Universidade da Beira Interior
Departamento de Informática

Nº 4984 9- P41: *Evaluation of the SOTA in Soft Biometrics for Visual Surveillance Settings*

Elaborado por:

Maria José Teixeira

Orientador:

Professor Doutor Hugo Proença

30 de junho de 2025

Agradecimentos

A conclusão deste trabalho, bem como o meu percurso académico, não teria sido possível sem o apoio incondicional das pessoas que sempre estiveram ao meu lado.

Em primeiro lugar, agradeço profundamente à minha família, pelo carinho, ternura e apoio constante ao longo desta jornada. Foram o meu alicerce nos momentos difíceis e a minha inspiração nos momentos de dúvida. Este trabalho é tanto meu quanto vosso.

Aos meus amigos, que estiveram, estão e espero que continuem sempre presentes, deixo um agradecimento eterno. Sem o vosso companheirismo, incentivo e partilhas ao longo do caminho, não estaria aqui hoje. O vosso apoio foi fundamental para eu chegar até esta etapa.

Gostaria também de expressar a minha sincera gratidão ao meu orientador, Professor Doutor Hugo Proença, pela disponibilidade, orientação e pelas oportunidades de aprendizagem e crescimento que me proporcionou ao longo deste projeto. A sua orientação foi essencial para o desenvolvimento deste trabalho e para o aprofundamento dos meus conhecimentos.

Agradeço ainda à Universidade da Beira Interior (Universidade da Beira Interior (UBI)) e ao Centro de Formação Interação UBI Tecido Empresarial (CFIUTE) pelo suporte técnico e académico, bem como pelas oportunidades de aprendizagem e desenvolvimento que me foram proporcionadas ao longo do curso.

Por fim, agradeço a mim mesmo, por ter tido a coragem de continuar, por não desistir nos momentos difíceis e por acreditar que era possível chegar até aqui.

Muito obrigada a todos.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Excertos de Código

Acrónimos

API	Application Programming Interface
CFIUTE	Centro de Formação Interação UBI Tecido Empresarial
IA	Inteligência Artificial
SOTA	State of the Art
UBI	Universidade da Beira Interior
CNNs	Convoloutional Neural Networks
ResNet	Residual Network

Capítulo

1

Introdução

1.1 Enquadramento

A biometria suave (*Soft Biometrics*), ao passar do tempo, tem ganho relevância em sistemas de vigilância por desempenhar um papel crucial na identificação e rastreamento de indivíduos de forma não intrusiva, preservando a privacidade dos mesmos. O corpo humano, assim como os seus comportamentos, interações e/ou qualquer tipo de vestuário associado constituem fontes ricas de informação para identificação. Estes dados podem ser usados para reconhecimento de indivíduos e recuperação de padrões num grande acervo.

Ao contrário da biometria rígida, que recorre a métodos de impressões digitais ou digitalização da íris e que, por sua vez, exige interação direta e de dados de alta precisão, a biometria suave explora atributos descritivos como a altura, género, tatuagens, cor do cabelo ou padrões de vestuário, extraídos de imagens ou vídeos à distância. Estes traços são particularmente valiosos quando não se dispõe de dados biométricos de alta resolução, ou quando a sua recolha não é viável. Na prática, aplicações a nível de segurança pública e investigações forenses beneficiam dessa abordagem, pois permitem filtrar e acelerar processos de reconhecimento sem comprometer direitos individuais.

Além disso, a biometria suave tem aplicações em outras áreas, como marketing e análise de comportamento, onde características observáveis podem ser usadas para personalizar experiências ou compreender padrões de interação. No entanto, desafios como iluminação, pose e oclusões continuam a ser barreiras significativas para a implementação eficaz desses sistemas.

Por fim, este projeto visa avaliar o estado da arte (State of the Art (SOTA)) na classificação de atributos físicos observáveis, utilizando imagens de vigi-

lância adquiridas em condições desafiadoras. A implementação baseia-se nos modelos **Homogenus**, um *framework open-source* e no modelo **Attribute-Recognition** para a classificação de imagens de corpo inteiro, adaptado para extrair atributos como o género.

1.2 Motivação

Como estudante de Engenharia Informática, sempre me fascinou o potencial da Inteligência Artificial (IA) aliada à visão computacional para a resolução de desafios práticos. Este projeto, em particular, destacou-se ao demonstrar como técnicas inovadoras, como a deteção de pontos-chave corporais e modelos SOTA podem ser aplicados em cenários reais, reconhecendo características humanas mesmo em condições adversas.

A motivação central foi explorar como sistemas inteligentes interpretam dados imperfeitos e os transformam em informações úteis para a segurança pública.

Além disso, este projeto destaca-se por abordar questões éticas e sociais relacionadas à biometria suave, como privacidade e aceitação pública. A capacidade de identificar indivíduos sem interação direta levanta preocupações sobre o uso responsável dessas tecnologias, reforçando a necessidade de sistemas éticos e transparentes.

Este trabalho reforçou a minha visão de que a IA não se limita à teoria, demonstrando que é uma ferramenta ética e transformadora, capaz de operar em contextos complexos.

1.3 Objetivos

Esta proposta de projeto[1] centra-se na avaliação da SOTA em biometria suave para a vigilância visual, com foco em atributos como género, etnia, idade, altura e peso, validando modelos de classificação em contextos reais.

O objetivo é comparar métodos de classificação, identificando lacunas em cenários não controlados, visto que a maioria dos modelos SOTA são treinados em base a *datasets* ideais, falhando em cenários reais. Por isso, o propósito é testar os sistemas em grupos de dados diferentes, por exemplo, em imagens frontais, com variações nos ângulos, tiradas à distância, com diversos tamanhos e com diferentes resoluções.

Como resultado, o projeto visa obter a predição do atributo género acompanhada da sua classificação com métricas de confiança. As métricas de ava-

liação incluem precisão, *recall* e robustez em diferentes condições de vigilância.

1.4 Organização do Documento

De modo a refletir o trabalho feito, este documento encontra-se estruturado da seguinte forma:

1. O primeiro capítulo – **Introdução** – apresenta o projeto, a motivação para a sua escolha, o enquadramento para o mesmo, os seus objetivos e a respetiva organização do documento.

2. O segundo capítulo – **Estado da Arte** – descreve os conceitos-chave no âmbito deste projeto, bem como as tecnologias utilizadas durante o desenvolvimento da aplicação.

3. O terceiro capítulo – **Método Proposto** – detalha os modelos escolhidos.

4. O quarto capítulo – **Experiências e Resultados** – explica o processo de desenvolvimento, os testes realizados e os resultados obtidos.

5. O quinto capítulo – **Conclusões e Trabalho Futuro** – resume as principais conclusões e sugere possíveis extensões do trabalho.

Capítulo

2

Estado da Arte

2.1 Introdução

A biometria suave refere-se a características humanas observáveis que, embora não permitam a identificação inequívoca de um indivíduo por si só, fornecem informações complementares relevantes em sistemas de supervisão. Em contraste à biometria dura, que requer dados biométricos únicos e detalhados, os atributos suaves como género, idade, etnia, altura e peso podem ser estimados de forma não intrusiva, mesmo em condições de captação adversas.

Estes atributos são especialmente valiosos em contextos reais onde a qualidade da imagem é limitada e a cooperação do indivíduo não é garantida. Câmaras de videovigilância instaladas em espaços públicos frequentemente captam imagens com resolução reduzida e ângulos variáveis. Nestes casos, a análise de *soft biometrics* contribui significativamente para tarefas como seleção de suspeitos, filtragem por características visuais e apoio à re-identificação de indivíduos mediante múltiplas câmaras.

2.2 Aplicações em videovigilância

O uso de sistemas automáticos de videovigilância tem registado um crescimento expressivo, impulsionado pela crescente necessidade de segurança pública, prevenção de crimes e gestão eficiente de multidões em eventos de grande escala. Paralelamente, surgem novos desafios impostos por ambientes reais, tais como variações de iluminação, movimentos imprevisíveis, múltiplas posturas e oclusões parciais.

Neste cenário, os atributos suaves revelam-se cruciais para reforçar a robustez e a precisão dos algoritmos de vigilância inteligente. Com os avanços na visão computacional e na aprendizagem profunda (*deep learning*), permitiu-se extrair automaticamente uma gama diversificada de atributos demográficos a partir de imagens RGB, mesmo em situações não ideais.

Convoloutional Neural Networks (CNNs), tais como VGGNet[2], Residual Network (ResNet)[3] (incluindo as variantes amplamente usadas **ResNet-18** e **ResNet-50**) e MobileNet[4] têm sido utilizados na inferência de atributos, como género, idade e etnia. Esses modelos destacam-se pela sua capacidade de extrair características visuais relevantes com alta precisão e eficiência computacional. Adicionalmente, técnicas mais avançadas como estimativa de pose e segmentação semântica têm contribuído significativamente para análises mais complexas relacionadas à altura e estrutura física, ampliando assim as possibilidades na inferência de atributos corporais detalhados.

As principais aplicações de biometria suave em vigilância incluem:

- **Re-identificação de pessoas (*Person Re-identification*):** Reconhecimento de um mesmo indivíduo em diferentes câmaras, captado em momentos distintos. Essencial em redes com câmaras não sobrepostas, esta tarefa enfrenta obstáculos como variações de iluminação e pose.
- **Rastreamento em tempo real:** Atributos como cor da roupa, género e altura estimada são utilizados para acompanhar indivíduos em movimento, de forma contínua e não intrusiva, particularmente em ambientes como aeroportos, centros comerciais ou eventos desportivos.
- **Análise forense:** Em contextos com imagens degradadas ou ficheiros antigos, os atributos suaves permitem restringir o universo de correspondências com base em descritores aproximados (ex.: 'homem jovem com casaco preto'), complementando os métodos biométricos tradicionais

2.3 Principais características analisadas

Os atributos mais frequentemente analisados nos sistemas de *soft biométricos* incluem:

- **Género:** Geralmente classificado em categorias binárias (masculino/feminino), sendo uma das características mais estáveis e com maior taxa de acerto em modelos de supervisão.

- **Idade:** Estimada a partir de traços faciais e corporais, é sensível a expressões, iluminação, maquilhagem e ocultações, sendo frequentemente agrupada em faixas etárias (ex.: jovem, adulto, idoso) para maior robustez.

- **Altura e peso:** Características derivadas de proporções antropométricas, extraídas com base em *keypoints* 2D, pose estimada ou reconstrução 3D.

- **Etnia:** Inferida com base em padrões faciais, cor da pele e estrutura óssea.

- **Atributos secundários:** Incluem vestuário, acessórios (óculos, boné), tatuagens, entre outros. São particularmente úteis para re-identificação temporária, mesmo sendo suscetíveis a alterações no tempo.

2.4 Técnicas Utilizadas

A evolução das técnicas de classificação de atributos tem acompanhado os avanços na aprendizagem profunda e na representação multimodal do corpo humano. As abordagens mais comuns incluem:

- **Redes convolucionais CNNs**: Arquiteturas como **VGGNet**, **ResNet** e **MobileNet** são amplamente utilizadas na classificação direta de atributos visuais (género, idade, etnia), a partir de imagens RGB.

- **Modelos *multi-task:*** Permitem inferir múltiplos atributos simultaneamente, partilhando camadas convolucionais comuns e especializando-se em tarefas específicas em camadas finais.

- **Inferência a partir de *keypoints* 2D:** Pontos articulares extraídos por ferramentas como *OpenPose*[5] são usados como entrada para modelos de classificação com base em postura e proporção corporal.

- **Modelos de reconstrução 3D:** Abordagens como **SMPL** e **SMPLify-X**[6] permitem gerar representações tridimensionais paramétricas do corpo humano a partir de imagens e pontos-chave, aumentando a precisão dos atributos.

2.5 Trabalhos Relacionados

A tarefa de classificação de género em ambientes de vigilância visual apresenta diversos desafios técnicos, nomeadamente a baixa resolução das imagens, variações de escala, ângulos de visão não frontais e oclusões parciais, fatores comuns em contextos urbanos, aeroportuários ou capturas aéreas. Dada esta complexidade, os modelos do estado da arte SOTA podem ser agrupados consoante o tipo de entrada visual utilizada e a arquitetura de rede subjacente.

- **Modelos baseados na face:** Exemplificados pelo **BOSCH-Age-and-Gender-Prediction** [7], estes modelos utilizam imagens faciais recortadas como entrada e aplicam redes neuronais convolucionais profundas CNNs, para estimar atributos como género e idade. Apesar de apresentarem resultados satisfatórios em ambientes controlados, a sua eficácia degrada-se significativamente em condições adversas;
- **Modelos baseados em pontos-chave 2D:** Modelos como o **human-sex-classifier** [8] utilizam exclusivamente as coordenadas dos pontos-chave anatómicos extraídos por ferramentas como o *OpenPose*[5]. Estas representações estruturais permitem inferência com elevada velocidade e baixo custo computacional. No entanto, a ausência de informação visual detalhada como textura ou cor reduz a sua robustez em cenários com variações significativas de postura e silhueta;
- **Modelos com representação 3D:** O *framework* **Homogenus** [6] representa uma abordagem mais robusta, ao combinar imagens RGB de corpo inteiro com pontos-chave 2D e reconstrução tridimensional do corpo humano utilizando o modelo **SMPL**. Esta abordagem híbrida permite capturar a morfologia corporal de forma mais completa e precisa, mesmo na presença de oclusões, iluminação irregular ou ausência da face. A utilização de múltiplas pistas visuais torna este modelo especialmente adequado para aplicações em vigilância urbana ou aérea.
- **Modelos CNN baseados em imagem RGB de corpo inteiro:** Esta categoria inclui modelos que processam diretamente a imagem completa do corpo humano utilizando redes neuronais convolucionais profundas CNNs, sem recorrer a pontos-chave ou reconstrução tridimensional. Um exemplo é o modelo **Attribute-Recognition**[9], que utiliza uma arquitetura baseada em ResNet para inferir atributos visuais como género, idade, estilo de vestuário e presença de acessórios, a partir de imagens RGB de corpo inteiro. O modelo permite inferência rápida em lote, sem necessidade de segmentação prévia.

Outro exemplo é o modelo proposto por *rdev12/BOSCH* [10], que integra uma pipeline de deteção e classificação: primeiro, um módulo ***Faster* R-CNN** que identifica a região corporal da pessoa na imagem; em seguida, essa região é analisada por uma ResNet para estimar o género e a faixa etária. Este modelo mostra boa aplicabilidade em fluxos de vídeo, mas a dependência parcial da face pode limitar a robustez em certos contextos urbanos.

De forma resumida, a tabela 2.1 resume os principais modelos mencionados no estado da arte, comparando as suas características, vantagens e limitações.

Modelo	**Entrada Visual**	**Técnicas Utilizadas**	**Vantagens**	**Limitações**
BOSCH-Age-and-Gender-Prediction [7]	Imagem facial recortada	CNNs (ResNet)	Elevada precisão em condições controladas	Fraco desempenho com baixa resolução, pose variável ou face parcialmente ocluída
human-sex-classifier [8]	Keypoints 2D	Inferência baseada em pose corporal	Alta velocidade, baixo custo computacional	Sensível a variações de postura, ausência de informação visual detalhada
Homogenus [6]	Imagem RGB corpo inteiro + Keypoints 2D + Reconstrução 3D	Combinação multimodal com OpenPose e SMPL-X	Robusto a oclusões e pose variável, utiliza informação tridimensional completa	Elevada complexidade computacional, dependência da qualidade dos keypoints 2D
Attribute-Recognition [9]	Imagem RGB corpo inteiro	CNNs (ResNet) diretamente aplicadas sobre imagem completa	Simples implementação, eficiente em tempo real, adequado para baixa resolução	Pode ser menos robusto a poses extremas ou forte oclusão
rdev12/BOSCH [10]	Imagem corpo inteiro (Faster R-CNN + CNN facial)	Pipeline híbrido (deteção seguida de CNN)	Bom desempenho em fluxos contínuos de vídeo	Dependência parcial da face reduz robustez em ângulos desfavoráveis

Tabela 2.1: Comparação dos principais modelos para classificação de atributos suaves

2.6 Lacunas na Literatura e Posicionamento deste Pojeto

A revisão dos trabalhos relacionados permite identificar claramente várias lacunas importantes ainda existentes na literatura atual, especialmente no contexto da classificação de atributos suaves em videovigilância:

- **Dependência excessiva de informação facial:** Muitos modelos existentes ainda exigem imagens faciais bem definidas, que raramente estão disponíveis em ambientes reais de vigilância devido a ângulos desfavoráveis, baixa resolução e oclusões;

- **Fragilidade perante variações extremas de pose e oclusões:** As abordagens baseadas exclusivamente em pontos-chave ou imagens faciais são frequentemente pouco robustas em ambientes onde a postura varia significativamente;
- **Falta de avaliação sistemática em ambientes reais:** Embora existam diversos modelos SOTA, poucos estudos avaliam explicitamente a robustez destes modelos em condições reais de videovigilância.
- **Necessidade de integração multimodal eficiente:** Apesar de modelos multimodais como o **Homogenus** serem promissores, a sua complexidade computacional elevada limita o uso prático, destacando a necessidade de abordagens mais eficientes que preservem robustez.

Este projeto propõe uma avaliação dos modelos **Homogenus** e **Attribute-Recognition**, que representam duas abordagens complementares: uma baseada em reconstrução 3D e outra em CNNs diretamente sobre imagens RGB de corpo inteiro. A comparação destes dois modelos pretende fornecer recomendações concretas para a aplicação prática de biometria suave em cenários reais de videovigilância, destacando que método oferece melhor desempenho.

2.7 Conclusões

A revisão do estado da arte realizada permitiu identificar claramente as oportunidades e desafios no uso de biometria suave para aplicações em videovigilância. A maioria dos modelos existentes para a classificação de atributos, como o género, depende significativamente de informação facial nítida e de alta resolução. Fatores adversos tornam impraticável o uso destes modelos baseados exclusivamente na face em cenários reais.

Diante dessas limitações, modelos que operam diretamente com imagens de corpo inteiro apresentam-se como uma alternativa viável e promissora. Técnicas baseadas em inferência a partir de *keypoints* 2D e modelos tridimensionais como o **Homogenus**, ou abordagens que usam diretamente CNNs sobre imagens RGB completas como o **Attribute-Recognition**, têm vindo a demonstrar maior robustez e adaptabilidade a condições desafiadoras típicas de sistemas de vigilância. Enquanto métodos baseados exclusivamente em CNNs são eficientes, mas podem comprometer a robustez em cenários extremos de variação de pose e oclusões.

Capítulo

3

Método Proposto

3.1 Introdução

O presente capítulo descreve os modelos selecionados para a classificação do género a partir de pontos-chave 2D, imagens RGB de corpo inteiro no contexto de monitorização por vídeo.

A escolha fundamentou-se na disponibilidade de modelos pré-treinados e na robustez frente a condições adversas de aquisição de imagem. Foram considerados dois modelos com abordagens distintas:

- **Homogenus**: Modelo baseado na reconstrução tridimensional do corpo humano a partir da extração de pontos-chave 2D das imagens utilizadas.

- **Attribute-Recognition**: Modelo baseado em redes neuronais convolucionais CNNs aplicadas diretamente à imagem RGB.

3.2 Objetivo da Implementação

As implementações descritas neste capítulo procuram avaliar e comparar o desempenho dos modelos **Homogenus** e **Attribute-Recognition** na classificação do género a partir de imagens de corpo inteiro em contextos reais de vigilância visual.

3.3 Homogenus

3.3.1 Arquitetura e Funcionamento do Modelo

O modelo **Homogenus** (*Human Image Gender Classifier*) é uma arquitetura multimodal concebida para a classificação de género com base em imagens de corpo inteiro, sendo particularmente adequada para contextos de videovigilância onde a face da pessoa se encontra desfocada ou completamente ausente. Para contornar a ausência de informação facial, o sistema baseia-se na morfologia corporal, recorrendo à imagem RGB da pessoa e aos pontos-chave 2D extraídos.

Para funcionar corretamente, o modelo requer como dados de entrada uma imagem RGB de corpo inteiro da pessoa alvo. Esta imagem deve permitir a deteção clara da postura corporal e das articulações principais, de forma a viabilizar a extração precisa dos pontos-chave 2D. A qualidade da imagem e a visibilidade do corpo são fatores determinantes para o sucesso das etapas subsequentes de reconstrução e classificação.

Portanto, a arquitetura do **Homogenus** assenta em três componentes principais:

1. **Deteção de pontos-chave 2D:** Utilizando a ferramenta **OpenPose** [11], são extraídos pontos articulares do corpo humano em duas dimensões, os quais são armazenados em ficheiros `(.json)`. Estes representam a estrutura esquelética da pessoa observada e servem como entrada para o modelo.

2. **Reconstrução 3D com SMPLify-X:** Os pontos 2D extraídos alimentam o algoritmo **SMPLify-X** [6], que ajusta os parâmetros do modelo tridimensional **SMPL-X**, via otimização iterativa. O resultado é uma malha 3D que representa a forma corporal, descrita por dois vetores: $\boldsymbol{\beta}$ (relativo à forma do corpo) e $\boldsymbol{\theta}$ (relativo à pose).

3. **Classificação de género:** Os vetores tridimensionais extraídos são normalizados e concatenados, servindo como entrada para uma **rede neuronal densa (MLP)** que realiza a inferência da classe de género. O modelo considera três classes: **masculino**, **feminino** e **neutro**.

A decisão final é determinada pela confiança da predição: se o valor ultrapassar um limiar definido, (`accept_threshold`), é atribuída a classe correspondente; caso contrário, o modelo devolve a classe `neutro`, sinalizando

incerteza. Adicionalmente, o sistema pode armazenar as probabilidades individuais associadas às classes masculino e feminino, o que permite decisões forçadas ou análise posterior mais aprofundada.

Esta *framework* foi concebida para funcionar preferencialmente com imagens onde o corpo esteja visível na sua totalidade ou maioritariamente, uma vez que a robustez do modelo depende da qualidade da deteção dos pontos-chave e da integridade da reconstrução 3D. A conceção modular permite dissociar a extração de características estruturais da tarefa de classificação, aumentando significativamente a fiabilidade em cenários adversos como imagens captadas de ângulos não ortogonais.

Detalhes Técnicos

O **SMPL-X** é um modelo paramétrico tridimensional avançado do corpo humano que estende os modelos anteriores, como o **SMPL** e o **SMPL-H**. Incorpora não apenas o tronco e membros, mas também as **mãos** e o **rosto**, oferecendo uma representação mais completa e expressiva da forma humana.

Este modelo representa o corpo com base em dois vetores principais:

- $\boldsymbol{\beta}$: vetor de forma que codifica características antropométricas como proporção dos ombros, cintura e comprimento dos membros;
- $\boldsymbol{\theta}$: vetor de pose que descreve a orientação das articulações;

A reconstrução é feita por ajuste dos parâmetros (β, θ), para minimizar a diferença entre os pontos-chave 2D detetados pelo **OpenPose** e os pontos-chave projetados pela malha tridimensional do modelo.

O **SMPL-X** produz uma malha tridimensional que pode ser renderizada ou usada como entrada para tarefas de inferência. No **Homogenus**, os vetores de `forma` e `pose`, extraídos da malha, são concatenados e normalizados antes de serem passados a uma rede neural densa, composta por múltiplas camadas totalmente conectadas, com funções de ativação **ReLU**. A camada final é uma *softmax*, com três neurónios correspondentes às classes: masculino, feminino e neutro.

A classificação é realizada com base na maior probabilidade resultante da função *softmax*. Caso essa probabilidade não ultrapasse um limiar de confiança pré-definido (`threshold`), o modelo devolve a classe neutro, adotando assim uma abordagem conservadora para casos de ambiguidade ou incerteza.

O código disponibilizado no repositório original refere-se apenas à fase de inferência, pressupondo que o modelo já foi treinado com vetores (β, θ) anotados com a classe de género correspondente.

Implementação Prática

Para garantir a reprodutibilidade e compatibilidade com o modelo **Homogenus**, foi configurado um ambiente virtual com as bibliotecas e versões específicas, detalhadas a seguir:

- **Criação do ambiente virtual com Python 3.7:** garante compatibilidade com as dependências legadas do modelo:

```
python3.7 -m venv homogenus_env
source homogenus_env/bin/activate
```

- **Clonagem do repositório Homogenus.**
- **Instalação das dependências listadas no ficheiro de requisitos:**

```
pip install -r requirements.txt
```

- Caso ocorra alguma incompatibilidade com o TensorFlow listado, pode ser necessário forçar a instalação da versão compatível com o código (ex: `tensorflow==1.15.2`), tal como indicado abaixo:

```
pip install tensorflow==1.15.2
```

  O TensorFlow 1.15.2 foi escolhido por ser a versão mais recente compatível com a 1.13.1, originalmente usada pelo **Homogenus**.

- **Download dos pesos pré-treinados e ficheiros SMPL-X:**

  Os pesos do modelo Homogenus e os ficheiros SMPL-X devem ser descarregados do site oficial[6].Extraí-se os **pesos** pré-treinados do ficheiro zip para:

```
homogenus/trained_models/tf/
```

3.3.2 Funcionamento dos Filtros Internos do Homogenus

Durante o processo de análise do método, foi possível identificar o funcionamento interno dos **filtros de validação do Homogenus**, os quais afetam diretamente se uma imagem é processada e se uma predição final é registada. Entre os critérios implementados no código, destacam-se os seguintes:

Filtro 1: Confiança mínima dos keypoints

```
if not should_accept_pose(pose, human_prob_thr=0.5):
    continue
```

Excerto de Código 3.1: Filtro de confiança mínima dos keypoints

- **Função:** verifica se a pose da pessoa tem pelo menos um número mínimo de keypoints com confiança superior a 0.5.
- **Objetivo:** garantir que apenas pessoas com detecção de pose suficientemente confiável sejam processadas pelo modelo.

Filtro 2: Tamanho mínimo do recorte (crop)

```
if cropped_image.shape[0] < 200 or cropped_image.shape[1] <
    ↪ 200:
    continue
```

Excerto de Código 3.2: Filtro de tamanho mínimo do crop.

- **Função:** descarta imagens cujo crop resultante da pose seja inferior a 200 pixels em largura ou altura.
- **Objetivo:** garantir que o modelo receba imagens com qualidade mínima suficiente para realizar a inferência com precisão.

Filtro 3: Probabilidade mínima para aceitação da predição

- **Parâmetro:** `accept_threshold = 0.9`

```
if gender_prob > accept_threshold:
    # Aceita o genero predito
    color = 'green'
else:
    # Marca como neutral se a confianca for baixa
    gender_pd = 'neutral'
    color = 'grey'
```

Excerto de Código 3.3: Filtro para aceitação da predição.

- **Função:**O modelo apenas aceita a predição de género quando a probabilidade associada supera o limiar de 0,9. Nesses casos, o resultado é visualmente assinalado com uma *bounding box* de cor verde sobre a imagem. Caso contrário, o género é classificado como `neutral`, sendo

a imagem retornada com uma *bounding box* a cinzento, indicando incerteza na predição.

- **Objetivo:** evitar classificações com baixa confiança, aumentando a robustez e fiabilidade do sistema.

Parâmetro adicional: Margem de recorte

- **Parâmetro:** `crop_margin = 0.08`
- **Função :** adiciona uma margem de 8% ao redor da pose ao efetuar o crop da imagem.
- **Objetivo:** garantir que todos os *keypoints* relevantes fiquem incluídos na imagem recortada.

Mecanismo de Inferência

No contexto do código, a confiança (*confidence*) representa a probabilidade atribuída pelo modelo à classe predita.

- `probs_ob` é um vetor com duas probabilidades: uma para `male` e outra para `female`.
- `gender_id` indica qual das classes tem maior probabilidade.
- `gender_prob` corresponde ao valor da probabilidade da classe com maior valor.

A confiança da predição é comparada ao limiar definido por *threshold* 3.3. Se o valor ultrapassar esse limiar, a predição é aceite. Caso contrário, é classificada como `neutral`.

No entanto, como o objetivo deste projeto é obter apenas duas classes: *masculino* ou *feminino*, foi realizada uma modificação no *script* de inferência para que, além da predição do género, o ficheiro `(.json)` inclua também as probabilidades atribuídas a cada classe. Dessa forma, nos casos em que a predição seja `neutral`, é possível comparar diretamente as probabilidades associadas a `homem` e `mulher`, selecionando aquela com maior valor como classe final.

Caso seja obtida uma predição `neutral` sem probabilidades associadas, isso significa que os filtros mencionados anteriormente permitiram ao modelo rejeitar dados de entrada de baixa qualidade (pose incerta, imagem pequena ou predição pouco confiante), contribuindo para uma inferência mais fiável e robusta no contexto de videovigilância.

3.3.2.1 Vantagens e Limitações

Este modelo apresenta diversas vantagens que o tornam particularmente adequado para aplicações em cenários de videovigilância:

- **Independência de feições faciais:** A inferência baseada na morfologia corporal permite lidar eficazmente com imagens onde a face está oculta, desfocada ou não orientada para a câmara;
- **Classificação não binária:** A inclusão da classe *neutro* permite ao sistema expressar incerteza de forma explícita, contribuindo para maior fiabilidade em ambientes com ruído ou ambiguidade;
- **Reconstrução precisa da forma corporal:** O uso do **SMPL-X** fornece uma representação tridimensional rica e realista da estrutura física do indivíduo, o que permite capturar indicadores visuais relevantes como proporção do tronco, largura dos ombros, entre outros;

Contudo, o modelo também apresenta algumas limitações que devem ser consideradas:

- **Dependência da qualidade dos keypoints:** A precisão da reconstrução 3D depende fortemente da qualidade dos pontos-chave gerados pelo **OpenPose**. Ocultações ou ruído nos pontos-chave podem comprometer todo o pipeline;
- **Elevado custo computacional:** A reconstrução tridimensional com o **SMPLify-X** é computacionalmente intensiva, tornando o modelo pouco adequado para aplicações em tempo real;
- **Necessidade de imagens com corpo visível:** O desempenho degrada-se significativamente quando o corpo está parcialmente fora da imagem ou se a postura da pessoa dificulta a deteção de pontos-chave relevantes;

3.4 Attribute-Recognition

3.4.1 Descrição técnica do modelo.

O modelo *Attribute-Recognition*[9] é uma implementação leve e eficaz de classificação *multi-label*, desenvolvida em *Pytorch*, treinada para reconhecer atributos físicos de pessoas a partir de imagens RGB de corpo inteiro. A sua arquitetura baseia-se numa rede convolucional profunda (**ResNet-18** ou, alternativamente, **ResNet-50**), que extrai características diretamente das imagens,

sem requerer pontos-chave corporais, nem segmentação ou reconstrução tridimensional, como no modelo mencionado em 3.3.

Os atributos previstos incluem categorias como género, faixa etária, comprimento e cor da roupa, tipo de vestuário, comprimento do cabelo, e presença de acessórios como mochilas, chapéus ou óculos. O modelo foi treinado com o *dataset* **Pedestrian Attribute Recognition-WACV'24**[12][13][14], que resulta da fusão de três conjuntos de dados públicos amplamente utilizados na literatura, a saber: **Market-1501**[15], **PA-100K**[16] e **PETA**[17], com mais de 130 mil imagens anotadas com atributos visuais.

Este volume de dados permitiu ao modelo adaptar-se a diferentes variações de pose, iluminação, vestuário e resolução.

Arquitetura utilizada

O modelo, como mencionado anteriormente, está implementado com base em duas possíveis arquiteturas: **ResNet-18** ou **ResNet-50**, que são redes neuronais convolucionais CNNs e fazem parte da família ResNet:

- **ResNet-18**: constituída por 18 camadas, incluindo camadas convolucionais, camadas de normalização e camadas de ativação **ReLU**. É uma rede relativamente leve em comparação com a **ResNet-50**, o que a torna adequada para aplicações em tempo real e dispositivos com recursos limitados.

- **ResNet-50**: composta por 50 camadas, é uma rede mais profunda e complexa que a **ResNet-18**. Esta oferece maior capacidade de representação e é capaz de capturar características mais complexas nos dados de entrada.

Ambas as arquiteturas utilizam **blocos residuais**, que facilitam o treino de redes profundas ao permitir que a informação flua diretamente entre camadas, evitando o problema de degradação do gradiente.

No contexto deste projeto, foram utilizadas ambas as arquiteturas, de modo a comparar a eficiência e os resultados obtidos.

Implementação Prática

Tal como descrito na Secção 3.3.1, foi necessário configurar um novo ambiente virtual para a execução do modelo **Attribute-Recognition**, a fim de evitar possíveis conflitos entre dependências.

- **Criação do ambiente virtual:**

```
python3.8 -m venv attribute_env
source attribute_env/bin/activate
```

Excerto de Código 3.4: Criação do Ambiente Viertual.

- **Clonagem do repositorio**;
- **Instalação das dependências:** Para evitar problemas com a GPU, decideu-se instalar as dependências manualmente para usarem a CPU.
  - **Torch==2.1.0**
  - **opencv-python==4.8.0.76**
  - **torchvision==0.16.0**
  - **numpy==1.23.5**
  - **scikit-learn==1.2.2**
  - **tqdm**
- ***Download* dos pesos do modelo**

  Para realizar a inferência com o modelo **Attribute-Recognition**, foi necessário utilizar os **pesos previamente treinados** disponibilizados pelo autor do repositório.

```
models/
ResNet18_best\_model.pth
ResNet50_best\_model.pth
```

Excerto de Código 3.5: Arquitetura base do modelo

3.4.1.1 Funcionamento do Modelo

Tal como no modelo anterior 3.3.1, o modelo **Attribute-Recognition** também segue uma sequência estruturada de etapas para processar o conjunto de imagens e realizar a inferência. Abaixo, descreve-se o pipeline completo:

Preparação dos dados de entrada

Todas as imagens correspondem a capturas de corpo inteiro, provenientes do mesmo conjunto de dados, com características morfológicas visíveis.

A principal diferença está no **pré-processamento**. Enquanto o **Homogenus** requer extração de pontos-chave e reconstrução 3D, o modelo **Attribute-Recognition** opera diretamente sobre as imagens RGB, sem qualquer anotação estrutural adicional. Portanto, a *pipeline* é mais simples:

- A imagem RGB é carregada com **OpenCV** e convertida de BGR para RGB(caso necessário).
- A imagem é redimensionada para 224×224 píxeis, conforme o tamanho esperado pela arquitetura ResNet.
- Os valores dos píxeis são normalizados com a médidas e desvios padrão do *dataset* **ImageNet**.
- A imagem é convertida num tensor **PyTorch**.

Extração de características com ResNet

A imagem é processada por uma rede convolucional pré-treinada da família **ResNet**, responsável por extrair automaticamente as características visuais mais relevantes:

- As primeiras camadas extraem características de baixo nível, como bordas e texturas;
- As camadas intermédias identificam padrões regionais, como partes do corpo e/ou tipo de roupa;
- As últimas camadas produzem um **vetor de características de alto nível**, ou seja, uma representação numérica compacta da imagem, deixando de ser representada como uma matriz de pixeis RGB e passa a ser convertida num vetor numérico que contém as características mais relevantes da pessoa retratada.

Este vetor é passado para uma camada totalmente conectada (densa), que produz um vetor de *logits*, ou seja, *scores* numéricos que ainda não representam probabilidades.

Conversão dos *logits* em probabilidades

O vetor de *logits* é transformado em probabilidades reais no intervalo $[0,1]$, através da função de ativação **sigmoid**, aplicada individualmente a cada atributo.

Essa função permite que cada atributo seja tratado como uma tarefa de classificação binária independente (presente vs. ausente), visto que uma pessoa pode possuir múltiplos atributos simultaneamente.

Classificação binária dos atributos

Após a obtenção das probabilidades, aplica-se um limiar de 0,5 para decidir a presença ou ausência de cada atributo:

- Se a probabilidade ≥ 0.5, o atributo é considerado presente;
- Caso contrário, é considerado ausente;

```
predicted_results = predicted_probs[0] > 0.5
pos = np.where(predicted_results == 1)[0]
labels = label_col[pos]
```

Excerto de Código 3.6: Filtro binário por limiar de 0.5.

Por exemplo, no caso do atributo *Gender*, se o valor da probabilidade associada à classe *female* for 0,81, o modelo interpreta a predição como `feminino`, por estar acima do limiar de 0,5. Já um valor como 0,24 é assumido como `masculino` por exclusão, dado que o modelo não fornece diretamente uma probabilidade binária para a classe *male*

Suporte para inferência em múltiplas imagens

O ficheiro `inference.py` foi adaptado para permitir a inferência em lote de todas as imagens contidas numa pasta. Além disso, os resultados dos atributos reconhecidos são exportados para um ficheiro `.csv`, facilitando a análise posterior.

Este fluxo é ideal para aplicações em videovigilância e análise forense, devido à sua compatibilidade com *pipelines* automáticas e escaláveis.

3.4.1.2 Vantagens e Limitações

Este modelo apresenta diversas vantagens que o tornam particularmente adequado para aplicações em cenários de videovigilância:

- **Simplicidade e leveza:** O modelo é mais leve que abordagens baseadas em reconstrução 3D, podendo ser executado em tempo real ou em sistemas com poucos recursos computacionais;

- **Sem pré-processamento estrutural:** Dispensa a deteção de pontos-chave, segmentação ou reconstrução corporal;
- **Capacidade multi-label:** Permite inferir género, idade e diversos outros atributos visuais em simultâneo, aumentando o valor informativo de cada inferência;
- **Robustez a variações visuais:** Mostra bom desempenho mesmo com pose variada, iluminação não uniforme e resoluções médias, o que o torna ideal para aplicações em videovigilância de larga escala.

Contudo, o modelo também apresenta algumas limitações que devem ser consideradas:

- **Sensível a oclusões severas:** Como depende apenas da imagem RGB, o modelo pode falhar em condições de forte oclusão ou quando partes importantes do corpo não estão visíveis;
- **Sem interpretação morfológica explícita:** Ao contrário do modelo3.3, este modelo não utiliza reconstrução estrutural do corpo, o que pode limitar a interpretação contextual de atributos físicos em certos cenários;
- **Inferência binária do género:** A classe de género funciona por exclusão (ex: baixa probabilidade para `femenino` implica `masculino`), o que limita a representação de morfologias ambíguas ou não binárias.

3.5 Critérios de Seleção das Metodologias

Durante a pesquisa por modelos pré-treinados adequados aos objetivos do projeto, mais especificamente para a classificação do género, constatou-se que a maioria das abordagens existentes, como o *FairFace*[18], se concentra na análise facial e requer imagens limpas e bem enquadradas do rosto. Esses modelos, embora eficazes em contextos controlados, mostram-se pouco adequados às exigências do presente estudo, que visa operar sobre imagens realistas e com limitações típicas de sistemas de videovigilância.

Adicionalmente, os modelos identificados durante a revisão da literatura 2.5 não disponibilizavam publicamente o seu código-fonte ou os pesos pré-treinados, o que inviabilizava a sua utilização prática. Outros modelos, apesar de relacionados com atributos físicos, não realizavam diretamente a classificação do género, como o *BodyNet*[19], focando-se em tarefas auxiliares, como a deteção da pose ou o reconhecimento de ações.

Embora o **Homogenus** ofereça vantagens significativas, a sua implementação revelou-se desafiante, como observado na Secção 3.3.2.1. Face a essas limitações, foi necessário procurar uma alternativa que permitisse uma implementação mais rápida e direta. Nesse contexto, foi identificado o modelo **Attribute-Recognition**, que se revelou particularmente vantajoso por recorrer a redes neuronais convolucionais aplicadas diretamente sobre imagens RGB, sem necessidade de etapas prévias de extração de pontos-chave. A abordagem baseada em *deep learning* deste segundo modelo facilitou consideravelmente o processo de integração, mantendo um desempenho competitivo na tarefa de classificação de género.

Ambos os modelos foram desenvolvidos em *Python*, recorrendo a bibliotecas como *PyTorch* e *TensorFlow*. A acessibilidade ao código-fonte e aos modelos pré-treinados contribuiu para a seleção dos mesmos.

3.6 Tecnologias e Ferramentas Utilizadas

Para a implementação e realização de testes com os modelos, foram utilizadas diversas tecnologias e ferramentas que facilitaram o desenvolvimento e a execução do pipeline. As principais ferramentas utilizadas foram:

- **Python 3.7 e 3.8**:[20] Linguagem de programação amplamente utilizada em projetos de inteligência artificial devido à sua simplicidade, sintaxe clara e vasta oferta de bibliotecas como *TensorFlow*, *PyTorch*, *NumPy* e *OpenCV*, que facilitam o desenvolvimento e teste de modelos de aprendizagem automática e visão computacional.

- **ResNet-18 e ResNet-50**: As ResNet são redes neuronais profundas que utilizam conexões residuais para facilitar o treino de modelos com muitas camadas. Estas arquiteturas são amplamente utilizadas em tarefas de visão computacional, incluindo classificação de imagens e reconhecimento.

- **OpenPose**: [11]Ferramenta utilizada para a extração dos pontos-chave 2D das imagens de entrada, essencial para o pipeline do modelo **Homogenus**.

- **SMPLify-X**:[6] *Framework* utilizado para reconstrução tridimensional do corpo humano com base nos *keypoints* extraídos.

- **PyTorch**:[21] Biblioteca de *deep learning* em código aberto amplamente usada para construir e treinar redes neuronais. Oferece uma Application

Programming Interface (API) intuitiva em *Python* e é especialmente popular em ambientes de investigação. Neste projeto, foi utilizada para implementar o modelo **Attribute-Recognition**, recorrendo também às bibliotecas `torch` e `torchvision` para operações de rede e pré-processamento de imagens.

- **TensorFlow**:[22]Biblioteca de código aberto desenvolvida pela **Google Brain** para tarefas de aprendizagem automática e *deep learning*. No presente projeto, foi utilizado para a construção e inferência do modelo **Homogenus**, aplicado ao reconhecimento de género a partir de imagens de corpo inteiro.
- **NumPy e Pandas**: Bibliotecas utilizadas para manipulação de arrays e análise de dados tabulares.
- **Matplotlib**: Ferramenta utilizada para geração de gráficos e visualizações, como a matriz de confusão e métricas de desempenho.
- **Google Colab**: Ambiente utilizado para execução do modelo **OpenPose**[11], para a extração dos pontos-chave das imagens.
- **Github**: Sistema de controle de versão utilizado para clonar repositórios e gerir o código-fonte.

Estas ferramentas foram escolhidas pela sua robustez e ampla utilização na comunidade científica, garantindo a reprodutibilidade e eficiência do projeto.

3.7 Conclusões

Este capítulo apresentou os dois modelos selecionados para a classificação de género em contextos de videovigilância: **Homogenus** e **Attribute-Recognition**, destacando-se pelas suas abordagens complementares.

O **Homogenus**, baseado em reconstrução 3D com pontos-chave 2D e no modelo paramétrico **SMPL-X**, oferece maior precisão morfológica, sendo adequado para cenários sem informação facial. No entanto, o seu *pipeline* apresenta elevada complexidade computacional e dependência da qualidade dos dados de entrada.

Por outro lado, o **Attribute-Recognition**, destaca-se pela leveza e facilidade de integração. A inferência direta sobre imagens RGB permite maior escalabilidade, embora com menor robustez em casos de oclusão ou ambiguidades visuais.

Ambos os modelos foram escolhidos pela disponibilidade pública, documentação acessível e adequação ao domínio de videovigilância. A diversidade metodológica permite uma comparação empírica entre estratégias baseadas em morfologia 3D e atributos visuais diretos.

Capítulo

4

Experiências e Resultados

4.1 Introdução

Este capítulo apresenta as experiências realizadas e os resultados obtidos com dois modelos distintos de classificação de género em imagens de corpo inteiro, no contexto de videovigilância: o **Homogenus**, baseado em reconstrução tridimensional, e o **Attribute-Recognition**, baseado em redes convolucionais profundas, avaliando-se as variantes **ResNet-18** e **ResNet-50**.

As experiências foram conduzidas sobre um conjunto de dados realista, previamente filtrado para garantir a qualidade dos exemplos analisados. Para cada modelo, descreve-se o processo de preparação dos dados, as etapas do *pipeline* de inferência e os critérios de validação utilizados.

A avaliação foi realizada com base em métricas padronizadas de desempenho, como taxa de acerto, precisão, *recall* e *F1-score*, complementadas por análises gráficas e matrizes de confusão. A comparação entre os modelos permite identificar os pontos fortes e limitações de cada abordagem, bem como as implicações práticas da sua aplicação em cenários reais.

4.2 Modelo Homogenus

4.2.1 Preparação dos Dados de Entrada

Para a realização dos testes com o modelo **Homogenus**, foi disponibilizado um conjunto de dados pelo orientador, que consiste numa pasta principal com dois subconjuntos distintos de vídeos: um conjunto *Indoor* e outro *Outdoor*. Cada subconjunto apresenta uma estrutura e nomenclatura específicas, contendo informação embutida no nome de cada ficheiro.

- **Indoor:** os ficheiros encontram-se no formato de vídeo (`.mp4`). A nomenclatura segue o padrão:

```
ID_GENDER_AGE_HEIGHT_WEIGHT_ETHNICITY_TYPEDATA
```

Excerto de Código 4.1: Nomenclatura dos Ficheiros Indoor

Onde `TYPEDATA` pode ser:

  - `L` – vista lateral esquerda
  - `R` – vista lateral direita
  - `F` – vista frontal

Exemplo de nome de ficheiro: `0001_0_3_2_1_0_L`

- **Outdoor:** neste conjunto, os dados encontram-se em formato de vídeo (`.mp4`), sendo necessário um processo de extração de frames para utilização no modelo. A nomenclatura dos vídeos codifica múltiplos atributos, como data, hora, localização, aparência e contexto da cena, seguindo o padrão:

```
ID\_day\_month\_year\_hour\_minute\_altitude\_horizontal-
    ↪ distance\_angle\_location\_...
```

Excerto de Código 4.2: Nomenclatura dos Ficheiros *Outdoor*

Exemplo: `0004_23_10_2024_09_18_5.8_10_30_3_0_1_1_0_-1_-1_6_0_0_6_0_178_1.mp4`

De modo a tornar os dados *outdoor* compatíveis com o modelo **Homogenus**, foi desenvolvido um pequeno *script* em *Python* que percorre todas as subpastas da diretório de vídeos e extrai os *frames* de cada um, guardando as imagens em formato (`.jpg`) numa pasta específica, neste caso, a pasta *Frames*. Estes *frames* constituem as imagens de entrada para o *pipeline* de inferência, sendo processadas individualmente.

Posteriormente, foi criado um segundo *script* para guardar as *labels* dos vídeos *Indoor* num ficheiro (`.csv`), a ser utilizado futuramente na comparação dos resultados obtidos pelo modelo.

4.2.2 Primeira Experiência

Inicialmente, foi testada a utilização do modelo **OpenPose** oficial, disponível no repositório da CMU [23, 24, 25, 26]. No entanto, verificaram-se dificuldades técnicas durante o processo de clonagem e configuração, nomeadamente na impossibilidade de descarregar os pesos — mesmo manualmente — como: `pose_iter_584000.caffemodel`, `pose_iter_116000.caffemodel`, `pose_iter_102000.caffemodel`, que corresponde aos pesos do corpo, mãos e cara essenciais para o funcionamento do modelo.

Como alternativa, foi utilizada a implementação em **PyTorch** [25, 26], a qual permitiu extrair os *keypoints 2D* das imagens. Os resultados foram guardados em ficheiros no formato (`.json`), com estrutura semelhante à utilizada pelo **OpenPose**, mas com algumas diferenças notáveis: os pontos-chave da face e das mãos não estavam presentes, o que limita a riqueza da representação corporal.

Para testar o *pipeline* do **Homogenus**, foram selecionadas aleatoriamente **10 imagens** da pasta *Frames* e extraíram-se os respetivos pontos-chave 2D. Para assegurar a qualidade dos dados de entrada, foi implementado um **filtro automático** no ficheiro `homogenus_infer.py` do modelo, de forma a ignorar imagens que não tivessem ficheiros `.json` válidos, ou seja, ficheiros nos quais o **OpenPose** não detetou pessoas. No entanto, apenas **4 imagens** geraram ficheiros `.json` com um número suficiente de pontos-chave detetados.

Posteriormente, foi utilizada essa mesma pasta e os seus ficheiros (`.json`) correspondentes como entrada no modelo **Homogenus**. O modelo devolve, para cada pessoa detetada, uma imagem com a *bounding box*, juntamente com a predição do género e as probabilidades associadas às classes *male* e *female*. Além disso, devolve os ficheiros (`.json`) atualizados, contendo os seguintes campos adicionais:

- `gender_pd` – género predito (masculino, feminino ou neutro)
- `male_prob` – probabilidade associada à classe "homem"
- `female_prob` – probabilidade associada à classe "mulher"

4.2.3 Resultados da Primeira Experiência

Dados de entrada

- **Imagens de Entrada:**

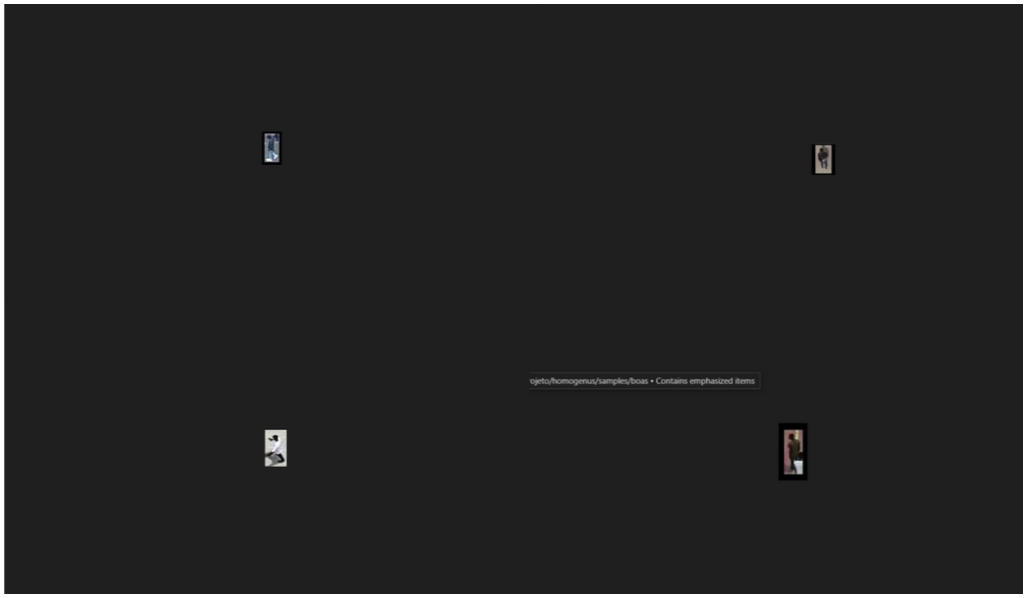

Figura 4.1: Imagens de entrada do primeiro teste.

- ***Pontos-Chave de Entrada:***

```
{"version": 1.3, "people": [{"person_id": [-1], "
    ↪ pose_keypoints_2d":[0, 0, 0, 22.48, 27.50, 0.655,
    ↪ 28.19, 25.90, 0.618,
30.13, 35.95, 0.112, 0, 0, 0, 17.46, 28.98, 0.494,
17.23, 38.57, 0.225, 16.54, 48.73, 0.164, ...]
```

Excerto de Código 4.3: Exemplo de pontos-chave 2D de entrada.

Dados de saída

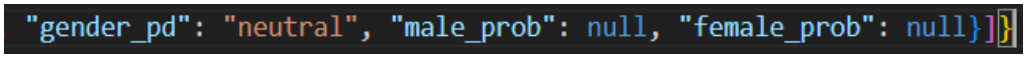

Figura 4.2: Dados do ficheiro Json.

Interpretação

Durante os testes iniciais, foram utilizadas quatro imagens de corpo inteiro, ilustradas na Figura 4.1. Cada imagem foi processada com a implementação

em **PyTorch** do **OpenPose**[5], de onde foram extraídos os pontos-chave 2D e logo armazenados em ficheiros `.json`. No entanto, verificou-se que, mesmo com esses ficheiros gerados, o modelo **Homogenus** [6] não executou a predição completa do género. As imagens resultantes foram visualmente idênticas às originais, sem qualquer *bounding box*, e os ficheiros associados indicavam probabilidades nulas para ambas as classes4.2. Este comportamento indica que os dados de entrada não passaram os filtros de validação da *pipeline* interna.

Como mencionado3.3.2, o modelo inclui vários filtros rigorosos antes da fase de inferência. A interpretação destes dados4.3 revela que diversos pontos apresentam coordenadas nulas (`0, 0`) e confianças muito baixas (como 0.112 ou 0.164), ou seja, pontos-chave com confianças inferiore a 0 5. Isto indica deteção incompleta ou ruidosa da pose corporal. Pontos essenciais como a cabeça, pescoço ou ombros estão ausentes ou mal definidos, comprometendo o sucesso da etapa seguinte de recorte e inferência.

Adicionalmente, a implementação utilizada não inclui os modelos para detetar **mãos** e **face** (ao contrário da versão original do **OpenPose**), o que limita ainda mais a qualidade e abrangência dos *keypoints* extraídos. Este facto, combinado com o reduzido tamanho das imagens de entrada, contribuiu para que o modelo rejeitasse imagens pequenas, neste caso, imagens **inferior a 200x200 px**, ainda na fase inicial do pipeline.

Em suma, a ausência de *bounding boxes* nas imagens de saída e a inexistência de probabilidades de género indicam que as instâncias foram rejeitadas por não cumprirem os critérios mínimos estabelecidos pelo modelo. Esta análise valida o correto funcionamento dos filtros de segurança do sistema, que impedem inferências em condições de entrada incerta ou duvidosa.

4.2.4 Segunda Experiência

Face ao tempo limitado para efetuar análises profundadas e aos resultados preliminares da primeira experiência4.2.2 com o modelo **Homogenus**, optou-se por um segundo teste com foco em imagens de melhor qualidade, utilizando o **OpenPose** no ambiente Google Colab[11] completamente funcional.

Seleção de Imagens de Melhor Qualidade

Com base na experiência anterior, foi observado que imagens de baixa resolução e/ou demasiado pequenas são rejeitadas pelos filtros internos do **Homogenus**. Por isso, foi criado um *script* para selecionar as 10 000 imagens com melhor resolução do conjunto de dados original, a pasta *Frames*. A nomenclatura dessas imagens foram guardados num ficheiro (`.txt`) e, posteriormente,

foi utilizado um *script* em *Python* para guardar essas imagens em específico numa nova pasta chamada *boas*.

Exemplo das imagens de *Input*

Figura 4.3: Imagens de entrada da segunda experiência.

Extração dos pontos-chave com OpenPose (Colab)

Como mencionado em 4.2.4, utilizou-se uma versão funcional do **OpenPose** via Google Colab[11] para processar as 10 000 imagens selecionadas. Destas, apenas 6 052 geraram ficheiros `.json` com pontos-chave, indicando que as restantes não continham detecções válidas.

Filtragem das Imagens e keypoints

Das 10,000 imagens inicialmente disponíveis, apenas 6,052 resultaram na deteção de pessoas com *keypoints*. Dada a acentuada diminuição de imagens com *keypoints* válidos, optou-se por aplicar um filtro adicional, com o objetivo de melhorar a qualidade dos dados utilizados e, assim, obter um melhor desempenho nas fases seguintes da análise. Para isso, foi desenvolvido um *script* em *Python* que percorre o conjunto das imagens e os respetivos ficheiros `.json`, retendo apenas os exemplos cuja pose corporal apresenta *keypoints* de qualidade aceitável. Apenas as amostras que satisfazem estes

critérios são copiadas para uma pasta de saída, garantindo que as análises posteriores se baseiam em dados mais fiáveis e representativos.

Funcionamento do *script*:

Função `is_good_pose`

Esta função recebe os *keypoints* de uma imagem onde foi detetada uma pessoa. Verifica se uma pose é considerada "boa", retornando `True` apenas se existirem pelo menos 10 *keypoints* com confiança superior a 0,5. A confiança corresponde ao terceiro valor de cada *keypoint*, na estrutura `(x, y, confiança)`.

Função `crop_bbox_from_keypoints`

Esta função calcula o *bounding box* (caixa envolvente) dos *keypoints* válidos, considerando apenas os que têm confiança superior a 0.5. Retorna a largura e altura da região detetada.

Loop Principal

O *loop* principal do *script* percorre todos os ficheiros `(.json)` na pasta indicada:

- Ignora ficheiros que não sejam `(.json)` ou que não contenham pessoas detetadas.
- Para cada pessoa, verifica se cumpre os critérios de qualidade através das funções descritas em 4.2.4 e em 4.2.4 .
- Se a pose for válida e o *bounding box* tiver pelo menos 100x100 píxeis, copia a imagem e o `.json` correspondente para as pastas de saída.
- Apenas uma pessoa válida por imagem é considerada.

Ao final desta filtragem, foram mantidas **3 017 imagens** e respetivos ficheiros `(.json)`, os quais serviram como entrada para o modelo.

4.2.5 Resultados da Segunda Experiência

Como dados de entrada foram utilizadas as seguintes imagens 4.3 e os respetivos pontos-chave em formato `(.json)`. Após o processamento, o modelo devolve novas imagens com as respetivas *bounding boxes* desenhadas, bem

como ficheiros (`.json`) atualizados. Assim, os resultados das predições do modelo são extraídos diretamente desses ficheiros (`.json`) 4.5 gerados durante a execução.

Os ficheiros contêm, além dos pontos-chave detetados, a predição do género (`gender_pd`) e as probabilidades associadas às classes `masculino` (`male_prob`) e `feminino` (`female_prob`).

Para facilitar a análise posterior e a visualização dos resultados, todas as informações relevantes foram organizadas num ficheiro (`.csv`), gerado por meio de um *script* automatizado.

Este *script* realiza os seguintes passos:

1. Lê os ficheiros `.json` que resultaram após a execução do modelo;
2. Extrai o `ID` a partir do nome do ficheiro, que corresponde ao primeiro campo da nomenclatura do ficheiro;
3. Extrai de cada ficheiro `.json` a predição de género e as probabilidades associadas;
4. Em casos de predição `neutral` com probabilidades associadas, é selecionada automaticamente a classe com maior valor:

```
if genero_predito == "neutral" and prob_masculino is not None
    ↪  and prob_feminino is not None:
     genero_predito = "male" if prob_masculino >=
    ↪ prob_feminino else "female"
```

Excerto de Código 4.4: Forçar classificação binária em predições neutras com base nas probabilidades.

5. Ignora ficheiros com predições `neutral` que não contêm probabilidades válidas associadas.
6. Guarda os dados válidos no ficheiro `resultados_outdoor.csv`. O ficheiro está organizado por colunas. Coluna para o ID, género predito, probabilidade de ser `masculino` e probabilidade de ser `femenino`.

Resultados Visuais do Modelo

Figura 4.4: Output da segunda experiência.

"gender_pd": "neutral", "male_prob": 0.3487938642501831, "female_prob": 0.6512061357498169}]}

Figura 4.5: Ficheiros jsons da egunda experiência.

Verificação dos Resultados

Para comparar os resultados preditos com os dados reais, realizou-se a junção de dois ficheiros `.csv` já existentes:

- `labels_videos.csv`: ficheiro que contém os géneros reais (0 para masculino, 1 para femenino).
- `resultados_outdoor.csv`: ficheiro que contém os géneros preditos pelo modelo.

A comparação foi feita com o seguinte procedimento:

1. Os dados foram fundidos pela coluna `ID`.

2. O valor de género real foi mapeado para `male/female`.

3. Adicionou-se uma nova coluna `Acertou` com valor booleano para indicar acertos/falhas.

Em suma, das 3 017 imagens, apenas **1 314** imagens é que passaram todos os filtros com sucesso.

Resultados Finais

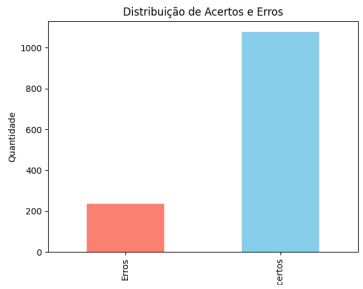

Figura 4.6: Distribuição acertos e erros.

Total de Imagens	Acertos	Erros	Taxa de acertos
1.314	1.077	237	81,96%

Tabela 4.1: Resultados da primeira experiência do Homogenus.

Interpretação

Dos 10 000 exemplos iniciais, 6 052 geraram pontos-chave válidos. Após a aplicação de filtros externos de qualidade, baseados na confiança dos pontos detetados e na dimensão mínima da *bounding box* (100×100 píxeis), obteve-se um subconjunto de 3 017 imagens consideradas adequadas para inferência.

Contudo, nem todas estas imagens foram efetivamente processadas pelo modelo **Homogenus**. Isso deve-se à aplicação de critérios internos mais exigentes, nomeadamente a rejeição de imagens cuja área recortada seja inferior a 200×200 píxeis. Assim, várias instâncias aprovadas externamente acabaram por ser descartadas internamente.

Além disso, nos casos em que o modelo classificava a imagem como `neutral`, a classe com maior probabilidade foi considerada como predita. Este procedimento permitiu maximizar a cobertura da análise sem comprometer a integridade dos dados.

No total, 1 314 imagens geraram uma predição binária válida (`masculino ou feminino`), constituindo a base para a avaliação do desempenho. A taxa de acerto obtida foi de **81,96%.**

Análise Gráfica

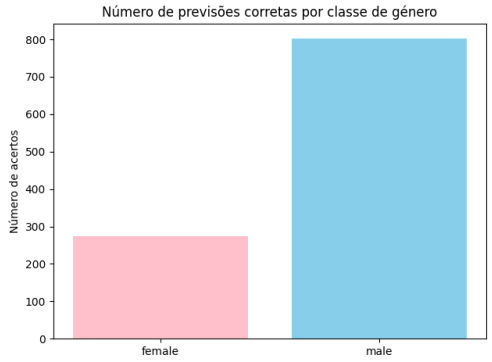

Figura 4.7: Número de previsões por classe de género.

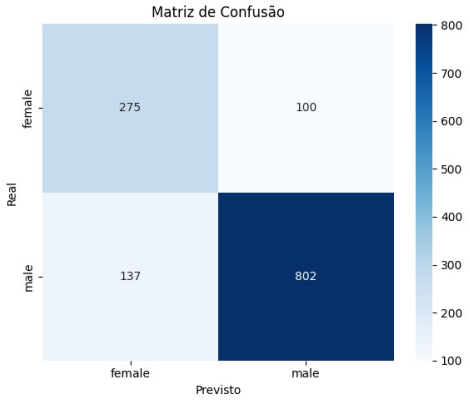

Figura 4.8: Matriz de confusão da primeira experiência.

Interpretação

A avaliação do modelo **Homogenus** sobre um total de 1 314 imagens revelou um desempenho global satisfatório, especialmente para a classe `male`. De acordo com a matriz de confusão, foram corretamente classificadas 802 imagens de indivíduos do género masculino e 275 do género feminino. Por outro lado, verificaram-se 137 classificações incorretas de indivíduos do género masculino como `female`, e 100 classificações erradas da classe `female` como `male`.

Estes resultados evidenciam uma assimetria no desempenho entre as classes, com vantagem para a predição da classe maioritária. Esta discrepância pode ser justificada pela distribuição desequilibrada do conjunto de teste, composto por 939 imagens da classe `male` e apenas 375 da classe `female`. Importa também sublinhar que, apesar da aplicação de filtros para garantir a qualidade dos dados de entrada, o desbalanceamento estrutural do conjunto de imagens permaneceu.

Adicionalmente, é relevante notar que o modelo foi utilizado unicamente em regime de inferência, com recurso aos pesos pré-treinados fornecidos pelos autores. Logo, o desempenho observado reflete não só as capacidades da arquitetura, mas também possíveis enviesamentos herdados dos dados utilizados durante o treino original.

Métricas por clase

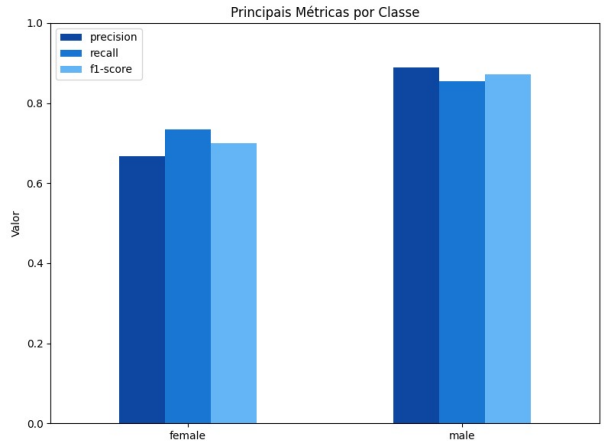

Figura 4.9: Métricas por género.

O gráfico de barras apresenta as principais métricas de desempenho *precision*, *recall* e *f1-score* para cada uma das classes consideradas pelo modelo:

- **Classe Feminina:**
  - **Precisão**: quando o modelo prevê que alguém é do género feminino, acerta em cerca de 67% dos casos. Ou seja, há uma taxa considerável de falsos positivos (pessoas classificadas como *female* que na verdade são *male*).
  - ***Recall***: O modelo consegue identificar 73% das pessoas realmente do género *female*. Ou seja, perde 27% dessas pessoas (falsos negativos).
  - ***F1-score***: representa o equilíbrio entre precisão e o *recall*. Um valor médio, indicando que o modelo tem dificuldades em reconhecer esta classe.
- **Classe Masculina:**
  - **Precisão**: O modelo erra pouco ao prever *male* — as previsões são 89% corretas.
  - ***Recall***: identifica corretamente 85% das pessoas realmente do género masculino.
  - ***F1-score***: ~0,87, mostra que o modelo é consistente e equilibrado ao lidar com a classe maioritária.

4.3 Modelo Attribute-Recognition

4.3.1 Experiência Utilizando ResNet-18

Nesta experiência, utilizámos imagens que foram previamente filtradas para o modelo **Homogenus** 4.2.4, num total de **3,017 imagens**.

Como dados de entrada, foi unicamente necessário o conjunto de imagens sobre o qual se pretendia testar os resultados. Após a execução, obtiveram-se saídas do tipo ilustrado na Figura 4.5.

```
Predicted results: {
    'labels': array(['Age-Adult', 'Gender-Female',
                     'LowerBody-Color-Black',
                     'LowerBody-Type-Trousers&Shorts'], dtype='<
    ↪ U30'),
    'prob': array([0.64786081, 0.61053316, 0.63944295,
    ↪ 0.85024354])
}
```

Excerto de Código 4.5: Saída predita pelo modelo.

- **labels:** representa os atributos inferidos como presentes na imagem, ou seja, aqueles cuja probabilidade foi superior a 0.5.
- **prob:** apresenta a respetiva probabilidade associada a cada atributo inferido.

A validação consistiu na comparação entre os valores reais do género, contidos no ficheiro `labels_videos.csv`, e os valores preditos pelo modelo, registados no ficheiro `resultados.csv`, gerado automaticamente após a execução do `inference.py`.

Para tal, procedeu-se à fusão dos dois ficheiros, utilizando o campo `ID` como chave. O género real foi obtido a partir da coluna `Género` e mapeado para *male* ou *female*.

Em relação ao género predito, é importante salientar que o modelo apenas devolve explicitamente o atributo `Gender-Female` quando este está presente com probabilidade superior a 0,5. Caso contrário, nenhum género é retornado e, por exclusão, assume-se o género *male*. Esta lógica foi implementada no processo de extração dos resultados inferidos.

Adicionalmente, com o auxílio de um pequeno *script* em *Python*, foi gerado o ficheiro `comparacao_final.csv`, que resume as predições realizadas pelo modelo e a sua correspondência com os valores reais, permitindo uma

avaliação quantitativa e qualitativa do seu desempenho. Para facilitar a análise, foi criada uma nova coluna, `Acertou`, que indica se a predição coincidiu com o género real.

4.3.2 Resultados da Primeira Experiência

Resultados finais

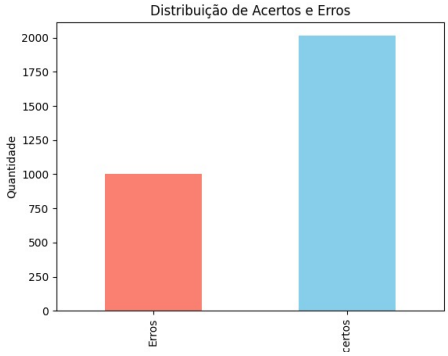

Figura 4.10: Distribuição de acertos e erros.

Total de Imagens	Acertos	Erros	Taxa de acertos
3.017	2.014	1.003	66,76%

Tabela 4.2: Resultados após execução do modelo Attribute-Recognition.

Interpretação

A análise da Figura 4.10 e da Tabela 4.2 mostra que o modelo **Attribute-Recognition** atingiu uma taxa de acerto de **66,76%**, com 2.014 classificações corretas em 3.017 imagens. Embora razoável, este desempenho revela limitações na tarefa específica de classificação binária de género.

Isso pode ser explicado pelo facto de o modelo ter sido treinado para prever múltiplos atributos em simultâneo, e não exclusivamente o género. Além disso, foram utilizados pesos pré-treinados, sem qualquer re-treinamento neste projeto, limitando a adaptação ao novo conjunto de dados.

O modelo apenas retorna atributos com probabilidade superior a 0,5, o que pode levar à sub-representação da classe *female* se as probabilidades forem sistematicamente baixas, mesmo que visivelmente corretas. Esta abordagem pode amplificar erros, especialmente em dados não equilibrados.

Análise Gráfica

Previsões por classe de Género

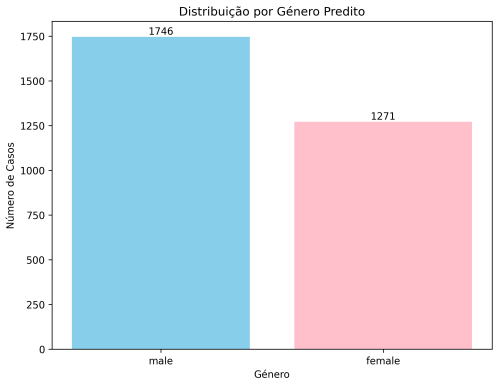

Figura 4.11: Número de previsões por classes.

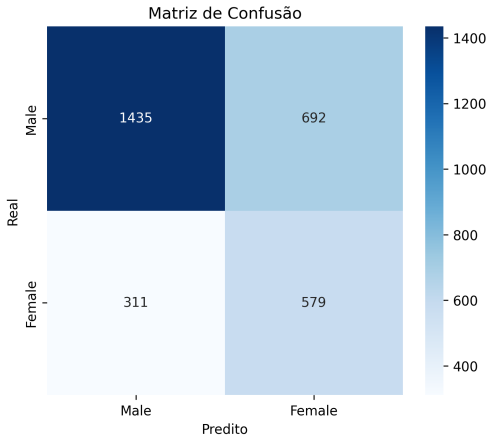

Figura 4.12: Matriz de confusão utilizando resnet-18.

Interpretação

A análise dos gráficos presentes nas figuras 4.11 e 4.12 mostram que o modelo classificou 1.746 imagens como *male* e 1.271 como *female*, valores próximos da distribuição real (2.127 homens e 890 mulheres). A matriz de confusão indica um melhor desempenho na classe `masculina`, com 1.435 acertos, contra 579 na classe `femenina`. O modelo errou mais ao classificar mulheres (692 erros), o que pode ser explicado pela sua lógica de inferência: se a probabilidade do atributo `Gender-Female` não ultrapassa 0,5, assume-se o género *male*. Isso tende a penalizar a classe minoritária em casos de incerteza.

Métricas por classe

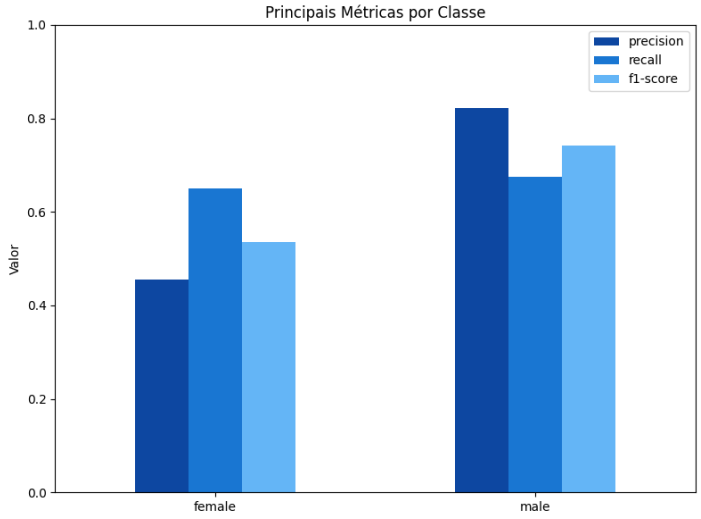

Figura 4.13: Métricas por classe.

Este gráfico compara três métricas importantes para cada género:

Classe Feminina:

- **Precisão:** 0,46 (46% das predições femeninas estão corretas);
- ***Recall:*** 0,65 (65% das mulheres reais foram identificadas);
- ***F1-Score:*** 0,54 (média harmónica entre precisão e *recall*)

Classe Masculina:

- **Precisão:** 0,83 (83% das predições masculinas estão corretas);
- ***Recall:*** 0,67 (67% dos homens reais foram identificados);
- ***F1-Score:*** 0,74 (média harmónica entre precisão e *recall*);

4.3.3 Experiência Utilizando ResNet-50

Nesta segunda experiência, avaliou-se a arquitetura **ResNet-50** na tarefa de classificação de género, utilizando exatamente o mesmo procedimento aplicado à **ResNet-18**: mesma função de pré-processamento, conjunto de imagens e lógica de extração do atributo `Gender`.

A única alteração consistiu na substituição do ficheiro de pesos, recorrendo à versão previamente treinada da **ResNet-50**, disponibilizada no repositório original. Esta troca foi efetuada apenas ajustando o caminho no comando para correr o *script* de inferência, mantendo-se inalterada toda a *pipeline*.

Resultado Finais

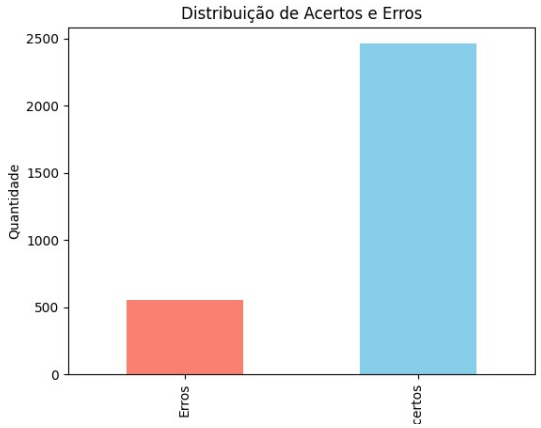

Figura 4.14: Distribuição de acertos e erros com o modelo ResNet-50.

Imagens	Acertos	Erros	Taxa de Acertos
3.017	2.461	556	81,57%

Tabela 4.3: Resultados após execução do modelo ResNet-50.

Interpretação

Consoante a análise da tabela 4.3, a utilização da arquitetura **ResNet-50** resultou numa melhoria significativa no desempenho da tarefa de classificação de género, com uma taxa de acerto de **81,57%**, correspondente a 2.461 predições corretas num total de 3,017 imagens.

Este resultado demonstra que o aumento da profundidade da rede, em comparação com a **ResNet-18**, contribuiu para uma melhor capacidade de generalização do modelo, mesmo sem qualquer ajustamento adicional ou fase de re-treino. O número reduzido de erros — 556 casos — reforça a maior aptidão da **ResNet-50** para captar padrões discriminativos mais robustos, essenciais para a tarefa em análise.

A diferença entre o número de acertos e erros, visível na figura, ilustra claramente esta melhoria, evidenciando o impacto positivo da maior complexidade arquitetónica no desempenho global do modelo.

Análise Gráfica

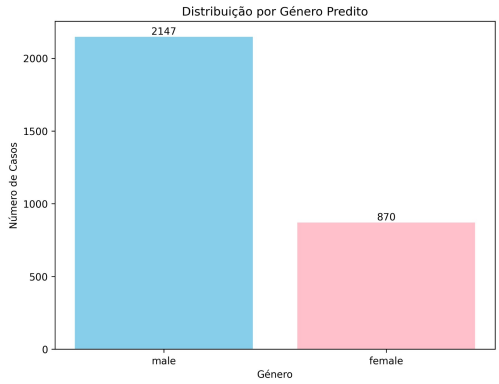

Figura 4.15: Distribuição de previsões por género ResNet-50.

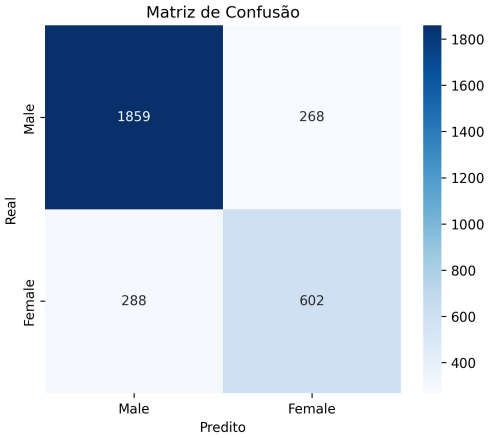

Figura 4.16: Matriz de confusão da ResNet-50.

Interpretação

A análise das Figuras 4.15 e 4.16 demonstra que a arquitetura **ResNet-50** alcançou uma taxa de acerto de **81,57%**, superando a **ResNet-18**.

Observa-se uma predominância de predições para a classe *male* (2.147), em linha com o desequilíbrio do conjunto de dados utilizado, que contém 2.127 imagens de homens e apenas 890 de mulheres. Esta desproporção poderá ter afetado negativamente a capacidade do modelo em identificar corretamente indivíduos do género *female*, como ilustrado pela matriz de confusão.

Os resultados evidenciam o potencial das redes mais profundas na generalização, mesmo sem re-treino. No entanto, salientam também a necessidade de dados equilibrados para evitar enviesamentos na classificação.

Métricas por classe

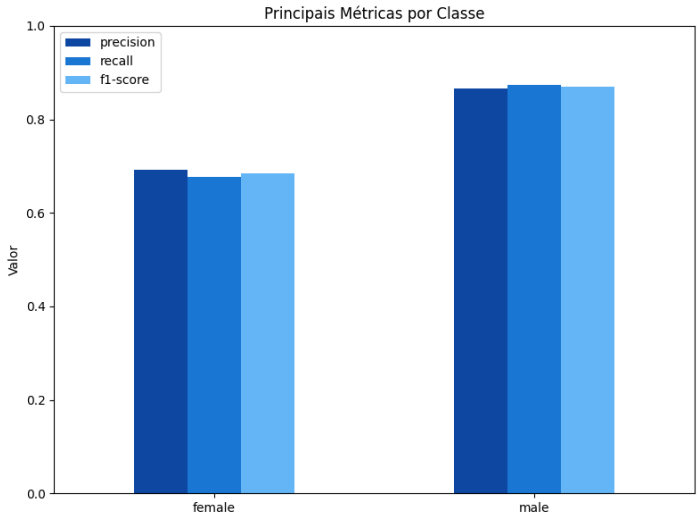

Figura 4.17: Métricas por clase.

Interpretação

A Figura 4.17 apresenta uma análise detalhada das métricas de desempenho do modelo **ResNet-50** por classe, especificamente, para os géneros `female` e `male`:

Classe Feminina:

- **Precisão:** 0,68 (68% das predições femininas estão corretas);
- ***Recall:*** 0,69 (69% das mulheres reais foram identificadas);
- ***F1-Score:*** 0,68 (média harmónica entre precisão e *recall*)

Classe Masculina:

- **Precisão:** 0,87 (87% das predições masculinas estão corretas);
- ***Recall:*** 0,87 (87% dos homens reais foram identificados);
- ***F1-Score:*** 0,87 (média harmónica entre precisão e *recall*);

A classe `male` apresenta valores elevados e consistentes em todas as métricas: **precisão**, **revocação (recall)** e **f1-score**, todas próximas ou superiores a 0.87. Estes resultados indicam que o modelo consegue não só identificar corretamente a maioria dos homens, como também tende a não cometer muitos falsos positivos nesta classe. Isto sugere uma elevada confiança do modelo na classificação de indivíduos do género masculino.

Por outro lado, a classe `female` apresenta valores inferiores: a precisão situa-se em torno de 0.69, a revocação ligeiramente abaixo (≈ 0.68), e o *f1-score* acompanha esta tendência.

Este comportamento é comum em cenários de **desequilíbrio de classes**, onde o modelo tende a otimizar o desempenho para a classe maioritária. Assim, embora o desempenho global do modelo (taxa de acerto global) seja satisfatório, a análise por classe revela a necessidade de abordar este viés, seja por re-amostragem, aplicação de técnicas de balanceamento, ou ajustes no processo de treino.

4.4 Conclusões

Neste capítulo, foram avaliados três modelos aplicados à tarefa de classificação de género em imagens de corpo inteiro: **Homogenus**, **Attribute-Recognition com ResNet-18** e **Attribute-Recognition com ResNet-50**. A análise considerou métricas como taxa de acerto, precisão, *recall* e *f1-score*, com foco na capacidade de cada modelo em distinguir corretamente indivíduos do género `masculino` e `feminino`.

O modelo **Homogenus**, baseado em reconstrução tridimensional do corpo, destacou-se pela sua capacidade de incorporar informação estrutural, alcançando uma taxa de acerto de **81,96%**. Obteve bons resultados, especialmente na classe masculina, mas mostrou alguma limitação na predição de indivíduos do género feminino, possivelmente devido à menor representação desta classe no conjunto de dados.

O modelo **Attribute-Recognition com ResNet-18**, por outro lado, apresentou o pior desempenho geral (**66,76%**), com dificuldades notórias na identificação correta de mulheres. Isso deve-se, em parte, à arquitetura mais simples da rede e ao facto de o modelo ser *multi-label*, o que dispersa o foco da tarefa de classificação de género.

A versão com **ResNet-50** demonstrou uma melhoria expressiva, atingindo **81,57%** de taxa de acerto. A maior profundidade da rede contribuiu para uma representação visual mais robusta, com melhores resultados tanto na identificação de homens quanto de mulheres. Ainda assim, observou-se um número superior de erros na classe `female`, reflexo do desequilíbrio no conjunto

de teste (com mais imagens de homens) e da ausência de qualquer re-treino adaptado ao contexto.

Em resumo, os resultados evidenciam que modelos pré-treinados mais profundos podem atingir bom desempenho mesmo sem afinação, mas continuam sensíveis a dados não balanceados. A classe *female* foi consistentemente mais difícil de prever, o que destaca a importância de equilibrar os dados e, eventualmente, ajustar os modelos ao domínio específico.

A tabela seguinte 4.4 resume as principais métricas de desempenho:

Métrica	**Homogenus**	**AttrRecog (ResNet-18)**	**AttrRecog (ResNet-50)**
Total de Imagens	1.314	3.017	3.017
Taxa de acerto	81,96%	66,76%	81,57%
Precisão (Masculino)	89%	83%	87%
Recall (Masculino)	85%	67%	87%
F1-Score (Masculino)	87%	74%	87%
Precisão (Feminino)	67%	46%	68%
Recall (Feminino)	73%	65%	69%
F1-Score (Feminino)	70%	54%	68%

Tabela 4.4: Comparação de desempenho entre os modelos.

Capítulo

5

Conclusões e Trabalho Futuro

5.1 Conclusões Principais

O projeto permitiu explorar e comparar diferentes abordagens para a classificação de género em imagens de corpo inteiro, no contexto de videovigilância, com foco nas características de *soft biometrics*. Os modelos avaliados foram o **Homogenus**, baseado na reconstrução tridimensional do corpo humano, e o **Attribute-Recognition**, com as arquiteturas **ResNet-18** e **ResNet-50**, que operam diretamente sobre imagens RGB.

A principal conclusão é que o modelo **Homogenus** apresentou um excelente desempenho, mesmo com um conjunto reduzido de imagens, atingindo uma taxa de acerto de 81,96%. Este desempenho pode ser atribuído à incorporação de informação estrutural, mediante pontos-chave e malhas 3D, permitindo uma representação corporal mais robusta. No entanto, este modelo demonstrou dificuldades em processar imagens de péssima resolução, com dimensões muito reduzidas ou capturadas sob ângulos não convencionais — como, por exemplo, imagens obtidas a partir de cima, com ângulos de 90 graus, possivelmente recolhidas por drones —, uma vez que o *pipeline* depende da extração de pontos-chave confiáveis e de qualidade suficiente para alimentar a reconstrução tridimensional.

Por outro lado, o modelo **Attribute-Recognition** demonstrou como abordagens puramente visuais podem beneficiar do aumento da profundidade da arquitetura. A transição de **ResNet-18** para **ResNet-50** resultou num salto significativo de desempenho, com a taxa de acerto a subir de 66,76% para 81,57%. Este resultado mostra que redes profundas são capazes de extrair padrões visuais mais ricos e discriminativos, aproximando-se do desempenho do modelo **Homogenus**. Contudo, não foi possível testar o modelo **Attribute-**

Recognition com imagens de baixa qualidade ou resolução muito reduzida, dado que se optou por manter o conjunto de dados inalterado em ambas as experiências, visando uma comparação mais direta dos seus desempenhos.

Conclui-se que modelos que utilizam informação estrutural, como o **Homogenus**, têm elevado potencial mesmo com conjuntos de dados pequenos, desde que as imagens respeitem critérios mínimos de qualidade e estrutura. Por sua vez, o modelo **Attribute-Recognition** com **ResNet-50** apresenta-se como uma solução competitiva para aplicações em tempo real, graças à sua simplicidade, desempenho equilibrado e menor dependência de pré-processamentos complexos. Por fim, destaca-se que o desempenho dos modelos está diretamente relacionado com a qualidade dos dados, a arquitetura utilizada e a representação adotada para captar as características relevantes ao problema.

5.2 Trabalho Futuro

Embora o presente trabalho tenha cumprido o objetivo proposto — avaliar o desempenho de modelos pré-treinados na tarefa de classificação de género em contextos de videovigilância, algumas etapas relevantes ficaram por explorar e poderão ser consideradas em trabalhos futuros.

Uma das principais oportunidades de aprofundamento será a realização de **fine-tuning** do modelo **Attribute-Recognition**, ajustando os seus pesos às características específicas do conjunto de dados utilizado neste projeto. Esta etapa poderá permitir ganhos significativos de desempenho, nomeadamente na precisão e equilíbrio entre as classes

Bibliografia

[1] Hugo Proença. Evaluation of the sota in soft biometrics for visual surveillance settings: Project proposal. `https://github.com/nghorbani/homogenus`, 2024. Universidade da Beira Interior.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE transactions on pattern analysis and machine intelligence*, volume 43, pages 172–186, 2021.

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. Acedido de junho de 2025.

[7] R. Dev. Bosch age and gender prediction, 2020. `https://github.com/rdev12/BOSCH-Age-and-Gender-Prediction`.

[8] Kristijan Bartol. Human-sex classifier, 2021. `https://github.com/kristijanbartol/human-sex-classifier`.

[9] D. Sabarinathan. Multi-attribute classification from full-body images. `https://github.com/dsabarinathan/attribute-recognition`, 2022. Acedido de junho de 2025.

[10] rdev12. Bosch age and gender prediction using faster r-cnn and resnet. `https://github.com/rdev12/BOSCH-Age-and-Gender-Prediction`, 2021. Acedido de junho de 2025.

[11] Google Colab User. Openpose - google colab notebook. `https://colab.research.google.com/drive/15YDdKzLiFBnt4JPTNxPWnrg-Lx6Ybh-7`, 2025. Acessado em junho de 2025.

[12] Andreas Specker, Mickael Cormier, and Jürgen Beyerer. Upar: Unified pedestrian attribute recognition and person retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[13] Mickael Cormier, Andreas Specker, Julio C. S. Jacques Junior, Lucas Florin, Jürgen Metzler, Thomas B. Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge: Pedestrian attribute recognition and attribute-based person retrieval – dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 166–175, January 2023.

[14] Mickael Cormier, Andreas Specker, Julio C. S. Jacques Junior, Lennart Moritz, Jürgen Metzler, Thomas B. Moeslund, Kamal Nasrollahi, Sergio Escalera, and Jürgen Beyerer. Upar challenge 2024: Pedestrian attribute recognition and attribute-based person retrieval - dataset, design, and results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 359–367, January 2024.

[15] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[16] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017.

[17] Y. Deng, P. Luo, C.C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *ACM Multimedia*, 2014.

[18] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. Acedido de junho de 2025.

[19] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. Acedido de junho de 2025.

[20] Microsoft Corporation. Python 3.7 - microsoft store, 2018. Acedido em junho de 2025.

[21] IBM Corporation. Pytorch: Uma biblioteca de aprendizado profundo, 2023. Acedido em junho de 2025.

[22] Martín Abadi et al. Tensorflow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org/versions/r1.15`, 2015. Software available from `https://www.tensorflow.org/`.

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

[24] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[25] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[26] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.