oclusão parcial, auto-oclusão, planos de fundos confusos e outros complicadores [Li et al., 2014]. A ReID-LP, que é o foco desse trabalho, é caracterizada pela ocorrência de um intervalo entre as capturas das imagens maior que aquele que caracteriza o ReID de curto-prazo. [Zhang et al., 2018; Hasan and Babaguchi, 2016]. Trata-se de um intervalo sem duração específica, embora seja, em geral, superior a um dia, no qual a pessoa observada pode ter trocado de roupa ou sofrido pequenas mudanças na aparência (tal como o surgimento de uma barba ou o uso de óculos escuros). Esse maior intervalo entre as capturas de imagens representa um desafio adicional, pois as cores das roupas e texturas, que são características muito utilizadas para reidentificação, não podem ser usadas.

Para reidentificar pessoas em longo prazo, é necessário encontrar características que não sejam baseadas apenas em aparências, pois as pessoas podem ter trocado de roupas ou estar carregando acessórios diferentes entre as capturas de imagens [Zhang et al., 2018]. Características como a cor da roupa, por não serem constantes, não ajudam no caso de reidentificação de longo prazo.

A principal proposta desta dissertação consiste na utilização de GEIs, combinadas com características da face, ambas obtidas à distância, sem que a pessoa se submeta ativamente à captura de imagens. As GEIs, calculadas por meio da média de silhuetas

sobrepostas, representam o padrão de movimentação da pessoa ao caminhar. Tal padrão é bastante idiossincrático, além de não variar significativamente para uma mesma pessoa [Ju Han et al., 2006]. Argumentamos que, embora as imagens da face, quando obtidas lateralmente e a distância, não sejam suficientes para obter uma boa identificação da pessoa, tais imagens potencializam a acurácia da inferência, quando combinadas com as GEIs.

Do ponto de vista computacional e de forma resumida, o problema da ReID consiste em comparar a imagem de uma determinada pessoa com imagens de outras pessoas armazenadas em uma galeria. Na comparação, atribuem-se *scores* a cada item da galeria, correspondentes às probabilidades de semelhança com a imagem de referência. O elemento com score mais alto, identifica quem é a pessoa na imagem de referência. Esse processo é amplamente utilizado em problemas de classificação multiclasse [Har-Peled et al., 2002; Bucak et al., 2009].

O objetivo deste trabalho é comprovar a eficácia da combinação de redes neurais, GEIs e reconhecimento facial na tarefa de reidentificar pessoas em sequências de imagens obtidas em diferentes dias, por múltiplas câmeras. A ReID, seja ela em curto ou longo prazo, é um tema de pesquisa com potencial de aplicação em diversas áreas, como por exemplo robótica e inteligência artificial, autenticação de usuários, segurança patrimonial e segurança pública, entre outras. Ademais, a ReID vem capturando cada vez mais o interesse dos pesquisadores e foi elencada em relatório recente do *Institute for Scientific Information* como uma “fronteira da ciência”. Podemos perceber a relevância do ReID pelo aumento expressivo na quantidade de artigos relacionados ao tema, conforme exemplificado na Tabela 1 [Wu et al., 2019].

2010	2011	2012	2013	2014	2015	2016	2017	2018
1	1	1	3	6	14	19	26	46

Tabela 1 – Quantidade anual de artigos relacionados a reidentificação de pessoas apresentados em três das principais conferências da área: CVPR, ICCV e ECCV [Wu et al., 2019].

Os avanços científicos na tarefa de reidentificação de pessoas vêm da aplicação de novas técnicas para extrair as características (*feature extraction*) que distingam da melhor forma possível a pessoa na imagem ou vêm do desenvolvimento de novos algoritmos para comparar a imagem de referência, a *probe image*, com as da galeria

(*distance metric)* [Karanam et al., 2019]. Nesse trabalho, duas técnicas de extração de características são combinadas para viabilizar uma reidentificação robusta de pessoas em longo prazo: GEI e reconhecimento facial. A primeira consiste em capturar a forma da pessoa, por meio da sobreposição de silhuetas extraídas de sequências de imagens. A segunda, o método de identificação mais frequentemente utilizado, consiste em identificar pessoas por sua aparência facial . Entretanto, ao contrário do trabalho apresentado em [Jayaraman et al., 2020], no contexto desta dissertação, as imagens não são obtidas por câmeras frontais e sim por câmeras posicionadas em ângulos oblíquos e à média distância. Isoladamente, na tarefa específica de ReID-LP, tanto a GEI quanto o reconhecimento facial dificilmente produziriam resultados expressivos mas, utilizadas em conjunto, as duas técnicas apresentam resultados semelhantes ou superiores aos trabalhos de referência.

1.1- Publicações

O autor desta dissertação publicou, ao longo da pesquisa, dois artigos em periódicos:

1. Felipe X. Viana, Gabriel M. Araujo, Milena F. Pinto, **Jefferson Colares**, Diego B. Haddad. "*Aerial Image Instance Segmentation Through Synthetic Data Using Deep Learning*", *Learning & Nonlinear Models*, pp. 35-46, Out. 2020.

2. Lucas P. R. da Silva, José V. G. de Souza, **Jefferson Colares**, Amaro A. de Lima, Diego B. Haddad. "*Bias-compensated FX-LMS algorithm*", *Electronics* Letters, 56, (24), pp. 1340-1343, Nov. 2020.

1.2- Estrutura da Dissertação

O restante da presente dissertação é organizado em capítulos, a seguir descritos: O Capítulo 2 descreve um referencial teórico, o qual oferece mais detalhes sobre as tecnologias mobilizadas neste trabalho. O Capítulo 3 disserta sobre outros trabalhos

relacionados. O Capítulo 4 apresenta a metodologia utilizada na pesquisa e o Capítulo 5 condensa os resultados desta dissertação. Por fim, o Capítulo 6, de conclusões, articula as considerações finais e trabalhos futuros.

2- Referencial Teórico

Como anteriormente dito, esta dissertação une duas técnicas de extração de características: reconhecimento de GEIs e reconhecimento facial. Tais características combinadas alimentam uma rede neural, projetada de forma a possibilitar a reidentificação. Este capítulo descreve estas duas técnicas principais, as redes neurais e mais outras utilizadas neste trabalho, de modo a elucidar as ferramentas consideradas para o objetivo final de ReID-LP.

2.1- Redes neurais

Dotar máquinas com a capacidade de aprendizado autônomo é um dos principais temas de pesquisa científica em computação há muitos anos. As redes neurais artificiais são uma de suas principais ramificações, tendo sido tema de pesquisa teórica desde os anos 60 [Uttley, 1966]. O estudo dessas redes obteve grandes avanços a partir dos anos 80 e mais recentemente, com o trabalho de pesquisadores como Yann LeCun e Yoshua Bengio [LeCun and Bengio, 1998]. [Wang, 2003] apresenta todos os conceitos-chave e discorre sobre como ocorre o aprendizado nestas redes.

O estudo da visão computacional consiste em fazer com que o computador perceba significado em imagens. Esse ramo de estudo encontrou nas redes neurais uma poderosa ferramenta. Utilizando as redes neurais, é possível encontrar o significado das imagens de entrada com relativa confiabilidade através da extração informações de alto nível. O fato de que redes neurais de múltiplas camadas são capazes de aproximar qualquer função matemática “bem comportada” é uma garantia de seu poder e expressividade [Hornik, 1991].

Mais recentemente, observou-se que as camadas iniciais de uma rede neural, após o treinamento, se transformam em extratores de características (*feature extractors*) [LeCun and Bengio, 1998], o que significa que a utilização dessas redes para o reconhecimento de imagens pode dispensar a necessidade de desenvolver manualmente esses extratores

por meio de conhecimento *a priori*. A extração automática de *features* com redes neurais, além de mais fácil de implementar, é não raro mais eficiente do que os métodos manuais [Seide et al., 2011].

A adição da operação de convolução nas camadas iniciais da rede neural deu origem às redes neurais convolucionais (CNNs, do inglês *convolutional neural networks*), que possibilitaram um salto de qualidade na capacidade de reconhecer (classificar) objetos em imagens. A operação matemática de uma CNN é brevemente descrita a seguir, *baseando-se fortemente na formulação de [Wu et al., 2017]*. Uma CNN é composta por uma série de estágios encadeados. Cada um destes estágios pode ser decomposto em operações sucessivas de diferentes camadas, as quais incluem um banco de filtros (responsáveis por efetuar a operação de convolução espacial), uma função de ativação não-linear, uma camada de *pooling* espacial e, por vezes, uma camada de normalização.

Uma camada de convolução (*convolution layer*) pode ser derivada de uma camada completamente conectada (*fully connected*) por meio de duas simplificações, responsáveis por reduzir o número de parâmetros ajustáveis. A primeira destas simplificações impõe localidade espacial, de modo que os neurônios são conectados apenas a regiões locais da camada imediatamente pregressa. Por essa razão, este tipo de camada é também chamado de camada localmente conectada. O próximo passo de simplificação compartilha os mesmos pesos sinápticos ao longo de todas as localizações espaciais. Com estes passos, temos por fim uma camada de convolução. Como funções de ativação não-lineares, podemos utilizar a função logística ($f(x) = \frac{1}{1+e^{-x}}$), a hiperbólica ($f(x) = \tanh(x)$) e a *rectified linear unit* (ReLU, descrita por $f(x) = \max(0, x)$), entre outras.

Em termos matemáticos, assumindo a adoção da ReLU e sabendo que $\boldsymbol{H}_0$ armazena os dados de entrada originais, a l-ésima camada de convolução $\boldsymbol{H}_l$ de uma rede neural convolucional pode ser descrita como

$$\boldsymbol{H}_l = \text{pool}\left(\max(0, \boldsymbol{W}_l \otimes \boldsymbol{H}_{l-1} + \boldsymbol{b}_l)\right), \tag{1}$$

onde $\otimes$ denota a operação de convolução, $\boldsymbol{H}_{l-1}$ é a entrada obtida da camada imediatamente anterior, $\boldsymbol{H}_0$ armazena os dados de entrada originais, $\boldsymbol{W}_l$ contém diversos filtros espaciais e $\boldsymbol{b}_l$ contém os *biases* (ou termos polarizadores) compartilhados entre as diferentes localizações espaciais. Note que a cada mapa de característica (*feature map*)

em $\boldsymbol{H}_l$ corresponderá um filtro em $\boldsymbol{W}_l$, bem como uma entrada em $\boldsymbol{b}_l$ (ou seja, o *bias*).

Uma rede neural para classificação de imagens pode ser "alimentada" com uma quantidade suficiente de amostras e, a partir delas, aprender a extrair automaticamente as *features* relevantes para reconhecer objetos ou animais com grande acurácia, como demonstrado em [Krizhevsky et al., 2012].

A utilização de redes neurais na tarefa específica de reidentificação de pessoas em longo prazo exige um *dataset* com exemplos apropriados para esse fim, a saber, imagens de pessoas capturadas em diferentes momentos, utilizando diferentes roupas e acessórios. Utilizando tal dataset para o treinamento, a rede seria capaz de perceber *features* relevantes, que sejam idealmente independentes de aspectos que se modifiquem. A dificuldade em obter *datasets* com quantidade e variedade adequadas de exemplos é uma das maiores dificuldades deste tema de pesquisa [Wu et al., 2019]. Alguns dos mais relevantes trabalhos na área precisaram preparar seus próprios *datasets* e estes, até o presente momento, não estão disponíveis para a comunidade científica, particularmente para o problema considerado nesta dissertação se debruça, a reidentificação *de longo prazo*.

2.1.1 Transfer Learning

Uma forma de contornar o problema da indisponibilidade de *datasets* reside ma adoção de estratégias de *transfer learning*, as quais são, muitas vezes, capazes de garantir um bom desempenho sem exigir custosos esforços de anotação manual de dados[1]. O procedimento de transferência de aprendizado emula o sistema visual humano ao aplicar conhecimento *a priori* em domínios correlatos quando novas tarefas são necessárias no novo domínio [Shao et al., 2015].

Métodos de aprendizagem de máquina tradicionais assumem que os dados de treinamento e de teste são oriundos do mesmo espaço de características e, portanto, apresentam a mesma distribuição. Seja um *domínio* $\mathcal{D}$, o qual consiste de dois componentes: um espaço de características $\mathcal{X}$ e uma distribuição de probabilidade marginal $P(X)$, onde $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Tomado como exemplo o caso de algum algoritmo de aprendizado

[1] A discussão a seguir é *fortemente baseada* em [Pan and Yang, 2010].

de máquina que almeja estimar a nota que um especialista daria a amostras de café, cada elemento x_i (para $i \in \{1, \ldots, n-1\}$) corresponderia ao vetor de características referente à i-ésima amostra de café no conjunto de dados de aprendizagem.

Um domínio pode ser inteiramente especificado pelo conjunto $\mathcal{D} = \{\mathcal{X}, P(X)\}$, o que significa que dois domínios podem ser diferentes se apresentarem espaços de características distintos ou, no caso em que estes são idênticos, se suas distribuições $P(X)$ diferirem. A partir de um determinado domínio $\mathcal{D} = \{\mathcal{X}, P(X)\}$, uma *tarefa* $\mathcal{T}$ consiste de dois componentes:

- Um espaço de rotulagem $\mathcal{Y}$;
- Uma função objetivo de predição $f(\cdot)$.

Assim, pode-se denotar, sem ambiguidade, uma tarefa T por $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, na qual a função $f(\cdot)$ normalmente é desconhecida. No exemplo supracitado da classificação automática de amostras de café, a função $f(\cdot)$ mapearia um vetor de características $x_i \in X$ em um valor (seja este escalar ou vetorial) $y_i \in \mathcal{Y}$, onde y_i seria a nota (possivelmente numérica) associada à i-ésima amostra de café. Num problema de classificação binária, y_i poderia assumir dois valores, como “falso” e “verdadeiro”.

Normalmente, as abordagens de *transfer learning* admitem a existência de dois domínios: um relacionado ao conjunto de dados da fonte ($\mathcal{D}_\mathsf{S}$, sendo S usado para denotar *source*) e outro relacionado ao conjunto de dados do alvo ($\mathcal{D}_\mathsf{T}$, sendo T usado para denotar *target*). Assim, os *dados no domínio da fonte* são dados por:

$$D_S = \{(x_{\mathsf{S}_1}, y_{\mathsf{S}_1}), \ldots, (x_{\mathsf{S}_{n_s}}, y_{\mathsf{S}_{n_s}})\}, \tag{2}$$

onde assume-se que o conjunto de dados da fonte contém n_s instâncias. Similarmente, os *dados no domínio do alvo* encontram-se no conjunto

$$D_T = \{(x_{\mathsf{S}_1}, y_{\mathsf{S}_1}), \ldots, (x_{\mathsf{S}_{n_T}}, y_{\mathsf{S}_{n_T}})\} \tag{3}$$

onde n_T é o conjunto de instâncias do conjunto de dados no alvo.

Tais conceitos nos permitem definir o procedimento de *transfer learning*:

Definição. Dado um domínio da fonte $\mathcal{D}_\mathsf{S}$, uma tarefa de aprendizado respectiva $\mathcal{T}_\mathsf{S}$, um domínio do alvo $\mathcal{D}_\mathsf{T}$ e uma tarefa de aprendizado respectiva $\mathcal{T}_\mathsf{T}$, um procedimento de transferência de aprendizado tem como objetivo melhorar o aprendizado da função

preditiva alvo $f_{\mathsf{T}}(\cdot)$ em $\mathcal{D}_T$ usando o conhecimento em $\mathcal{D}_{\mathsf{S}}$ e $\mathcal{T}_{\mathsf{S}}$, onde $\mathcal{D}_{\mathsf{S}} \neq \mathcal{D}_{\mathsf{T}}$ e/ou $\mathcal{T}_{\mathsf{S}} \neq \mathcal{T}_{\mathsf{T}}$.

Um dos tipos de *transfer learning* é o *multitask learning* [Pan and Yang, 2010], no qual uma única rede[2] é treinada para mais de uma tarefa ao mesmo tempo, potencializando a capacidade de generalização para ambas as tarefas. Por exemplo, pode-se treinar a rede neural alimentando-a simultaneamente com imagens de faces, para reconhecimento facial, e com GEIs, para reconhecimento através da forma de caminhar. Observa-se que frequentemente a rede se vale de um aprendizado para potencializar o outro, melhorando sua capacidade de generalização nos dois domínios.

O sucesso de um modelo de *machine learning* pode ser avaliado de grosso modo, por sua acurácia. A acurácia do modelo, por sua vez, é limitada pela qualidade do *dataset* utilizado no treinamento, já que é a partir do próprio conjunto de dados que a rede neural aprende as características relevantes para classificar os objetos.

A qualidade de um *dataset* é determinada por dois fatores: pela quantidade de elementos diversos (isto é, sua heterogeneidade) que ele contém; e pela qualidade de suas anotações. No caso de um dataset para visão computacional, as anotações incluem a delimitação da área da imagem onde se encontra o objeto de interesse então, a acurácia com que esta área é delimitada determina a qualidade da anotação.

Um dos maiores problemas para o desenvolvimento de redes neurais eficientes é a pouca disponibilidade de *datasets* de qualidade. Como vimos, uma das formas de contornar esse problema é aplicar a técnica de *transfer learning*, pois enquanto o aprendizado tradicional busca aprender as tarefas do zero, o *transfer learning* busca transferir o aprendizado de tarefas anteriores para tarefas onde há menos disponibilidade de dados [Pan and Yang, 2010]. Desta forma é possível conseguir que o modelo apresente bons resultados mesmo no cenário onde há menos dados.

Grosso modo, para se utilizar *transfer learning*, é necessário treinar a primeira rede para uma determinada tarefa, utilizando o *dataset* inicial D_{S} (com mais informações). Nessa primeira etapa é utilizado um algoritmo de treinamento para ajustar os parâmetros do modelo. Em seguida, substituímos a última camada da rede neural por uma nova, mais apropriada para desempenhar a tarefa no novo domínio. Finalmente, reutilizamos todas as camadas ocultas aprendidas nos dados no domínio da fonte D_{S} e efetuamos novo treinamento dessa camada, usando o dados no domínio do alvo D_{T} [Lu et al., 2015a].

[2]Embora em tese diversos algoritmos de aprendizagem de máquina possam ser utilizados em *transfer learning*, doravante restringiremos a discussão a redes neurais.

Modalidade	Áreas Correlatas	Rótulos na Entrada	Rótulos no alvo	Tarefas
Indutiva	Multitarefa	Disponível	Disponível	Reg., Clas.
Indutiva	Auto-aprendizado	Não disponível	Disponível	Reg., Clas.
Transdutiva	Adaptação de domínio	Disponível	Não disponível	Reg., Clas.
Não supervisionada	-	Não disponível	Não disponível	Agr. e RD.

Tabela 2 – Diferentes modalidades de transferência de aprendizado, onde “Reg.” denota regressão, “Clas.” denota classificação, “Agr.” significa agrupamento (*clusterização*) e “RD” denota redução de dimensionalidade. Adaptado de [Pan and Yang, 2010].

Num nível mais intuitivo, digamos que, por exemplo, haja um *dataset* muito vasto de imagens de maçãs, mas que a tarefa desejada é a classificação de kiwis, para a qual não temos muitas imagens disponíveis. Nesse caso, o procedimento do transfer learning consiste em treinar uma rede neural para classificar maçãs e depois reutilizá-la inteira na tarefa de classificação de kiwis, adotando os parâmetros obtidos no treinamento original. Para que a rede original possa ser útil na nova tarefa, sua última camada é substituída, e treinada com o *dataset* de kiwis. Ao final, a classificação de kiwis pode ser realizada com maior acurácia do que se fosse utilizado apenas o *dataset* de kiwis, porque a rede neural treinada com maçãs ainda é capaz de extrair características que são relevantes para a classificação de kiwis.

As técnicas de *transfer learning* se apresentam em três modalidades distintas: o primeiro é o *transfer learning* indutivo. Nele, a tarefa de origem é diferente da tarefa destino, independente se os domínios são iguais ou não. No segundo tipo, o *transfer learning* transdutivo, a tarefa é a mesma na origem e no destino, mas os domínios são diferentes. No terceiro tipo, o *transfer learning* não-supervisionado, a tarefa no destino é diferente, mas apresenta relação com a tarefa na origem, enquanto os domínios não contém rótulos, tanto na origem quanto no destino (veja Tab. 2).

A forma como o *transfer learning* será empregado neste trabalhado será detalhada no Cap. 4. Em resumo, ao treinar as redes neurais utilizando *datasets* especializados em faces e GEIs, pretendemos aplicá-las a outros *datasets*, desenvolvidos para ReID de pessoas, como extratores de características. Essa operação caracteriza um *transfer learning* na modalidade transdutiva, já que a tarefa, predição, estimada pela rede neural convolucional, é preservada.

2.2- Gait Energy Image

Uma das principais aplicações do reconhecimento de pessoas em longo prazo são os sistemas de monitoramento e segurança. Dotar um sistema de vigilância com a capacidade de reconhecer suspeitos ou pessoas de interesse em imagens captadas pela rede de câmeras pode representar um grande avanço na área. Diferentemente de outras aplicações, como as de controle de acesso por exemplo, os sistemas de vigilância não podem contar com dados fornecidos antecipadamente e voluntariamente pela pessoa de interesse, como digitais, fotos da face ou outros tipos de dados biométricos. Por esse motivo, o ReID de longo prazo só pode contar com *soft biometrics*, dados biométricos que podem ser obtidos sem a colaboração explícita do sujeito.

O modo de caminhar é uma característica biométrica individual que pode ser usada com sucesso para reidentificar pessoas. Por muitos anos, diversas abordagens vêm sendo desenvolvidas para, a partir de sequências de imagens, extrair de forma cada vez mais eficiente as características que distinguem a marcha de cada pessoa. Essas abordagens são classificadas em dois grupos [Singh et al., 2018]: *model-based* e *model-free*.

As abordagens baseadas em modelos, as *model-based*, buscam obter parâmetros como o tamanho, posição, ângulo e distância relativa de membros e articulações do corpo humano para gerar um modelo que represente a marcha individual de cada pessoa [Han et al., 2017]. Já as abordagens *model-free* são baseadas em aparência e apenas segmentam a figura humana na imagem, gerando silhuetas das quais são extraídas as características para distinguir os indivíduos. [Yao et al., 2019]. Uma dessas abordagens utiliza *Gait energy images* (GEIs).

As GEIs são imagens obtidas calculando-se a média de silhuetas sobrepostas. Uma silhueta é obtida a partir de uma imagem de uma pessoa e contém apenas a área da imagem ocupada pelo contorno da pessoa preenchida em cor branca, contra um fundo totalmente preto. Convertendo-se uma sequência de imagens da pessoa caminhando em silhuetas e sobrepondo-as de modo a formar uma única imagem, obtemos uma representação do caminhar da pessoa, onde as partes com maior movimentação, como as pernas e braços, são representadas em cor branca mais fraca, como que transparentes, enquanto as áreas que se movimentam menos ao caminhar, como o tronco e a cabeça,

são representadas em um branco mais forte, quase sólido. Essa representação é a GEI. A Figura 1 mostra diversas silhuetas obtidas a partir de cada frame de duas sequências de vídeo. No último quadro de cada linha linha, há uma GEI, que é formada pelo cálculo do valor médio dos pixels das silhuetas.

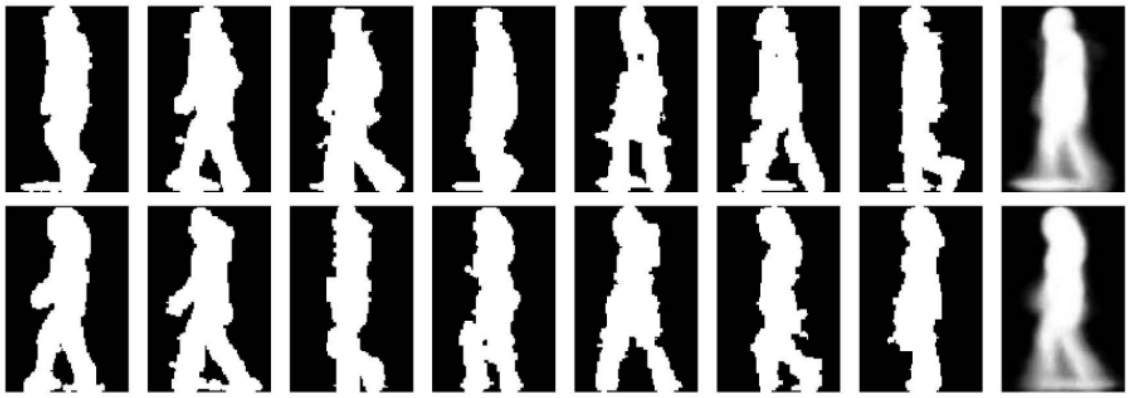


Figura 1 – Exemplo de como as silhuetas são combinadas para formar a *gait energy image* (GEI), na última coluna [Ju Han et al., 2006].

As GEIs combinam em uma única imagem múltiplas poses da pessoa durante o ciclo de caminhada. Desta maneira, a informação temporal do ciclo é perdida, mas a individualidade do movimento de cada pessoa é preservada. A CGI (*Chrono Gait Image*), uma variante da GEI proposta em [Wang et al., 2010], codifica a informação temporal em cores distintas para cada etapa do ciclo. Avaliações realizadas em [Wu et al., 2017] sugerem que a preservação da informação temporal pode produzir melhores resultados.

2.2.1 Extração de GEIs

Inicialmente, obtém-se as sequências de imagens de onde são extraídas as silhuetas. Elas podem ser capturadas por câmeras de vídeo comuns (RGB), câmeras infra-vermelho (RGB-IR) ou mesmo câmeras 3D (RGB-D). Entretanto, este último tipo de câmera captura também a profundidade, sendo por isso mais utilizado em abordagens *model-based*, como a apresentada em [Yao et al., 2019]. No presente trabalho, buscando aumentar a aplicabilidade da técnica proposta, são analisadas apenas imagens obtidas por meio de câmeras comuns (RGB), como as que são comumente utilizadas em sistemas de vigilância e monitoramento. Embora tal escolha torne o problema mais desafiador, tem

como vantagem aumentar o escopo de aplicações da solução proposta.

Depois de capturar as imagens, ainda como parte do pré-processamento, é necessário gerar as silhuetas das figuras humanas nelas encontradas. Para isso, pode ser aplicada a técnica de segmentação semântica por meio de uma rede neural treinada para reconhecer uma classe: pessoa.

Uma única sequência de imagens pode conter frames suficientes para originar diversas GEIs. O *frame* em que as pernas encontram-se mais próximas uma da outra marca o início do "*gait cycle*". As duas ocorrências seguintes dessa distância máxima marcam o meio e o final do ciclo completo, respectivamente. Todas as silhuetas de um ciclo de marcha completo são sobrepostas para compor cada GEI.

Após a segmentação, é necessário fazer a normalização e o alinhamento. A normalização consiste em redimensionar as silhuetas de modo que todas tenham a mesma altura, porém mantendo a largura proporcional. O alinhamento, por sua vez, refere-se ao alinhamento horizontal das silhuetas, utilizando como referência o centroide de sua parte superior [Sarkar et al., 2005]. Ao final desta etapa, tem-se uma sequência de silhuetas alinhadas que pode ser utilizada para a obtenção da GEI. Em termos matemáticos, a GEI pode ser definida como:

$$\text{GEI}(x, y) = 1/s \sum_{t-1}^{s} F^t(x, y), \tag{4}$$

onde s denota o número total de *frames* em um ciclo de marcha e F^t representa a silhueta em um momento t e o par ordenado (x, y), os pixels em cada frame da silhueta.

Ao sobrepor as silhuetas, diversos cálculos podem ser aplicados. Da aplicação da média entre as imagens, obtém-se a GEI [Liu and Sarkar, 2004; Ju Han et al., 2006]. No entanto, outras abordagens mais sofisticadas podem ser utilizadas, como veremos adiante.

Além da GEI, vários outros métodos foram desenvolvidos para extrair *features* de silhuetas, como por exemplo, a GEnI (Gait Entropy Image) [Bashir et al., 2010], que calcula a entropia em vez da média das silhuetas; o SGEI (Skeleton GEI), que estima a posição de pontos de interesse nas silhuetas formando esqueletos e deles extrai GEIs, com o objetivo de minimizar o impacto das roupas e acessórios na acurácia do modelo [Yao et al., 2019] e a CGI (*Chrono Gait Image*), que busca preservar a informação temporal do ciclo de marcha utilizando cores diferentes para os contornos das silhuetas [Wang et al.,

2010] e mais inúmeros outros, descritos detalhadamente em [Singh et al., 2018].

O desempenho das abordagens do tipo *model-free* é afetado por variações nas roupas, acessórios e calçados. A Fig. 2 exemplifica alguns tipos de variância que uma GEI de um mesmo indivíduo pode sofrer. Entretanto, o fator que mais afeta a acurácia dos modelos é a variação no ângulo de visão da câmera no momento da captura das silhuetas. As técnicas utilizadas para criar modelos *view-invariant* são muito variadas mas, em geral, se dividem em: abordagem generativa baseada em modelos de transformação da visão (*view transformation model*, VTM) [Wu et al., 2018; Zheng et al., 2011], abordagens baseadas em galerias multi-visão, invariância baseada em características e abordagens baseadas em sub-espaços invariantes a ângulo de visão.

A maioria das abordagens citadas utiliza GEIs obtidas por câmeras posicionadas próximo ao chão. Neste trabalho, consideramos que as câmeras podem estar posicionadas mais alto, como em geral ocorre com câmeras de vigilância.

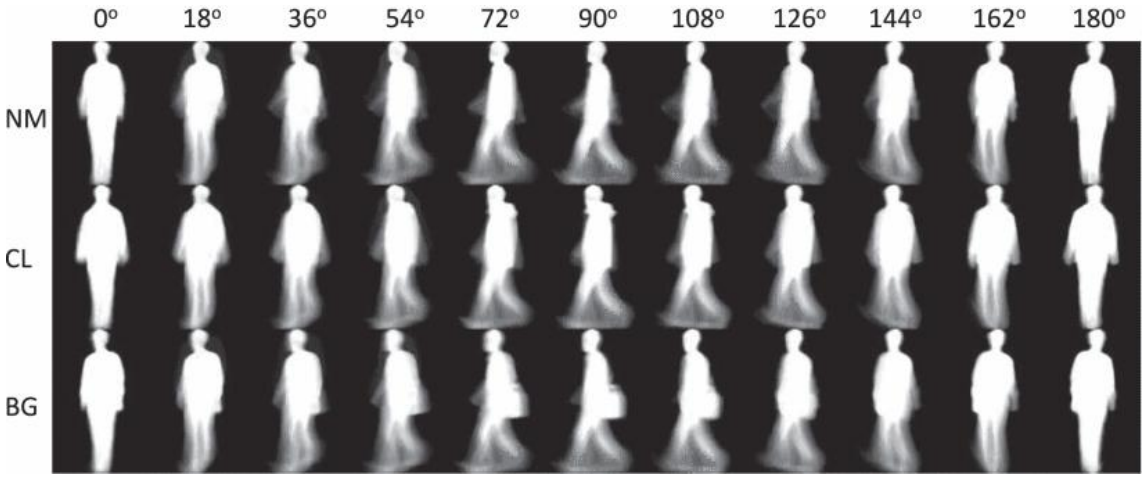


Figura 2 – Variância de GEIs de um mesmo indivíduo em múltiplos ângulos de visão: Na primeira linha, caminhando normalmente, na segunda, usando um casaco, na terceira, carregando uma bolsa [Wu et al., 2017].

Os principais *datasets* utilizados para reconhecimento de GEIs são obtidos em ambiente controlado e contêm pouca variação em quantidade de pessoas. A tabela 3, adaptada de [Xu et al., 2017], lista os datasets. Os datasets que apresentam maior quantidade de indivíduos não possuem muita variação nos ângulos de captura [Iwama et al., 2012; Xu et al., 2017].

Dataset	Indivíduos	Horas de vídeo
SOTON Small database	12	-
SOTON Large database	115	-
SOTON Temporal	25	-
USF *dataset*	122	-
CASIA *dataset* A	20	-
CASIA *dataset* B [Yu et al., 2006]	124	-
CASIA *dataset* C	153	-
OU-ISIR Treadmill *dataset* A	34	0,9
OU-ISIR Treadmill *dataset* B	68	4,6
OU-ISIR Treadmill *dataset* C	200	8,3
OU-ISIR Treadmill *dataset* D	185	0,6
OULP *dataset* [Iwama et al., 2012]	4.016	8,1
OULP Age *dataset* [Xu et al., 2017]	63.846	44,3
TUM GAID *dataset* [Hofmann et al., 2014]	305	$2{,}4^3$

Tabela 3 – Principais *datasets* para reconhecimento de GEI, com correspondentes quantidades de identidades e horas de vídeo disponíveis.

2.3- Reconhecimento de Faces

Valendo-se do fato de que a face humana contém rica informação biométrica [Yan and Song, 2021], o reconhecimento facial consiste em comparar uma imagem de entrada, que contém um rosto com outras imagens de rostos armazenadas em uma galeria e identificar quais imagens da galeria correspondem ao mesmo indivíduo da imagem de entrada [Lu et al., 2015b]. Em geral, o processo divide-se em três etapas: detecção e extração da face, extração de *features* e o reconhecimento propriamente dito através da comparação das features de duas imagens [Jayaraman et al., 2020]. A técnica de reconhecimento facial tem atraído considerável atenção na comunidade de visão computacional devido à crescente demanda por aplicações biométricas [Sang et al., 2021], em especial por evitar a violação de sistemas tradicionais de segurança, tais como senhas [Plichoski et al., 2021].

Conforme mencionado, é muito comum que a primeira etapa do processo de reconhecimento seja a detecção da face na imagem. Em outras palavras, cumpre construir uma *bounding box* delimitando a região da imagem onde supostamente se encontra a face que se deseja reconhecer. Um exemplo de bounding box é mostrado na figura 3. A área da imagem circunscrita pela *bounding box* é então armazenada para processamento adicional; o restante da imagem costuma ser descartado. Na etapa

extração de *features*, rotineiramente utiliza-se um algoritmo de *machine learning* para aprender quais características da foto são as mais relevantes para a identificação da pessoa [Taghipour-Gorjikolaie et al., 2020]. Em seguida é realizado o alinhamento da face, de modo a facilitar a identificação. No alinhamento, a imagem de uma face é movimentada e deformada de forma a se adequar a um padrão [Zhu et al., 2019]. Em muitos casos, este alinhamento demanda a detecção dos pontos-chave ou pontos fiduciais (*landmark-points*. Ver a Fig.3), como olhos, sobrancelhas e boca, em um vídeo ou em uma imagem. O alinhamento é um dos mais pesquisados tópicos na área de visão computacional nas últimas décadas [Hoang et al., 2021]. O alinhamento pode melhorar significativamente a eficiência de diversos sistemas de processamento automático de vídeo, incluindo os de reidentificação de pessoas [Bansal et al., 2017]. Embora inicialmente as técnicas de alinhamento tenham se baseado em estruturas pictóricas (por exemplo, detecção do centro do olho utilizando filtros gaussianos) ou em extensões mais sofisticadas capazes de modelar grandes mudanças de aparência, posteriormente houve uma convergência para a adoção de métodos de regressão em cascata [Hoang et al., 2021] e redes neurais convolucionais [Sun et al., 2013].

A etapa final do processo de reconhecimento facial reside na classificação ou comparação com as imagens da galeria, que é o reconhecimento, propriamente dito.

Figura 3 – Exemplo de delimitação da área da imagem contendo a face utilizando *bounding box*.

O tema reidentificação de pessoas remete imediatamente à técnica de reconhecimento facial, pois a imagem da face da pessoa pode ser utilizada para sua identificação. Entretanto, para a aplicação proposta neste trabalho, é necessário levar em consideração que as imagens são obtidas utilizando câmeras de vigilância comuns e, por esse motivo, temos:

- Imagens das faces obtidas através de câmeras de vigilância comuns, usualmente posicionadas à distância, em condições não-controladas (*"in the wild"*);
- Imagens das faces com pouca nitidez e em tamanho reduzido;
- Grande variância nas condições de iluminação e no ângulo de captura da imagens;
- Longo intervalo entre duas capturas de imagens, a ponto de a pessoa a ser reconhecida ter mudado de aparência ou envelhecido.

À primeira vista, a técnica de reconhecimento facial não aparenta ser muito adequada a esse tipo de aplicação, já que cenários sobre os quais o controle é limitado são desafiadores [Wu et al., 2021; Hu et al., 2021] e mudanças de perspectiva são um dos principais fatores a reduzir o desempenho de algoritmos de reconhecimento de faces [Zhang et al., 2021]. No entanto, há estudos que mostram que imagens de tamanho tão pequeno quanto 30 *pixels* podem ser utilizadas para obter diversas informações biométricas sobre o indivíduo analisado, como por exemplo, sexo, idade ou etnia [Bobeldyk and Ross, 2019]. Assim sendo, nesta dissertação optamos por utilizar o reconhecimento facial como técnica auxiliar no reconhecimento de pessoas em longo prazo, respeitando suas limitações. Uma dessas limitações é o fato de que a maioria das técnicas de reconhecimento de imagens é otimizada para faces presentes em imagens e não em vídeos. No caso da utilização de vídeos, podem ocorrer problemas de borramento e perda de foco [Ding and Tao, 2018]. Ademais, parte significativa dos trabalhos, quando contemplam alguma variação (como pose ou iluminação) à qual as faces podem estar sujeitas, enfoca uma variação particular e assume a ausência das outras, numa estratégia de "dividir para conquistar" que prejudica o projeto de um classificador robusto e genérico [Geng et al., 2008].

Como mencionado anteriormente, a utilização de uma rede neural convolucional para o reconhecimento facial elimina a necessidade de desenvolver manualmente os extratores de características, uma vez que as camadas iniciais da rede aprenderão a realizar essa tarefa durante a etapa de treinamento. Entretanto, por essa mesma razão, o desempenho final da rede depende da base de dados utilizada para o seu treinamento. Um conjunto de dados com pouca variação não será suficiente para que a rede neural tenha uma boa generalização. É fundamental, portanto, treinar a rede utilizando um bom conjunto de dados, ou ainda, treiná-la usando mais de um *dataset*.

Antes que se possa executar a atividade de reconhecimento facial, é necessário localizar a posição do rosto na imagem. Essa atividade é conhecida como detecção [Zafeiriou et al., 2015]. Existem diversos modelos de redes neurais especializados na detecção de diferentes tipos de objetos, animais e plantas. No caso específico de detecção facial, o modelo mais amplamente difundido, por sua eficiência computacional e desempenho, é conhecido como “Viola-Jones” [Viola and Jones, 2004]. Esse modelo, bem como a maioria de seus similares, é bastante eficiente na detecção de faces em posição frontal. Entretanto, para este trabalho, o modelo de detecção utilizado deve ser capaz de detectar faces em diversos ângulos; em outras palavras, deve ser invariante à pose (*pose-invariant*).

É interessante notar que, podendo ser avistado em diferentes dias, o indivíduo em análise poderá estar ora mais próximo, ora mais distante da câmera. Por este motivo, durante o processamento, quando as áreas das imagens contendo as faces detectadas forem redimensionadas para o tamanho padrão de entrada da rede, isso resultará em imagens com resoluções distintas representando o mesmo indivíduo. Dessa forma, a capacidade de detectar e reconhecer faces em imagens com múltiplas resoluções (*multi-resolution*) mostra-se bastante relevante para o bom desempenho geral do modelo utilizado.

No contexto desta dissertação, em que estaremos utilizando redes neurais em todas as etapas do processo, é importante que o modelo utilizado para detecção facial seja treinado com um *dataset* que contenha imagens de pessoas capturadas em múltiplos ângulos. Desejamos que a rede adquira uma habilidade de generalização suficiente para que possa ser posteriormente empregada no reconhecimento de pessoas em longo prazo.

2.3.1 Combinação de Reconhecimento facial e *Gait Energy Image*

O objetivo é combinar as saídas de duas redes neurais especializadas e utilizá-las como entrada para uma nova rede neural, esta sim, especializada em reidentificação de pessoas. Para o reconhecimento de face, é necessário localizar e limitar a área contendo a face de cada indivíduo em cada um dos *frames* da sequência. No entanto, para obter

uma única GEI, múltiplos *frames* precisam ser combinados. Como resultado, temos para cada indivíduo múltiplas imagens de faces e apenas uma GEI. Para as faces utilizamos a técnica da detecção, que consiste em extrair da imagem original um segmento retangular que contém a face, como mostrado na Figura 4. Já para a obtenção das silhuetas, é necessário utilizar segmentação, que é uma técnica que identifica os pixels que compõem a figura humana, isolando-os do restante da imagem ao redor [Tu et al., 2006], conforme pode ser visto na Figura 5.

Figura 4 – Exemplo de detecção facial, adaptado de [Guo et al., 2019]. As faces detectadas são mostradas com um quadrado vermelho. O algoritmo falha em detectar as faces em ângulos oblíquos, como as duas faces no canto inferior direito.

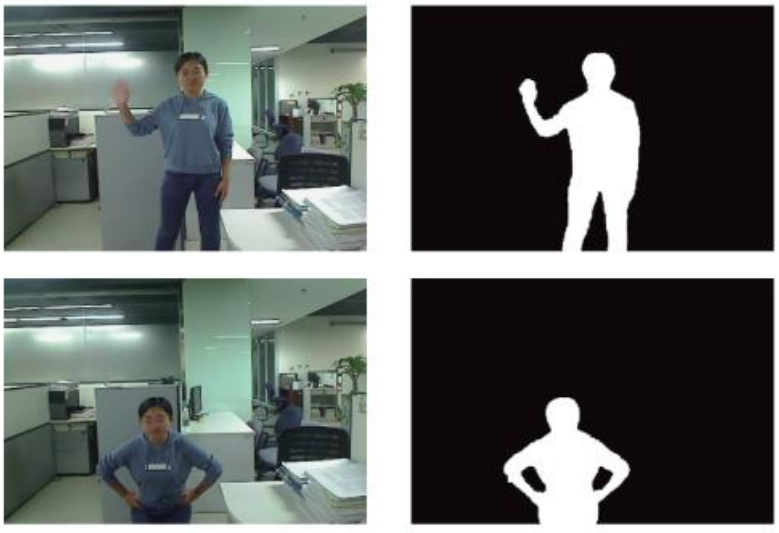


Figura 5 – Exemplos de segmentação. Adaptado de [Zhao et al., 2019].

2.3.2 Datasets para ReID

A combinação das características da face e modo de caminhar foi avaliada na tarefa de reconhecimento de pessoas em longo prazo. Para tanto, é necessário utilizar *datasets* específicos para o ReID em longo prazo. Há diversos *datasets* disponíveis para treinamento e avaliação de modelos de ReID de pessoas. Para o objetivo proposto por esse trabalho, que é reidentificar pessoas em imagens geradas por diferentes câmeras, em momentos distintos, é importante que o *dataset* contenha:

- sequências de vídeo;
- sequências obtidas por múltiplas câmeras;
- múltiplas roupas por indivíduo;
- mesmos indivíduos filmados em várias ocasiões.

Dentre os principais *datasets* de domínio público para ReID, os que melhor satisfazem essas condições são os descritos na Tabela 4, sendo que, desses, o mais adequado à avaliação é o TUM GAID [Hofmann et al., 2014]. O dataset TUM GAID, até onde foi possível pesquisar, é o único dataset de domínio público que contempla trocas de roupas e captura de dados em múltiplas ocasiões. Uma amostra de seus *frames* pode ser vista na figura 6. Existem dois outros *datasets* com essas características, porém não estão disponíveis para pesquisa. No caso de [Schumann and Monari, 2014], devido a restrições de confidencialidade.

Dataset	Indivíduos	Sequências	Câmeras	MR	MO
PRID2011 [Hirzer et al., 2011]	200	400	2	não	não
iLIDS-VID [Wang et al., 2014]	300	600	2	não	não
Airport [Karanam et al., 2019]	9.651	n.d.	6	não	não
RPIField [Zheng et al., 2018]	112	6.577	12	não	não
TUM GAID [Hofmann et al., 2014]	305	3370	1	sim	sim

Tabela 4 – Principais *datasets* para ReID de pessoas em domínio público, com correspondentes quantidades de identidades e sequências, número de câmeras e atributos (MR = múltiplas roupas, MO = múltiplas ocasiões).

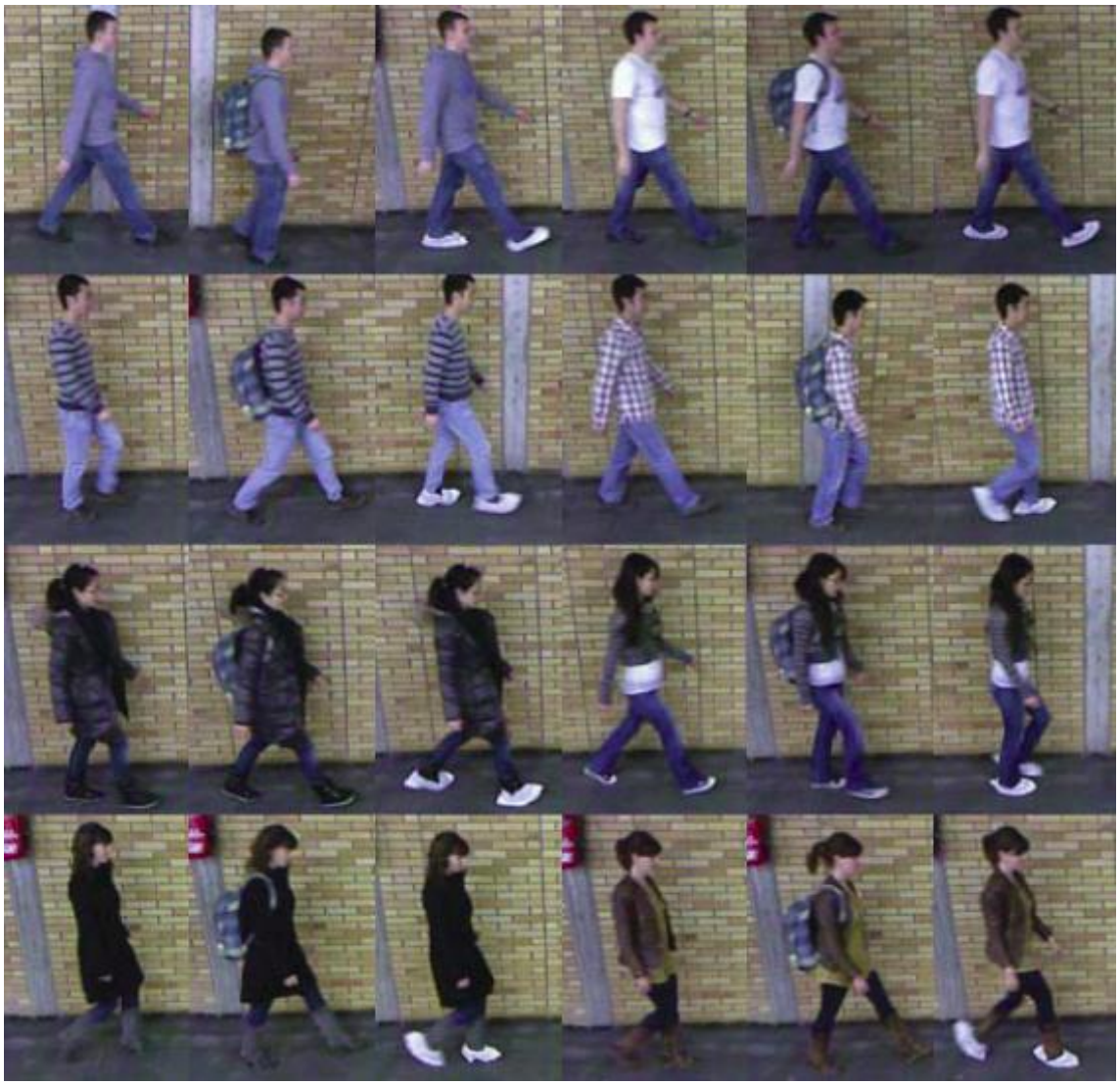


Figura 6 – Amostra do dataset TUM GAID. Cada linha apresenta imagens de um mesmo indivíduo, usando diferentes roupas e calçados [Hofmann et al., 2014]

3- Trabalhos Relacionados

Como descrito no Capítulo 2, este trabalho combina diversas técnicas já conhecidas com o objetivo de aumentar a acurácia na reidentificação de pessoas em longo prazo. Neste capítulo, são examinados alguns trabalhos correlatos. Por propósitos didáticos, este capítulo encontra-se dividido nas seguintes seções: reidentificação de curto prazo, reidentificação de longo prazo, modo de caminhada e reconhecimento de face e, por fim, comentários finais.

3.1- Reidentificação de Curto Prazo

Na maior parte das aplicações de ReID de interesse, é necessário detectar a pessoa de prova de modo automático. Normalmente, tal processo é feito de duas formas distintas: por meio de subtração de *background*, ou por meio de abordagens baseadas em reconhecimento de padrões [Figueira et al., 2016].

Mesmo assumindo uma delimitação precisa da pessoa na imagem, restam ainda os problemas de variação de iluminação, de perspectiva, das respostas diferentes às cores em diferentes câmeras e de oclusão, os quais existem mesmo quando os indivíduos mantém suas vestimentas ao longo do tempo de captura das imagens [Lee et al., 2017]. O problema da oclusão foi tratado em [Yang et al., 2019], que extrai *features* espaciais multidimensionais de representação discriminativa da imagem do pedestre por meio de LSTMs (*Long Short-Term Memory*), um tipo particular de rede neural que preserva a informação temporal. Uma estratégia similar foi adotada por [Song et al., 2020], referência que introduziu o conceito de imagem chave derivada da sequência de imagens extraídas de um único indivíduo. Em [Guo et al., 2020b], investigou-se a melhora de desempenho de métodos de re-ranqueamento (*reranking*) através da análise da relação entre as amostras mais próximas. A referência [Li et al., 2015] elaborou um sistema baseado em máquinas de vetores suporte latentes para mapear as relações entre características de baixo e médio nível das vestimentas dos indivíduos,ou seja, utilizando explicitamente as roupas

como variáveis descritivas dos indivíduos.

A referência [Zhu et al., 2020] trata do problema de geral de reidentificação de objetos, enfocando, num contexto da *Internet de Veículos*, pessoas e automóveis. Tal reidentificação é melhorada mediante a adoção de *features* profundas aprendidas de modo parcialmente alinhado (DLPA, do inglês *deeply learned part-aligned features*) e do método de incorporação sensível à pose [Sarfraz et al., 2018], que contempla informação fina e grosseira referente à pose dos pedestres para apresentar um modelo de reidentificação suficientemente discriminativo. Outra abordagem de extração de características é proposta em [Lee et al., 2017], com informações regionais do corpo sendo capturadas por modelos de misturas de gaussianas dos histogramas de cores, de modo a representar o modo de cor dominante de cada indivíduo.

O fato de a extração de *features* ser limitada devido ao ângulo de captura poder ser muito distinto entre câmeras diferentes é contornado em [Guo et al., 2020a], por meio de redes adversárias generativas (GANs), que produzem imagens dos pedestres de frente, de lado e de trás (ou seja, ocorre uma geração de múltiplas perspectivas de cada indivíduo), enquanto um discriminador multi-classes determina se a imagem de entrada é uma imagem real ou inventada.

A literatura acerca do problema de reidentificação de curto prazo é muito extensa. No entanto, como essa dissertação tem como objetivo a reidentificação de longo prazo, as conclusões da literatura de curto prazo podem não ajudar. Por esse motivo, outras referências relevantes para curto prazo não serão detalhadas.

3.2- Reidentificação de longo prazo

Embora o tema de reidentificação de pessoas venha atraindo cada vez mais atenção, são poucos os trabalhos que tratam especificamente do tema da reidentificação em longo prazo (*long-term person re-identification*), possivelmente devido à extrema dificuldade do problema. A reidentificação de pessoas é uma tarefa da área de visão computacional, já que busca dar sentido a imagens obtidas através de câmeras. Já a reidentificação em longo prazo não pertence *necessariamente* a essa área, já que as *features* mais adequadas a essa tarefa podem estar mais relacionadas à area de

biometria, que pode utilizar outros tipos de sensores para capturá-las. Em geral, as técnicas de reidentificação podem ser categorizadas entre as que focam na representação (ou obtenção de *features* apropriadas) [Gray and Tao, 2008]; no aprendizado de métricas de distância [Köstinger et al., 2012; Li and Wang, 2013]; ou as que tratam de ambos ao mesmo tempo [Liu et al., 2019].

Mesmo em artigos de *surveys*, a reidentificação de longo prazo é pouco mencionada [Satta, 2013; Bedagkar-Gala and Shah, 2014; Wu et al., 2019; Karanam et al., 2019]. Apesar disso, é possível encontrar alguns trabalhos que buscam soluções que podem ser aplicadas na solução desse problema utilizando a visão computacional e dispositivos bastante especializados. [Munaro et al., 2014], por exemplo, utiliza câmeras 3D de baixo custo para obter dados de profundidade e com eles criar uma representação tridimensional da forma de caminhar (*gait*) dos indivíduos, Esta dissertação também utiliza o gait de forma bastante parecida com a proposta neste trabalho, com exceção do tipo de câmera. Os testes conduzidos em tiveram o objetivo de reidentificar pessoas no curto prazo, mas a técnica utilizada é adequada para ReID em longo prazo. Diversos outros trabalhos exploram a utilização de câmeras RGB-D, que capturam cores e profundidade na tarefa de reidentificação de pessoas, mas essa tecnologia tem mais sucesso quando aplicada à identificação de ações em vídeos [Pala et al., 2019; Han et al., 2017; Bondi et al., 2018; Hasan and Babaguchi, 2016].

A referência [Zhang et al., 2020] aborda a reidentificação de longo prazo por meio do aprendizado de uma representação híbrida de *features* através de uma rede chamada SpTSkM. A SpTSkM incorpora um fluxo espaço-temporal e um fluxo relacionado ao movimento do esqueleto. Enquanto este último fluxo opera em esqueletos 3D normalizados por meio de redes convolucionais em grafos, o primeiro opera diretamente em sequências de imagens, almejando o aprendizado de padrões de identidade relacionados à estrutura geométrica do corpo e ao seu movimento.

[Martinel et al., 2017] introduz um *framework* de reidentificação distribuída de pessoas que incorpora as informações trazidas pela topologia da rede e pela estrutura do ambiente. Estas informações podem ser exploradas para melhorar o processo de reidentificação. Neste caso, as trajetórias dos diversos indivíduos que se encontram na área sob monitoramento geram diferentes graus de correlação entre pares de câmeras. Os diferentes graus de correlação podem ser usados para melhorar a acurácia das técnicas de ReID. Embora este artigo não trate especificamente de ReID de longo prazo,

a ideia principal deste artigo é extensível para esta configuração.

Um dos principais entraves à pesquisa sobre reidentificação em longo prazo é a baixa disponibilidade de *datasets* para treinamento de redes neurais. Os poucos datasets existentes apresentam uma quantidade muito pequena de exemplos rotulados para o treinamento. Para ser útil ao treinamento de reID em longo prazo, é importante que o dataset contenha imagens capturadas em várias ocasiões, de modo que os indivíduos estejam usando roupas e acessórios diferentes e, eventualmente, até cortes de cabelos e estilos de barba diferentes. Em [Barbosa et al., 2018], optou-se por criar um *dataset* sintético para treinamento, o SOMAset, e o modelo desenvolvido obteve desempenho expressivo quando testado com *datasets* populares. Outra forma original de contornar esse problema foi adotada em [Gou et al., 2016]. Nesse artigo os autores selecionaram, a partir de um *dataset* popular, apenas as imagens das pessoas utilizando roupas totalmente pretas. Desta forma, a rede neural foi forçada a encontrar *features* discriminantes de cada indivíduo que não incluíssem as cores das roupas. Essa abordagem assemelha-se bastante à de utilizar silhuetas e *gait energy images*, que, por serem em preto-e-branco, não carregam informação de cores para alimentar a rede neural.

3.3- Combinação do Modo de caminhada e reconhecimento de face

A principal proposta desta dissertação consiste na combinação de características de modo de caminhada e de face para a reidentificação de longo prazo com câmeras comuns. Tais características parecem, a princípio, ofertar informações discriminativas (mesmo que insuficientes para a resolução total do problema) para o processo de ReID, já que são bastante específicas para cada indivíduo. Outras representações, tais como *features* densas invariantes, poderiam ser utilizadas [Tan et al., 2018]. Como o campo da ReID de longo prazo é ainda pouco explorado, não é claro qual estratégia resultaria em um melhor desempenho.

Outros trabalhos já investigaram a fusão de *gait* e *face recognition* com o propósito de reidentificar pessoas, sendo por conseguinte mais similares ao foco desta dissertação, No entanto, não foi possível encontrar nenhum trabalho que, como essa dissertação, utilizasse redes neurais para fazer reidentificação de pessoas em longo prazo. [Eisenbach

et al., 2012] menciona considera viável tal possibilidade, embora não a explore. [Liu and Sarkar, 2007] comprovou que a combinação das duas características produz resultados melhores, porém utilizou um modelo de reconhecimento facial treinado com um *dataset* de apenas 70 identidades, e é restrito à ReID de curto prazo. [Xing et al., 2015] propõe a fusão da identificação de *gait* e de face a nível de característica, mas utiliza um conjunto de dados composto de 40 identidades obtidas da combinação do *dataset* CASIA-B [Yu et al., 2006] com o *dataset* de faces ORL [AT&T Laboratories Cambridge, 1994] como entrada para o treinamento. [Imani et al., 2019] propõe uma abordagem geral de representação esparsa combinando aprendizagem de dicionários com tensores, que contempla simultaneamente os problemas de representação esparsa de características e a combinação de diferentes níveis de *features*.

O primeiro esforço para o *reconhecimento pelo modo de caminhar* em visão computacional foi apresentado no início dos anos 1990 [Wagg and Nixon, 2004]. Atualmente, normalmente reconhece-se que modo da caminhada e a dinâmica da caminhada são consideradas informações essenciais para técnicas biométricas baseadas em *gait* [Zongyi Liu and Sarkar, 2006]. No entanto, as condições nas quais o reconhecimento por gait é passível de solução não sejam completamente compreendidas ou caracterizadas [Sarkar et al., 2005].

Dentre os trabalhos que escolhem o modo de caminhar como principal característica para reidentificação, destaca-se [Song et al., 2019]. Este é um trabalho que introduz um novo modelo de rede neural "ponta-a-ponta" que encapsula todas as etapas do processo de reconhecimento de pessoas através de *gait*, ao invés de tratá-las como atividades separadas. A arquitetura de rede utilizada em [Song et al., 2019] é mostrada e comparada na Figura 7.

A GaitNet representa o estado-da-arte em reconhecimento de pessoas através de *gait*, ultrapassando o desempenho das outras abordagens nos *datasets* CASIA-B e SZU RGB-D Gait, onde atinge a acurácia máxima de 97,1% em rank-1. A expressão "rank-n" denota que a identidade da pessoa em análise encontra-se entre as n primeiras propostas pelo algoritmo de ReID, de acordo com as probabilidades por ele calculadas. Deste modo, um acerto em rank-1 indica que a pessoa proposta pelo algoritmo correspondência era de fato a pessoa analisada. Ou seja, indica que o algoritmo reconheceu corretamente a pessoa; um acerto em rank-2 indica que o algoritmo errou a pessoa, mas ela estava entre as duas com mais alta probabilidade de correspondência; acerto em rank-3, indica que a

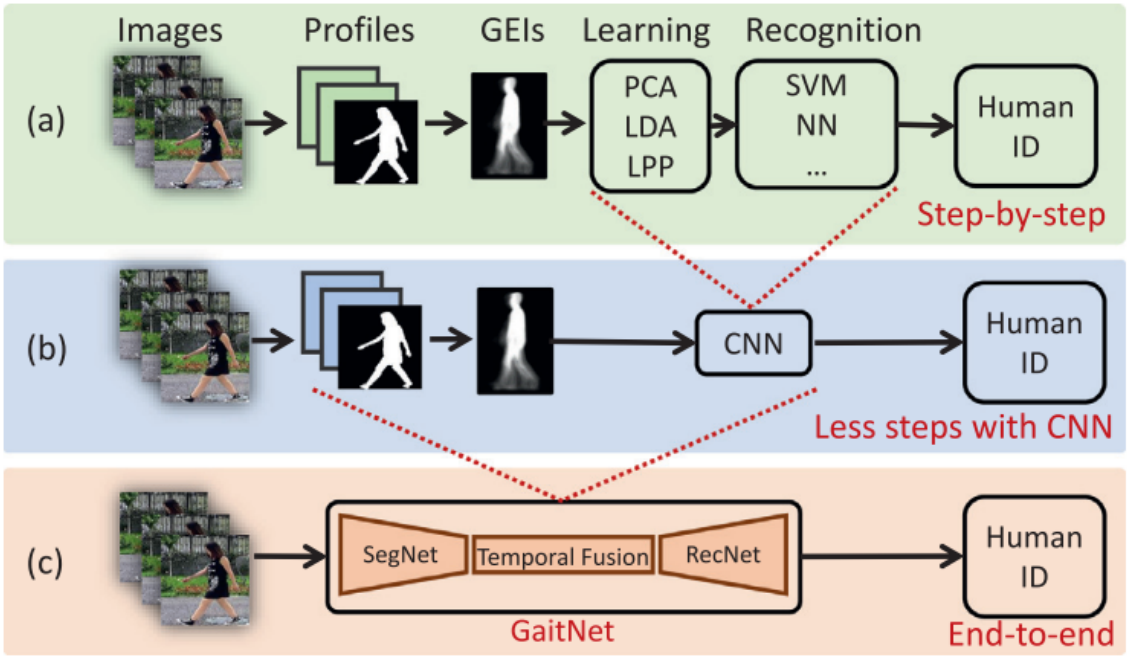


Figura 7 – Comparação das arquiteturas de soluções para reconhecimento de pessoas através de *gait*. Na primeira linha, encontra-se o processo manual de extração de *features*; na segunda linha, temos a utilização de rede neural convolucional e na terceira, o modelo integrado GaitNet [Song et al., 2019].

pessoa correta estava entre as 3 com mais alta probabilidade, e assim por diante. Como esta métrica é muito empregada na literatura referente à reidentificação, esta dissertação a emprega como métrica de desempenho.

3.4- Comentários Finais

Este capítulo descreveu alguns trabalhos correlatos na literatura. Infelizmente, a maior parte das conclusões é restrita ao problema de reidentificação de curto prazo. Após a contextualização fornecida neste capítulo, iremos no próximo detalhar a metodologia de pesquisa efetuada nesta dissertação.

4- Metodologia de Pesquisa

Como já mencionado, este trabalho tem o objetivo de comprovar a eficácia da combinação de redes neurais, GEIs e reconhecimento facial na tarefa de reidentificar pessoas em sequências de imagens obtidas em diferentes dias, por múltiplas câmeras. Este capítulo descreve os métodos utilizados para alcançar o objetivo desejado. Adicionalmente, é descrita a metodologia de testes e é realizada uma comparação com abordagens propostas em outros trabalhos.

4.1- Visão Geral do Experimento

A Figura 8 apresenta uma visão geral do experimento desenvolvido para testar a hipótese proposta. O experimento consiste em:

1. Implementar e treinar uma rede neural convolucional especializada na reidentificação de GEIs;
2. Implementar e treinar uma rede neural convolucional especializada no reconhecimento facial a distância;
3. Implementar e treinar uma terceira rede neural especializada na reidentificação de pessoas que utiliza o aprendizado adquirido pelas duas redes anteriores por meio da técnica de *transfer learning*.

O procedimento de reidentificação, seja ele de faces ou de pessoas, de modo geral é composto de três etapas principais: detecção, extração de *features* e classificação [Prihasto et al., 2018]. Estas três etapas podem ser combinadas em uma única rede neural, como em Song et al. [2019], ou podem ser executadas separadamente, utilizando múltiplas redes neurais especializadas e combinando a saída de uma rede como entrada de outra. O presente trabalho opta pela abordagem em etapas separadas, de modo a conseguir tirar proveito dos *datasets* disponíveis em maior abundância para cada tarefa

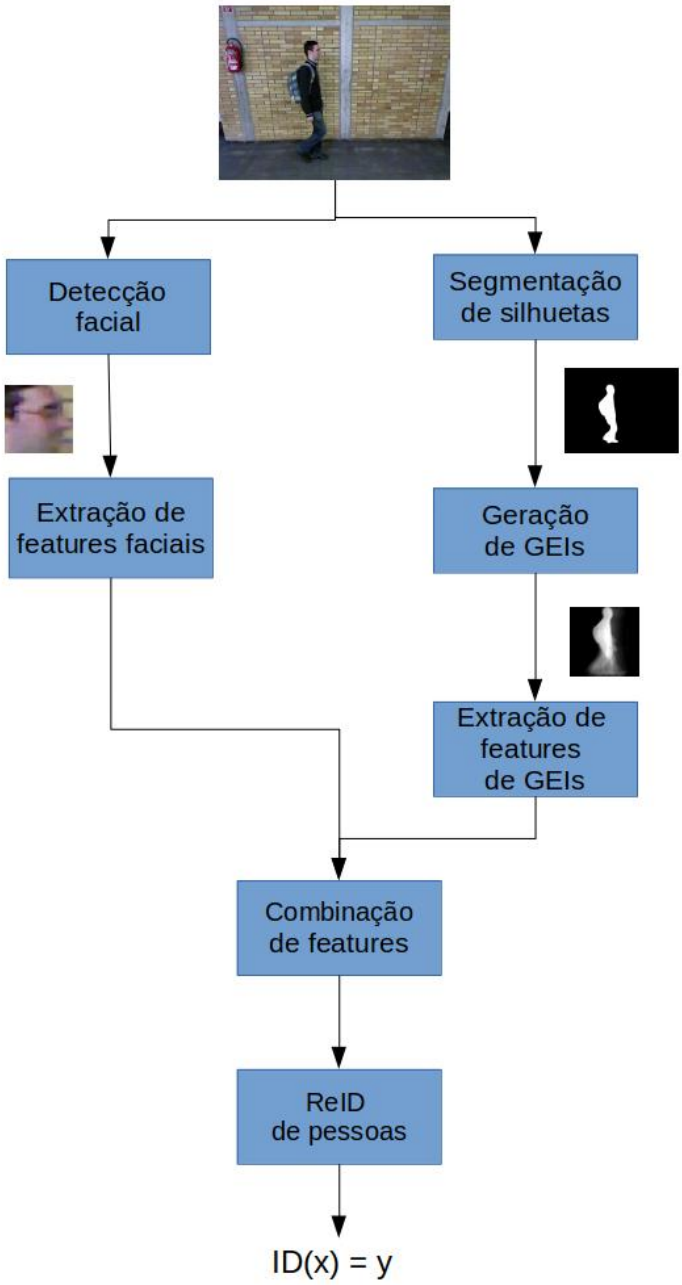


Figura 8 – Visão geral do experimento desenvolvido para testar a hipótese proposta.

isolada. Dessa maneira, em nosso experimento teremos as seguintes atividades isoladas em módulos:

- Detecção de faces;
- Extração de *features* de faces;
- Segmentação de figuras humanas;

- Geração de GEIs;
- Extração de *features* de GEIs;
- Combinação das *features* das faces e GEIs;
- Classificação ou ReID de pessoas.

As seções a seguir apresentam em maior detalhe cada módulo do experimento.

4.2- Datasets

O treinamento de redes neurais para executar uma tarefa específica depende da existência de *datasets* apropriados. Neste trabalho, dois *dataset*s são utiliados:

- **TUM GAID**, contém vídeos de pessoas caminhando e é utilizado para ReID de pessoas. É o principal *dataset* para o trabalho, pois é o utilizado para treinar e avaliar o desempenho da rede de ReID em longo-prazo.
- **VGGFace2**, contém milhares de faces e é utilizado para reconhecimento facial.

4.2.1 Dataset TUM GAID

O TUM GAID [Hofmann et al., 2014] é o principal *dataset* utilizado neste trabalho. Ele contém 264.857 imagens, de 305 indivíduos em 3 variações de situação (normal, carregando bolsa e com diferentes calçados), além de variação temporal.

Além das imagens RGB, o TUM GAID também fornece dados de profundidade, obtidos com uma câmera 3D. Os dados de profundidade foram desconsiderados no desenvolvimento deste trabalho, pois pretendemos investigar uma solução que possa ser empregada na maioria dos sistemas de segurança atualmente em vigência, que são desprovidos de câmeras 3D.

A variação temporal no *dataset* é restrita a 32 indivíduos, cujas imagens foram capturadas em duas ocasiões, com três meses de intervalo, usando roupas, penteados e

sapatos diferentes. Na segunda ocasião, a temperatura estava 15 ºC mais alta, então as roupas utilizadas são significativamente diferentes, como pode ser visto na Figura 6.

Para a avaliação de desempenho de ReID em longo prazo, utilizamos as sequências adicionais dos 32 indivíduos capturadas na segunda ocasião. Deste modo, para cada indivíduo do TUM GAID, temos os seguintes diretórios, cada um contendo uma sequência:

- **b01, b02, b03 e b04** - Carregando bolsa ou mochila.
- **n01, n02, n03, n04, n05 e n06** - Caminhando normalmente.
- **n07, n08, n09, n10, n11 e n12** - Caminhando normalmente (capturadas na segunda ocasião, disponíveis para apenas 32 identidades).
- **s01, s02, s03 e s04** - Variação no tipo de calçado.

Para as atividades neste trabalho, utilizamos apenas as sequências do tipo 'Normal' (n01 a n06), ou seja, desprezamos as variações de sapatos e bolsas ou mochilas carregadas. As sequencias nos diretórios n07 a n12 são reservadas para serem utilizadas apenas na avaliação do desempenho da rede em longo prazo.

4.2.2 Dataset VGGFace2

O *dataset* VGGFace2 [Cao et al., 2018] é um dos mais adequados ao propósito deste trabalho, pois possibilita o treinamento de modelos com elevada acurácia em tarefas relacionadas à detecção e ao reconhecimento facial. Este *dataset* contém 3,31 milhões de imagens de 9.131 indivíduos coletadas em diferentes idades, obtidas em ambiente não controlado, reproduzindo as condições que seriam encontradas no mundo real. Um exemplo das imagens deste *dataset* pode ser visto na Figura 9.

4.3- Particionamento de Dados

Todos os treinamentos que foram realizados com o *dataset* TUM GAID obedeceram ao seguinte esquema de particionamento:

Figura 9 – Exemplos de imagens do *dataset* VGGFace2. Cada linha contém fotos de um mesmo indivíduo.

- Treino: 70%
- Validação: 15%
- Teste: 15%

Para obter os resultados de reidentificação em longo prazo, foram utilizados apenas os dados correspondentes à segunda sessão de filmagem, ocorrida 3 meses após a primeira. Estes dados foram reservados apenas para teste e não foram utilizados em nenhum momento durante o treinamento, para que desta forma se pudesse reproduzir um cenário real de ReID em longo prazo.

4.4- Reconhecimento Facial

4.4.1 Detecção de Faces

Na solução proposta, a primeira etapa do processamento consiste em detectar em cada *frame* de cada sequência a região onde se encontra o rosto do indivíduo. Nesta etapa, o sistema proposto irá construir *bounding boxes* para delimitar a região de interesse. Embora seja este o procedimento mais amplamente adotado na literatura, alguns métodos alternativos são por vezes utilizados, tais como listras que abarcam toda a dimensão horizontal [Bird et al., 2005] e a adoção de uma grade regular, a partir da qual as *features* de cor e de textura são extraídas de blocos com sobreposição [Köstinger et al., 2012]. No entanto, tais técnicas também exigem procedimentos de segmentação, responsáveis por obter uma máscara que separa a região do indivíduo da região de fundo [Tao et al., 2015]. O procedimento de segmentação pode utilizar informações espaciais e temporais para obter desempenho robusto, no caso de sinais de vídeo [Gheissari et al., 2006].

Para realizar o procedimento de determinação das bounding boxes, é utilizado o modelo de rede neural MTCNN [Zhang et al., 2016], pré-treinado com o *dataset* VGG-Face2. Este modelo é capaz de analisar as imagens do *dataset* e detectar faces de pessoas, mesmo que não estejam olhando diretamente para a câmera.

O MTCNN, além de localizar as partes da imagem que contém faces também

consegue obter a posição de pontos específicos do rosto (*facial landmarks*), como por exemplo, olhos, ponta do nariz e cantos da boca. A estrutura do MTCNN é composta por 3 redes: P-Net, R-Net e O-Net. A primeira processa a imagem inteira e propõe diversas *kiibounding boxes* candidatas de forma rápida; a segunda faz um refinamento da saída da rede anterior, descartando as *bounding boxes* propostas que não contém faces; por fim, a terceira rede refina o resultado e determina a posição da *bounding boxes* finais e dos *facial landmarks*. A Figura 10 oferece uma representação visual das etapas do processamento do modelo MTCNN.

Nenhum treinamento adicional ou modificação foi realizado para adequar o modelo MTCNN ao escopo deste trabalho. Ainda assim, é esperado que ele obtenha uma boa capacidade de generalização graças à grande quantidade de exemplos disponíveis no VGGFace2.

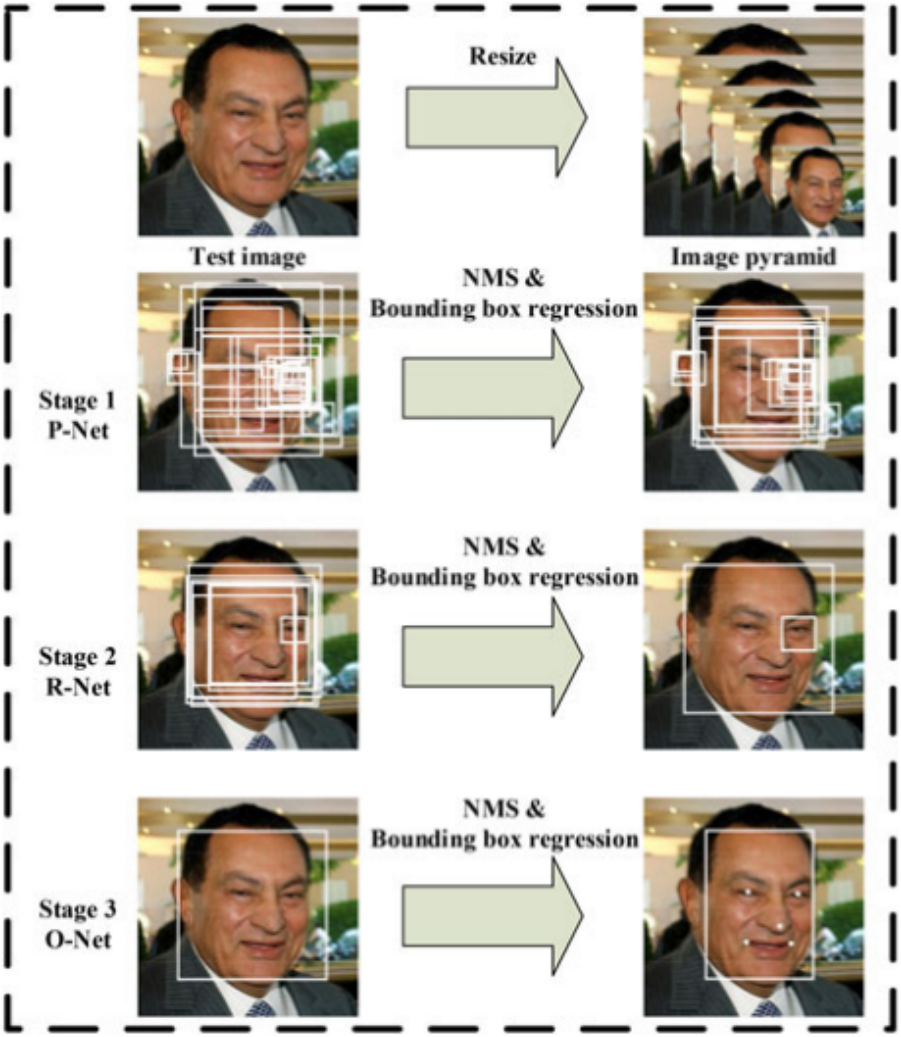


Figura 10 – Detalhamento das etapas do processamento do modelo MTCNN, em imagem extraída de [Zhang et al., 2016].

Com as coordenadas obtidas pelo MTCNN, é feito o recorte da parte da imagem original que contém a face da pessoa, de forma que a saída esperada deste módulo é uma nova imagem, contendo apenas a face da pessoa. Para cada imagem original fornecida é criado um novo arquivo com a face segmentada, conforme mostrado na Figura 11.

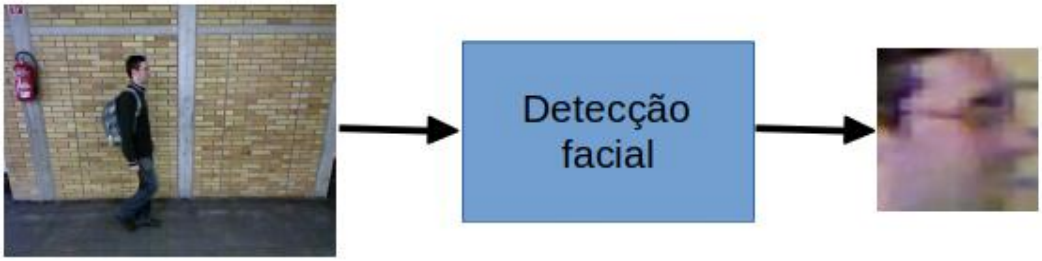


Figura 11 – Exemplo de entrada e saída do módulo de detecção facial.

4.4.2 Extração de características de faces

Enquanto o componente de detecção tem o único objetivo de obter a parte da imagem original que contém a face, este segundo componente tem atribuições mais complexas: O modelo de rede neural utilizado nessa etapa precisa ser capaz de receber as imagens das faces e aprender as características que a permitam diferenciar um indivíduo de outro. Ou seja, o objetivo nessa etapa é fazer com que a rede neural aprenda intuitivamente uma métrica que aproxime os rostos com o mesmo rótulo, isto é, pertencentes à mesma pessoa, e afaste os demais.

Para avaliar o modelo em sua capacidade de extrair *features* relevantes, é necessário treiná-lo como um classificador para reconhecimento facial. Isto é, dada uma imagem de um rosto, o modelo precisa avaliar probabilisticamente se esse rosto corresponde a uma das pessoas previamente armazenadas em uma galeria. Uma boa acurácia nessa classificação indica que temos um bom modelo. Para usar o modelo na aplicação desejada, as camadas da rede neural utilizadas para fazer essa classificação são removidas de modo que ele funcione como um extrator de *features*.

O conjunto de dados utilizado é um elemento de fundamental importância para o sucesso do projeto, pois deve conter quantidade suficiente de amostras de cada uma das características que a rede neural precisa aprender a distinguir. Por exemplo, para o ReID em longo prazo, o *dataset* precisa conter amostras da mesma pessoa utilizando roupas diferentes.

Uma vez treinado, este modelo de rede neural, já sem a camada de classificação, receberá como entrada a imagem da face de uma pessoa e produzirá como saída um

vetor contendo características que a individualizam. Este vetor é utilizado adiante, como entrada para a tarefa de reidentificação de pessoas em longo prazo. Ou seja, temos um extrator e codificador (*encoder*) de características faciais.

O modelo de rede neural convolucional selecionado para executar as atividades descritas acima foi o FaceNet [Schroff et al., 2015], pré-treinado com o *dataset* VGGFace2 [Cao et al., 2018]. O número de classes na última camada *fully connected* do modelo foi modificado para refletir as 305 classes do *dataset* TUM GAID. A camada *fully connected* foi retreinada utilizando o *dataset* TUM GAID.

O modelo FaceNet é uma implementação de uma rede residual Resnet com camadas de convolução paralelas, de acordo com a arquitetura Inception. Para este trabalho foi utilizada uma Resnet-50, adptada de [Esler, 2020]. O treinamento da última camada do FaceNet deu-se em 20 épocas, com acurácias parciais medidas a cada época, mostradas na figura 12.

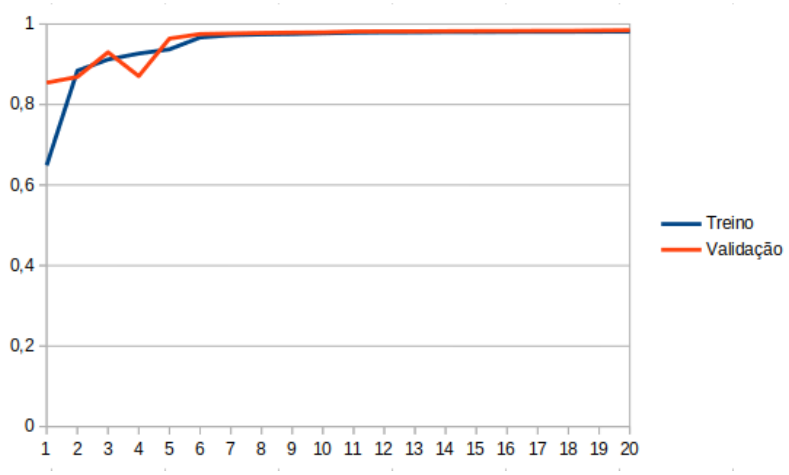


Figura 12 – Dados do treinamento da rede FaceNet, para reconhecimento facial, no *dataset* TUM GAID. No eixo horizontal, as épocas. No eixo vertical, a acurácia (rank-1)

Embora o objetivo do trabalho não seja fazer reconhecimento facial, o desempenho deste modelo é importante pois serve como base de comparação para avaliação do modelo proposto por este trabalho, indicando se a fusão de características de face e *gait* pode apresentar resultados melhores que apenas o reconhecimento facial na tarefa de reidentificar pessoas. A acurácia alcançada no conjunto de testes, para ReID de curto-prazo, foi de 96,66%, em rank-1.

Após obter resultados satisfatórios na classificação, o modelo adaptado foi utilizado como um extrator de *features*, de modo a produzir vetores representando cada uma das identidades do *dataset*. Em resumo, esta etapa do modelo proposto recebe como entrada

as diversas faces extraídas de cada uma dos frames no *dataset* TUM GAID e gera, para cada face, um vetor correspondente, de 512 bytes. A Figura 13 mostra um esquema do módulo de extraçao de *features* faciais e sua respectiva saída.

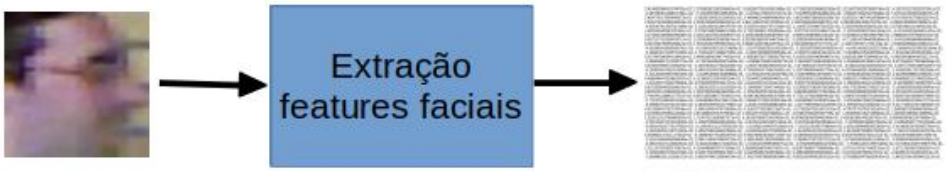


Figura 13 – Esquema do módulo de extração de features. A entrada é uma imagem da face, recortada do frame original. A saída, um vetor de 512 posições, contendo uma representação da imagem de entrada.

4.5- Reconhecimento de *Gait*

4.5.1 Segmentação de figuras humanas

A segunda parte do trabalho ora proposto consiste em obter o vetor de *features* que representa a forma de caminhar da pessoa. Ele é obtido em três etapas:

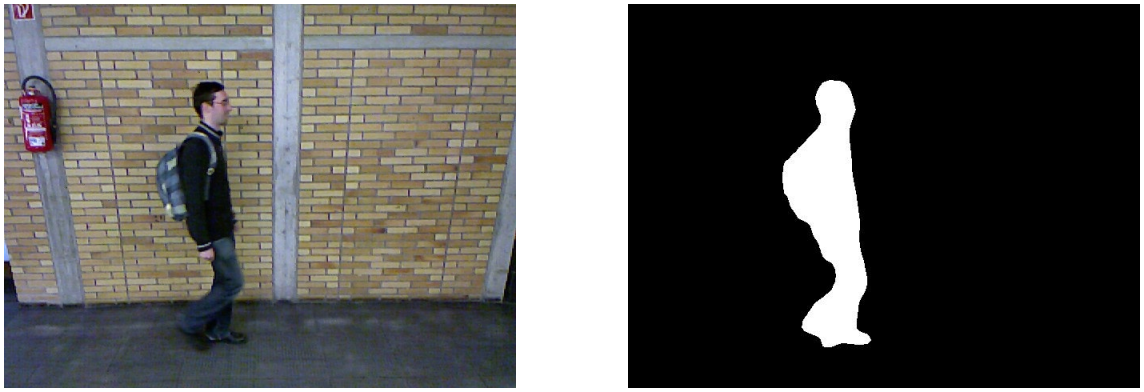
1. Segmentação
2. Geração de GEIs
3. Extração de *features* de GEIs

4.5.2 Segmentação de silhuetas

A etapa de segmentação recebe como entrada as sequências de *frames* do vídeo com a pessoa caminhando, processa essas imagens individualmente e, para cada uma delas, fornece como saída uma imagem de mesmo tamanho da imagem original,

contendo apenas a silhueta da figura humana segmentada em branco, com o fundo preto. Utilizando uma sequência de imagens como entrada, obteremos uma sequência de mesmo comprimento como saída.

Um modelo de segmentação consagrado, o DeepLabV3 [Chen et al., 2017] foi utilizado para gerar as silhuetas. Um exemplo de silhueta gerada pode ser visto na Figura 14.



(a) Frame original (b) Silhueta

Figura 14 – Exemplo de entrada e saída do módulo de segmentação de silhuetas.

4.5.3 Geração de GEIs

A GEI é a sobreposição das silhuetas obtidas na segmentação. Quatro atividades principais compõem essa etapa: alinhamento, normalização, delimitação do ciclo de *gait* e o cálculo da média dos valores dos *pixels.* Nenhuma dessas atividades demanda o uso de redes neurais ou qualquer outro algoritmo de aprendizado. Os cálculos são realizados com base na posição dos *pixels* da extremidade da figura delimitada na silhueta. Para o alinhamento e normalização, são utilizados os *pixels* dos extremos superior e inferior e o centroide da metade superior da silhueta. Já para a delimitação do início e fim do ciclo de marcha, é utilizada a distância entre os *pixels* extremos (esquerda e direita) do quarto inferior da imagem. As silhuetas alinhadas são utilizadas para obter o valor médio de cada *pixel* da imagem. A figura final obtida é a GEI. A Figura 15 mostra o esquema do módulo de geração de GEIs, com sua respectivas entrada e saída.

Embora uma única sequência possa conter vários ciclos de marcha que podem

Figura 15 – O módulo de geração de GEIs recebe uma sequência de silhuetas como entrada e fornece uma GEI como saída.

originar várias GEIs, neste trabalho, optou-se por utilizar todos os *frames* disponíveis em cada sequência para gerar as GEIs. Portanto, é gerada apenas uma GEI para cada sequência de vídeo. A justificativa para esta abordagem é aproveitar melhor a quantidade limitada de frames disponíveis, que se tornou escassa pois, para fazer o alinhamento, é preciso que a silhueta inteira esteja dentro da imagem e, por isso, muitos dos *frames* originais, como os da figura 16, são descartados.

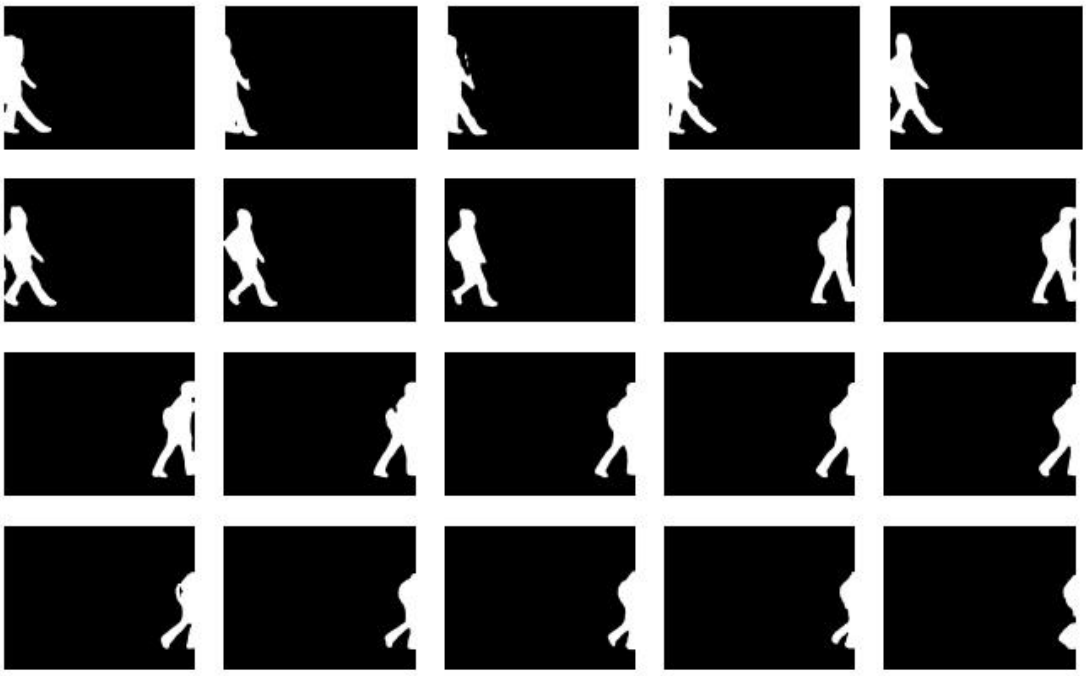
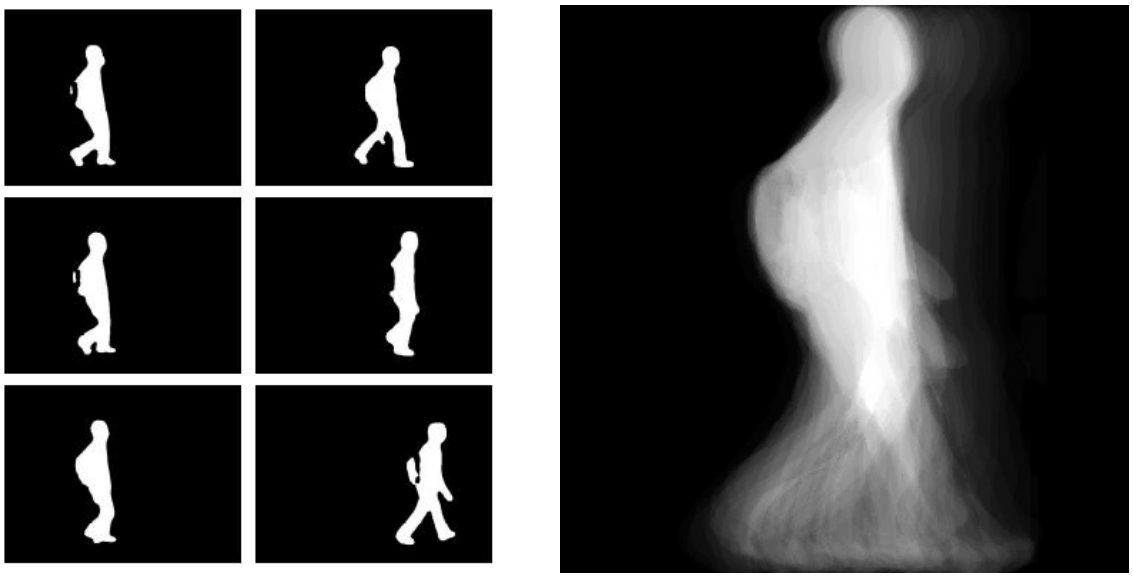


Figura 16 – Exemplos de frames descartados por não enquadrarem a silhueta completa.

A etapa de geração de GEIs recebe como entrada uma sequência contínua de silhuetas obtida na etapa anterior e fornece como saída uma GEI, conforme a figura 17. O

número total de GEIs corresponde ao número total de sequências disponíveis no *dataset*.



(a) Sequência de silhuetas (b) GEI

Figura 17 – Exemplo de entrada e saída do módulo de geração de GEIs.

4.5.4 Extração de *features* de GEIs

O módulo de extração de *features* de GEIs executa a tarefa de aprendizado que possibilita reconhecer uma pessoa através de sua GEI e é implementado por meio de uma rede neural convolucional que recebe uma GEI como entrada e fornece um vetor de características como saída, conforme a Figura 18. Esta seção fornece mais detalhes sobre esta implementação.

A GEI é uma representação da forma de caminhar de cada pessoa. O objetivo desta etapa do trabalho é extrair, de cada uma das GEIs, *features* que possam ser utilizadas para reconhecer cada indivíduo.

[Min et al., 2019b] utiliza uma rede neural de 10 camadas para fazer o reconhecimento de pessoas através de sua forma de caminhar. Nesta dissertação, para obter as *features*, reproduzimos a rede neural do trabalho supracitado, cuja arquitetura pode ser vista em detalhes na Figura 19. Apenas uma modificação foi realizada para adequar esta rede neural ao propósito desta pesquisa: O número de classes de saída foi alterado de

Figura 18 – O módulo de extração de *features* de GEIs recebe GEIs como entrada e fornece um vetor de características de 305 posições como saída.

1024 para 305, correspondentes ao número de indivíduos no *dataset* TUM GAID para possibilitar a avaliação do desempenho da rede durante o treinamento.

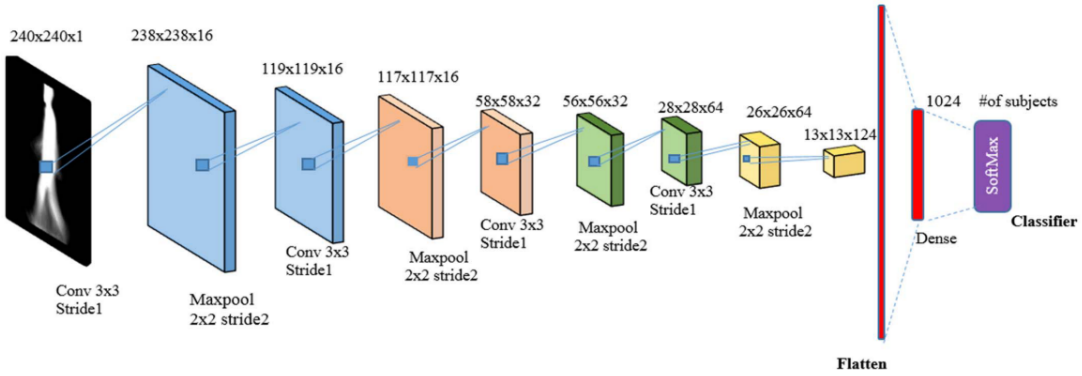


Figura 19 – Arquitetura de rede proposta em [Min et al., 2019a]

A saída da rede é um vetor de 305 posições. Passando este vetor por uma função ArgMax, obtém-se a predição da identidade da pessoa, que é utilizada para avaliar o desempenho da rede. Uma vez treinada e apresentando um bom desempenho, esta rede é utilizada como um extrator de características. Ela é alimentada com cada uma das GEIs obtidas na etapa anterior de modo a produzir vetores de características que representam os indivíduos. Assim como no caso das faces, o reconhecimento de pessoas utilizando apenas suas GEIs não é o objetivo desta dissertação, mas a acurácia apresentada por este módulo pode auxiliar na avaliação do desempenho do modelo multimodal. O extrator de características das faces e o extrator de características das GEIs são parte da rede para ReID, detalhada na próxima seção. Para utilizar o extrator de características das GEIs na rede para ReID, cada um dos vetores obtidos pelo extrator de características das GEIs é salvo em um arquivo.

O *dataset* de GEIs também é de fundamental importância para que o modelo apresente boa acurácia. Para os objetivos deste trabalho, o próprio TUM GAID [Hofmann et al., 2014] é a melhor opção, pois possibilita a avaliação do ReID tanto em curto quanto em longo prazo. No entanto, para diminuir a complexidade, são utilizadas apenas as sequências de imagens de indivíduos caminhando normalmente (n01 a n12). Ou seja, variações causadas por diferentes calçados ou objetos carregados não são explicitamente contempladas pelo modelo. Com isso, para cada indívíduo no *dataset*, são geradas 6 GEIs, correspondentes às sequências n01 a n06. Para 32 dos indivíduos, temos 6 GEIs adicionais, correspondentes às sequências n07 a n12, que são utilizadas para teste de desempenho de reconhecimento em longo prazo.

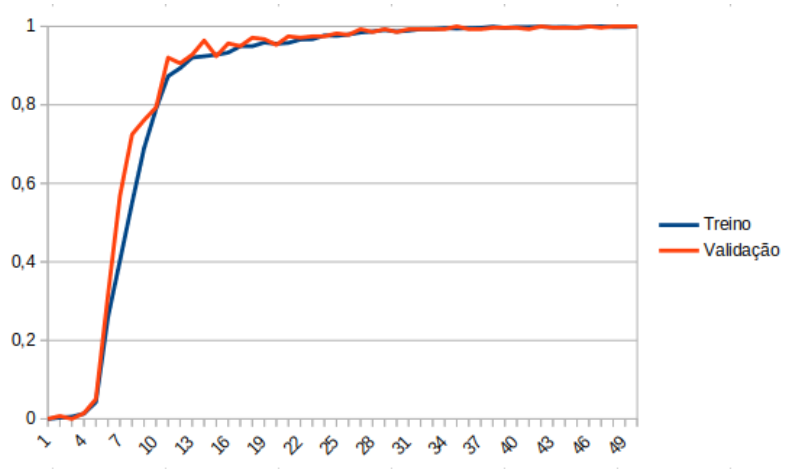
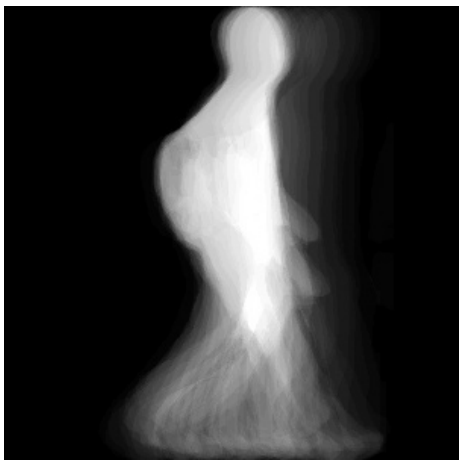


Figura 20 – Dados do treinamento da rede de reconhecimento de GEI,com *dataset* TUM GAID. No eixo horizontal, as épocas. No eixo vertical, a acurácia (Rank-1)

Conforme mencionado, após o treinamento, a rede é utilizada como um extrator de *features* de GEIs. Este extrator de *features* recebe como entrada uma GEI e produz como saída um vetor de características de 305 bits. Um trecho de um vetor de características referente à Figura 21(a) é mostrado na Figura 21(b).

4.6- ReID de pessoas

Esta é a terceira e mais importante etapa do trabalho, a que leva ao objetivo final. Tanto na Seção 4.4.2 quanto na Seção 4.5.4 foram mostradas redes para a extração de características relevantes para a reidentificação de pessoas. No caso da rede da Seção



(a) GEI

-1.70508638024330139 2e-01, -1.45758911967277526 9e-01, -1.05904871597886085 5e-02, -1.49349123239517211 9e-01,
1.12540960311889648 4e-01, -2.07813948392868042 0e-01, 1.45837208256125450 1e-02, 3.71648930013179791 e-03,
8.88385027647018432 6e-02, -2.14965209364891052 2e-01, -1.68125048279762268 1e-01, -1.92378178238868713 4e-01,
-5.74723482131958007 8e-02, -5.93964159488677978 5e-02, 8.63771587610244751 0e-02, -1.47213235497467467 04e-01,
-3.77135388553142547 6e-02, 1.99478328227996826 2e-01, 3.18799853324890136 7e-01, -1.92615300416946411 1e-01,
5.75263686478137970 0e-02, 7.11799273267388343 8e-03, -1.49119913578033447 3e-01, -6.03559315204620361 3e-02,
4.68300469219684600 8e-03, 2.65447348356246948 2e-01, -1.11630082130432128 9e-01, -7.46023282408714294 4e-02,
-4.68475464731454849 2e-03, 1.08619024977087974 5e-02, -7.87900164723396301 3e-02, -7.65336900949478149 4e-02,
9.27760824561119079 6e-02, 4.20802682638168335 0e-01, 9.33817252516746521 0e-02, -2.16976270079612731 9e-01,
-5.10266646742820739 7e-02, 1.80166989564895629 9e-01, -4.89284992218017578 1e-02, -1.98820941150188446 0e-02,
-6.96932077407836914 1e-02, 3.57882112264633178 7e-01, -1.36851608753204345 7e-01, 1.80830359458923339 8e-02,
7.73055255413055419 9e-02, -8.00929144024848938 0e-02, -1.90502300858497619 6e-01, 3.19892447441816330 0e-03,
-9.84933525323867797 9e-02, -1.90783187747001647 9e-01, 2.68490850925455566 e-01, 2.13051557540893554 7e-01,
-6.14068834243202209 e-02, 1.64839282631874084 5e-01, -6.75843982025980949 4e-03, 4.98361319303512573 2e-01,
7.60243237018585205 1e-02, -7.32280015945434570 3e-02, 3.71219277381896972 7e-01, -1.30839467048654019 5e-01,
1.88865773379802703 9e-02, -9.81331169605255127 0e-02, -2.09055662155151367 2e-01, -4.18387167155742645 3e-02,
1.16760022938251495 4e-01, 1.64308100938796997 1e-01, 2.91530042886734008 8e-01, 1.63813844323158264 2e-01,
-2.30305522680282592 8e-01, 1.91104203462600708 0e-01, -1.34957104921340942 4e-01, -3.54690775275230407 7e-02,
-6.88996240496635437 0e-02, 1.38841405510902404 8e-01, -1.34514923555946350 1e-01, 1.77633717656135559 1e-01,
1.16579242050647735 6e-01, -6.32713884115219116 2e-02, 5.30814290046691894 5e-01, -5.51644191145896911 6e-02,
8.59426241368055343 6e-03, 2.10756123065948486 3e-01, -5.24918884038925170 9e-02, 1.37775167822837829 6e-01,
-6.67749047279357910 2e-02, -2.63512015342712402 3e-01, 4.84104864299237328 e-02, 3.95102202892303466 8e-01,
4.89881932735443115 2e-02, 1.33482545614242553 7e-01, -2.13497374206781387 3e-02, 3.70179116725921630 9e-02,
-3.89209985733032266 6e-02, 6.60995960235595703 1e-01, 7.57021531462666937 26e-02, -3.56215909123420715 3e-02,
-1.56268283727478485 11e-01, 2.05648079514503479 0e-01, 3.48570607602596283 0e-02, -8.87720063328742981 0e-02,
-1.52395576238632202 1e-01, 1.56557232141494751 0e-01, 3.21508318185806274 4e-01, 3.46746370196342468 3e-02,
6.61991760134696960 4e-02, -2.09383945912122726 4e-03, -8.17096531391143798 8e-02, -3.34595918655395507 8e-01,
1.82022035121917724 6e-01, 1.01896747946739196 8e-01, -9.09932851791381835 9e-02, 3.62184941768646240 2e-02,
-1.14732049405574798 6e-01, 1.31127893924713134 8e-01, 1.31289996206760046 5e-02, 1.01965010166168212 9e-01,
4.52093482017517089 8e-02, -3.08827832341194152 8e-02, -2.57701337337493896 5e-01, -1.05179704725742340 1e-01,
2.23754584789276123 0e-01, -1.81374415755271911 6e-01, 1.91742151975631739 e-01, 2.79191523790359497 1e-01,
-1.81504011154174804 7e-01, -3.88981252908706650 e-02, -6.46895542740821838 4e-02, -1.63813978433609008 8e-01,
-6.48122504353523254 4e-02, 6.17191120982170105 0e-02, -1.15570332854986190 8e-03, -9.09536257386207580 6e-02,
4.58637028932571411 e-02, -8.92406255006790161 1e-02, 1.11667618155479431 2e-01, -4.95124012231826782 2e-02,
1.85691848397254943 8e-01, 5.88382221758365631 1e-02, -1.96169559980708923 3e-01, -4.04515862464904785 2e-02,

(b) Vetor de *features* (parte)

Figura 21 – Exemplo de entrada (a) e saída (b) do módulo de extração de *features* de GEIs.

4.4.2, o objetivo é extrair características de faces, e na rede da Seção4.5.4, características da forma de caminhar. Nesta seção é mostrado como essas características (*features*) podem ser combinadas e utilizadas para identificar pessoas em longo prazo.

As atividades nesta etapa são:

1. Combinação dos vetores de características
2. Reidentificação em longo prazo

4.6.1 Combinação dos vetores de características

Como mostrado na Figura 22, o módulo de combinação de características combina em um único vetor, próprio para o ReID de pessoas em longo prazo, as *features* de face e de *gait*, estas últimas obtidas a partir das GEIs. No início desta etapa de combinação têm-se um diretório para cada uma das 305 identidades no *dataset*. Em cada diretório há arquivos .gei com os vetores de *features* de *gait* e arquivos .fac, que armazenam vetores de *features* de face.

Como discutido anteriormente, de cada individuo são produzidas, no máximo, 6 GEIs, enquanto a quantidade de recortes contendo faces pode chegar a 500. Para combinar as *features* sem gerar um desequilíbrio, sem descartar dados e, ao mesmo tempo

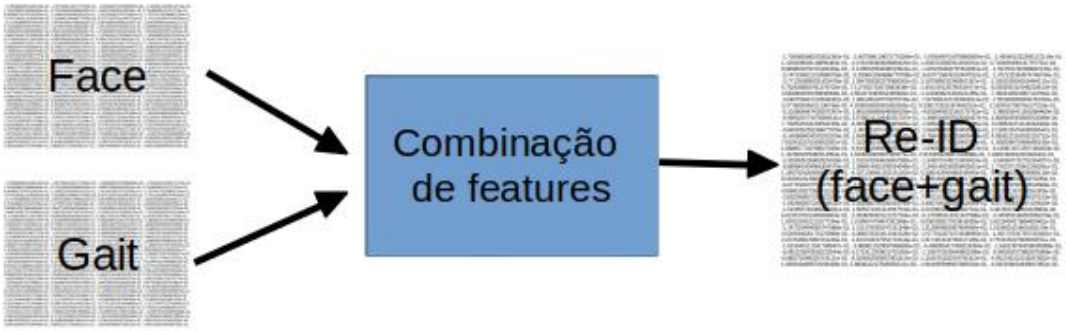


Figura 22 – No modulo de combinação de *features*, os vetores com características de face são combinados com os de *gait*, gerando novos vetores.

potencializando os dados disponíveis, optou-se pela estratégia de combinar cada um dos arquivos contendo *features* de *gait* com cada um dos arquivos contendo *features* de faces. Para cada uma destas correspondências, é feita uma concatenação simples dos dois vetores. Este produto cartesiano gera uma massa de dados de aproximadamente 3.000 exemplos para cada identidade no *dataset*, em um total de 429.300 arquivos, o suficiente para treinar uma rede neural com sucesso. Os novos vetores de *features* resultantes da combinação descrita acima são armazenados em 305 diretórios, correspondentes a cada indivíduo no *dataset*.

4.6.2 Reidentificação em longo prazo

A rede neural projetada para fazer a reidentificação propriamente dita é composta de duas camadas totalmente conectadas (*fully-connected*) entre duas camadas de normalização, conforme mostrado na Figura 23. A função desta rede é apenas fazer uma predição de classe a partir dos dados obtidos por meio dos extratores de *features*. A saída da rede é um vetor de probabilidades de 305 posições. Este vetor é utilizado como entrada para uma função ArgMax para obter a predição da classe, completando o processo de reidentificação de pessoas.

A rede de reidentificação é treinada com as *features* concatenadas geradas na etapa anterior e organizadas em diretórios, como um *dataset*, com as mesmas 305 classes do *dataset* original, porém com muito mais exemplos de treinamento, gerados pelo módulo de combinação de features. Para aproveitar a abundância de dados, o

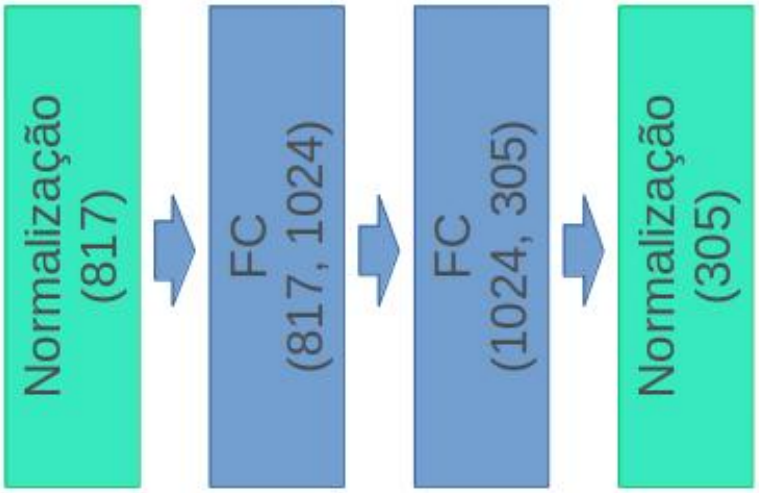


Figura 23 – Modelo de rede neural para re-identificação de pessoas.

treinamento da rede de ReID foi feito em apenas 10 ciclos de treinamento, sendo que em cada um destes ciclos, foram processados 18.782 lotes de 16 imagens. Para a avaliação do desempenho de classificação de pessoas em curto prazo, foi utilizado um subconjunto deste *dataset* contendo 15% das amostras, que foi previamente reservado para esta finalidade. Uma vez que a rede foi treinada e demonstrou um desempenho satisfatório usando o *dataset* de curto prazo, a rede de reidentificação pode ser utilizada (e avaliada) no reconhecimento em longo prazo.

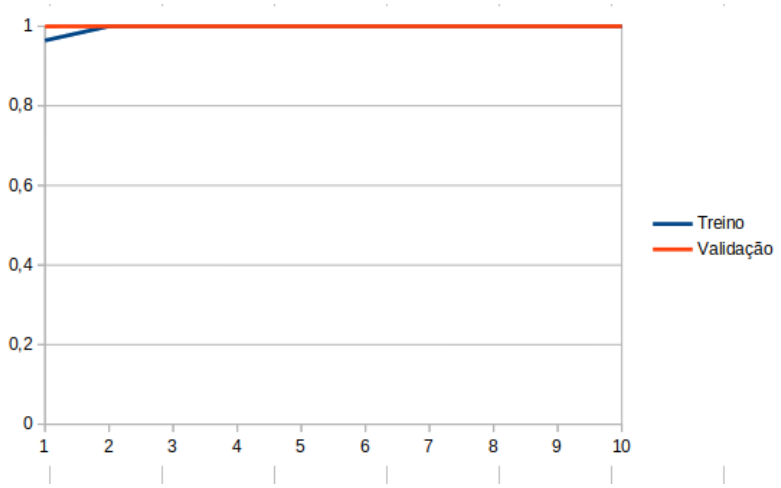


Figura 24 – Dados do treinamento da rede de ReID de pessoas. Foram utilizados 18.782 lotes de 16 imagens a cada época de treinamento. No eixo horizontal, as épocas. No eixo vertical, a acurácia (Rank-1)

Para avaliar o desempenho do ReID em longo prazo, são utilizadas as imagens dos diretórios n07 a n12, disponíveis para apenas 32 indivíduos do *dataset* original.

Estas imagens passaram pelo mesmo tratamento que as demais imagens no TUM GAID, mas ficaram armazenadas em um diretório à parte, sendo utilizadas apenas para esta avaliação. Desta forma, simulamos um cenário real de reidentificação, onde as pessoas são observadas novamente depois de alguns meses.

5- Resultados

Após a execução dos procedimentos descritos na sessão anterior, foram executados testes para comprovar a viabilidade da combinação de *features* de *gait* e face na atividade de reconhecimento de pessoas em longo prazo. Neste capítulo são apresentados os resultados destes testes, sendo tecidos alguns comentários a respeito deles.

5.1- ReID em curto prazo

Embora o modelo tenha sido desenvolvido com o objetivo de reidentificar pessoas em longo prazo, ele também apresentou bons resultados no reconhecimento em curto prazo como esperado. O desempenho do reconhecimento utilizando os dois modais (*gait* e face) melhorou em relação ao desempenho das duas *features* individuais, que já era alto. Este resultado demonstra que a combinação de múltiplas características para aprimorar o desempenho na tarefa de reidentificação de pessoas pode resultar em melhor acurácia. A Tabela 5 traz os números obtidos em cada caso. A conjugação de técnicas de reconhecimento de face e de *gait* garantiu, neste caso mais simplificado e para a base de dados considerada, uma acurácia de 100% (já em rank-1). Este resultado não implica necessariamente que a abordagem proposta seja realmente interessante no caso de ReID de longo prazo, pois o ReID de longo prazo é um problema de complexidade bastante superior.

	Rank-1	Rank-5	Rank-10	Rank-20
Face	0,9666	0,9748	0,9784	0,9824
Gait	0,9203	0,9614	0,9711	0,9769
ReID	1,000	1,000	1,000	1,000

Tabela 5 – Comparação do desempenho dos modelos de reconhecimento facial, de reconhecimento por GEI e de re-identificação (combinação das duas *features*) no cenário de curto prazo.

5.2- ReID em longo prazo

Na reidentificação em longo prazo, não são esperados resultados tão expressivos quando os do curto prazo, já que a mudança de roupas pode afetar bastante a formação das GEIs, especialmente no caso do *dataset* TUM GAID, onde roupas mais pesadas são utilizadas em função da mudança brusca de temperatura entre as duas ocasiões.

Como vimos, a reidentificação de longo prazo é um problema extremamente desafiador, que ainda se encontra longe de ser resolvido, mesmo com métodos situados no estado da arte. Mesmo assim, os resultados obtidos podem ser considerados promissores. Os dados na Tabela 6 mostram uma acurácia de 40% em rank-1. Comparando-se com o reconhecimento de face isoladamente (cuja acurácia supera a do ReID baseado apenas em *gait*), temos um ganho de aproximadamente 6 pontos percentuais em rank-1 e de 4 pontos percentuais em rank-20.

	Rank-1	Rank-5	Rank-10	Rank-20
Face	0,3463	0,5732	0,6610	0,7215
Gait	0,1406	0,3281	0,4115	0,5104
ReID	0,4009	0,5776	0,6772	0,7608

Tabela 6 – Comparação do desempenho dos modelos de reconhecimento facial, de reconhecimento por GEI e de re-identificação (combinação das duas *features*) *em longo prazo*.

Os resultados detalhados nas Tabelas 5 e 6 são bastante expressivos, mas foram obtidos em um cenário controlado, que desconsidera questões como ângulos de captura de imagem, variações na iluminação, calçados, bolsas e mochilas carregadas. Todos estes são fatores que tendem a afetar negativamente o desempenho.

5.3- Robustez em condições diversas

O dataset TUM GAID possibilita avaliar a robustez do modelo com relação a variações nos calçados, roupas e objetos carregados. Estas variações foram desprezadas nos testes anteriores, onde só foram utilizadas as sequências caminhada em condição normal, tanto para o longo quanto para o curto prazo.

O reconhecimento de pessoas utilizando GEIs tende a não ter um desempenho muito bom quando há fortes variações na silhueta. Um casaco pesado ou uma blusa com capuz produzem silhuetas muito diferentes de uma camiseta, por exemplo, como mostra a Figura 25.

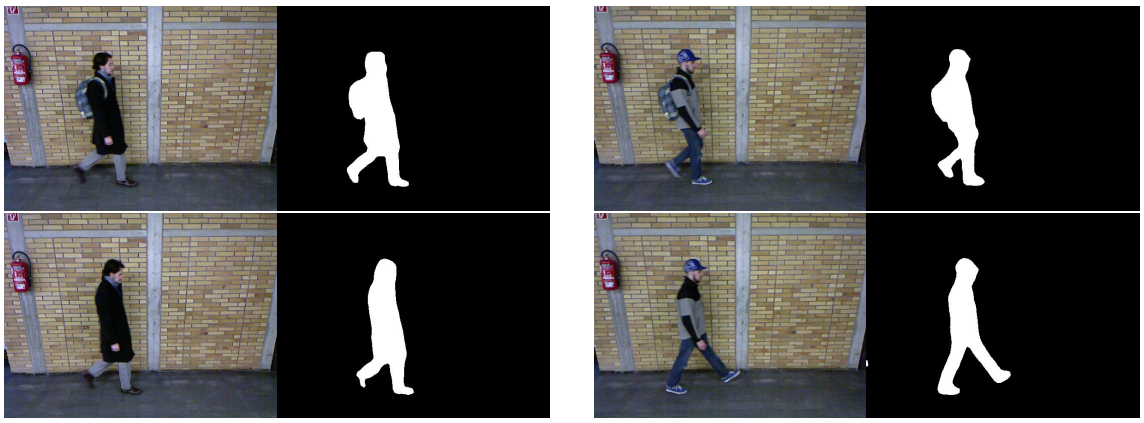


Figura 25 – Exemplos de variações na aparência de um mesmo indivíduo causada por roupas e acessórios, com suas respectivas silhuetas.

O desempenho do modelo diante de tais variações pode ser melhorado com o acréscimo de informações obtidas da face. De fato, os resultados na Tabela 7 mostram que mesmo diante desta grande variação, a combinação de face e *gait* produz bons resultados. Os resultados mostrados na Tabela 7 são melhores do que aqueles obtidos no modelo treinado apenas com imagens na condição normal, graças ao aumento na variância dos exemplos fornecidos à rede neural.

A repetição dos experimentos utilizando toda a base de dados (desta vez sem excluir os cenários em condições “não-normais”) apresentou resultado superior no reconhecimento em longo prazo. A Tabela 7 apresenta uma comparação com os testes anteriores, que utilizaram apenas a parte de dados correspondente à condição ”normal”. Os números obtidos mostram que o modelo proposto é robusto em relação a variações em roupas e acessórios. Observa-se um ganho de acurácia em rank-1 de quase 6 pontos percentuais, quando múltiplas condições são utilizadas no treinamento. Apenas no caso de rank-20 não observou-se tal tendência. O rank-20, no entanto, representa um caso menos crítico, visto que o desempenho já se encontra em torno de 75%.

	Rank-1	Rank-5	Rank-10	Rank-20
Re-ID apenas condição normal	0,4009	0,5776	0,6772	0,7608
Re-ID múltiplas condições	0,4592	0,6227	0,6900	0,7472

Tabela 7 – Comparação do desempenho dos modelos treinados apenas com sequências em condição normal e com todas as condições *em longo prazo*.

5.4- Comentários Finais

Os resultados mostrados nesse Capítulo indicam que a combinação de características de face e de *gait* pode obter resultados melhores que os das técnicas de reidentificação baseadas apenas em face ou apenas em *gait*, seja em termos de reidentificação em longo prazo, seja em termos de reidentificação em curto prazo. Em razão da dificuldade de se obter bases de dados para reidentificação de longo prazo, só foi possível testar a metodologia proposta em um subconjunto de dados de uma base específica. Foi feito um tratamento da base TUM GAID para garantir um resultado quantitativo, capaz de indicar a viabilidade da técnica proposta; caso contrário, ficaríamos restritos a avaliações pouco realistas (restritas à configuração de curto prazo).

A técnica proposta, ainda que vantajosa, não garante ainda resultados bons o suficiente para embasar uma aplicação comercial. De todo modo, pode ajudar pesquisadores que, neste momento, tentam encontrar uma solução que apresente um bom desempenho para o ReID de longo prazo.

6- Conclusão

O problema de reidentificação de pessoas é uma área da biométrica que vem atraindo crescente atenção ao longo dos anos, dada a sua extensa aplicabilidade. No relatório *Global Research*, produzido em setembro de 2020 pelo *Institute for Scientific Information*, a reidentificação de pessoas figura como uma “fronteira de pesquisa em *machine learning*”, cuja solução ainda encontra-se muito distante das soluções vigentes.

Nos primórdios das técnicas de reidentificação, algoritmos procedurais e avaliação em pequena escala foram predominantes. Mais recentemente, a adoção de bases de dados extensas e métodos de *deep learning* vêm permitindo a tais métodos angariar uma acurácia inédita, no caso da reidentificação de curto prazo. No entanto, tal desempenho ainda está longe de ser obtido no problema de reidentificação de longo prazo, o qual ainda carece de bases de dados suficientemente ricas para uma avaliação robusta das diferentes soluções propostas.

Os métodos de ReID podem ser classificados em diversas formas, quanto ao uso ou não de sensores multimodais, se são restritos a curto ou são aplicáveis ao longo prazo, se utilizam uma base de dados anotada ou não. Esta dissertação trata de uma configuração entre as mais desafiadoras (e mais relevantes, em termos práticos) destas soluções: o de reidentificação em longo prazo.

Mesmo para as técnicas de ReID de curto prazo, os resultados reportados em ambientes não controlados não se aproximam da acurácia obtida por seres humanos. Isso significa que a resolução do problema tratado nesta dissertação ainda encontra-se em aberto, a despeito do intenso esforço a ele dedicado nos últimos anos. No caso da reidentificação de pessoas de longo prazo, ainda há um problema adicional, a baixa disponibilidade de *datasets* disponíveis para treinamento e avaliação de soluções.

Essa dissertação propõe a fusão de características relativas ao modo de caminhar e de face. Investigamos se essa fusão pode ser aplicada à resolução do problema do ReID em longo prazo. Extensa pesquisa bibliográfica indica que a utilização de redes neurais para a fusão de características biométricas tem o potencial de prover ótimos resultados quando aplicada ao problema da reidentificação de pessoas em longo prazo. O presente trabalho busca quantificar esse potencial, com o objetivo de viabilizar uma

solução nova para o problema de reidentificação de longo prazo. Na base de dados considerada, a rede de reidentificação com fusão de características obtém um ganho de acurácia em rank-1 de aproximadamente 5 pontos percentuais, em comparação com a adoção isolada de técnicas de reconhecimento de faces no estado da arte. Conclui-se que tal combinação de características possa ser aplicada para realizar o ReID em longo prazo.

Referências Bibliográficas

AT&T Laboratories Cambridge (1994). The Database of Faces.

Baabou, S., Bremond, F., Fradj, A. B., Farah, M. A., and Kachouri, A. (2018). A Study of Person Re-Identification System. *American Journal of Engineering Research (AJER)*, 7(7):210–218.

Bansal, A., Castillo, C., Ranjan, R., and Chellappa, R. (2017). The Do's and Don'ts for CNN-Based Face Verification. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2545–2554.

Barbosa, I. B., Cristani, M., Caputo, B., Rognhaugen, A., and Theoharis, T. (2018). Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Computer Vision and Image Understanding*, 167(December):50–62.

Bashir, K., Xiang, T., and Gong, S. (2010). Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060.

Bedagkar-Gala, A. and Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286.

Bird, N. D., Masoud, O., Papanikolopoulos, N. P., and Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177.

Bobeldyk, D. and Ross, A. (2019). Predicting Soft Biometric Attributes from 30 Pixels: A Case Study in NIR Ocular Images. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 116–124. IEEE.

Bondi, E., Pala, P., Seidenari, L., Berretti, S., and Del Bimbo, A. (2018). Long term person re-identification from depth cameras using facial and skeleton data. In Wannous, H., Pala, P., Daoudi, M., and Flórez-Revuelta, F., editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10188 LNCS, pages 29–41, Cham. Springer International Publishing.

Bucak, S. S., Mallapragada, P. K., Jin, R., and Jain, A. K. (2009). Efficient multi-label ranking for multi-class learning: Application to object recognition. In {*IEEE*} *12th International Conference on Computer Vision,* {*ICCV*} *2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 2098–2105. {IEEE} Computer Society.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 67–74.

Chen, L. C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv*.

Ding, C. and Tao, D. (2018). Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1002–1014.

Eisenbach, M., Kolarow, A., Schenk, K., Debes, K., and Gross, H. (2012). View Invariant Appearance-Based Person Reidentification Using Fast Online Feature Selection and Score Level Fusion. In *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 184–190.

Esler, T. (2020). Face recognition using pytorch. `https://github.com/timesler/facenet-pytorch`.

Figueira, D., Taiana, M., Nascimento, J. C., and Bernardino, A. (2016). A Window-Based Classifier for Automatic Video-Based Reidentification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(12):1736–1747.

Geng, X., Zhou, Z., and Smith-Miles, K. (2008). Individual Stable Space: An Approach to Face Recognition Under Uncontrolled Conditions. *IEEE Transactions on Neural Networks*, 19(8):1354–1368.

Gheissari, N., Sebastian, T. B., and Hartley, R. (2006). Person Reidentification Using Spatiotemporal Appearance. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1528–1535.

Gou, M., Zhang, X., Rates-Borras, A., Asghari-Esfeden, S., Camps, O., Sznaier, M., Mengran Gou Xikang Zhang, A. R.-B. S. A.-E. O. C., Sznaier, M., Gou, M., Zhang,

X., Rates-Borras, A., Asghari-Esfeden, S., Camps, O., Sznaier, M., Mengran Gou Xikang Zhang, A. R.-B. S. A.-E. O. C., and Sznaier, M. (2016). Person Re-identification in Appearance Impaired Scenarios. In Richard C. Wilson, E. R. H. and Smith, W. A. P., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–48. BMVA Press.

Gray, D. and Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In Forsyth, D., Torr, P., and Zisserman, A., editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg. Springer Berlin Heidelberg.

Guo, G., Wang, H., Yan, Y., Zheng, J., and Li, B. (2019). A fast face detection method via convolutional neural network. *Neurocomputing*, 395:128–137.

Guo, J., Pang, Z., Yu, M., Xie, P., and Liu, D. (2020a). A Novel Pedestrian Reidentification Method Based on a Multiview Generative Adversarial Network. *IEEE Access*, 8:181943–181954.

Guo, R., Li, C.-G., Li, Y., Lin, J., and Guo, J. (2020b). Density-adaptive kernel based efficient reranking approaches for person reidentification. *Neurocomputing*, 411:91–111.

Han, F., Reily, B., Hoff, W., and Zhang, H. (2017). Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding*, 158:85–105.

Har-Peled, S., Roth, D., and Zimak, D. (2002). Constraint Classification for Multiclass Classification and Ranking. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, {NIPS} 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 785–792. {MIT} Press.

Hasan, M. and Babaguchi, N. (2016). Long-term people reidentification using anthropometric signature. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person Re-Identification by Descriptive and Discriminative Classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*.

Hoang, V. T., Huang, D. S., and Jo, K. H. (2021). 3-D Facial Landmarks Detection for Intelligent Video Systems. *IEEE Transactions on Industrial Informatics*, 17(1):578–586.

Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., and Rigoll, G. (2014). The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation*, 25(1):195–206.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hu, C.-H., Yu, J., Wu, F., Zhang, Y., Jing, X.-Y., Lu, X.-B., and Liu, P. (2021). Face illumination recovery for the deep learning feature under severe illumination variations. *Pattern Recognition*, 111.

Imani, Z., Soltanizadeh, H., and Orouji, A. A. (2019). Multidimensional Sparse Representation for Multishot Person Reidentification. *IEEE Sensors Letters*, 3(12):1–4.

Iwama, H., Okumura, M., Makihara, Y., and Yagi, Y. (2012). The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. *IEEE Transactions on Information Forensics and Security*, 7(5):1511–1521.

Jayaraman, U., Gupta, P., Gupta, S., Arora, G., and Tiwari, K. (2020). Recent development in face recognition. *Neurocomputing*.

Ju Han, Bir Bhanu, Han, J., and Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322.

Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., and Radke, R. J. (2019). A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):523–536.

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and

Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

LeCun, Y. and Bengio, Y. (1998). Convolutional Networks for Images, Speech, and Time-Series. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, chapter Convolutio, pages 255–258. MIT Press, Cambridge, MA, USA.

Lee, Y., Chen, S., Hwang, J., and Hung, Y. (2017). An Ensemble of Invariant Features for Person Reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):470–483.

Li, A., Liu, L., Wang, K., Liu, S., and Yan, S. (2015). Clothing Attributes Assisted Person Reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):869–878.

Li, W. and Wang, X. (2013). Locally Aligned Feature Transforms across Views. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

Liu, Z., Lu, H., Ruan, X., and Yang, M. H. (2019). Person Reidentification by Joint Local Distance Metric and Feature Transformation. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10):2999–3009.

Liu, Z. and Sarkar, S. (2004). Simplest representation yet for gait recognition: Averaged silhouette. *Proceedings - International Conference on Pattern Recognition*, 4(Eia 0130768):211–214.

Liu, Z. and Sarkar, S. (2007). Outdoor recognition at a distance by fusing gait and face. *Image and Vision Computing*, 25(6):817–832.

Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., and Zhang, G. (2015a). Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23.

Lu, J., Liong, V. E., Zhou, X., and Zhou, J. (2015b). Learning Compact Binary Face Descriptor for Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):2041–2056.

Martinel, N., Foresti, G. L., and Micheloni, C. (2017). Person Reidentification in a Distributed Camera Network Framework. *IEEE Transactions on Cybernetics*, 47(11):3530–3541.

Min, P. P., Sayeed, S., and Ong, T. S. (2019a). Gait recognition using deep convolutional features. *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, pages 2–6.

Min, P. P., Sayeed, S., and Ong, T. S. (2019b). Gait Recognition Using Deep Convolutional Features. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–5. IEEE.

Munaro, M., Basso, A., Fossati, A., Van Gool, L., and Menegatti, E. (2014). 3D reconstruction of freely moving persons for re-identification with a depth sensor. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4512–4519. IEEE.

Pala, P., Seidenari, L., Berretti, S., and Bimbo, A. D. (2019). Enhanced skeleton and face 3D data for person re-identification from depth cameras. *Computers & Graphics*, 79:69–80.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Plichoski, G. F., Chidambaram, C., and Parpinelli, R. S. (2021). A face recognition framework based on a pool of techniques and differential evolution. *Information Sciences*, 543:219–241.

Prihasto, B., Choirunnisa, S., Nurdiansyah, M. I., Mathulaprangsan, S., Chu, V. C. M., Chen, S. H., and Wang, J. C. (2018). A survey of deep face recognition in the wild. *2016 International Conference on Orange Technologies, ICOT 2016*, 2018-Janua:76–79.

Sang, X., Lu, H., Zhao, Q., Zhang, F., and Lu, J. (2021). Nonconvex regularizer and latent pattern based robust regression for face recognition. *Information Sciences*, 547:384–403.

Sarfraz, M. S., Schumann, A., Eberle, A., and Stiefelhagen, R. (2018). A Pose-Sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 420–429.

Sarkar, S., Phillips, P. J., Liu, Z., Vega, I. R., Grother, P., and Bowyer, K. W. (2005). The humanID gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177.

Satta, R. (2013). Appearance Descriptors for Person Re-identification: a Comprehensive Review. *arXIv*.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Schumann, A. and Monari, E. (2014). A soft-biometrics dataset for person tracking and re-identification. *11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014*, pages 193–198.

Seide, F., Li, G., Chen, X., and Yu, D. (2011). Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, volume 2018-Janua, pages 24–29. IEEE.

Shao, L., Zhu, F., and Li, X. (2015). Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034.

Singh, J. P., Jain, S., Arora, S., and Singh, U. P. (2018). Vision-based gait recognition: A survey. *IEEE Access*, 6:70497–70527.

Song, C., Huang, Y. Y., Huang, Y. Y., Jia, N., and Wang, L. (2019). GaitNet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96:106988.

Song, W., Chen, C., Zhao, Q., and Liu, F. (2020). Spatial–temporal representation for video re-identification via key images. *IET Computer Vision*, 14(6):399–406.

Sun, Y., Wang, X., and Tang, X. (2013). Deep Convolutional Network Cascade for Facial Point Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483.

Taghipour-Gorjikolaie, M., Sadri, J., and Razavi, S. M. (2020). Deep adaptive feature enrichment. *Expert Systems with Applications*, 162.

Tan, S., Zheng, F., Liu, L., Han, J., and Shao, L. (2018). Dense Invariant Feature-Based Support Vector Ranking for Cross-Camera Person Reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(2):356–363.

Tao, D., Jin, L., Wang, Y., and Li, X. (2015). Person Reidentification by Minimum Classification Error-Based KISS Metric Learning. *IEEE Transactions on Cybernetics*, 45(2):242–252.

Tu, Z., Yuille, A., and Chen, X. (2006). Image Parsing: Segmentation, Detection, and Recognition. *Towards Category-Level Object Recognition*, 63(2):545–576.

Uttley, A. M. (1966). The transmission of information and the effect of local feedback in theoretical and neural networks. *Brain Research*, 2(1):21–50.

Viola, P. and Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154.

Wagg, D. K. and Nixon, M. S. (2004). On automated model-based extraction and analysis of gait. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 11–16.

Wang, C., Zhang, J., Pu, J., Yuan, X., and Wang, L. (2010). Chrono-Gait Image: A Novel Temporal Template for Gait Recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6311 LNCS, pages 257–270. Springer-Verlag Berlin Heidelberg.

Wang, S.-C. (2003). Artificial Neural Network. In *Interdisciplinary Computing in Java Programming*, volume 725, chapter Artificial, pages 81–100. Springer US, Boston, MA.

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8692 LNCS(PART 4):688–703.

Wu, C., Zhang, Y., and Song, Y. (2018). Multi-view gait recognition using 2D-EGEI and NMF. *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis, ISBA 2018*, 2018-Janua:1–8.

Wu, D., Zheng, S.-J. J., Zhang, X.-P. P., Yuan, C.-A. A., Cheng, F., Zhao, Y., Lin, Y.-J. J., Zhao, Z.-Q. Q., Jiang, Y.-L. L., and Huang, D.-S. S. (2019). Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371.

Wu, F., Jing, X.-Y., Feng, Y., Ji, Y.-M., and Wang, R. (2021). Spectrum-aware discriminative deep feature learning for multi-spectral face recognition. *Pattern Recognition*, 111.

Wu, Z., Huang, Y., Wang, L., Wang, X., and Tan, T. (2017). A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226.

Xing, X., Wang, K., and Lv, Z. (2015). Fusion of Gait and Facial Features using Coupled Projections for People Identification at a Distance. *IEEE Signal Processing Letters*, 22(12):2349–2353.

Xu, C., Makihara, Y., Ogi, G., Li, X., Yagi, Y., and Lu, J. (2017). The OU-ISIR Gait Database comprising the Large Population Dataset with Age and performance evaluation of age estimation. *IPSJ Transactions on Computer Vision and Applications*, 9(1).

Yan, H. and Song, C. (2021). Multi-scale deep relational reasoning for facial kinship verification. *Pattern Recognition*, 110.

Yang, W., Yan, Y., and Chen, S. (2019). Adaptive deep metric embeddings for person re-identification under occlusions. *Neurocomputing*, 340:125–132.

Yao, L., Kusakunniran, W., Wu, Q., Zhang, J., Tang, Z., and Yang, W. (2019). Robust gait recognition using hybrid descriptors based on Skeleton Gait Energy Image. *Pattern Recognition Letters*, (xxxx).

Yu, S., Tan, D., and Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *Proceedings - International Conference on Pattern Recognition*, 4:441–444.

Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1–24.

Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Zhang, P., Wu, Q., Xu, J., and Zhang, J. (2018). Long-Term Person Re-identification Using True Motion from Videos. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 494–502. IEEE.

Zhang, P., Xu, J., Wu, Q., Huang, Y., and Ben, X. (2020). Learning Spatial-temporal Representations over Walking Tracklet for Long-term Person Re-Identification in The Wild. *IEEE Transactions on Multimedia*, page 1.

Zhang, Z., Liang, R., Chen, X., Xu, X., Hu, G., Zuo, W., and Hancock, E. R. (2021). Semi-Supervised Face Frontalization in the Wild. *IEEE Transactions on Information Forensics and Security*, 16:909–922.

Zhao, X., Tang, F., and Wu, Y. (2019). Real-time human segmentation by BowtieNet and a SLAM-based human AR system. *Virtual Reality & Intelligent Hardware*, 1(5):511–524.

Zheng, M., Karanam, S., and Radke, R. J. (2018). RPIfield: A new dataset for temporally evaluating person re-identification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2018-June:1974–1976.

Zheng, S., Zhang, J., Huang, K., He, R., and Tan, T. (2011). Robust view transformation model for gait recognition. In *2011 18th IEEE International Conference on Image Processing*, pages 2073–2076. IEEE.

Zhu, J., Huang, J., Zeng, H., Ye, X., Li, B., Lei, Z., and Zheng, L. (2020). Object Reidentification via Joint Quadruple Decorrelation Directional Deep Networks in Smart Transportation. *IEEE Internet of Things Journal*, 7(4):2944–2954.

Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2019). Face Alignment in Full Pose Range: A 3D Total Solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92.

Zongyi Liu and Sarkar, S. (2006). Improved gait recognition by gait dynamics normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):863–876.