Reconhecimento Facial Multiespectral em Ambiente Não Controlado

Pedro Afonso Roque Martins

Dissertação para obtenção do Grau de Mestre em

Engenharia Eletrotécnica e de Computadores

Orientadores: Professor Doutor Alexandre José Malheiro Bernardino
Professor Doutor José Silvestre Serra da Silva

Júri:

Presidente: Professora Doutora Teresa Maria Sá Ferreira Vazão Vasques
Orientador: Professor Doutor Alexandre José Malheiro Bernardino
Vogais: Professor Doutor Carlos Jorge Andrade Mariz Santiago
Tenente-Coronel Henrique Martins dos Santos Cunha

Novembro de 2021

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Agradecimentos

A presente dissertação de mestrado não poderia chegar a bom porto sem o precioso apoio de várias pessoas e instituições. Esta simboliza uma das últimas etapas de um percurso que já vai em 6 anos na Academia Militar. Correndo o risco de injustamente não mencionar algum dos contributos, pretendo expressar os meus agradecimentos nas linhas seguintes.

Aos meus orientadores, Professor Alexandre Bernardino e Professor José Silvestre Silva agradeço por todo o tempo despendido para aconselhamento e orientação, e pela compreensão e empenho demonstrados ao longo desta etapa.

À Academia Militar pelo contributo que teve na minha formação e no meu desenvolvimento enquanto académico e soldado e, pela disponibilidade em garantir as melhores condições possíveis ao desenvolvimento deste trabalho.

A todos os meus amigos e camaradas da Academia Militar, e em especial ao José Almeida, Pedro Marques, Rodolfo Rodrigues, Fátima Martins e Luís Pires pelas incalculáveis horas de convívio. Apesar de a vida de internato não ser a mais aprazível, a amizade genuína que partilho convosco fez valer a pena.

A toda a minha família, aos que cá estão e aos que já partiram, e em especial aos meus pais e à minha irmã, que me apoiam em todas as derrotas e a quem dedico todas as minhas vitórias. Aos meus pais, que com espírito de abnegação me proporcionaram uma infância rica e feliz, ao lado da minha irmã, e me transmitiram os valores éticos necessários para enveredar nesta carreira. Vós representais tudo aquilo que aspiro ser.

Uma última palavra de gratidão para a Lu, pela sua paciência, compreensão e carinho, ao longo de mais um ano entre muitos. Sou grato por estares ao meu lado em toda esta jornada do ensino superior, pois o abraço com que sempre me recebes permite-me ultrapassar as adversidades com maior tranquilidade.

A todos o meu profundo e sincero agradecimento!

Pedro Afonso Roque Martins

Resumo

Neste trabalho é proposto um sistema de reconhecimento facial multiespectral num ambiente não controlado, com o objetivo de identificar ou autenticar identidades (pessoas) através das suas imagens faciais.

Os sistemas de reconhecimento facial em ambiente não controlado têm demonstrado uma melhoria contínua do desempenho através de soluções cada vez mais complexas e exigentes. Contudo, a maioria está limitada à utilização de apenas uma banda espectral. A utilização de imagens multiespectrais permite recolher informações que não são passíveis de obter na banda do visível quando existem determinadas oclusões (por exemplo, nevoeiro e materiais plásticos) e em ambientes com pouca ou nenhuma luminosidade. O trabalho proposto utiliza as pontuações obtidas nas diferentes bandas espectrais a fim de tomar uma decisão final conjunta na identificação. A avaliação dos diferentes métodos para cada tarefa permitiu selecionar os mais adequados para um sistema de reconhecimento facial multiespectral num ambiente não controlado.

Os resultados experimentais obtidos em *Rank-1* na base de dados multiespectral TUFTS foram de 99,5% e 99,6% com variação de pose e variação de expressão, respetivamente, e de 100,0% na base de dados CASIA NIR-VIS 2.0, indiciando que a utilização de imagens multiespectrais em ambiente não controlado é vantajosa quando comparada com a utilização de imagens de banda espectral única.

Palavras-Chave: ambiente não controlado, fusão de pontuações, reconhecimento facial multiespectral, redes neuronais profundas.

Abstract

This work proposes a multispectral face recognition system in an uncontrolled environment, aiming to identify or authenticate identities (people) through their facial images.

Facial recognition systems in an uncontrolled environment have shown continuous performance improvement through increasingly complex and demanding solutions. However, most are limited to the use of only one spectral band. The use of multispectral images makes it possible to collect information that is not obtainable in the visible band when certain occlusions exist (e.g., fog and plastic materials) and in low or no light environments. The proposed work uses the scores obtained in the different spectral bands to make a joint final decision in identification. The evaluation of different methods for each task allowed selecting the most suitable ones for a multispectral face recognition system in an uncontrolled environment.

The experimental results obtained in Rank-1 in the TUFTS multispectral database were 99.5% and 99.6% with pose variation and expression variation, respectively, and 100.0% in the CASIA NIR-VIS 2.0 database, indicating that the use of multispectral images in an uncontrolled environment is advantageous when compared with the use of single spectral band images.

Keywords: deep neural networks, multispectral face recognition, on the wild, score fusion.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Siglas e Acrónimos

AUC	Área Debaixo da Curva ROC
CCFF-GAN	*Cycle-Consistency Face Frontalization GAN*
CNN	*Convolutional Neural Network*
DA-GAN	*Dual-Agent GAN*
DCGAN	*Deep Convolutional GAN*
DCNN	*Deep Convolutional Neural Network*
DR-GAN	*Disentangled Representation-learning GAN*
DSFD	*Dual Shot Face Detector*
FFWM	*Flow-based Feature Warping Model*
FNM	*Face Normalization Model*
GAN	*Generative Adversarial Network*
IV	Infravermelho
LWIR	*Long-wavelength infrared*
MDNDC	*Multiple Deep Networks with scatter loss and Diversity Combination*
MFM	*Max-Feature-Map*
MWIR	*Mid-wavelength infrared*
NIR	*Near-Infrared*
PIE	Pose-Iluminação-Expressão
ROC	*Receiver operating characteristic*
S3FD	*Single Shot Scale-invariant Face Detector*
SSD	*Single Shot Multibox Detector*
SWIR	*Short-wavelength infrared*
TAF	Taxa de Aceitação Falsa
TAV	Taxa de Aceitação Verdadeira
UV-GAN	*GAN for UV Completion*
WCNN	*Wasserstein Convolutional Neural Network*

1. Introdução

O sentido da Visão sempre desempenhou um papel fundamental, sendo que é através dele que o ser humano obtém até 80% das informações do meio em que se encontra. Este sentido permite observar perigos, identificar objetos, e reconhecer pessoas. Esta última função é fundamental para o ser humano enquanto ser social, pois é o que permite diferenciar o nível de confiança que pode dar a determinada pessoa, estando na base de construção das comunidades.

Tal é a importância desta função, que se tornou um dos principais tópicos de investigação com o aparecimento da aprendizagem automática, permitindo assim que máquinas consigam incorporar esta capacidade biológica. Os atuais sistemas de reconhecimento facial que operam no domínio do visível atingiram um nível significativo de maturidade, sendo possível observar a sua ampla utilização nos dias que correm, desde mecanismos de segurança para desbloquear aparelhos eletrónicos como smartphones e computadores pessoais, até sistemas de controlo de população, onde a China é líder.

1.1. Motivação

Cenários de ambiente não controlado, como motins e manifestações violentas, podem muitas vezes ser utilizados por criminosos e elementos de células terroristas para se movimentarem e causarem danos à Segurança Nacional, pois este tipo de ambiente adiciona dificuldades à sua deteção.

O autor, como engenheiro eletrotécnico militar da Guarda Nacional Republicana, com interesse na área de aprendizagem automática, e preocupado com questões de Segurança Nacional, vê neste tema uma forma de juntar ambos os interesses, aumentando o seu conhecimento técnico, enquanto tem como objetivo auxiliar as diferentes Forças e Serviços de Segurança na mitigação de ameaças à Segurança Nacional.

1.2. Dificuldades

O ambiente não controlado é principalmente caraterizado por [1]:

- Variedade de iluminação;
- Variedade de pose;
- Variedade de expressões faciais;
- Existência de oclusões.

Estas características colocam desafios aos sistemas de reconhecimento facial devido às múltiplas variações intrapessoais que estes proporcionam, o que dificulta identificar corretamente a identidade do indivíduo, tendo por base uma imagem colaborativa deste. A comunidade científica tem feito um grande esforço para enfrentar os desafios do ambiente não controlado, desenvolvendo soluções cada vez mais complexas para ultrapassar os vários obstáculos, como a variedade de iluminação e a existência de oclusões. No entanto, até à presente data, não existem soluções capazes de resolver todos os problemas do ambiente não controlado, estando as soluções muitas vezes circunscritas à resolução de uma ou duas das caraterísticas do ambiente não controlado.

1.3. Objetivos

Este trabalho tem como objetivo principal o desenvolvimento de um sistema de reconhecimento facial multiespectral em ambiente não controlado. Para a consecução deste objetivo, são exploradas as soluções utilizadas pelos atuais sistemas de reconhecimento e a avaliação dos benefícios da utilização de imagens multiespectrais. O motivo da utilização de imagens multiespectrais é o facto de permitirem obter um maior leque de caraterísticas do ambiente em redor, e consequentemente existir mais informação para identificar os rostos presentes nas imagens.

1.4. Estrutura da Dissertação

A presente dissertação encontra-se dividida em seis capítulos, e está organizada da seguinte forma:

- **Capítulo 1 – Introdução:** neste capítulo é descrita a motivação do trabalho apresentado nesta dissertação de mestrado, os objetivos e a estrutura da dissertação;
- **Capítulo 2 – Conceitos Base:** neste capítulo são explanados conceitos importantes, como o funcionamento de um sistema de reconhecimento facial e o que são imagens multiespectrais;
- **Capítulo 3 – Trabalhos Relacionados:** neste capítulo é feito um estudo do estado da arte sobre os métodos de reconhecimento facial multiespectral em ambiente não controlado e das bases de dados multiespectrais públicas;
- **Capítulo 4 – Metodologia:** neste capítulo é definida e proposta a metodologia com vista à consecução dos objetivos da dissertação;
- **Capítulo 5 – Resultados e Discussão:** neste capítulo são descritas as bases de dados multiespectrais utilizadas. São também feitas diversas experiências aos diversos m´módulos propostos na metodologia. Cada experiência e acompanhada pela sua respetiva análise e discussão;
- **Capítulo 6 – Conclusões:** neste capítulo são apresentadas as conclusões deste trabalho, consolidando assim os objetivos propostos. São também apresentados os possíveis trabalhos futuros.

2. Conceitos Base

Neste capítulo são explicados os diferentes conceitos abordados ao longo do trabalho a ser desenvolvido.

2.1. Deteção e Reconhecimento Facial

A determinação da existência ou não de uma face numa imagem, é feita através de algoritmos de deteção facial. Estes algoritmos utilizam aprendizagem automática de forma a identificar corretamente zonas na imagem que correspondem a faces e delinear estas áreas, para que o sistema possa processar apenas na face.

A deteção facial surge como um pré-processamento para diversas aplicações que necessitam da correta deteção da face, tal como detetar o número de indivíduos numa determinada imagem, a sua expressão facial e especificar a sua identidade. Esta última consiste no reconhecimento facial [2].

De forma geral, um sistema de reconhecimento facial geral pode ser descrito pelas seguintes fases:

1. Entrada: A primeira fase do sistema de reconhecimento facial consiste na entrada de imagens faciais.
2. Pré-processamento: O pré-processamento tem como objetivo primário verificar se existem faces na imagem, através de um algoritmo de deteção facial. No entanto, este pode ser mais completo, por exemplo eliminando ruído da imagem e facilitar a fase seguinte.
3. Extração de Caraterísticas: Esta fase extrai um conjunto caraterísticas da imagem facial, as quais podem ser a posição de marcos faciais, distância dos olhos ou mesmo as tonalidades da face.
4. Identificação ou Verificação (abordada na subsecção seguinte).

As três primeiras fases estão ilustradas na Figura 1:

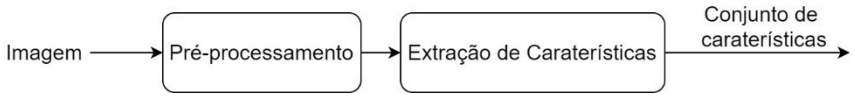

Figura 1 - Esquema resumo de obtenção de caraterísticas de uma imagem facial.

2.2. Verificação e Identificação

O reconhecimento facial tem duas aplicações principais: a verificação e a identificação. Enquanto a verificação consiste na habilidade de comparar duas faces e verificar se correspondem à mesma identidade ou não (comparação), a identificação consiste na capacidade de procurar se uma face corresponde a uma identidade presente numa galeria de imagens de identidades (procura) [2].

Para a realização de ambas as tarefas, é necessário possuir antecipadamente um registo da pessoa a identificar/verificar associado a um conjunto de caraterísticas. O processo de população da

base de dados está representado na Figura 2, em que cada pessoa terá na base de dados a sua identificação associada a um conjunto (ou vários) de caraterísticas obtidas de uma imagem facial dela.

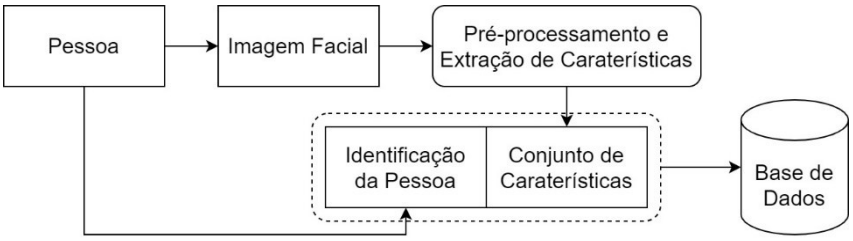

Figura 2 - Esquema de um registo numa base de dados de um sistema de reconhecimento facial.

Para efetuar a verificação, o utilizador deve captar uma imagem facial sua no momento, da qual se obterá um conjunto de caraterísticas. O utilizador deve também identificar-se de alguma forma requisitada pelo sistema, por exemplo com um número identificador. Com esta identificação prestada, o sistema irá obter da base de dados o conjunto (ou conjuntos) de caraterísticas associados a essa identificação. Assim, o conjunto de caraterísticas obtido da imagem facial prestada pelo utilizador será comparada com os conjuntos de caraterísticas associados à identidade que o utilizador afirma ser. Esta comparação, realizada através de funções de semelhança, por exemplo, serve para verificar se a pessoa é quem afirma ser ou não. Na Figura 3 está ilustrada a situação descrita.

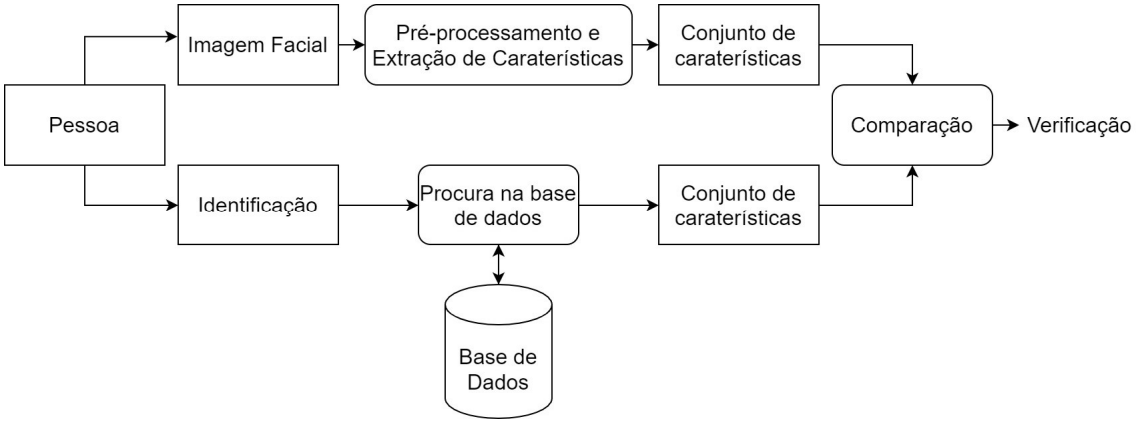

Figura 3 - Verificação num sistema de reconhecimento facial.

Para efetuar a identificação, o utilizador deve novamente captar uma imagem facial sua no momento, de forma a se obter um conjunto de caraterísticas. No entanto, diferente da situação da verificação, o identificador não irá fornecer a sua identificação. O objetivo do sistema de reconhecimento facial nesta situação é identificar a pessoa entre todas as identidades que possui na sua base de dados. Para tal, é feito uma comparação do conjunto de caraterísticas obtido da imagem facial prestada pelo utilizador com todos os conjuntos de caraterísticas presentes na base de dados, onde de cada comparação surge uma pontuação, de acordo com a sua semelhança. A identidade associada ao conjunto de caraterísticas da base de dados que obtiver maior pontuação servirá então para identificar a pessoa.

Na Figura 4 está ilustrado o processo de identificação num sistema de reconhecimento facial.

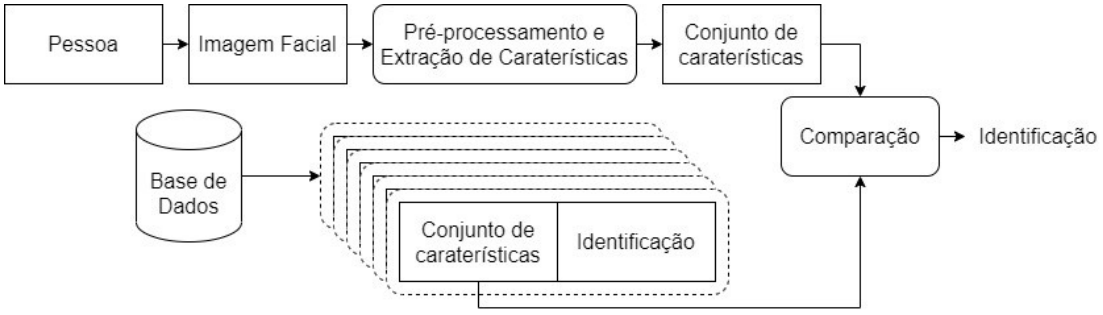

Figura 4 - Identificação num sistema de reconhecimento facial.

2.3. Métricas Utilizadas

De forma a aferir o desempenho de um sistema com outros, surge a necessidade de utilizar medidas padronizadas.

A métrica mais utilizada para a tarefa da verificação consiste na taxa de aceitação verdadeira (TAV), dado uma taxa de aceitação falsa (TAF) fixa. A TAV é definida como a percentagem de vezes que um sistema verifica corretamente uma reivindicação de identidade verdadeira.

$$TAV\ (\%) = \frac{VP}{VP + FN} \times 100 \tag{1}$$

onde VP é o número de verdadeiros positivos e FN é o número de falsos negativos.

A TAF, sendo o oposto da TAV, é a percentagem de vezes que um sistema verifica incorretamente uma reivindicação de identidade falsa. No caso de um controlo de acesso, esta verificação incorreta dá acesso a um utilizador não autorizado.

$$TAF\ (\%) = \frac{FP}{FP + VN} \times 100 \tag{2}$$

onde FP é o número de falsos positivos e VN é o número de verdadeiros negativos.

Na métrica TAV@TAF=*x*, o valor de *x* costuma ser de 1% e 0,1%. Para exemplificar, tomando *x*=1%, ter uma TAF@TAF=1% implica que o sistema de reconhecimento facial utilize um limiar (*threshold*, em inglês) na verificação de forma que, em 100 casos, o sistema concede acesso apenas a 1 indivíduo não autorizado. Quanto maior for o nível de segurança, menor deve ser o valor da TAF, sendo 0 o valor desejável (nunca se confere acesso não autorizado). No entanto, quanto menor o valor de TAF, maior deverá ser o limiar, o que leva a valores de TAV mais baixos. A visualização gráfica da relação da TAF com a TAV é uma curva Característica de Operação do Recetor (ROC, do inglês *Receiver operating characteristic*), onde pode-se auferir a Área Sob a Curva ROC (AUC, do inglês *Area Under the Curve*).

A métrica mais utilizada para a identificação é a precisão com *Rank-n,* com *n* a tomar os valores de 1 e 5. O *Rank-n* consiste na percentagem de vezes em que, dado uma imagem de uma face como entrada, o classificador obtém as *n* identidades mais prováveis e uma delas é a identidade correta. Assim, o cálculo do *Rank-1* consiste na percentagem de vezes em que o classificador identifica corretamente a pessoa presente na imagem facial.

$$Rank\text{-}1\ (\%) = \frac{IC}{IE} \quad (3)$$

onde IC é o número de identificações corretas e IE é o número de identificações efetuadas.
Já o *Rank-5* consiste nas vezes em que a identidade da pessoa presente na imagem facial está entre as 5 identidades a que o classificador atribui uma maior pontuação.

2.4. Ambiente Controlado e Não Controlado

O ambiente em que o reconhecimento facial é feito, pode ser num ambiente controlado ou não controlado. O ambiente controlado, também conhecido como reconhecimento com consentimento, é aquele em que o utilizador coopera no reconhecimento, facilitando-o através de uma postura correta e estática, bem como com boa iluminação.

No ambiente não controlado, o reconhecimento é dinâmico, sem que o utilizador coopere para a aquisição de uma imagem, o que dificulta bastante o processo de reconhecimento facial devido à diversidade do ambiente envolvente (p.e. baixa visibilidade) e de poses e expressões faciais [1].

2.5. Espectro do Visível e do Infravermelho

Para este trabalho, as bandas do espetro eletromagnético relevantes, são a do visível e do Infravermelho (IV). A banda do espetro IV é geralmente subdividida em 5 bandas: *Near-infrared* (NIR), *Short-wavelength infrared* (SWIR), *Mid-wavelength infrared* (MWIR), *Long-wavelength infrared* (LWIR) e *Far infrared*. Os intervalos espectrais de interesse estão expostos na tabela 1.

Tabela 1 - Comprimentos de onda dos intervalos espectrais, adaptado de [3].

Nome do Intervalo Espectral	Comprimento de Onda
Visível	380 – 750 nm
NIR	750 – 1400 nm
SWIR	1.4 – 3 μm
MWIR	3 – 8 μm
LWIR	8 – 15 μm

As bandas do IV também podem ser categorizadas quanto à forma de obtenção de imagens, as quais podem ser obtidas passivamente (sem necessidade de luminosidade) ou ativamente (necessidade de luminosidade). Assim tem-se:

- Bandas ativas – NIR e SWIR. Para obter imagens nestas bandas, é necessário o objeto receber iluminação, mesmo que escassa, pois é através da reflexão que a imagem é obtida. Tal facto faz com que sejam utilizadas em aparelhos de visão noturna. A banda NIR permite enfrentar as variações de iluminação, enquanto que o SWIR tem a vantagem de obter imagens através de fumo e nevoeiro [3].
- Bandas passivas – MWIR e LWIR. Ao contrário das ativas, estas permitem a obtenção de imagens apenas utilizando a radiação térmica emitida por um corpo, as quais são comumente conhecidas por imagens térmicas [3].

2.6. Vantagens das Imagens Multiespectrais em Ambiente Não Controlado

Existem fatores que influenciam o desempenho do reconhecimento facial no ambiente não controlado, como é o caso de alterações das expressões faciais, oclusões, variações de pose, e iluminação [1].

Relativamente à variedade de pose e de expressões faciais, as bases de dados do domínio do visível (cada vez mais completas e a utilização de sintetizadores de imagem) têm permitido contornar estas dificuldades. No entanto, existem dois pontos que se têm revelado mais difíceis de ultrapassar: variedade de iluminação e as oclusões.

Figura 5 - Diferenças provocadas pela variância da iluminação, retirado de [4].

Na Figura 5 exemplifica-se o problema da variedade da iluminação isolada das restantes caraterísticas do ambiente não controlado. Nesta está representada a mesma pessoa com a mesma pose e expressão facial, havendo apenas variação de iluminação. Mesmo para um observador humano, seria muito difícil identificar com certezas as imagens da 4ª à 6ª coluna, e poderia mesmo assumir que nas restantes imagens, existem identidades diferentes.

O principal problema neste caso é o facto de o ser humano estar limitado à observação do visível, assim como a câmara que tirou estas fotografias. A variedade de iluminação influencia uma perceção de alteração de expressões faciais e de uma vasta diversidade de oclusões, dificultando o reconhecimento facial.

Assim, consegue-se observar que a limitação ao domínio do visível pode acarretar consequências prejudiciais à correta identificação de uma face humana. Esta limitação é algo que o olho humano não consegue ultrapassar, pois a biologia circunscreveu-nos ao domínio do visível, mas o mesmo não se aplica a alguns animais (exemplo: abelhas) nem às máquinas, as quais podem captar intervalos espectrais do infravermelho. Tal possibilita captar caraterísticas onde o espetro do visível não consegue, como é exemplo a imagem anterior. Uma das direções que se tem tomado é a exploração da radiação IV, a qual permite ultrapassar desafios relativos à variedade de iluminação (Figura 6), bem como de algumas oclusões (Figura 7 e Figura 8).

Figura 6- Imagem multiespectral à noite (Visível à esquerda e IV à direita), retirado de [5].

Figura 7 - Imagem multiespectral com fumo (Visível à esquerda e IV à direita), retirado de [6].

Figura 8 - Imagem multiespectral com face coberta por um saco de plástico (Visível à esquerda e IV à direita).

A utilização de imagens de IV para reconhecimento facial automático não está isenta de desafios, sendo sensível às condições emocionais, físicas e de saúde do indivíduo, bem como do meio que o rodeia, não servindo como alternativa absoluta ao uso do espetro do visível, mas sim como complemento [7].

Outra dificuldade que surge é o baixo número de bases de dados públicas com imagens de ambos os intervalos espectrais e em ambiente não controlado [8]. Nestas bases de dados:

- existe um número reduzido de sujeitos;
- maioritariamente, cada indivíduo só participa numa sessão, não havendo os normais desafios que ocorreriam de diferentes sessões (envelhecimento, alteração do meio ambiente, emoções);
- as imagens de diferentes espetros por norma não são tiradas simultaneamente, havendo desalinhamento e não uma correspondência direta de imagens visível com o infravermelho.

3. Trabalhos Relacionados

O problema apresentado prende-se com dois aspetos. O primeiro, é o reconhecimento facial em ambiente não controlado, o qual por si só, já é desafiante. O segundo, é o reconhecimento facial multiespectral, ou seja, utilizar diferentes bandas espectrais no reconhecimento facial.

Nesta secção, faz-se uma breve revisão dos progressos realizados nestas duas áreas.

3.1. Reconhecimento Facial em Ambiente Não Controlado

Os sistemas de reconhecimento facial têm por norma dois passos principais: detetar uma face e fazer reconhecimento. Este problema já é alvo de estudo com a utilização de computadores desde o ano de 1966 [9] [10], sendo interessante salientar que os principais desafios enumerados por Bledsoe há 54 anos continuam como áreas ativas de investigação, devido aos fatores pose-iluminação-expressão (PIE), que representam "distrações" e que confundem ainda nos dias de hoje os sistemas de reconhecimento mais avançados [1]. Tais fatores estão presentes no ambiente não controlado.

Os sistemas de reconhecimento facial atuais utilizam por norma o espetro do visível, devido à facilidade da captação de imagens do visível, o que leva a que existam grandes bases de dados com faces humanas obtidas nesta banda do espetro eletromagnético. Estas bases de dados massivas são benéficas para as redes de aprendizagem profunda [11], onde a utilização de mais dados na fase de treino, leva a uma melhoria na fase de teste. No entanto, o espetro do visível tem as suas desvantagens, pois a utilização de apenas esta faixa do espetro eletromagnético é sujeita a oclusões provocadas pelas variações da iluminação, a gases e materiais sólidos que são "transparentes" a outras faixas do espetro (p.e. o fumo e o plástico no IV).

O ambiente não controlado, caraterizado fortemente pelos fatores PIE, surge como um problema para os atuais sistemas de reconhecimento.

3.2. Invariância à Pose

Nos últimos anos, os métodos robustos à variância da pose, centraram-se primeiramente em lidar com esse problema em bases de dados semi-controladas, isto é, num ambiente com as restantes variáveis fixas, como é o caso da iluminação e não existência de oclusões, com a utilização da base de dados Multi-PIE [12].

De acordo com [1], um grande passo foi dado rumo à solução deste tipo de problema, através da introdução de bases de dados gigantes para treinar as redes neuronais convolucionais profundas (DCNN, do inglês *Deep Convolutional Neural Networks*), alcançando uma performance tão boa quanto humanos em verificação na base de dados LFW [13]. No entanto, quando as imagens faciais apresentam variações mais extremas, como na base de dados IJB-A [2], apesar dos grandes avanços, apresentados na Figura 9, continua a haver problemas por resolver.

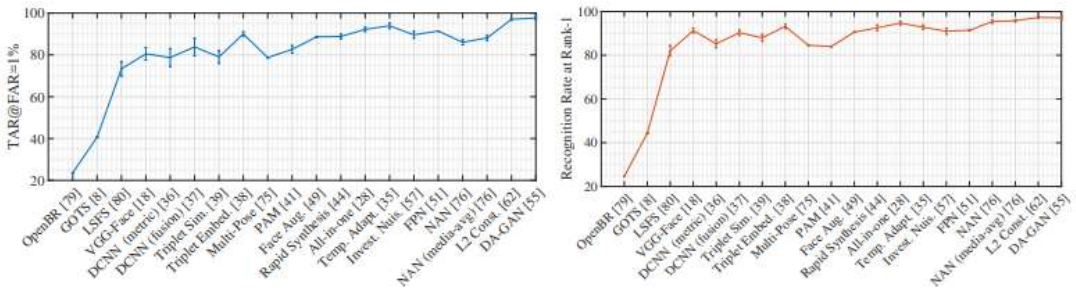

Figura 9 - Progresso no desafio IJB-A entre 2015 e 2018, retirado de [1].

Esta melhoria dos resultados foi alcançada com a aplicação de dois métodos de síntese de imagem: aumentação de um para vários, que consiste em gerar diferentes poses de uma face a partir de uma canónica, e normalização de vários para um, que consiste em normalizar qualquer pose da face para uma posição canónica [1].

A aumentação de um para vários consiste em gerar diferentes imagens com diferentes poses de uma face a partir de uma imagem, permitindo que o classificador aprenda diferentes representações da mesma identidade. Já a normalização de vários para um, que consiste em normalizar qualquer pose da face para uma posição canónica, tira complexidade ao classificador, que recebe imagens sempre na mesma pose, permitindo que o classificador faça apenas uma classificação de um para um, como se a imagem tivesse sido captada num ambiente controlado.

Artigos mais recentes utilizam redes adversárias generativas (GAN, do inglês *Generative Adversarial Networks*), introduzidas por Goodfellow *et al*. [14] e que possuem uma grande capacidade de criação de imagens como mostrado por [15], o que permite a sua utilização nos métodos acima indicados. Estas redes são caraterizadas pela utilização de um gerador e de um discriminador. O gerador é responsável por produzir amostras dado uma entrada, de forma que o discriminador não consiga discernir qual das amostras é real e qual é falsa (ver Figura 10).

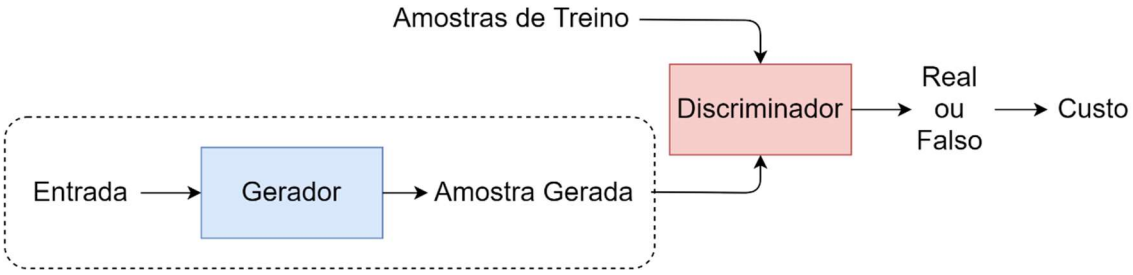

Figura 10 – Esquema do treino de uma GAN, a tracejado o processo de geração de amostras.

Desde o seu aparecimento na normalização de vários para um com a *Disentangled Representation-learning GAN* [16] (DR-GAN), que estas redes tomaram a dianteira na solução deste problema. Quanto à aumentação de um para vários, o seu poder de produção de imagens também lhes dá vantagem quando comparadas com outras redes, como é o caso da *Dual-Agent GAN* [17] (DA-GAN).

3.2.1. Normalização de Vários para Um

A normalização de imagens de vários para um é um problema extremo de síntese de imagem devido à natureza das diferenças de pose de uma face. Métodos recentes para a resolução deste problema utilizaram deformação de texturas locais 3D/2D [18] [19] [20] [21], aprendizagem profunda [16] [22] [23] [24] [25] e uma conjugação de ambos [26].

Cole *et al.* [20] introduz um método para sintetizar uma imagem frontal, de expressão neutra, aprendendo a gerar marcos faciais e texturas a partir de características extraídas de uma rede de reconhecimento facial. A codificação de caraterísticas é em grande parte invariável à iluminação, pose e expressão facial. Esta invariância permite treinar uma rede de descodificação utilizando apenas fotografias de expressão frontal e neutra. O descodificador decompõe então as faces em conjuntos esparsos de marcos faciais e mapas de textura, combinando-os através de uma operação de distorção de imagem. As imagens resultantes podem ser utilizadas para uma série de aplicações, tais como análise de atributos faciais, ou criação de um avatar 3D.

Em Huang *et al.* [22] é proposto uma estrutura GAN de duas vias, percebendo simultaneamente estruturas globais e detalhes locais, com supervisão percetiva, para síntese de imagens faciais com pose frontal a partir de uma única imagem facial de entrada. Os autores introduzem a função de custo adversarial, para guiar a síntese de imagem na criação de faces com pose frontal, a função de custo de simetria para explorar explicitamente a simetria, de forma a aliviar o efeito de auto-oclusão devido à variação da pose, e a função de custo de preservação de identidade, com o intuito de preservar as caraterísticas faciais à entrada e à saída.

Zhao *et al.* [23] argumentam que, ao invés de extrair diretamente caraterísticas invariantes para o reconhecimento, ou normalizar primeiro as imagens faciais para uma pose frontal antes da extração de caraterísticas, é mais desejável executar ambas as tarefas em conjunto. Desta forma, introduzem um modelo que unifica uma sub-rede responsável por passar a face uma pose frontal e uma sub-rede responsável por fazer o reconhecimento facial invariante à pose.

Qian *et al.* [24] propõem um modelo de normalização de face (FNM, do inglês *Face Normalization Model*) não supervisionado que codifica as imagens utilizando uma rede pré-treinada para extração de caraterísticas, sendo depois usadas para gerar imagens realistas através de uma GAN.

Um modelo de alta fidelidade invariante à pose (HF-PIM, do inglês High Fidelity Pose Invariant Model) é proposto por Cao *et al.* [26] (Figura 11). Estes autores normalizam a face para uma posição frontal através de um procedimento de deformação de fusão de texturas aproveitando um campo de correspondência denso proposto por Deng *et al.* [27] para interligar os espaços de superfície 2D e 3D. Além disso, propõem uma função de custo guiada por multi-percepção para resolver o desalinhamento entre as faces com pose frontal e de perfil, permitindo que o modelo utilize eficazmente múltiplas imagens durante o treino.

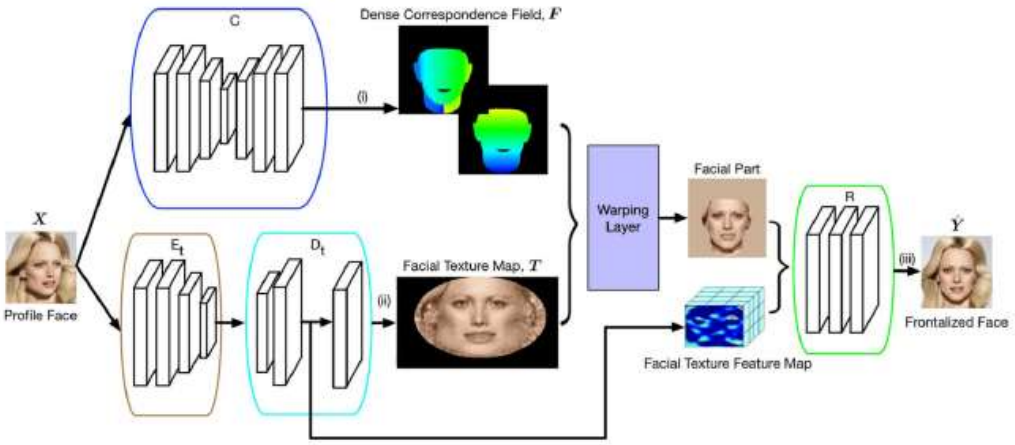

Figura 11 - Arquitetura de HF-PIM, retirado de [26].

Zhang *et al.* [25], com CCFF-GAN (do inglês *Cycle-Consistency Face Frontalization GAN*) propõem a utilização de uma rede semi-supervisionada. A CCFF-GAN, na fase de treino, utiliza inicialmente pares de imagens de faces da mesma pessoa que possuem apenas a variação de pose (presentes em grande número em bases de dados, como MULTI-PIE [12]) e posteriormente utiliza pares em ambiente não controlado (que existem em menor número), de forma a generalizar a rede para os diferentes fatores do ambiente não controlado.

3.2.2. Aumentação de Um para Vários

A aumentação de um para vários é outra abordagem para alcançar o reconhecimento facial independentemente da pose. Tran *et al.* [21] sintetizam diferentes poses através da modelação 3D e, em seguida, treinam uma rede neuronal convolucional (CNN, do inglês *Convolutional Neural Network*) para conseguirem fazer um reconhecimento facial com poses variadas. A síntese de diferentes poses tem como objetivo superar a escassez de dados de treino necessários para o treino da CNN.

A DA-GAN proposta por Zhao *et al.* [17], cria imagens 2D através de modelação 3D e depois refina as imagens 2D obtidas para serem o mais realistas possíveis, utilizando uma GAN, de forma a tentar preservar a identidade da face. Tal melhora a qualidade da imagem, o que beneficia o treino. Deste modo, a rede DA-GAN é utilizada também para aumentar os dados de treino.

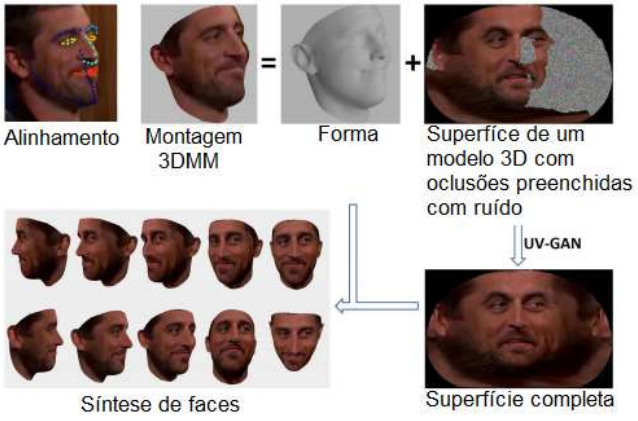

Figura 12 - Fluxograma da UV-GAN, adaptado de [27].

Deng *et al.* [27] propõem uma rede GAN para completar um modelo 3D facial (UV-GAN, do inglês *GAN for UV completion*), cujo fluxograma está na Figura 12. Neste trabalho, os autores treinam uma GAN para completar a superfície de um modelo 3D de imagens faciais extraídas de um ambiente não controlado. É demonstrado que ao anexar a forma da face com a superfície 3D completa, é possível gerar poses arbitrárias, permitindo assim aumentar as variações de pose para treinar modelos de reconhecimento facial.

3.3. Reconhecimento Facial Multiespectral

As imagens multiespectrais são imagens com diferentes bandas do espetro eletromagnético. As principais bandas utilizadas são as bandas do visível e IV. O IV tem sido usado principalmente como complemento ao espectro do visível, e que tem resultado num crescente interesse da comunidade científica no espetro do IV [1].

A utilização de diferentes bandas origina diferentes soluções de implementação do reconhecimento facial: múltiplas bandas *versus* múltiplas bandas (a imagem obtida é multiespectral e também existem imagens multiespectrais na base de dados), múltiplas bandas *versus* banda única (também chamada de reconhecimento facial heterogéneo, consiste em obter-se uma imagem de uma única banda, tendo disponíveis diferentes bandas na base de dados) e por fim banda única *versus* banda única (caso do reconhecimento na banda do visível, em que a banda da imagem captada é a mesma banda que existe na base de dados).

Os principais métodos de reconhecimento facial multiespectral podem ser divididos em três categorias: Síntese de Imagem, Fusão e Funções de Custo.

3.3.1. Síntese de Imagem

Os métodos de síntese de imagem, caraterizados por abordar o reconhecimento facial heterogéneo, permitem transformar uma imagem de uma banda espectral noutra, facilitando a comparação entre duas imagens. A principal vantagem de síntese de imagem é que permite passar uma imagem de qualquer banda espectral para a banda espectral do visível, consequentemente torna possível utilizar classificadores implementados para processar imagens do espetro do visível [28].

Zhang *et al.* [29] propõe um sintetizador de imagens baseado em GANs que consiste num gerador de fluxo múltiplo e num discriminador de várias escalas para sintetizar imagens do visível a partir de imagens LWIR, enquanto que Litvin *et al.* [30] utiliza uma CNN para o mesmo efeito.

Guei *et al.* [31] aplica uma GAN de convolução profunda (DCGAN, do inglês *Deep Convolutional GAN*), que aumenta o tamanho das imagens enquanto preserva detalhes faciais importantes. Cao *et al.* [32] propõe a resolução do reconhecimento facial heterogéneo através de aumentação de dados, proporcionando informação mais discriminatória da identidade.

Bae *et al*. [33] introduz um módulo de síntese de imagem com a utilização de uma cadeia de pré-processamento, uma GAN cíclica e uma rede siamesa. No mesmo trabalho é utilizado também um módulo de aprendizagem de caraterísticas, onde as imagens reais e as imagens sintetizadas são utilizadas para fazer pequenos ajustamentos a uma rede de classificação já pré treinada no domínio do visível.

He *et al.* [34] propõe um complemento facial multiespectral (CFC, do inglês *Adversarial Cross-spectral Face Completion)*, sintetizando imagens do visível a partir de imagens NIR através de GANs. Este utiliza uma componente de pintura para fazer um mapeamento das texturas do NIR para o visível, e outra componente de normalização de pose, através de um modelo apresentado por [27] para resolver eventuais diferenças de pose. Um processo de deformação é utilizado para juntar as duas componentes, resultando numa imagem sintetizada no visível com pose frontal.

3.3.2. Fusão

Os dois métodos de fusão mais relevantes são a fusão de caraterísticas e fusão de pontuação. No primeiro, é feita uma fusão de caraterísticas das diferentes bandas da imagem, permitindo extrair as caraterísticas mais relevantes das diferentes bandas, e juntá-las num vetor. O segundo método combina a pontuação obtida de cada classificador banda única *versus* banda única (p. e. um classificador que opera apenas na banda LWIR e outro que opera apenas na banda NIR).

Seal *et al.* [35] propõe fazer fusão de pontuação com imagens LWIR e Visível, com o objetivo de beneficiar de ambas as modalidades de imagens. O processo de fusão proposto é a soma ponderada da informação térmica e visível da face. Os fatores de ponderação atribuídas a cada banda são calculados consoante a taxa de acerto obtida no reconhecimento facial com apenas a utilização dessa banda. De seguida, realiza-se a fusão das pontuações obtidas em cada uma das bandas, da qual se obtêm as novas pontuações que são utilizadas para o processo de reconhecimento facial

Kanmani *et al.* [36] propõe a aplicação da fusão de caraterísticas de imagens faciais das bandas espectrais Visível e LWIR para melhorar a precisão do reconhecimento facial. São adotados três esquemas diferentes de fusão em que a informação facial é fundida pelos pesos ótimos obtidos por diferentes algoritmos de otimização. As duas primeiras abordagens de fusão operam no domínio da transformação de onda discreta de árvore dupla, enquanto a terceira opera no domínio da transformação de *Curvelet,* para efetuar uma decomposição da imagem, preservando as arestas ao longo das curvas.

3.3.3. Funções de Custo

Todos as redes neuronais possuem funções de custo para o momento do treino, de forma a atualizar os pesos da rede, no entanto, certas funções de custo têm-se revelado mais eficientes que outras para a classificação de imagens multiespectrais. Assim, este método tem como objetivo melhorar a classificação sem a necessidade de um pré-processamento, utilizando por norma redes neuronais profundas como classificadores e dando uma função de custo específica. São indicadas de seguida duas funções de custo que se revelam eficientes no reconhecimento facial multiespectral.

Hu *et al.* [37] propõem múltiplas redes profundas com função de custo de dispersão e combinação de diversidade (MDNDC, do inglês *Multiple Deep Networks with scatter loss and Diversity Combination*) para resolver o problema de reconhecimento facial heterogéneo. De forma a reduzir as variações dentro da mesma classe e aumentar as variações entre diferentes classes, utiliza a função de custo de dispersão, a qual pode colmatar a lacuna que advém da diferença de banda espectral ao mesmo tempo que preserva a informação da identidade. As múltiplas redes profundas são

responsáveis pela extração de caraterísticas. A utilização de uma estratégia de decisão conjunta, chamada de combinação de diversidade, permite ajustar, de forma adaptável, os pesos de cada rede profunda e tomar uma decisão de classificação conjunta.

He *et al.* [38] apresentam a distância de *Wasserstein* numa *Convolutional Neural Network* (WCNN). Esta distância permite reduzir o intervalo entre as bandas do NIR e do Visível, ao fazer a distância entre duas distribuições de probabilidade num dado espaço. Assim, a aprendizagem desta CNN visa alcançar a minimização da distância de *Wasserstein* entre a distribuição do NIR e a distribuição do Visível para uma representação invariante de características em imagens faciais heterogéneas. Na Figura 13 está exposta a arquitetura da WCNN.

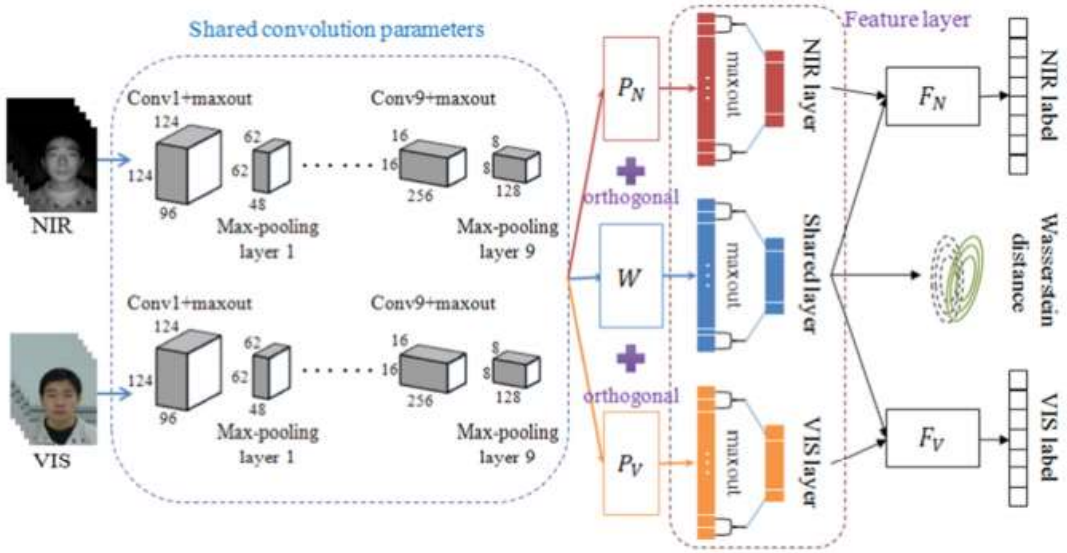

Figura 13 - Arquitetura da WCNN, retirado de [38].

3.4. Bases de dados

Nesta secção analisa-se as principais bases de dados utilizadas nos métodos de reconhecimento facial com variedades de pose no visível e nos métodos de reconhecimento facial multiespectral.

3.4.1. Reconhecimento Facial com Variedade de Poses no Visível

A base de dados Multi-PIE [12] é a maior base de dados para avaliar a síntese e reconhecimento facial em ambiente controlado. Ela permite uma avaliação graduada no que respeita à postura, iluminação, e variações de expressão. Esta base de dados contém 337 pessoas, com poses faciais entre os -90º e os +90º com um intervalo de 15º e 20 níveis de iluminação, possuindo 755.370 imagens, com anotações explícitas da pose, o que faz com que seja amplamente usada para construir modelos baseadas em redes neuronais profundas, tal como GANs. A base de dados Multi-PIE [12] é utilizada para tarefas de identificação (procura 1:N).

Tabela 2 - Resultados na tarefa de identificação (Rank-1 em %) na base de dados Multi-PIE [12].

Método	±90º	±75º	±60º	±45º	±30º	±15º
DR-GAN [16]	-	-	83,2	86,2	90,1	94,0
TP-GAN [22]	64,6	77,4	87,7	95,4	98,1	98,7
PIM [23]	86,5	95,0	98,1	98,5	99,0	99,3
FNM [24]	55,8	81,3	93,7	98,2	99,5	**100,0**
HF-PIM [26]	**92,3**	**96,4**	**99,1**	**99,9**	**100,0**	**100,0**
CCFF-GAN [25]	73,9	88,3	94,9	99,2	99,8	**100,0**

A base de dados LFW [13] é uma base de dados de referência para o reconhecimento facial. Contém 13.233 imagens de 5.749 pessoas e tem sido amplamente utilizada para avaliar o desempenho de síntese ou verificação de vários métodos em ambientes sem restrições. Uma vez que as imagens de faces da LFW são recolhidas da internet, contêm várias variações de pose, expressão e iluminação. A base de dados LFW é utilizada para tarefas de verificação (comparação 1:1).

Tabela 3 - Resultados na tarefa de verificação na base de dados LFW [13].

Método	Precisão (%)	AUC (%)
TP-GAN [22]	96,1	99,4
HF-PIM [26]	**99,4**	**99,9**
CCFF-GAN [25]	99,2	99,8

A base de dados IJB-A [2] contém não só imagens estáticas como também *frames* de vídeo de 500 indivíduos, com 5.397 imagens e 2.042 vídeos que são divididos em 20.412 *frames*, tendo cerca de 11 imagens e 4 vídeos por pessoa. A captura de imagens a partir de um ambiente o*n the wild* (não controlado), ajuda a evitar o que as poses frontais sejam mais frequentes que as restantes, o que acontece com a LFW [13]. Esta possui protocolos para avaliação tanto de tarefas de verificação (comparação 1:1) como de identificação (procura 1:N).

Tabela 4 - Resultados obtidos nas tarefas de identificação e Verificação (em %) na base de dados IJB-A [2].

	Identificação		Verificação	
Método	***Rank*-1**	***Rank*-5**	**TAV @TAF=0,01**	**TAV @TAF=0,001**
Tran *et al.* [21]	76,2	89,7	87,0	60,0
DR-GAN [16]	85,5	94,7	77,4	53,9
TP-GAN [22]	48,6	59,3	31,5	9,2
DA-GAN [17]	**99,0**	**99,5**	**98,9**	**97,3**
FNM [24]	96,0	98,6	93,4	83,8
HF-PIM [26]	95,3	89,9	95,3	89,9
CCFF-GAN [25]	98,1	98,9	84,1	72,3

Nas Tabelas 2, 3 e 4 estão representados os resultados obtidos nestas três bases de dados pelos métodos estudados.

Relativamente às bases de dados utilizadas para treino, devido ao grande número de imagens em ambiente não controlado, normalmente são a MS-Celeb-1M [39], a CASIA WebFace [40], MegaFace [41] e VGGFace2 [42].

3.4.2. Reconhecimento Facial Multiespectral

Tabela 5 – Resumo das bases de dados mais estudadas em cada banda, adaptado de [43].

Nome	Ano	Banda espectral	Nº de pessoas	Imagens / pessoa	Nº de imagens	Melhor *Rank*-1 (%)	Caraterísticas
CASIA NIR-VIS 2.0 [44]	2013	Visível NIR	725	24	17580	99,4 [33]	Não tem relação 1 para 1 (visível - NIR). Diferentes variações de luminosidade, expressão, pose e distância.
Oulu CASIA NIR-VIS [45]	2009	Visível NIR	80	36	2880	99,9 [34]	Com variações de luminosidade (3) e expressões faciais (6).
USTC-NVIE [46]	2010	Visível LWIR	215	162	34830	97,4 [47]	Com variações de luminosidade (3), pose (9) e expressões faciais (3).
TINDERS [48]	2009	Visível NIR SWIR	48	26	1255	97,8 [49]	Com variação de expressões faciais (3).
NIST-Equinox [50]	2007	Visível SWIR MWIR LWIR	95	-	-	88,6 [51]	Com variações de luminosidade (3) e expressões faciais (3).
TUFTS [8]	2020	Visível NIR LWIR	113	86	9100	94,5 [52]	Com variação de pose (9), e expressões faciais (5).

As bases de dados utilizadas no reconhecimento facial multiespectral são apresentadas na **Erro! A origem da referência não foi encontrada.**, adaptada de [43]. Estas bases de dados multiespectrais são aquelas que foram mais usadas pela comunidade científica, possuindo caraterísticas do ambiente não controlado. De acordo com [43], as bases de dados alvo de maior atenção são aquelas que possuem a banda NIR e a LWIR, existindo poucas bases de dados com as bandas intermédias do SWIR e do MWIR.

3.5. Lacunas Encontradas

Apesar de existir numerosos artigos que abordam o reconhecimento facial multiespectral, poucos destes demonstram o seu poder no ambiente não controlado. Tal deve-se principalmente ao facto de que nas bases de dados atuais de imagens de faces multiespectrais, as variações das condições não são extremas, sendo por norma ambientes semi-controlados e não *on the wild*. A título de exemplo, a base de dados mais estudada na área do reconhecimento facial CASIA NIR-VIS 2.0 [43],

utiliza imagens em que a pose tem poucos desvios relativamente à posição frontal, o que não caracteriza fidedignamente o ambiente não controlado. Assim, o facto de as bases de dados multiespectrais serem pequenas (quando comparadas com as do visível) é ainda um entrave à melhoria da capacidade dos sistemas de reconhecimento facial multiespectral em ambiente não controlado.

Posto isto, o que este trabalho propõe é um sistema que integra as capacidades dos sistemas de reconhecimento facial em ambiente não controlado do visível ao nível da variação da pose e as capacidades dos sistemas de reconhecimento facial multiespectral para combater alguns tipos de oclusões (p.e. fumo e materiais plásticos) e a variação de iluminação.

4. Metodologia

O presente capítulo descreve a arquitetura do sistema de reconhecimento facial multiespectral, constituído por 3 etapas:

- Deteção e Alinhamento da Face
- Síntese de Imagem
- Reconhecimento Facial

Na Figura 14 é apresentado de uma forma simples o funcionamento geral do sistema de reconhecimento facial proposto, com a indicação dos passos realizados em cada uma das etapas.

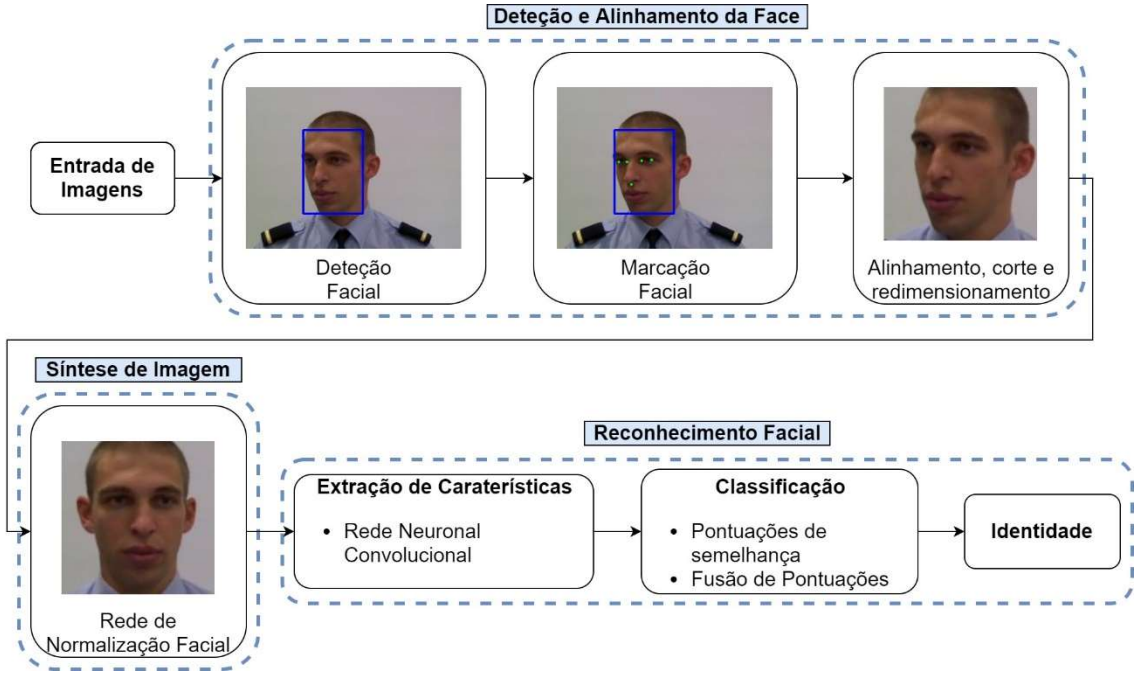

Figura 14 - Fluxograma do sistema de reconhecimento facial.

O primeiro passo consiste na aquisição de imagens multiespectrais, as quais podem ser obtidas através de uma combinação de equipamentos mono-espectrais (cada um recolhe imagens de apenas uma banda espectral) ou multiespectrais (recolhem imagens em diferentes bandas espectrais em apenas uma aquisição). Os equipamentos multiespectrais têm a vantagem de adquirir as imagens das diferentes bandas ao mesmo tempo e do mesmo ponto de observação. Desta forma, as imagens espectrais têm as mesmas condições de luminosidade e de pose do indivíduo que está a ser alvo de captação de imagem.

Após a obtenção de imagem, o módulo de Deteção e Alinhamento da Face (Secção 4.1) gera uma imagem facial em que os olhos são alinhados horizontalmente, centrada e com dimensões pré-definidas. Para a execução deste objetivo, é necessário detetar a presença das faces humanas e posteriormente detetar marcos importantes da face, como é o caso de olhos e nariz, permitindo um correto alinhamento da face e o recorte em torno desta.

O sistema de reconhecimento facial proposto apresenta uma etapa denominada de Síntese de Imagem (Secção 4.2), que visa a obter uma imagem facial frontal e com expressão neutra.

Na etapa seguinte surge o módulo de Reconhecimento Facial (Secção 4.3), onde é realizada a extração de caraterísticas da imagem facial a identificar, através de uma CNN. Tal permitirá identificar a pessoa presente na imagem facial, através da comparação com as imagens da base de dados de utilizadores, utilizando funções de semelhança e fusão de pontuação.

4.1. Deteção e Alinhamento da Face

A primeira fase do sistema de reconhecimento facial tem como objetivo detetar as faces presentes na imagem de entrada e identificar marcos faciais, de forma que, através de processamento de imagem, se obtenham imagens faciais centradas, alinhadas horizontalmente e com as mesmas dimensões. O funcionamento esperado deste módulo encontra-se apresentado na Figura 15. Esta fase revela-se de extrema importância para qualquer sistema de reconhecimento facial, pois serve para retirar grande parte da variabilidade da entrada para as restantes fases de processamento.

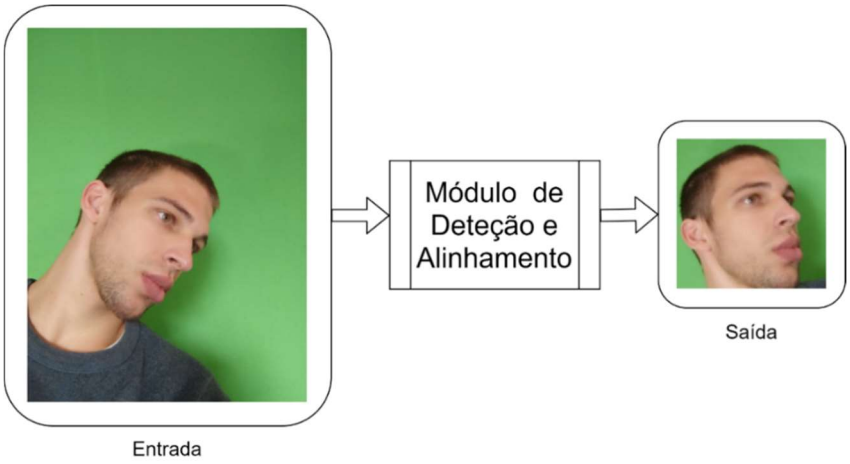

Figura 15 - Entrada e saída do módulo de Deteção e Alinhamento da Face.

O alinhamento horizontal da face pode ser visto como uma forma de normalização de dados. Tal como se pode normalizar um conjunto de vetores de caraterísticas centrando em zero ou colocar numa escala unitária antes do treino de um modelo de aprendizagem automática, também é comum alinhar os rostos antes de treinar um sistema de reconhecimento facial. O objetivo desta normalização espacial das imagens é remover o desalinhamento horizontal das imagens, inerente ao processo de aquisição de imagens em condições não controladas, o qual prejudica o reconhecimento facial.

Visto que os algoritmos de deteção facial detetam faces em retângulos sem rotação, é necessário detetar certos marcos faciais para poder aplicar uma rotação 2D à imagem de forma que a face fique alinhada no eixo horizontal, sendo por norma utilizada uma linha imaginária que passa pelos dois olhos.

Assim, como entrada, temos uma imagem, na qual devem ser identificadas faces, extrair marcos faciais e por fim obter imagens em que a face esteja numa posição central. Este alinhamento, corte e redimensionamento permite a normalização espacial no tratamento das imagens.

O procedimento desta fase encontra-se apresentado no fluxograma da Figura 16, onde estão identificadas as principais tarefas:

- Deteção Facial
- Deteção dos marcos faciais
- Alinhamento, corte e redimensionamento

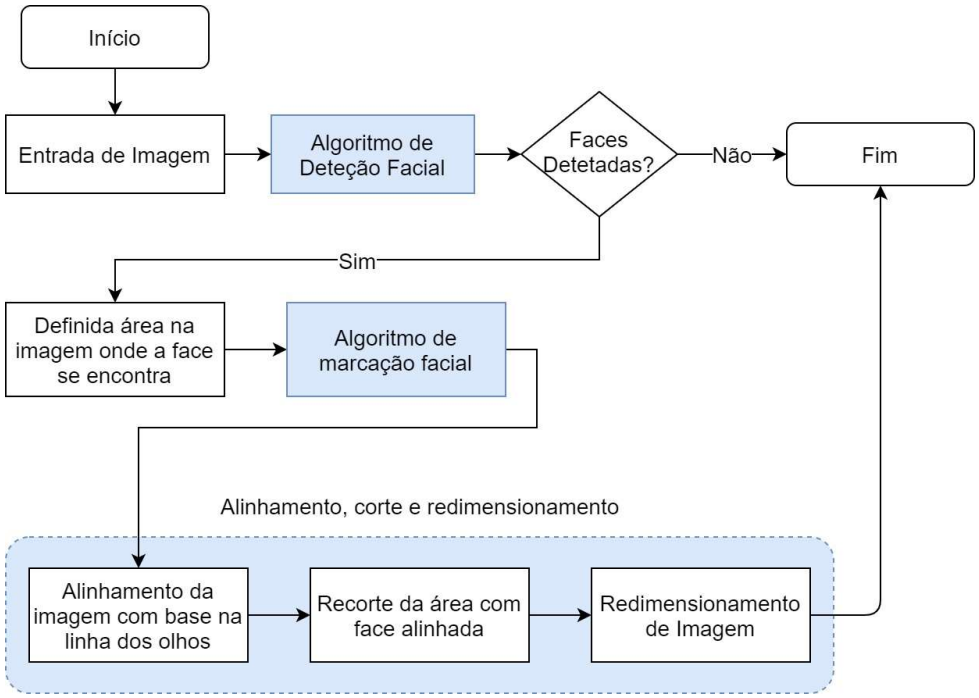

Figura 16 - Fluxograma do Módulo de Deteção e Alinhamento da Face com as tarefas identificadas a azul.

Com o objetivo de eleger o método mais eficiente na deteção facial e extração de marcos faciais, são avaliados diferentes algoritmos para cada uma das tarefas.

4.1.1. Deteção Facial

A deteção facial é um primeiro passo essencial para muitos sistemas de reconhecimento e análise facial [53] [54], onde, apesar de grandes avanços terem sido realizados, os fatores inerentes ao ambiente não controlado fazem com que uma deteção facial precisa e eficiente continue a ser um desafio em aberto [55].

Para esta tarefa são analisados três modelos pré-treinados baseados no detetor de múltiplas regiões de passagem simples (SSD, do inglês *single shot multibox detector*), uma arquitetura de aprendizagem profunda proposta por *Liu et al.* [56] para deteção de objetos. A ideia base do SSD consiste em gerar pontuações para a presença de cada categoria de objetos em cada caixa predefinida e produzir ajustes na caixa para corresponder à forma do objeto. Além disso, o SSD combina predições de múltiplos mapas de características com diferentes resoluções para lidar com objetos de diferentes tamanhos. O SSD é simples em relação a outros métodos, pois encapsula todos os cálculos numa única rede, com um bom balanço entre eficiência e a precisão [55].

Neste trabalho foram testados o detetor de faces invariante à escala de passagem simples[1] (S3FD, do inglês *Single Shot Scale-invariant Face Detector*) de *Zhang et al.* [57], a rede neuronal profunda de deteção facial da OpenCV[2] [58] e o detetor de face de passagem dupla[3] (DSFD, do inglês *Dual Shot Face Detector*) de *Li et al.* [59].

A rede neuronal profunda de deteção facial da OpenCV [58] faz uso de um *Caffe model* baseado no SSD que usa a arquitetura ResNet-10 [60] como espinha dorsal. A rede disponibilizada pela OpenCV foi treinada numa base de dados pública, no entanto não é indicada qual. A rede S3FD apresenta contribuições para enfrentar melhor as variações de escala com uma única rede neuronal profunda. Em primeiro lugar, apresenta uma estrutura de deteção de rostos em escala, para lidar bem com diferentes escalas de rostos. Em segundo lugar, a taxa de recolha de caras com dimensões reduzidas é melhorada através de uma estratégia de correspondência de âncoras de compensação em escala.

Na rede DSFD os seus autores desenvolveram um módulo de melhoria de características, onde estendem o detetor de múltiplas regiões de passagem simples para passagem dupla. Isto permite a utilização de informações de diferentes níveis e assim obter caraterísticas mais robustas e discriminàveis. Nesta rede são também utilizado dois conjuntos diferentes de âncoras para encontrar de forma mais efetiva as características. Tanto o DSFD como o S3FD foram treinados originalmente com o conjunto de treino da base de dados WIDER FACE [61], com 12880 imagens no visível, possuindo um elevado grau de variabilidade em escala, pose e oclusões.

Todos estes modelos têm um funcionamento similar. Estes devolvem as áreas retangulares da imagem em que o detetor acredita que existam faces, e níveis de confiança atribuídos pelo detetor para cada face. Caso o algoritmo detete *n* áreas retangulares em que acredita que existam faces, devolverá estas *n* áreas retangulares, cada uma com um determinado nível de confiança. Tal permite obter uma imagem para cada uma das faces presentes.

4.1.2. Deteção dos Marcos Faciais

Após a deteção facial é possível utilizar algoritmos de marcação facial (ver Figura 17) com o intuito de identificar automaticamente a localização de marcos chave de uma face numa imagem, ou vídeo. A deteção destes marcos revela-se fundamental para diversos métodos de análise facial, tal como reconhecimento da expressão facial e estimativa da pose da cabeça [62].

[1] Acedido em *https://github.com/1adrianb/face-alignment.*
[2] Acedido em *https://github.com/opencv/opencv/tree/master/samples/dnn/face_detector.*
[3] Acedido em *https://github.com/1adrianb/face-alignment.*

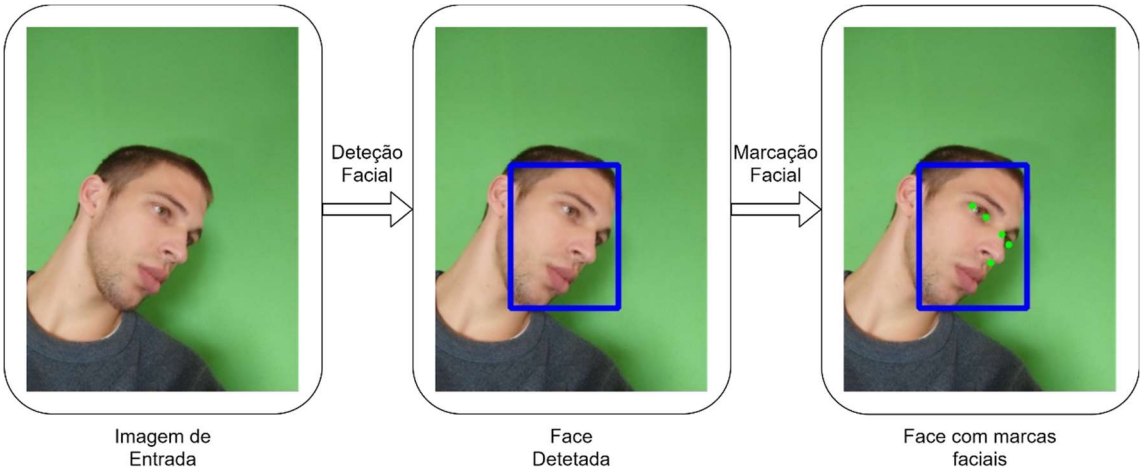

Figura 17 - Esquema do funcionamento da deteção e marcação facial.

Neste trabalho, faz-se o estudo de três modelos pré-treinados de marcação facial, baseados em redes neuronais, disponibilizados pela biblioteca DLIB[4] [63] e pela implementação oficial da rede 2D-FAN[5] de *Bulat et Tzimiropoulos* [64]. Os marcos faciais detetados por estas redes estão representados na Figura 18.

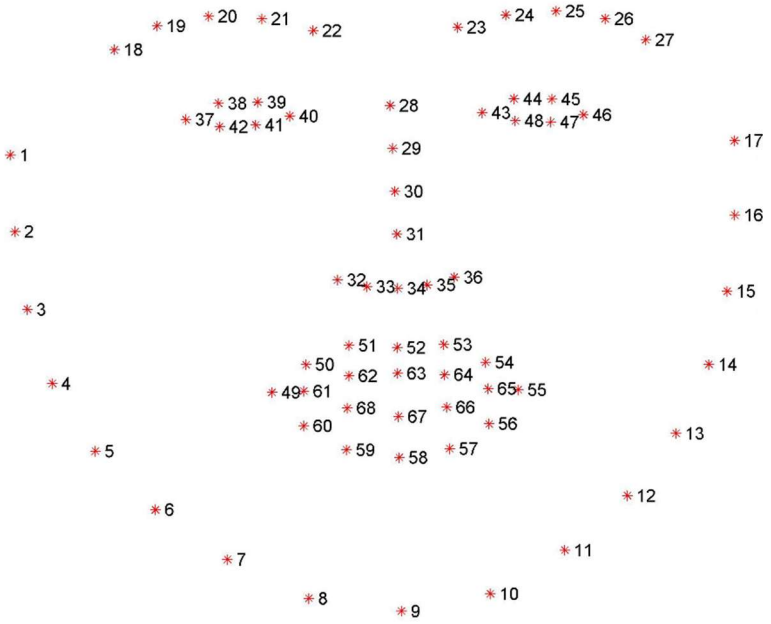

Figura 18 - Localização e numeração de 68 marcos faciais.

A biblioteca DLIB dispõe de dois marcadores faciais, um com 68 marcações faciais e outro com 5 marcações faciais. O algoritmo que devolve 68 marcos faciais da biblioteca DLIB é uma adaptação de *Khazemi et Sullivan* [65] e foi treinada na base de dados iBUG 300-W [66], composta por 7764 imagens no espetro do visível, com 68 marcos faciais anotados. A rede neuronal que devolve apenas 5 marcos faciais foi treinada numa base de dados com 7189 faces, criada especificamente para o efeito

[4] Acedido *em https://github.com/davisking/dlib-models.*
[5] Acedido em *https://github.com/1adrianb/face-alignment.*

pelo autor da biblioteca DLIB. Esta rede neuronal identifica a parte inferior do nariz e os cantos dos olhos (marcos 34, 37,40,43 e 46 da Figura 18).

Bulat e Tzimiropoulos disponibilizam uma rede de marcação facial com 68 marcações faciais a duas dimensões (2D-FAN, do inglês *Face Alignment Network*) [64], a qual é baseada na arquitetura Hour-Glass de *Newell et al.* [67], utilizada para estimar a pose humana. Esta rede foi treinada na base de dados 300-W-LP-2D [19], com 61225 imagens faciais no espetro do visível, geradas sinteticamente de forma a cobrir um maior espetro de poses, cada uma com 68 marcos faciais anotados.

4.1.3. Alinhamento, Corte e Redimensionamento

A subfase de alinhamento, corte e redimensionamento de imagem utiliza os dados obtidos pelas fases anteriores e produz uma imagem com a face alinhada e centrada através de rotação, recorte e redimensionamento da imagem.

Para este efeito, depois de recebido, como *input,* a área em que a face se encontra, bem como a localização dos marcos faciais, é calculado o centro de cada olho, ou centroide, através da média das coordenadas dos marcos faciais que o envolvem, como exposto na Figura 19.

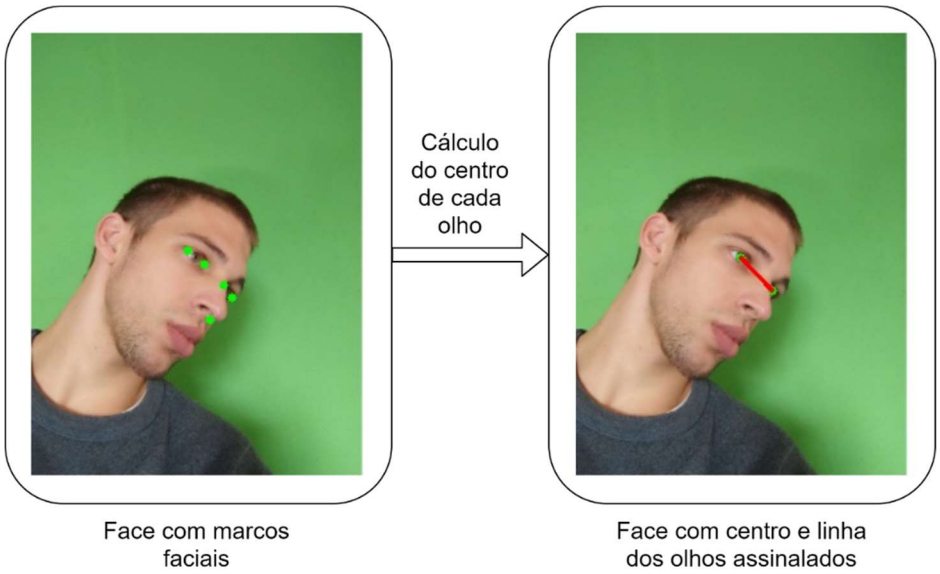

Figura 19 - Esquema de obtenção da linha dos olhos.

Obtida a posição do centro dos olhos, calcula-se o ângulo formado entre a reta definida pelas coordenadas dos centroides e o eixo horizontal, utilizando a expressão matemática do arco de tangente.

$$\alpha = \arctan\left(\frac{p2_y - p1_y}{p2_x - p1_x}\right) \tag{4}$$

onde α é o ângulo formado entre a linha imaginária dos olhos e o eixo horizontal e $p1$ e $p2$ são os pontos em que estão localizados centroides.

De seguida, calcula-se a matriz de rotação

$$M = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) & c_x(1-\cos(\alpha)) + c_y \sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) & c_y(1-\cos(\alpha)) + c_x \sin(\alpha) \end{bmatrix} \tag{5}$$

onde M é a matriz de rotação e c é o ponto intermédio entre os pontos $p1$ e $p2$, calculado por:

$$c = \frac{p2 - p1}{2} \tag{6}$$

Com a matriz de rotação aplica-se uma transformação afim:

$$\begin{bmatrix} p'_x \\ p'_y \end{bmatrix} = M \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \tag{7}$$

onde p é um ponto de duas dimensões da imagem original e p' é o ponto transformado. No fim desta etapa tem-se a linha dos olhos na horizontal, estando assim a face alinhada. Por fim, recorta-se a imagem utilizando as posições dos pontos obtidos pelo modelo de deteção facial (canto superior esquerdo e canto inferior direito da área em que acredita que existe uma face) e redimensiona-se a imagem. A dimensão final deve ser tal que sirva como entrada para os processos seguintes.

Na Figura 20 encontra-se um esquema resumo do procedimento adotado na subfase de alinhamento, corte e redimensionamento.

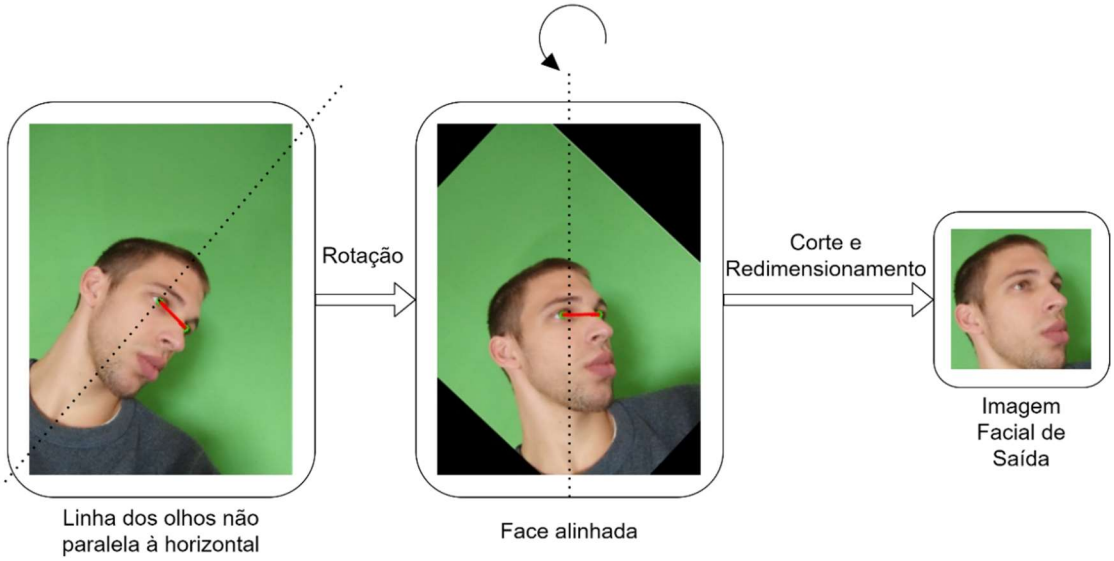

Figura 20 - Esquema resumo do alinhamento, corte e redimensionamento.

4.2. Síntese de Imagem

De forma a ultrapassar os problemas associados à aquisição de imagem em ambiente não controlado, como é o caso da variação da iluminação, existência de oclusões e variedade de poses, utiliza-se um módulo de síntese de imagem. Este módulo visa sintetizar (criar) uma imagem de uma face com pose frontal a partir de uma imagem de uma face não frontal.

Para exemplificar o comportamento esperado, na Figura 21 observa-se à entrada do módulo de síntese de imagem uma imagem facial numa pose não-frontal, obtendo-se à saída uma imagem facial frontal, preservando as caraterísticas da identidade. Assim, pretende-se que a imagem obtida facilite a obtenção das caraterísticas da identidade presente na imagem facial.

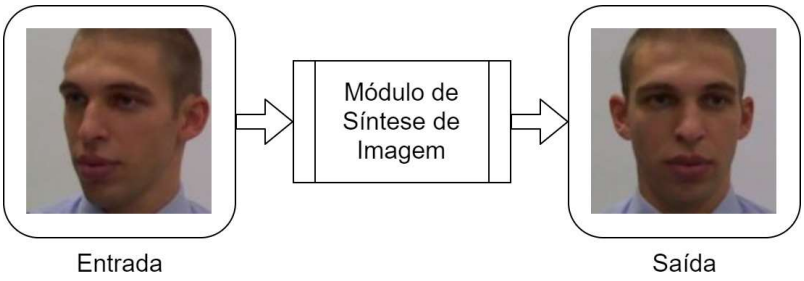

Figura 21 - Entrada e saída do módulo de síntese de imagem.

Para se sintetizar uma face com vista frontal a partir de uma imagem de uma pose não-frontal é necessária uma grande quantidade de pares de imagens de faces: uma com pose não-frontal e outra com pose frontal, da mesma pessoa. A utilização destes pares de imagens permitem que o modelo possa aprender esta tarefa de forma supervisionada. Para a normalização da pose facial para a pose frontal foram analisadas os modelos pré-treinados FNM[6] [24] e FFWM[7] (do inglês *Flow-based Feature Warping Model*) [68].

FNM é uma GAN da qual se destacam duas novidades. Em primeiro lugar, esta apresenta uma rede especializada em obter caraterísticas faciais para construir o gerador e fornecer a capacidade de preservar a identidade facial. Em segundo lugar, são utilizados discriminadores faciais para refinar texturas locais. Os seus autores afirmam que este modelo produz uma face em pose canónica sem expressão, o que melhora diretamente o desempenho de um sistema de reconhecimento facial.

O método de normalização da pose facial do modelo FFWM consiste no uso de um modelo de deformação, com o objetivo de sintetizar imagens frontais realistas com preservação de iluminação. Para a síntese de imagem frontal, apresenta um Módulo de Deformação, de forma a reduzir a discrepância de pose ao nível das caraterísticas faciais, e assim preservar mais detalhes de imagens de perfil. Para a fase de treino, o FFWM faz uso de pares de imagens de faces: uma com pose não-fr+ontal e outra com pose frontal, da mesma pessoa nas mesmas condições, imagens estas obtidas da base de dados Multi-PIE [12]. De forma diferente, o FNM utiliza imagens de faces não pares, isto é, que não são da mesma pessoa.

Com o intuito de utilizar imagens faciais não pares, o FNM utiliza imagens com pose frontal da base de dados Multi-PIE e imagens com pose não-frontal da base de dados CASIA-WebFace [40]. Deve-se frisar que estas bases de dados possuem apenas imagens no espetro do Visível. Assim, o FNM consiste num modelo de normalização da pose facial com o objetivo de descobrir padrões na forma de normalizar uma face para uma pose frontal, e, ao contrário do FFWM, não considera a preservação da iluminação.

4.3. Reconhecimento Facial

Este último módulo tem como objetivo identificar a pessoa presente numa imagem facial de entrada. Para este efeito, é necessário então executar duas tarefas: extração de caraterísticas e

[6] Acedido em *https://github.com/mx54039q/fnm.*
[7] Acedido em *https://github.com/csyxwei/FFWM.*

classificação. A primeira tarefa é realizada com um modelo de aprendizagem profunda que faz uso de redes neuronais convolucionais.

Para a tarefa da classificação, segue-se uma metodologia de aprendizagem de uma só vez (do inglês *one-shot learning*), obtendo pontuações de semelhança para cada banda espectral. De seguida, utilizando um método de fusão de pontuações, obtém-se valores de pontuação de semelhança combinada, sendo que a identidade predita será a que possuir maior pontuação de semelhança combinada. Desta forma, esta fase segue o procedimento apresentado no fluxograma da Figura .

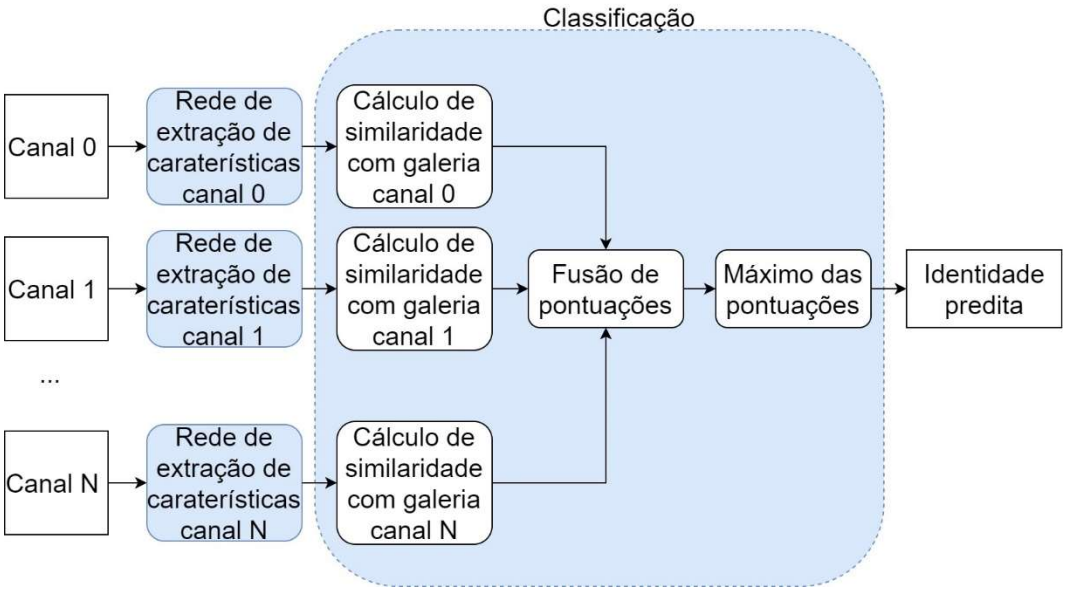

Figura 22 - Fluxograma do Módulo de Reconhecimento Facial com as tarefas de Extração de Caraterísticas e Classificação identificadas a azul.

4.3.1. Extração de Caraterísticas

Esta tarefa consiste em realizar a extração do conjunto de caraterísticas representativas de uma imagem facial através de uma CNN. A rede de extração de caraterísticas utilizada é a versão da Light CNN[8] [69] com 29 camadas convolucionais (Light CNN-29). A estrutura da Light CNN emprega uma extensão à função de ativação *maxout*, chamada de *Max-Feature-Map* (MFM) [69] em cada camada de convolução, como alternativa à ReLU. De acordo com [69], quando comparado com a ReLU, cujo limiar é aprendido a partir dos dados de treino, o MFM adota uma relação de concorrência para ter uma melhor capacidade de generalização e ser aplicável em diferentes distribuições de dados.

A arquitetura da Light CNN-29, detalhada na Tabela 6, contém 29 camadas convolucionais (conv), 4 camadas de agrupamento (pool) e uma camada totalmente ligada (fc).

[8] Acedido em *https://github.com/AlfredXiangWu/LightCNN*.

Tabela 6 - Arquitetura do modelo Light CNN-29, adaptado de [69].

Tipo	Tamanho do Filtro /Passada, enchimento	Tamanho de Saída	# Parâmetros (10^3)
Conv1	5 x 5/1, 2	128 x 128 x 96	2,4
MFM1	-	128 x 128 x 48	-
Pool1	2 x 2/2	64 x 64 x 48	
Conv2_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix}$ x 1	64 x 64 x 48	82
Conv2a	1 x 1/1	64 x 64 x 96	4,6
MFM2a	-	64 x 64 x 48	-
Conv2	3 x 3/1, 1	64 x 64 x 192	165
MFM2	-	64 x 64 x 96	-
Pool2	2 x 2/2	32 x 32 x 96	-
Conv3_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix}$ x 1	32 x 32 x 96	662
Conv3a	1 x 1/1	32 x 32 x 192	18
MFM3a	-	32 x 32 x 96	-
Conv3	3 x 3/1, 1	32 x 32 x 384	331
MFM3	-	32 x 32 x 192	-
Pool3	2 x 2/2	16 x 16 x 192	
Conv4_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix}$ x 1	16 x 16 x 192	3981
Conv4a	1 x 1/1	16 x 16 x 384	73
MFM4a	-	16 x 16 x 192	-
Conv4	3 x 3/1, 1	16 x 16 x 256	442
MFM4	-	16 x 16 x 128	-
Conv5_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix}$ x 1	16 x 16 x 128	2356
Conv5a	1 x 1/1	16 x 16 x 256	32
MFM5a	-	16 x 16 x 128	-
Conv5	3 x 3/1, 1	16 x 16 x 256	294
MFM5	-	16 x 16 x 128	-
Pool4	2 x 2/2	8 x 8 x 128	
fc1	-	512	4194
MFM_fc1	-	256	
Total	-	-	12673

É de notar que a Light CNN-29 utiliza também a estrutura de blocos residuais, inspirado em [60], cada um com duas camadas convolucionais, representado na Figura 23. A estrutura de blocos residuais permite às redes neuronais possuir um maior número de camadas convolucionais sem o problema da degradação de precisão no treino e com menos perda de informação ao longo da rede (devido ao desvanecimento do gradiente) [60].

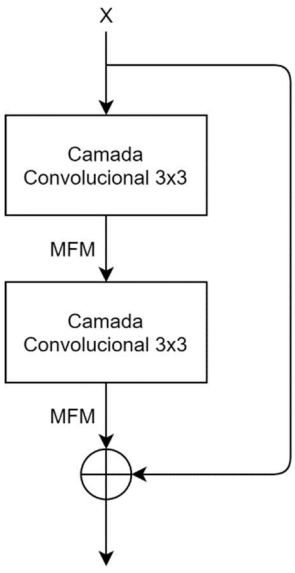

Figura 23 - Arquitetura dos blocos residuais empregues na Light CNN-29, adaptado de [69].

A Light CNN-29 recebe como entrada uma imagem em escala cinza com o tamanho de 128×128 *pixels* e produz à saída da primeira camada totalmente ligada um vetor de 256 dimensões. Este vetor consiste no conjunto de caraterísticas representativas da imagem facial. A Light CNN-29 foi treinada nas bases de dados de imagens faciais do espetro do visível CASIA-WebFace [40] e MS-Celeb-1M [39], contabilizando cerca de $5{,}1\times10^{6}$ imagens faciais de 80013 identidades diferentes. Assim, para o treino desta rede, foi adicionada uma camada totalmente ligada à sua saída de 80013 dimensões, de forma a proceder à classificação.

De forma a utilizar esta rede para a extração de caraterísticas em espetros diferentes do visível, é utilizada a transferência de aprendizagem. De acordo com [54], diversos modelos para reconhecimento biométrico são baseados em transferência de aprendizagem quando as bases de dados são limitadas. Assim, deve-se utilizar o modelo da Light CNN-29 já com os pesos obtidos pelo treino nas bases de dados do visível e proceder-se a um ajustamento fino (*fine-tune* do inglês) com as bases dados de imagens faciais nos espetros diferentes do visível.

No final da fase de extração de caraterísticas, são gerados B vetores de 256 dimensões, sendo B o número de bandas espectrais em que a imagem facial foi obtida.

4.3.2. Classificação

Após a realização da tarefa de extração de um conjunto de caraterísticas representativo de uma imagem facial, surge então a necessidade de classificar este conjunto na classe correspondente à identidade da pessoa presente na respetiva imagem facial. Para este efeito é utilizada uma técnica de aprendizagem de uma só vez. Esta técnica permite à Light CNN-29 aprender a generalizar novas classes com apenas um exemplo identificado para cada classe. Assim como o ser humano consegue reconhecer novas classes com poucos exemplos identificados, é também de grande interesse que o sistema de reconhecimento facial aprenda a classificar novas classes com uma quantidade limitada de exemplos.

O processo de classificação aplicado pela técnica de aprendizagem de uma só vez consiste em comparar uma imagem de entrada com uma imagem de cada classe presente no conjunto de suporte e determinar qual destas apresenta uma maior semelhança com a imagem de entrada [70]. Este conjunto de suporte é constituído por um exemplo de cada classe devidamente identificado a que o classificador tem acesso. No caso deste trabalho, o conjunto de suporte de cada banda espectral é composto por uma imagem facial frontal nessa mesma banda espectral de cada identidade passível de ser classificada nessa mesma banda.

Na Figura 24 está exemplificada a forma como se processa o cálculo da semelhança entre cada par de imagens. As funções mais utilizadas para calcular a semelhança entre dois vetores de caraterísticas são a distância euclidiana e a similaridade de cosseno.

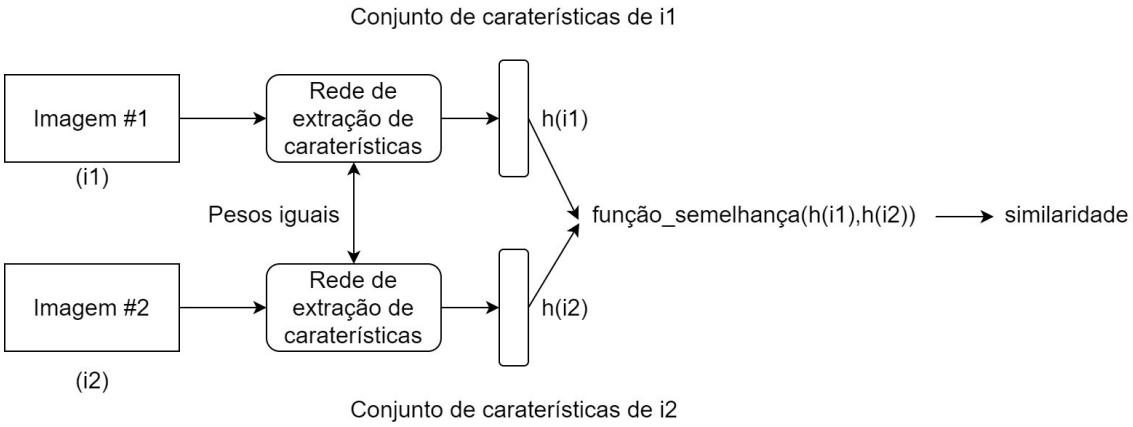

Figura 24 - Cálculo da pontuação de semelhança entre duas imagens.

A distância euclidiana, tal como o nome sugere consiste em encontrar a distância entre dois pontos, *x e y* num espaço-*n*, utilizando o teorema de Pitágoras, sendo que quanto menor a distância entre dois pontos, maior a semelhança entre estes.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{8}$$

A similaridade de cosseno consiste em encontrar o valor cosseno do ângulo formado entre estes dois pontos. Quanto maior o valor do cosseno, maior é a sua semelhança.

$$sim(x,y) = \cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|} \tag{9}$$

Os classificadores baseados em redes neuronais, como aquele que a Light CNN-29 utiliza para treinar, só permitem classificar um conjunto de determinadas classes, com as quais a rede foi treinada. Tal faz com que, de cada vez que é inserida uma nova classe, seja necessário treinar o classificador de novo, assim como também requer que esta nova classe tenha um conjunto suficiente de dados para a rede poder aprender a sua representação. A utilização da técnica de aprendizagem de uma só vez permite que não seja necessário este novo treino do classificador, estando assim aptos para adicionar novas classes e retirar outras a qualquer momento. Para além deste último ponto, esta técnica também permite classificar dados numa classe que apenas tenha um exemplo na galeria.

Ao receber diferentes conjuntos de caraterísticas da fase de extração de caraterísticas, obtém-se as pontuações de semelhança para cada banda espectral. Assim, recebendo uma imagem numa determinada banda espectral, calcula-se a semelhança com as N imagens faciais no grupo de suporte da mesma banda, onde N é o número de identidades. Após obter os valores de semelhança para cada identidade nas diferentes bandas espectrais, é feito uma fusão das pontuações obtidas, inspirado em [71]. Desta forma são calculadas as pontuações de semelhança nas B bandas espectrais para cada identidade na base de dados, que são combinadas usando um algoritmo de fusão de pontuação para obter as pontuações finais. Estas pontuações finais são utilizadas para identificar a melhor correspondência com a amostra, através de:

$$S_{ic} = \sum_{b=1}^{B} S_{ib} W_b \quad (10)$$

Aqui, S_{ic} é a pontuação combinada para cada identidade i e S_{ib} é a pontuação obtida por cada banda b para cada identidade i. W_b é o peso associado a cada banda.

Os pesos associados a cada banda são valores fixo, determinados pela precisão de *Rank-1* dada em (3) e obtida quando realizada a classificação apenas com essa banda [72]. Desta forma, a banda que obtenha por norma as pontuações de semelhança mais fiáveis para classificar, terá um peso maior na fusão de pontuações.

A classificação é então feita escolhendo a identidade i do conjunto de suporte que tenha maior pontuação de semelhança combinada com a identidade de prova:

$$predição = máx(S_{ic}) \; \forall i \in [1, \dots, N] \quad (11)$$

5. Resultados e Discussão

Neste capítulo são apresentados e analisados os resultados produzidos pela metodologia apresentada no capítulo anterior. Nas tarefas em que foram apresentadas mais do que uma técnica, estas são alvo de comparação, de forma a eleger a técnica que mais se adequa para a realização da tarefa em causa.

O presente capítulo está dividido em 4 Secções, iniciando-se pela apresentação das bases de dados multiespectrais. Na secção seguinte apresenta-se um estudo comparativo das diferentes técnicas utilizadas para as tarefas de deteção facial e deteção dos marcos faciais. Depois, é avaliado o desempenho dos modelos pré-treinados de normalização da pose, de forma a selecionar o modelo a utilizar. Por último, é feita a avaliação da rede de extração de caraterísticas, bem como a utilização da técnica de aprendizagem de uma só vez e fusão de pontuações. Ainda na última secção é efetuada a avaliação do desempenho geral da metodologia proposta, em várias bases de dados multiespectrais.

5.1. Bases de Dados

A presente secção tem como objetivo apresentar as bases de dados multiespectrais utilizadas durante o desenvolvimento desta dissertação de mestrado: a CASIA NIR-VIS 2.0 [44], a TUFTS [8] e a IRIS [73], as quais foram selecionadas por apresentarem imagens faciais multiespectrais com caraterísticas de ambiente não controlado.

Nestas bases de dados, cada imagem corresponde a uma identidade específica. No entanto, em algumas imagens foram detetadas mais do que uma face, pelo que foi necessário efetuar uma limpeza das bases de dados após a fase de deteção e alinhamento da face, tal que não existam diferentes pessoas associadas à mesma identidade (pessoa).

Para as fases de treino e teste do sistema de reconhecimento facial, foi necessário organizar as bases de dados por diretorias segundo a identidade da pessoa. Assim, todas as imagens pertencentes à mesma identidade foram inseridas na diretoria respetiva a essa identidade.

5.1.1. CASIA NIR-VIS 2.0

A base de dados multiespectral CASIA NIR-VIS 2.0 [44], construída pela *Chinese Academy of Sciences Institute of Automation* (CASIA), possui imagens faciais de 725 indivíduos nas bandas espectrais do NIR e visível, de idades variadas, sendo na sua maioria alunos da CASIA. Saliente-se que as imagens das diferentes bandas espectrais não foram obtidas nas mesmas condições. Esta base de dados é utilizada por diversos investigadores para a obtenção de resultados na tarefa de reconhecimento facial heterogéneo (NIR-VIS).

Na Figura 25 estão ilustradas algumas imagens da base de dados multiespectral CASIA NIR-VIS 2.0 [44].

Figura 25 - Imagens da base de dados CASIA NIR-VIS 2.0 [44]*, onde cada coluna corresponde a imagens faciais da mesma pessoa em diferentes bandas espectrais.*

Algumas imagens possuem em "plano de fundo" outras pessoas e, na fase de deteção e alinhamento, as faces de todas as pessoas são detetadas. Mas, de acordo com os criadores da base de dados, cada imagem corresponde apenas a uma identidade, pelo que as faces das pessoas em segundo plano são descartadas. Exemplos dos casos suprarreferidos estão presentes na Figura 26.

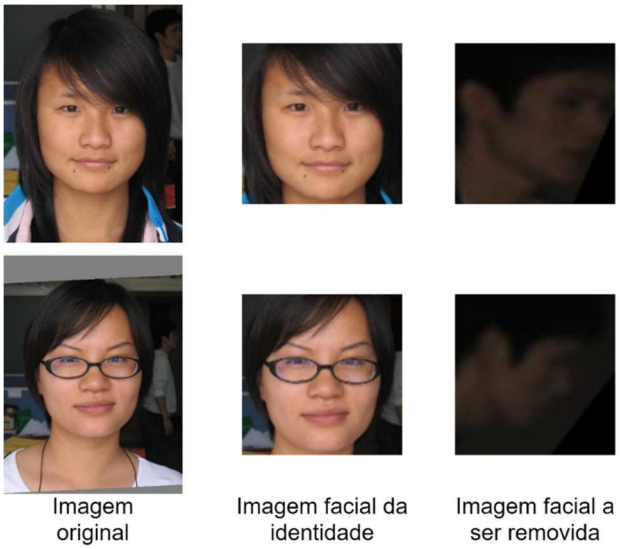

Figura 26 - Exemplos de imagens com duas pessoas na base de dados CASIA NIR-VIS 2.0 [44].

5.1.2. TUFTS

A base de dados multiespectral TUFTS [8], construída pela *Tufts University*, dos Estados Unidos da América, possui imagens faciais nas bandas espectrais do LWIR, NIR e visível, de 113 indivíduos de mais de 15 países, possuindo assim uma grande diversidade étnica, sendo alunos e *staff* da *Tufts University,* bem como familiares destes.

Esta base de dados apresenta imagens faciais com variações na pose no plano horizontal entre os -60º e os +60º, diferentes níveis de luminosidade e diferentes expressões faciais. Estas

caraterísticas tornam esta base de dados indicada para avaliar a capacidade do sistema de reconhecimento facial em ambiente não controlado.

No que diz respeito às imagens com variação na pose, as imagens multiespectrais são obtidas nas bandas do Visível, NIR e LWIR. Quanto às imagens com diferentes expressões faciais, a face está numa pose frontal, e foram obtidas nas bandas do Visível e do LWIR. Nas Figuras 27 e 28 estão ilustradas algumas imagens da base de dados multiespectral TUFTS [8].

Figura 27 - Imagens com variação da pose da base de dados TUFTS [8].

Figura 28 - Imagens com variação da expressão da base de dados TUFTS [8].

Nesta base de dados também foi necessário descartar as imagens faciais das pessoas em segundo plano.

A base de dados TUFTS está subdividida em diversas diretorias, de acordo com as caraterísticas das imagens presentes. Assim, nos testes realizados foram utilizadas em separado as diretorias com imagens com variação de pose facial e com variação da expressão facial, denominadas daqui em diante por TUFTS-Pose, e TUFTS-Expressão, respetivamente.

5.1.3. IRIS

A base de dados multiespectral IRIS [73] possui imagens faciais de 30 indivíduos nas bandas espectrais do LWIR e visível, com variações na pose no plano horizontal entre os -90º e os +90º,

expressões faciais e variação de iluminação. Na Figura 29 estão ilustradas algumas imagens da base de dados multiespectral IRIS.

Figura 29 - Ilustração de imagens da base de dados IRIS, onde cada coluna corresponde a imagens faciais da mesma identidade em diferentes bandas espectrais.

A base de dados IRIS é utilizada apenas para o treino da rede de extração de caraterísticas.

5.2. Deteção e Alinhamento da Face

Nesta secção é feito um estudo de diferentes algoritmos para a realização das tarefas de deteção facial e marcação facial. Visto as caraterísticas inerentes ao ambiente não-controlado (variações de pose, de iluminação, de expressões faciais e existência de oclusões) é importante aferir a robustez dos diferentes métodos para a consecução das tarefas presentes no módulo de deteção e alinhamento da face. Assim, são apresentados os resultados obtidos por diferentes algoritmos para as tarefas de deteção facial e marcação facial.

Para avaliação quantitativa, devido a não existirem bases de dados multiespectrais com etiquetagem para avaliar o desempenho dos algoritmos de deteção facial e deteção de marcos faciais, foi realizada uma etiquetagem para as caixas delimitadoras na base de dados TUFTS-Pose. No entanto, não foi realizada a etiquetagem dos marcos faciais, pelo que apenas se obtiveram resultados numéricos para os algoritmos de deteção facial. Para além da realização desta avaliação quantitativa, com o objetivo de identificar qual o método mais indicado para cada uma tarefas, foram também realizados testes qualitativos para observar visualmente o comportamento dos diferentes métodos. O procedimento adotado para os testes qualitativos está apresentado na Figura 30, que permite mostrar no fim de cada tarefa os resultados obtidos pelos métodos utilizados.

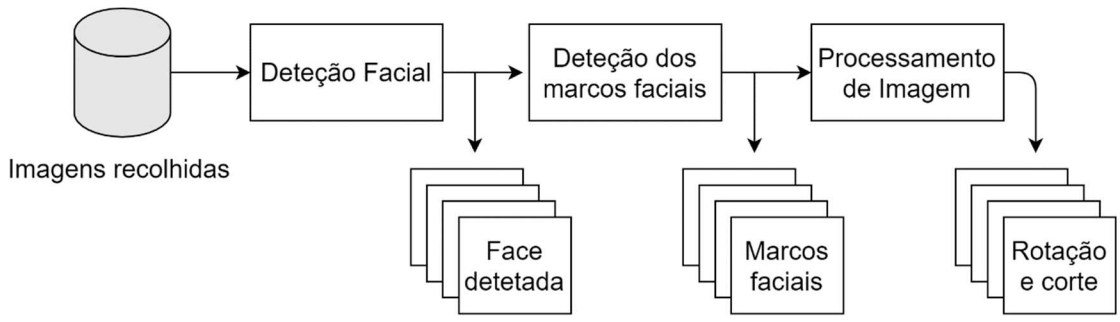

Figura 30 - Sequência dos passos a seguir para obtenção de resultados das diferentes tarefas

Para a realização dos testes qualitativos, fez-se uso do equipamento de aquisição de imagem disponibilizado pela Academia Militar para a obtenção de imagens faciais. Na Figura 31 estão presentes as imagens obtidas.

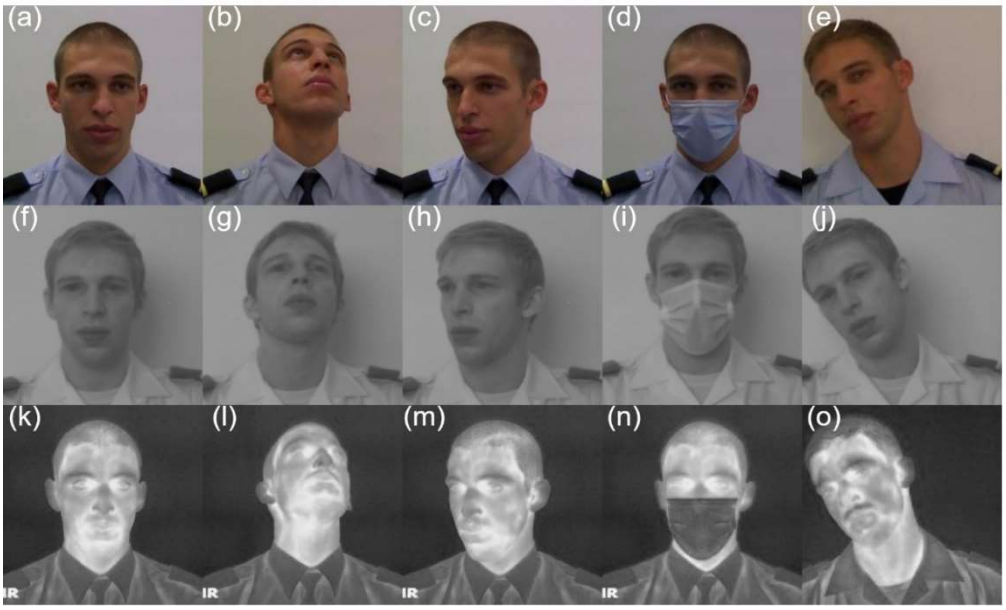

Figura 31 - Imagens utilizadas nos testes qualitativos do módulo de deteção e alinhamento de face nas bandas espectrais do Visível, NIR e LWIR.

À medida que existe um afastamento da pose frontal, menos são as caraterísticas faciais passíveis de se observar, o que aumenta a dificuldade para os algoritmos de deteção facial e deteção de marcos faciais. O mesmo ocorre quando existe algum tipo de máscara facial, passível de provocar oclusão das caraterísticas faciais. Desta forma, para o estudo qualitativo da capacidade dos algoritmos empregues num ambiente não controlado, foram obtidas as imagens com caraterísticas distintas – (i) olhar em frente, (ii) olhar para cima a aproximadamente 30º, (iii) olhar para o lado a aproximadamente 45º, (iv) olhar em frente com uma máscara cirúrgica e (v) olhar em frente com a cabeça inclinada a aproximadamente 20º.

As imagens faciais obtidas encontram-se nas diferentes bandas espectrais: (i) Visível, (ii) NIR, e (iii) LWIR, de forma a verificar a capacidade de generalização dos métodos para as diferentes bandas espectrais. Note-se que estes métodos foram apenas treinados na banda espectral do Visível. Assim, os testes realizados servem para auferir a capacidade do presente módulo na deteção de face e deteção de marcos faciais em imagens multiespectrais em ambiente não controlado. É de salientar que as imagens do Visível e do LWIR foram obtidas simultaneamente por um equipamento multiespectral, o que faz com que o único fator que difere é a banda espectral em que a imagem é captada. As imagens do NIR foram captadas com um equipamento diferente, mono-espectral, podendo haver ligeiras variações na iluminação ou na pose facial, quando comparadas com as imagens do Visível e do LWIR

5.2.1. Deteção Facial

Como descrito na Secção 4.1.1, os modelos pré-treinados aplicados na tarefa de deteção facial foram o S3FD de *Zhang et al.* [57], a rede neuronal profunda de deteção facial da OpenCV [58] e o

DSFD de *Li et al.* [59]. É de frisar que todos estes modelos foram treinados nos trabalhos originais em bases de dados na banda espectral do Visível, o que levou à necessidade de as validar previamente numa base de dados multiespectral, como a TUFTS-Pose, para verificar se continuam a funcionar corretamente com espetros diferentes do Visível.

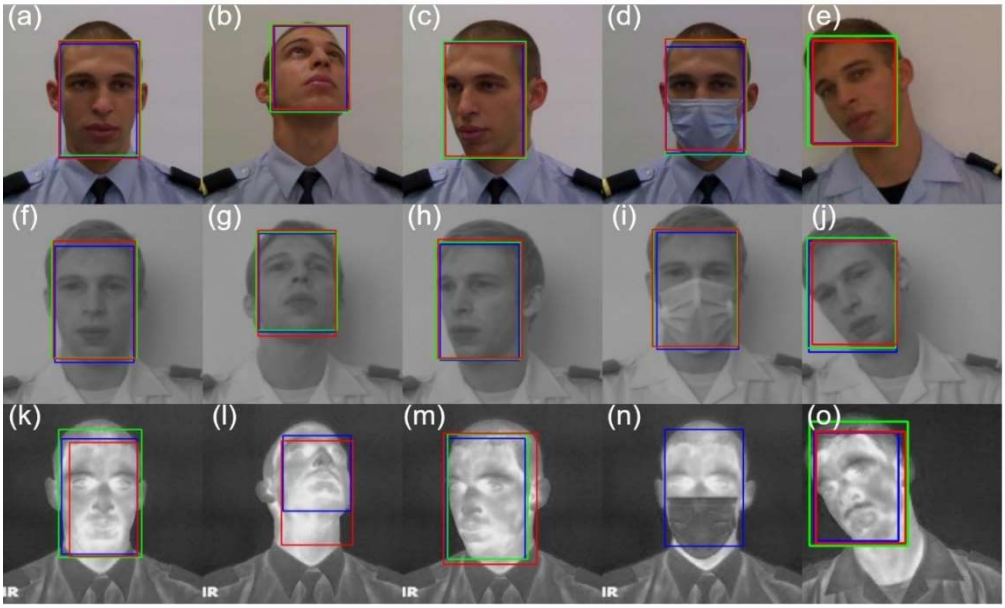

Figura 32 - Resultados obtidos pelos modelos pré-treinados de deteção facial nas diferentes bandas espectrais. S3FD-vermelho, DSFD-azul, OpenCV-verde.

Relativamente aos resultados qualitativos, apresentados na Figura 32, na banda espectral do Visível todos os modelos produziram resultados semelhantes, o que era de esperar uma vez que todos foram treinados em bases de dados da banda espectral do Visível, sem haver nenhuma falha de deteção facial. Na banda espectral NIR, todos os algoritmos detetaram corretamente as faces. Uma possível justificação para este desempenho similar deve-se à proximidade das bandas espectrais do NIR e do Visível no espetro eletromagnético. Na banda espectral LWIR, são observadas falhas pelos modelos da OpenCV (Figura 32l e 32n) e S3FD (Figura 32n). Estas imagens possuem um grau de dificuldade maior que as restantes por duas razões: (i) são imagens obtidas na banda espectral do LWIR, que é consideravelmente diferente da banda espectral do Visível e (ii) a pose algo incomum da Figura 32l e a oclusão de caraterísticas faciais provocadas pela máscara cirúrgica na Figura 32n. Além disso, quando os modelos da OpenCV e S3FD detetam as faces na banda espectral do LWIR, há uma variação na área retangular delimitadora da face, quando comparada com a banda espectral visível. O modelo DSFD manteve os mesmos resultados, sendo um bom indicador da sua capacidade de detetar faces mesmo na banda espectral LWIR.

Os resultados quantitativos encontram-se na Tabela 7, utilizando a base de dados TUFTS-Pose, onde se observa que a rede da OpenCV produz resultados inferiores aos outros algoritmos, especialmente nas bandas espectrais infravermelhas. Na comparação de resultados entre os modelos S3FD e DSFD, observam-se resultados muito semelhantes na banda espectral do visível e do NIR. No entanto, os resultados na banda espectral do LWIR do DSFD são cerca de 8 pontos percentuais melhores. Observa-se que o DSFD mantém uma precisão muito elevada para as diferentes

bandas espectrais, sendo o melhor algoritmo para a deteção facial num sistema de análise facial multiespectral.

Tabela 7 - Resultados (em %) obtidos na tarefa de deteção facial na base de dados TUFTS [8] *com variação da pose.*

	Banda espectral		
Método	**VIS**	**NIR**	**LWIR**
OpenCV [58]	99,2	90,4	77,7
S3FD [57]	**99,9**	**100,0**	90,8
DSFD [59]	**99,9**	**100,0**	**98,8**

5.2.2. Deteção dos Marcos Faciais

Na presente secção são avaliadas as redes pré-treinados de deteção facial de 5 e 68 marcos da biblioteca DLIB e a 2D-FAN.

Os resultados da rede DLIB de 68 marcos faciais estão apresentados na Figura 33. Para as imagens da banda espectral do Visível, a rede de 68 marcos faciais da biblioteca DLIB tem um bom desempenho, com apenas algumas falhas na Figura 33c ao nível da marcação do olho direito. Na banda do NIR observa-se uma falha na marcação do nariz na Figura 33g e uma marcação bastante incorreta em diversos marcos faciais na Figura 33h. Uma causa possível deste comportamento é o facto deste modelo de deteção de marcos faciais ter sido treinado num conjunto de dados sem grandes variações ao nível da pose. A rede revela ainda mais dificuldades na banda espectral do LWIR, falhando a marcação facial dos olhos em todas a imagens.

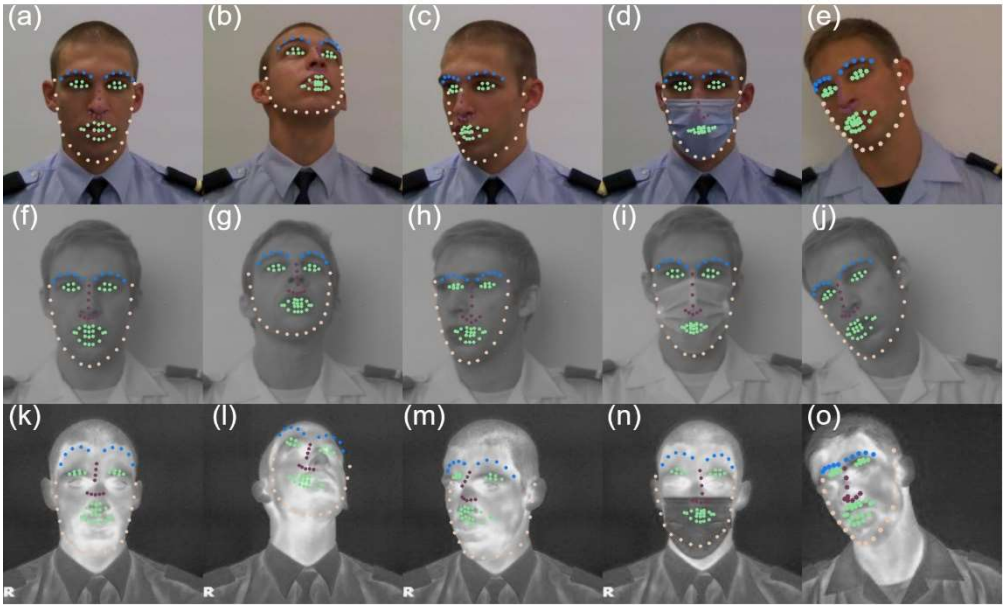

Figura 33 - Resultados obtidos pela rede pré-treinada DLIB de 68 marcos faciais nas diferentes bandas espectrais.

Os resultados obtidos pela rede DLIB de 5 marcos faciais encontram-se na Figura 34. Esta apresenta resultados positivos nas bandas espectrais do Visível e do NIR, onde identifica corretamente as posições dos cantos dos olhos nas diferentes imagens, falhando ligeiramente a marcação da parte inferior do nariz nas Figuras 34b e 34g. No entanto, na banda espectral do LWIR, é observável que a rede DLIB não consegue identificar a maioria dos marcos faciais, levando a que esta não realize nenhuma marcação correta dos olhos nesta banda espectral.

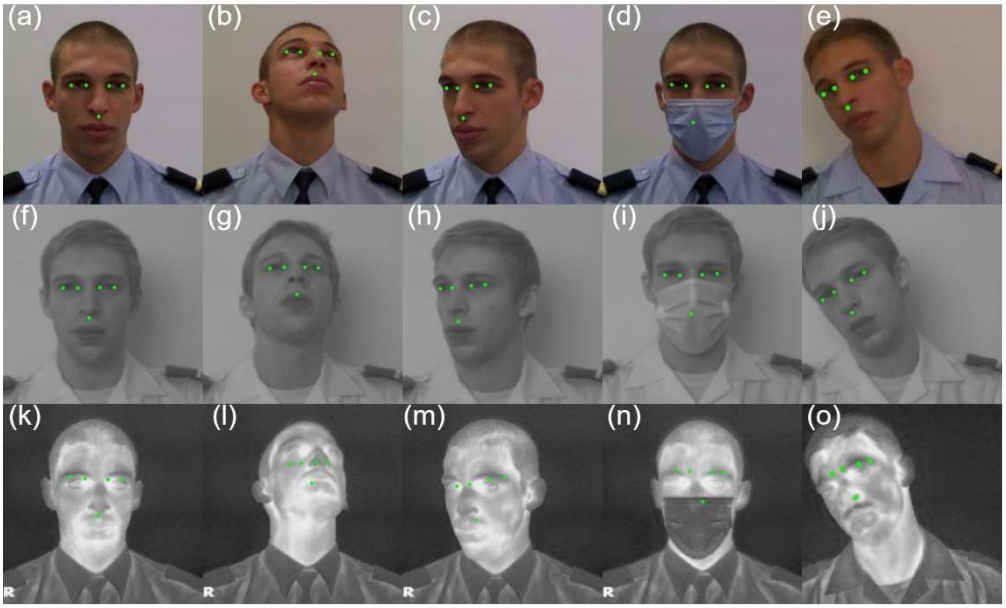

Figura 34 - Resultados obtidos pela rede pré-treinada DLIB de 5 marcos faciais nas diferentes bandas espectrais.

Quanto à rede 2D-FAN, os resultados apresentados na Figura 35 revelam uma boa extração de marcos faciais em qualquer das bandas espectrais do NIR e do Visível. Em relação às imagens do LWIR, quando não existe máscaras faciais os resultados são bastante semelhantes aos obtidos no Visível. Já com a existência de máscaras faciais, a rede 2D-FAN revela marcações faciais bastante erradas na Figura 35n, com os marcos faciais que não estão à vista a serem espalhados pela imagem. Na Figura 35l é observável uma falha na marcação do olho direito, ainda que não muito longe da localização real.

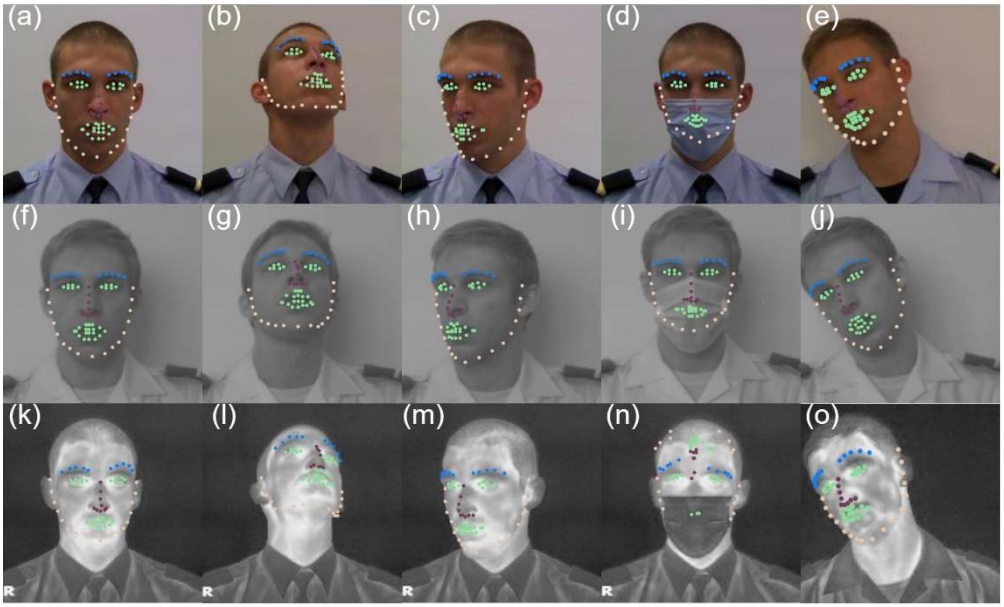

Figura 35 - Resultados obtidos pela rede pré-treinada 2D-FAN nas diferentes bandas espectrais.

Resumindo os resultados obtidos pelo teste qualitativo, na banda do Visível e do NIR, a rede DLIB de 68 marcos faciais apresenta resultados não positivos nas imagens com maior variação da pose, enquanto as restantes redes obtiveram bons resultados. A rede 2D-FAN foi a que obteve melhores resultados, principalmente nas imagens da banda espectral do LWIR. Na imagem com máscara cirúrgica na banda espectral do LWIR (Figura 35n), apesar de aparentemente possuir o pior resultado de entre as 3 redes, é aquela que faz uma marcação dos olhos mais próxima da realidade.

Atendendo às considerações anteriores, decidiu-se utilizar a rede 2D-FAN em detrimento das restantes, devido a dois fatores: (i) apresenta resultados positivos com variação da pose facial e (ii) é a única capaz de produzir resultados positivos na banda espectral LWIR.

5.2.3. Alinhamento, Corte e Redimensionamento

Após a deteção da face com o modelo pré-treinado DSFD e a deteção de marcos faciais com o modelo pré-treinado 2D-FAN, teve lugar a fase de alinhamento, corte e redimensionamento, que alinhou a linha imaginária dos olhos de todas as faces detetadas com a horizontal, centrou as faces nas imagens e redimensionou-as para o mesmo tamanho. Esta normalização espacial das imagens faciais ajuda um sistema de reconhecimento facial multiespectral num ambiente não controlado, onde as faces podem ser apresentadas em várias poses diferentes. Os resultados produzidos pelo módulo de deteção e alinhamento da face encontram-se na Figura 36. O efeito de alinhamento é percetível na pose facial mais à direita.

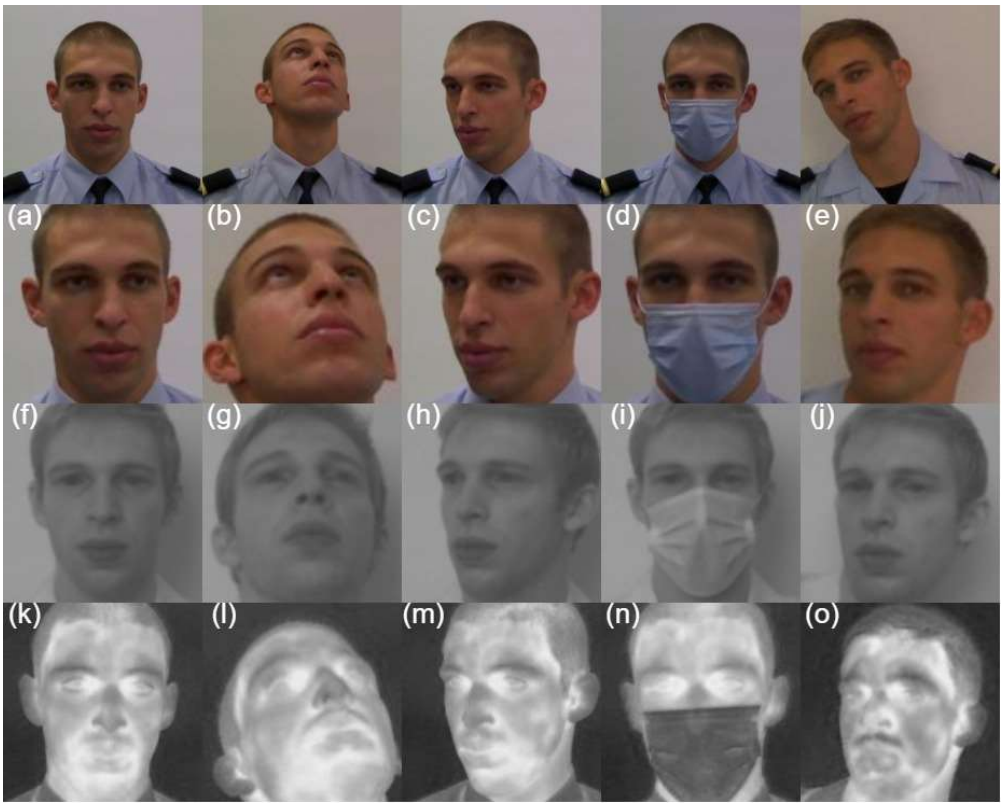

Figura 36 - Resultados obtidos pelo módulo de deteção e alinhamento da face proposto nas diferentes bandas espectrais. As imagens superiores (sem numeração) são as originais na banda espectral do visível.

5.3. Síntese de Imagem

O módulo de síntese de imagem tem como objetivo verificar a vantagem de se normalizar a face para uma pose frontal em sistemas de reconhecimento facial multiespectral. Foram avaliadas os modelos pré-treinados FNM [24] e FFWM [68]. Para uma avaliação qualitativa do funcionamento dos modelos, foram utilizadas as imagens da Figura 31, com exceção das Figuras 31e, 31j e 31o, as quais foram substituídas pelas imagens presentes na Figura 37. Esta alteração tem como objetivo analisar o comportamento dos modelos de síntese de imagem com variação da expressão facial.

Figura 37 - Imagens utilizadas nos testes qualitativos do módulo de síntese de imagem nas bandas espectrais do Visível, NIR e LWIR.

Para as todas as imagens utilizadas nas avaliações qualitativas e quantitativas, as imagens foram previamente processadas, de forma a ficarem devidamente centradas, alinhadas e dimensionadas. O modelo FFWM necessita de receber as imagens faciais onde as posições de determinados marcos faciais estejam sempre nas mesmas coordenadas, e por isso foi utilizado o módulo de deteção e alinhamento da face disponibilizado pelos autores do FFWM para obter os

resultados. As imagens utilizadas pelo modelo FNM foram processadas pelo módulo de deteção e alinhamento da face desenvolvido pelos autores do presente trabalho.

5.3.1. Seleção do Melhor Modelo

Na Figura 38 apresentam-se os resultados obtidos pelo modelo FFWM [68], com uma imagem em falta devido à falha de deteção da face pelo módulo disponibilizado pelos autores. Como é possível de observar, o seu desempenho tem uma queda acentuada à medida que se afasta do Visível. Analisando apenas a banda espectral do Visível e as imagens com variação de pose (Figura 38b e 38c), verifica-se uma boa normalização da pose na Figura 38c. No entanto quando a pessoa está a olhar para cima, na Figura 38b, o modelo produz uma face deformada por a imagem facial de entrada estar numa posição não perpendicular ao solo. A utilização exclusiva da base de dados Multi-PIE [12] no treino do modelo FFWM faz com que esta só tenha capacidade para proceder à normalização das faces onde a pose varie ao nível do plano horizontal. No que toca à variação de expressão facial, na Figura 38e verifica-se que o modelo FFWM não normaliza a expressão facial, pelo que é observável não demonstra apresentar qualquer vantagem a sua utilização neste caso.

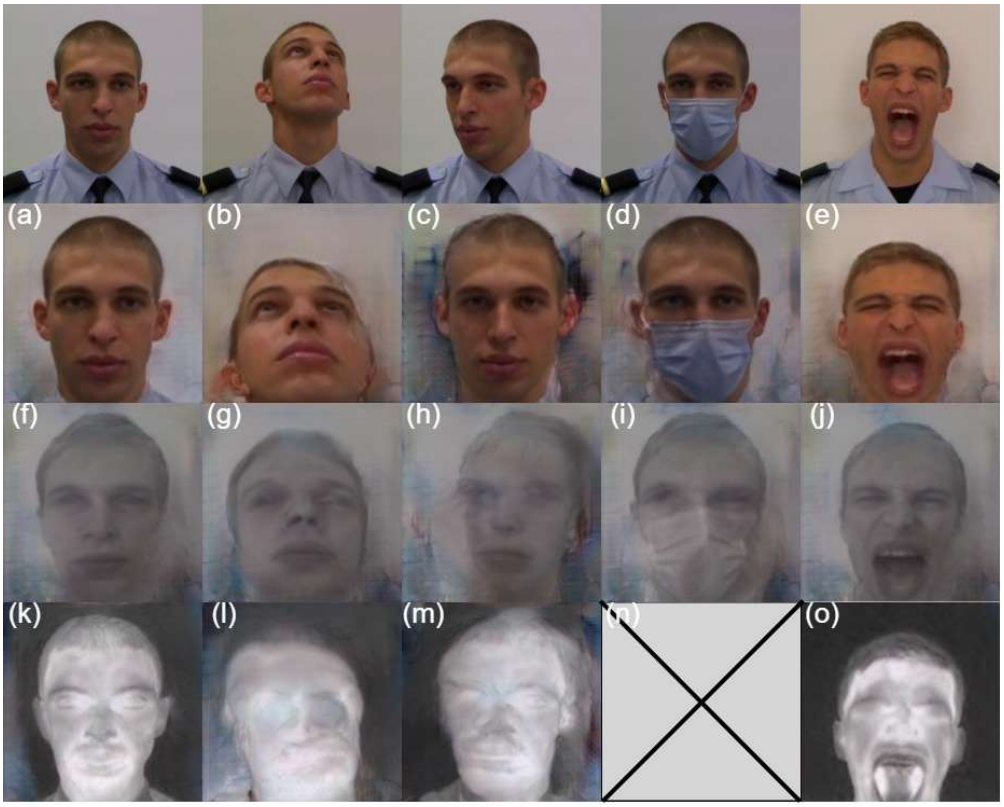

Figura 38 - Resultados obtidos pelo modelo pré-treinado FFWM na tarefa de normalização de vários para um nas diferentes bandas espectrais. As imagens superiores (sem numeração) são as originais na banda espectral do visível.

O modelo FNM apresenta resultados (Figura 39) mais satisfatórios na banda espectral NIR, onde as imagens faciais que providencia à sua saída são mais realistas e menos deformadas que os do modelo FFWM. É de notar que com o modelo FNM as identidades sofrem uma alteração, isto é, a

pessoa na imagem facial de saída parece ser diferente da pessoa presente na imagem facial de entrada. No entanto, a utilização de um extrator de caraterísticas faciais pela parte do modelo FNM permite manter as caraterísticas mais relevantes na imagem facial de saída. É relevante salientar que o modelo FNM não apenas normaliza a pose e expressão (ver Figura 39e), mas também elimina máscaras faciais (ver Figura 39d) e normaliza para a banda espectral do Visível. No entanto, esta normalização produz resultados não positivos com as imagens do LWIR, devido à diferença de domínio entre as bandas espectrais do LWIR e do Visível.

Figura 39 - Resultados obtidos pelo modelo pré-treinado FNM na tarefa de normalização de vários para um nas diferentes bandas espectrais. As imagens superiores (sem numeração) são as originais na banda espectral do visível. As imagens da 2ª, 3ª e 4ª linha são as imagens produzidas pelo modelo quando recebe como a entrada as imagens nas bandas espetrais do Visível, NIR e LWIR, respetivamente.

Atendendo às considerações anteriores, decidiu-se utilizar o modelo FNM em detrimento da FFWM, devido a dois fatores: (i) o modelo FFWM necessita de um módulo de deteção e alinhamento facial específico e que a face esteja perpendicular à horizontal, enquanto o modelo FNM é mais robusto a variações na pose na imagem de entrada e (ii) todas as imagens normalizadas pelo modelo FNM tendem a manter as proporções da face, sem as deformar, nas bandas espectrais do NIR e do Visível.

5.3.2. Avaliação do Modelo Selecionado

Para verificar a vantagem da normalização da face com o modelo FNM, foram executados testes onde se realizou a identificação com e sem normalização da face pelo modelo FNM [24], para cada uma das bandas espectrais. Foi utilizada a Light CNN-29 [69] para extração de caraterísticas e a identificação foi realizada com base na pontuação obtida pela similaridade de cosseno.

5.3.2.1. Normalização da Face

Os resultados apresentados na Tabela 8 foram obtidos com a utilização da base de dados TUFTS-Pose, a qual apresenta variações na pose facial. Estes resultados mostram que, no que respeita à classificação sem a utilização do modelo FNM, a utilização da banda espectral NIR produz melhores resultados que a banda do Visível em todas as métricas analisadas. Uma possível explicação é o facto de que as imagens obtidas na banda NIR não sejam tão afetadas pela variação de iluminação (devido à variação da pose), não provocando assim tantas oclusões como na banda do Visível, como é exemplo as imagens apresentadas na Figura 27.

Tabela 8 - Resultados (em %) obtidos na tarefa de reconhecimento facial com e sem o modelo pré-treinado de normalização da face FNM [24] *na base de dados TUFTS* [8] *com variação da pose.*

	***Rank*-1**		***Rank*-5**		**TAV @TAF=0,001**	
	s/ FNM	**c/ FNM**	**s/ FNM**	**c/ FNM**	**s/ FNM**	**c/ FNM**
VIS	80,3	**96,2**	91,0	**99,5**	60,8	**87,2**
NIR	98,3	**99,0**	99,5	**99,8**	90,4	**91,9**
LWIR	**41,8**	34,9	**58,2**	57,8	**28,7**	14,0

Note-se que as imagens do NIR são semelhantes a imagens obtidas no Visível em escala de cinzento, razão esta pela qual a Light CNN-29 consegue extrair corretamente as caraterísticas das imagens faciais obtidas na banda espectral do NIR. Também pela Tabela 8, observa-se que os resultados melhoram com a utilização do modelo FNM nas bandas espectrais do Visível e do NIR, com aumentos de desempenho no *Rank*-1 de 15,9% e 0,7% respetivamente. Nas restantes métricas, também se observa sempre melhores valores com a utilização do modelo FNM, o que demonstra que a aparente alteração de identidade nos testes qualitativos não tem um impacto negativo. Já os resultados na banda espectral do LWIR indicam que a utilização do modelo FNM não melhora o desempenho em nenhuma das métricas, obtendo-se resultados superiores sem a utilização desta, nomeadamente em 6,9 e 14,3 pontos percentuais nas métricas *Rank*-1 e TAV@TAF=0,001 respetivamente.

Devido à capacidade de normalização de expressão facial pelo modelo FNM, a TUFTS-Expressão foi utilizada para verificar se a normalização da expressão permitia à Light CNN-29 extrair caraterísticas faciais mais representativas.

Tabela 9 - Resultados (em %) obtidos na tarefa de reconhecimento facial com e sem o modelo pré-treinado de normalização da face FNM [24] *na base de dados TUFTS* [8] *com variação da expressão.*

	***Rank*-1**		***Rank*-5**		**TAV @TAF=0,001**	
	s/ FNM	**c/ FNM**	**s/ FNM**	**c/ FNM**	**s/ FNM**	**c/ FNM**
VIS	**99,6**	93,3	**100,0**	98,5	**99,4**	82,9
LWIR	**67,5**	42,7	**83,3**	48,2	**57,0**	23,9

Os resultados apresentados na Tabela 9 demonstram que os conjuntos de caraterísticas extraídos pela Light CNN-29 com as imagens originais, ou seja, sem haver a normalização de expressão facial, já são suficientemente representativos para permitir obter 99,6% e 67,5% para o *Rank-1* no Visível e no LWIR, respetivamente, e um TAV@TAF=0,001 de 99,4% e de 57 % na banda do Visível e do LWIR, respetivamente. A utilização do modelo FNM vem prejudicar a extração de caraterísticas e consequentemente os resultados, com especial incidência na banda espectral LWIR, onde o modelo FNM tem mais dificuldades em gerar imagens realistas.

Tabela 10 - Resultados (em %) obtidos na tarefa de reconhecimento facial com e sem o modelo pré-treinado de normalização da face FNM [24] *na base de dados CASIA NIR-VIS 2.0* [44].

	***Rank*-1**		***Rank*-5**		**TAV @TAF=0,001**	
	s/ FNM	**c/ FNM**	**s/ FNM**	**c/ FNM**	**s/ FNM**	**c/ FNM**
VIS	**99,9**	98,7	**100,0**	99,6	**100,0**	99,0
NIR	**99,3**	95,5	**99,8**	97,8	**98,7**	92,9

Ao utilizar a base de dados CASIA NIR-VIS 2.0 obtiveram-se os resultados apresentados na Tabela 10. Nesta tabela observamos que sem a utilização do modelo FNM consegue-se obter resultados perto dos 100% em todas as métricas. Quando utilizado o modelo, os resultados pioram, com especial ênfase na banda do NIR, onde baixam de 99,3% para 95,5% em *Rank-1* e de 98,7% para 92,9% em TAV@TAF=0,001.

Estes resultados podem ser explicados pelo facto de a base de dados CASIA NIR-VIS 2.0 não ter presente nas suas imagens do Visível as caraterísticas inerentes ao ambiente não-controlado, isto é, a face está frontal e com boa iluminação. Nas imagens NIR a face está também numa posição frontal, e a pouca luminosidade nestas imagens não provoca oclusões, devido à capacidade de se conseguir captar as caraterísticas do rosto neste ambiente menos luminoso com a banda espectral NIR. Estas caraterísticas fazem com que não exista qualquer vantagem em usar um modelo de normalização de face.

Tabela 11 - Resultados (em %) obtidos de Rank-1 na tarefa de reconhecimento facial com e sem o modelo pré-treinado de normalização da face FNM [24] *na base de dados TUFTS* [8] *com quantificação da variação da pose.*

		Variação da Pose(º)			
		± 60	**± 45**	**± 30**	**± 15**
VIS	**s/ FNM**	43,3	77,5	**100,0**	**100,0**
	c/ FNM	**87,4**	**97,7**	99,5	**100,0**
NIR	**s/ FNM**	93,4	**99,7**	**100,0**	**100,0**
	c/ FNM	**96,5**	99,4	**100,0**	**100,0**

Na Tabela 11 apresentam-se os resultados obtidos para *Rank-1* com a variação da pose quantificada. Os valores obtidos da banda do Visível mostram uma melhoria significativa na métrica

Rank-1 com a utilização do modelo FNM, resultando num aumento de 77,5% para 97,7% com variações de pose de 45º e de 43,3% para 87,4% com variações de pose de 60º. Na banda espectral do NIR, existe apenas uma melhoria quando a variação da pose é de 60º, passando de 93,4% para 96,5%. Os resultados obtidos comprovam a capacidade do modelo FNM relativamente à normalização da pose, onde quanto maior for a variação da pose, maior o benefício de a utilizar.

5.3.2.2. Normalização de Banda Espectral

Como é possível observar na avaliação qualitativa, o modelo FNM converte todas as imagens para o espetro do visível, em escala RGB (ver Figura 39). Devido a este facto, resolveu-se verificar a sua capacidade de auxiliar no reconhecimento facial heterogéneo NIR-visível, isto é, com imagens de prova na banda do NIR e as imagens na galeria na banda do visível.

Tabela 12 - Resultados (em %) obtidos na tarefa de reconhecimento facial com e sem o modelo pré-treinado de normalização da face FNM [24] *no desafio de reconhecimento facial heterogéneo NIR-visível, nas bases de dados TUFTS* [8] *com variação da pose e CASIA NIR-VIS 2.0* [44].

	***Rank*-1**		***Rank*-5**		**TAV @TAF=0,001**	
Base de Dados	**s/ FNM**	**c/ FNM**	**s/ FNM**	**c/ FNM**	**s/ FNM**	**c/ FNM**
TUFTS-Pose	95,4	**95,7**	98,9	**99,1**	85,4	**85,6**
CASIA NIR-VIS 2.0	**97,9**	81,4	**99,8**	95,28	**97,3**	80,8

Pelos resultados obtidos na Tabela 12, verificamos que a utilização do modelo FNM apenas apresenta melhorias a um sistema de reconhecimento facial heterogéneo NIR-visível quando existem poses desafiadoras, como na base de dados TUFTS-Pose. Isto deve-se principalmente à sua capacidade de normalizar uma face com variações na pose, e não o facto de converter corretamente uma imagem obtida na banda espectral NIR para o Visível.

No entanto, é interessante notar que a Light CNN-29 demonstra bons resultados nesta tarefa mesmo sem o módulo de síntese de imagem. Uma explicação para este facto é que a Light CNN-29, mesmo utilizando imagens do visível, converte-as de RGB para escala de cinza, resultando assim em imagens similares às que são obtidas na banda do NIR.

É de salientar também que, sem a utilização do módulo de síntese de imagem, os resultado NIR-visível são melhores que os visível-visível no desafio de variação da pose. Isto fortalece o argumento de que as imagens da banda NIR são mais robustas à variação de pose e iluminação, quando comparadas com as da banda do visível.

Tendo em consideração todos os resultados apresentados, nos restantes testes o modelo pré-treinado de normalização da face FNM é apenas aplicado às imagens faciais nas bandas espectrais do visível e do NIR na base de dados TUFTS-Pose.

5.4. Reconhecimento Facial

Na presente secção são apresentados os resultados obtidos pelos diferentes testes efetuados com respeito à rede de extração de caraterísticas e ao classificador. Inicialmente, foram efetuados

ajustamentos finos à rede de extração de caraterísticas, de forma que esta consiga extrair caraterísticas mais relevantes das imagens de banda espectral diferente do visível.

De seguida, foram realizados testes ao classificador para definir qual a melhor função para calcular a semelhança entre dois vetores de caraterísticas extraídos pela Light CNN-29, da qual resulta uma pontuação para cada identidade em cada banda espectral. As funções de semelhança testadas foram a distância euclidiana e a similaridade de cosseno.

Terminado este processo, foram definidos os pesos a aplicar às pontuações obtidas por cada banda espectral, procedendo assim à fusão de pontuações e posterior identificação da identidade, sendo esta a última etapa do sistema de reconhecimento facial multiespectral.

5.4.1. Treino e Avaliação da Rede de Extração de Caraterísticas

Nesta subsecção são apresentados os parâmetros utilizados para o treino da Light CNN-29, que foi responsável pela extração de caraterísticas, bem como os resultados obtidos em cada banda espectral. Nesta fase foi também aferida a melhor função de semelhança para calcular as pontuações. A Light CNN-29 foi implementada e treinada na linguagem *Python,* com a utilização da biblioteca *Pytorch*.

Para a fase de treino, e tendo em atenção os resultados apresentados na Secção 5.3., decidiu-se fazer apenas um ajustamento fino à rede de extração de caraterísticas da banda do LWIR. De forma a treinar a Light CNN-29 com identidades (pessoas) diferentes das de teste, utilizou-se as imagens da banda espectral do LWIR da base de dados IRIS. Foram utilizadas 16 imagens por lote, valor obtido por ser o maior valor com potência de base 2 e expoente inteiro que evitasse que a GPU ficasse sem memória durante a fase de treino.

Foram testados dois algoritmos de otimização, sendo estes o gradiente estocástico descendente (SGD, do inglês *stochastic gradient descent*) com momento (do inglês *momentum*) normal (SGD tradicional) e com momento de Nesterov (SGD com Nesterov), este último introduzido por Sutskever *et al.* [74] . A utilização do momento permite alcançar o mínimo mais depressa, ao mesmo tempo que tem em atenção à direção do gradiente. No momento de Nesterov, caso o termo do momento aponte na direção errada ou ultrapasse o mínimo, o gradiente ainda pode voltar para trás e corrigi-lo na mesma iteração. Para tal, utilizou-se os valores de 10^{-4} para a taxa de aprendizagem, devido a ser ajustamento fino, e para o decaimento do peso e 0,9 para o momento, valores sugeridos por Wu *et al.* [69].

Optou-se por fazer o ajustamento fino do modelo por 10 épocas para não haver um sobre ajuste do modelo ao conjunto de dados de treino, valor este sugerido também por Wu *et al.* [69]. Como função de custo, foi utilizada a entropia cruzada (do inglês *Cross-Entropy*) por ter como caraterística penalizar não só com base nas predições, mas também na confiança. Desta forma, penaliza de forma mais severa quando existe uma predição errada com bastante confiança, não deixando ao mesmo tempo de penalizar uma predição correta com pouca confiança.

Para a fase de treino foi necessário criar uma última camada ligada na Light CNN-29, responsável por proceder à classificação, com 30 dimensões, cada uma correspondendo a uma pessoa

da base de dados IRIS. Esta última camada, que converte o conjunto de caraterísticas de 256 dimensões obtido à saída da primeira camada totalmente ligada, emprega a função de ativação linear.

Com o objetivo de diminuir o sobre ajustamento do modelo e dar mais robustez à Light CNN-29 para previsões fora da amostra de dados de treino, foi também utilizado o *dropout* antes da última camada ligada, como proposto por Sristava *et al.* [75]. Esta função da biblioteca *Pytorch* elimina temporariamente, com probabilidade *p* (sendo *p*=50% no presente trabalho, como proposto por Wu *et al.* [69]), neurónios e respetivas ligações, a cada inserção de um novo vetor de dados na Light CNN-29 durante o processo de aprendizagem. Os valores utilizados em cada parâmetro podem ser consultados na Tabela 13.

Tabela 13 - Resumo dos valores utilizados em cada parâmetro para o treino da Light CNN-29 responsável pela extração de caraterísticas da banda LWIR.

Parâmetro	**Valor**
Tamanho do Lote	16
Algoritmo de Optimização	SGD (sem/com Nesterov)
Taxa de Aprendizagem	10-4
Momento	0,9
Número de Épocas	10
Função de Custo	Entropia Cruzada

Os resultados obtidos no processo de treino encontram-se indicados na Figura 40, com o custo e o *Rank-1* por época.

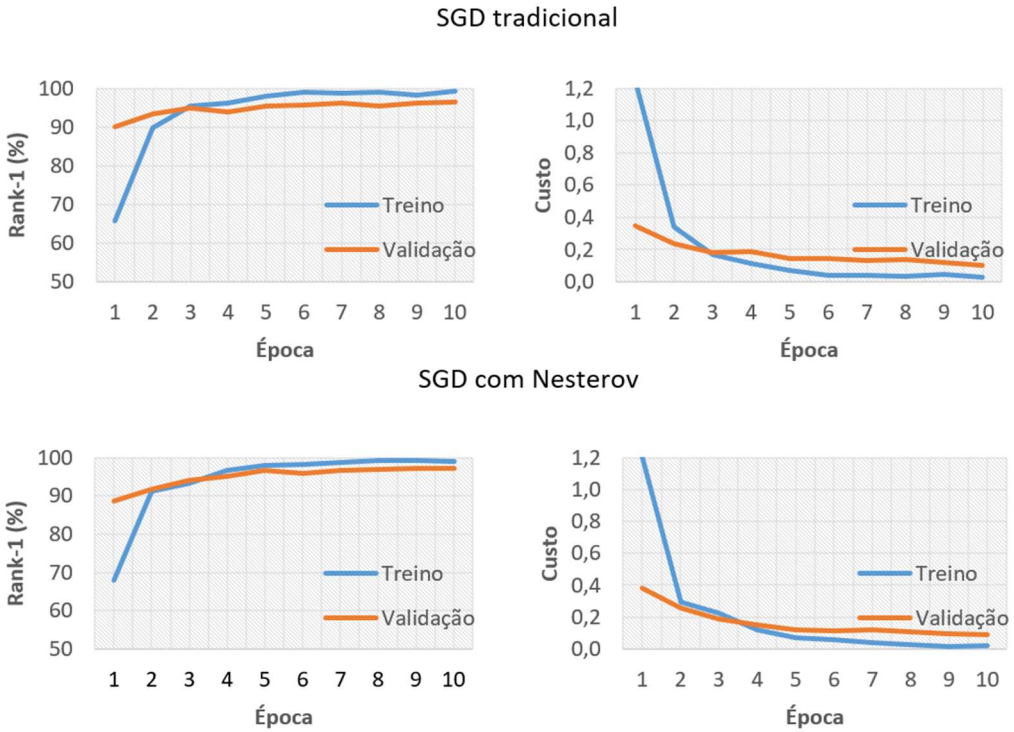

Figura 40 - Processo de ajuste fino da Light CNN-29 para a banda espectral LWIR da base de dados IRIS.

Na Figura 40 observa-se que na fase inicial o custo do treino é superior ao de validação, assim como o *Rank-1* do treino é inferior ao de validação. Este efeito é explicado por duas razões. A primeira razão é que os resultados do treino são medidos durante cada época, enquanto o custo e *Rank-1* da validação é medido após cada época. Isto leva a que, em média, os resultados do treino sejam obtidos meia época antes dos resultados da validação. A segunda razão é o facto de que a última camada totalmente ligada ainda não tinha recebido nenhum treino. Isto provoca, em especial na primeira época, a discrepância nos resultados obtidos pelo treino e pela validação.

Posto isto, e analisando com mais detalhe as curvas de custo e *Rank-1*, verifica-se que os resultados obtidos são similares, utilizando o momento de Nesterov ou não. No entanto, é possível observar que quando é utilizado o momento de Nesterov, as linhas de custo e *Rank-1* sofrem menos flutuações, devido às caraterísticas do momento de Nesterov que permitem alterar o valor de gradiente na própria iteração. No entanto, este não demonstra melhorias significativas, pois o número de épocas utilizado é suficiente para a correção destas flutuações.

Terminada a fase de treino, a última camada totalmente ligada da Light CNN-29, responsável pela classificação, deixa de ser utilizada. O objetivo do treino consiste em que a Light CNN-29 aprenda a extrair caraterísticas representativas das imagens faciais, e não apenas a classifica-las. Desta forma, a Light CNN-29 pode ser aplicada em outras bases de dados para extrair caraterísticas das imagens faciais, para servir como entrada para as funções de semelhança.

Todos os processos seguintes fazem uso do conjunto de caraterísticas de 256 dimensões obtido pela primeira camada totalmente ligada. Na Tabela 14 apresentam-se os resultados obtidos pelo modelo original e pelos modelos treinados na banda espectral do LWIR, utilizando como função de semelhança a similaridade de cosseno.

Tabela 14 – Resultados (em %) obtidos de Rank-1 na tarefa de reconhecimento facial pelos diferentes modelos para extração de caraterísticas da banda LWIR.

	Original	**SGD tradicional**	**SGD com Nesterov**
TUFTS-Pose	41,8	**55,5**	54,3
TUFTS-Expressão	67,5	**79,6**	75,9

Com os resultados obtidos observa-se que o ajustamento fino, apesar de ter sido realizado numa base de dados diferentes, permitiu à rede aprender a extrair caraterísticas mais representativas de imagens faciais da banda espectral LWIR. É também percetível que o modelo que obteve melhores resultados foi o SGD tradicional, sendo este o escolhido para os processos seguintes.

5.4.2. Funções de Semelhança e Fusão de Pontuação

Nesta fase existem então três Light CNN-29, cada uma delas responsável pela extração de caraterísticas de uma determinada banda. Apenas a Light CNN-29 responsável pela extração de caraterísticas da banda espectral do LWIR sofreu um ajustamento fino. Para se proceder à classificação, foi necessário encontrar a função de semelhança que mais se adequa à tarefa de

reconhecimento facial. Na Tabela 15 estão presentes os resultados obtidos com as duas funções de semelhança abordadas na Secção 4.4.

Tabela 15 - Resultados (em %) obtidos de Rank-1 na tarefa de reconhecimento facial com as funções de semelhança similaridade de cosseno (SCos) e Distância Euclidiana (DEuc), nas bases de dados TUFTS [8] *com variação da pose e da expressão e CASIA NIR-VIS 2.0* [44].

	TUFTS-Pose		**TUFTS-Expressão**		**CASIA NIR-VIS 2.0**	
	SCos	**DEuc**	**SCos**	**DEuc**	**SCos**	**DEuc**
VIS	**96,2**	95,3	**99,6**	99,4	**99,9**	99,8
NIR	**99,0**	96,6	-	-	**99,3**	99,1
LWIR	**55,5**	42,0	**79,6**	69,6	-	-

É possível observar pelos resultados obtidos que a função de semelhança *similaridade de cosseno* é aquela que obtém melhores resultados. Estes resultados estão de acordo com [76] e [77], que também afirmam que a utilização da métrica *similaridade de cosseno* é a mais indicada para a tarefa de reconhecimento facial. Uma explicação, dada por Liu *et al.* [77], é que esta vantagem advém da função de *similaridade de cosseno* estar relacionada com a regra de decisão de Bayes, relembrando que o classificador de Bayes é ótimo para minimizar o erro de classificação. Desta forma, decidiu-se utilizar a função de semelhança *similaridade de cosseno* para obter as diferentes pontuações em cada banda espectral.

É agora possível fazer uso das pontuações obtidas por cada banda espectral, para proceder à classificação final. Esta classificação, ao contrário das apresentadas anteriormente, é obtida utilizando as pontuações obtidas nas três bandas espectrais, e não apenas numa delas. Para tal, realiza-se uma fusão das pontuações obtidas, utilizando a equação ((10)). Foram realizados 2 estudos, onde os valores utilizados em cada um dos estudos para os pesos de cada banda constam na Tabela 16.

Tabela 16 - Valores de Wb a utilizar para cada banda espectral nos diferentes estudos.

	Estudo 1	**Estudo 2**
VIS	1,0	1,0
NIR	1,0	1,0
LWIR	1,0	0,7

No estudo 1 não são tidos em conta os resultados previamente obtidos do teste, dando assim o mesmo peso a todas as bandas espectrais. Desta forma, a pontuação final é obtida como uma média aritmética simples, assumindo que todas as bandas espectrais possuem a mesma capacidade de classificação.

Os valores de *Wb* do estudo 2 advêm da precisão média *Rank-1* de cada uma das bandas espectrais nos testes realizados nas bases de dados TUFTS-Pose, TUFTS-Expressão e CASIA NIR-VIS 2.0 (resultados obtidos com a função de similaridade de cosseno na Tabela 15), arredondados

às décimas. Assim, a pontuação final é obtida como uma média aritmética ponderada, onde cada banda apresenta diferentes pesos devido aos resultados obtidos.

Para esta fase foi necessário fornecer imagens faciais em diferentes bandas espectrais para cada classificação. Estando as bases de dados subdivididas em diretorias correspondentes a cada banda espectral, para cada classificação utilizou-se as imagens faciais com o mesmo nome de ficheiro, variando apenas na diretoria correspondente à banda espectral. Os nomes de ficheiro das imagens faciais do espetro do visível foram utilizados como suporte. Isto leva a que os resultados obtidos nesta fase possam variar relativamente aos apresentados anteriormente, quando analisado cada banda espectral em separado.

Nas Tabelas 17, 18 e 19 são apresentados os resultados derivados das pontuações obtidas por cada banda espectral e pela pontuação obtida pela fusão de pontuações das diferentes bandas espectrais.

Tabela 17 – Resultados (em %) obtidos na tarefa de reconhecimento facial, na base de dados TUFTS [8] *com variação da pose.*

	Rank					**TAV @TAF=0,001**
	1	**2**	**3**	**4**	**5**	
Estudo 1	99,4	**99,8**	99,9	**100,0**	**100,0**	90,5
Estudo 2	**99,5**	**99,8**	**100,0**	**100,0**	**100,0**	**93,5**
VIS	96,2	98,7	99,1	99,4	99,5	87,4
NIR	99,0	99,7	99,7	99,8	99,8	93,1
LWIR	55,6	62,2	66,7	69,9	72,6	30,5

Na Tabela 17 apresentam-se os resultados obtidos com a base de dados TUFTS-Pose. Nestes resultados observa-se que o estudo 2 obtém resultados superiores ao estudo 1, nas métricas de *Rank*-1 e *Rank*-3 por 0,1 pontos percentuais, e na métrica TAV@TAF=0,001 por 3 pontos percentuais. A superioridade dos resultados obtidos pelo estudo 2 face ao estudo 1 vem mostrar que o peso atribuído à banda espectral LWIR deve ser inferior ao peso atribuído às restantes, pois as caraterísticas obtidas na banda espectral LWIR são as menos representativas da identidade.

Analisando os resultados das diferentes bandas espectrais em separado, verificamos que a que obtinha os melhores resultados era a banda espectral do NIR, devido à sua robustez face à variação da iluminação, presente na base de dados TUFTS-Pose. Apesar desses bons resultados da banda NIR quando utilizada a solo, o estudo 2 conseguiu obter resultados superiores em todas as métricas, com especial ênfase no *Rank*-1 (de 99,0% para 99,5%) e na TAV@TAF=0,001 (de 93,1% para 93,5%). É de salientar que apenas os resultados obtidos com fusão de pontuações atingem a taxa de 100% de precisão nos *Rank*s avaliados (*Rank*-4 para o estudo 1 e *Rank*-3 para o estudo 2).

Tabela 18 - Resultados (em %) obtidos na tarefa de reconhecimento facial, na base de dados TUFTS [8] *com variação da expressão.*

	Rank					**TAV @TAF=0,001**
	1	**2**	**3**	**4**	**5**	
Estudo 1	**99,6**	**100,0**	**100,0**	**100,0**	**100,0**	98,7
Estudo 2	**99,6**	**100,0**	**100,0**	**100,0**	**100,0**	99,3
VIS	**99,6**	99,6	99,8	**100,0**	**100,0**	**99,4**
LWIR	79,6	86,3	88,5	90,4	91,6	54,9

Na Tabela 18 estão apresentados os resultados obtidos com a base de dados TUFTS-Expressão. Uma análise da tabela permite observar que os resultados de reconhecimento facial obtidos são melhores com a fusão de pontuações, onde ambos os estudos obtiveram o mesmo resultado que a banda espectral do Visível no *Rank-1* (99,6%), mas conseguem atingir um resultado superior no *Rank-2* (100% contra 99,6% da banda espectral do Visível). No entanto, o melhor resultado para o TAV@TAF=0,001 é obtido com a utilização de apenas a banda espectral do Visível, com 99,4%, enquanto o segundo melhor resultado foi obtido no estudo 2, com 99,3%.

Tabela 19 - Resultados (em %) obtidos na tarefa de reconhecimento facial, na base de dados CASIA NIR-VIS 2.0 [44].

	Rank					**TAV @TAF=0,001**
	1	**2**	**3**	**4**	**5**	
Estudo 1	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
VIS	99,9	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**
NIR	99,6	99,7	99,9	99,9	99,9	99,1

Na base de dados CASIA NIR-VIS 2.0 obteve-se os resultados apresentados na Tabela 19. Nesta, observa-se que o estudo 1 obteve um valor de 100% no *Rank-1,* enquanto ao utilizar as bandas espectrais do Visível e do NIR em separado, estas obtiveram 99,9% e 99,6% respetivamente, na mesma métrica. É de notar que não se realizou o estudo 2 para a base de dados CASIA NIR-VIS 2.0, pois a diferença entre o estudo 1 e o estudo 2 é o peso atribuído à banda espectral do LWIR, o qual esta não possui. Na métrica TAV@TAF=0,001, o estudo 1 iguala o resultado da banda espectral do Visível, com 100%.

Analisando todos os valores obtidos, concluiu-se que esta fusão de pontuações favorece principalmente os casos em que os resultados obtidos pelas diferentes bandas espectrais em separado eram menos satisfatórios. Olhando para os resultados das Tabelas 18 e 19, obtidos com as bases de dados TUFTS-Expressão e CASIA-NIR-VIS 2.0 respetivamente, observa-se que a banda espectral do Visível já obtém resultados bastante positivos em todas as métricas. Tal facto leva a que a fusão de pontuações não produza tantas melhorias. No entanto, apesar de se verificar uma descida dos

resultados no TAV@TAF=0,001 na Tabela 18, os resultados obtidos pela fusão de pontuações na sua generalidade foram superiores aos obtidos pelas bandas espectrais em separado. Os valores obtidos comprovam assim o benefício da utilização de imagens multiespectrais num sistema de reconhecimento facial.

6. Conclusão

Nesta dissertação foi proposto um sistema de reconhecimento facial multiespectral em ambiente não controlado, o qual tem como objetivo tomar uma decisão com o maior número de dados disponível, ou seja, utilizando as imagens faciais obtidas pelas diferentes bandas espectrais. O sistema é composto por três módulos: (i) deteção e alinhamento da face, (ii) síntese de imagem e (iii) reconhecimento facial.

Neste trabalho, várias técnicas foram implementadas, de forma a validar estas em diferentes bandas multiespectrais, pois todas estas foram treinadas em bases de dados do Visível, bem como analisar a influência ao nível das caraterísticas da imagem facial (pose, iluminação e expressão). Esta análise teve como objetivo selecionar a técnica mais adequada para cada módulo do sistema de reconhecimento facial proposto. Para a tarefa de deteção facial, três redes foram avaliadas qualitativamente e quantitativamente, o que permitiu concluir que a rede DSFD era a mais adequada, pois mantinha uma elevada precisão nas diferentes bandas espetrais. Para a tarefa de marcação facial, foram avaliadas qualitativamente três redes. Esta avaliação permitiu concluir que a rede 2D-FAN era aquela que melhor se adequava, devido à capacidade de identificar corretamente os marcos faciais tanto nas diferentes bandas espetrais, como para maior diversidade de poses faciais.

Para o módulo de síntese de imagem, foram analisados os modelos FFWM e FNM. Numa primeira fase, resultados qualitativos permitiram verificar que nenhum dos modelos produzia resultados fidedignos na banda espetral do LWIR (a identidade ficava irreconhecível). No entanto, o modelo FNM era o que produzia imagens faciais mais realistas para as bandas espetrais do Visível e do NIR, mantendo as proporções da face e as caraterísticas faciais mais relevantes. Uma posterior análise ao modelo FNM permitiu verificar a influência da variação da pose e da expressão facial no seu funcionamento, identificando assim as situações em que mais se adequa a sua utilização: um afastamento da pose frontal superior a 30º nas bandas espetrais do NIR e do Visível. Para uma variação de 45º na pose facial, o modelo FNM permite melhorar o resultado de *Rank-1* de 77,5% para 97,7% no Visível, ainda que obtenha um decréscimo de 0,3 pontos percentuais no NIR. Já para variações de 60º, a melhoria em *Rank-1* passa de 43,3% para 87,4% no Visível e de 93,4% para 96,5% no NIR. Estes resultados permitem concluir que: (i) quanto maior a variação da pose, maior a vantagem em utilizar o modelo FNM e (ii), as imagens do NIR permitem obter uma melhor classificação que as imagens do Visível, pois variação da pose pode acarretar a variações na iluminação, às quais a banda do NIR é mais resistente.

A extração dos conjuntos de caraterísticas das imagens faciais das diferentes bandas espectrais é realizada através da Light CNN-29 [69], sendo feita um ajustamento fino aos pesos da rede para a banda espectral do LWIR, visto esta ter sido treinada na banda espectral do Visível. Para a fase de classificação, é realizada a identificação nas diferentes bandas espectrais, cada uma produzindo diferentes pontuações para cada identidade. Uma fusão de pontuações é empregue de seguida, de forma a ser tomada uma decisão final na identificação ou verificação tendo em conta as pontuações obtidas nas diferentes bandas espectrais. Neste trabalho foram realizados dois estudos diferentes para a fusão de pontuações, onde no primeiro, todas as bandas espectrais têm o mesmo

peso, e no segundo, é dado como peso a cada banda espectral os resultados de *Rank-1* previamente obtidos nessa mesma banda. Os diferentes estudos permitiram identificar que: (i) a simples utilização das diferentes bandas espectrais para realizar a identificação/verificação é vantajosa (estudo 1) e (ii) uma média ponderada é benéfica quando os diferentes classificadores têm diferentes níveis de confiabilidade (estudo 2).

Na base de dados multiespectral TUFTS, com variação da pose e variação da expressão, os resultados obtidos em *Rank-1* pelo sistema proposto e com fusão de pontuações com média ponderada (estudo 2) foram de 99,5% e 99,6%, sendo os melhores resultados obtidos utilizando apenas uma banda espectral de 99,0% e 99,6%. Na métrica de TAV@TAF=0,001, os resultados obtidos pela média ponderada são de 93,5% e 99,3%, enquanto com apenas uma banda espectral se obteve 93,1% e 99,4%. Na base de dados CASIA NIR-VIS 2.0, a fusão de pontuações alcançou os resultados de 100,0% nas métricas *Rank-1* e TAV@TAF=0,001, onde sem a fusão de pontuação se obtém como melhor resultado 99,9% e 100,0% em *Rank-1* e TAV@TAF=0,001, respetivamente.

Como contribuições ao estado da arte, destaca-se a análise de diversas técnicas para diferentes tarefas. Esta análise permitiu: (i) apresentar um módulo de deteção e alinhamento facial eficiente para ser utilizado por qualquer sistema de análise facial multiespectral, (ii) identificar as situações em que o modelo FNM deve ser utilizado para normalizar imagens faciais e (iii) a seleção de uma função de semelhança e dos pesos a usar na fusão de pontuações, a fim de identificar identidades. Dos resultados experimentais, também se conclui que o sistema proposto permite obter resultados bastante elevados no reconhecimento facial multiespectral em ambiente não controlado, onde a utilização das pontuações obtidas em diferentes bandas espectrais permite, de uma forma geral, alcançar resultados superiores à utilização de apenas as pontuações obtidas por uma banda espectral.

6.1. Trabalho Futuro

Após o término de um projeto, existem sempre aspetos que podem ser explorados, pelo que existe um vasto leque de hipóteses de trabalho futuro. Assim, são destacadas algumas hipóteses que os autores consideram pertinentes:

Criação de uma base de dados multiespectral na Academia Militar. Através da dissertação de mestrado é observável as lacunas das bases de dados multiespectrais. A criação de uma base de dados multiespectral na Academia Militar, com a utilização das câmaras multiespectrais que esta dispõe (Visível, NIR, SWIR e LWIR), permitiria aumentar os dados disponíveis para o treino de diferentes algoritmos a operarem em bandas espectrais diferentes do visível. Além de ser benéfico para toda a comunidade científica, permitiria à Academia Militar fazer uso do sistema de reconhecimento facial proposto nesta dissertação para aumentar o controlo de segurança na unidade. O mesmo poderia ser estendido para as restantes unidades das forças armadas e de segurança.

Avaliar a viabilidade da implementação do presente sistema de reconhecimento facial na Guarda Nacional Republicana, a qual já possui drones com câmaras no espetro do visível e térmicas. Para este fim, além dos diferentes aspetos associados à parte técnica, dos quais se destacam a

capacidade de o algoritmo operar em tempo real, seria necessário fazer uma revisão das leis em vigor, de forma a entender em que casos seria possível a sua utilização.

Referências

[1] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, “Deep face recognition: A survey,” in *31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, 2018, pp. 471–478.

[2] B. F. Klare *et al.*, “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[3] R. Munir and R. A. Khan, “An extensive review on spectral imaging in biometric systems: Challenges & advancements,” *J. Vis. Commun. Image Represent.*, vol. 65, p. 102660, 2019.

[4] U. Tirosh, “Portrait Lighting Cheat Sheet Card - DIYPhotography,” 2008. https://www.diyphotography.net/portrait-lighting-cheat-sheet-card/ (accessed Oct. 27, 2021).

[5] D. D. Gundlach, “How thermal technologies improve facility security and workforce safety,” 2020. https://www.theengineer.co.uk/supplier-network/product/how-thermal-technologies-improve-facility-security-and-workforce-safety/ (accessed Oct. 27, 2021).

[6] G. G. Ung, “Hardcore Hardware: The Predator has nothing on the Zeus Pro 640 thermal imager,” 2015. https://www.pcworld.com/article/428311/hardcore-hardware-the-predator-has-nothing-on-the-zeus-pro-640-thermal-imager.html (accessed Oct. 27, 2021).

[7] W. Zhang, X. Zhao, J.-M. Morvan, and L. Chen, “Improving shadow suppression for illumination robust face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 611–624, 2018.

[8] K. Panetta *et al.*, “A Comprehensive Database for Benchmarking Imaging Systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, 2020.

[9] W. W. Bledsoe, “The model method in facial recognition,” *Panor. Res. Inc., Palo Alto, CA, Rep. PR1*, vol. 15, no. 47, p. 2, 1966.

[10] W. Bledsoe, “Man-machine facial recognition: Report on a large-scale experiment, panoramic research,” *Inc, Palo Alto, CA*, vol. 2, 1966.

[11] Y. Sun, X. Wang, and X. Tang, “Deep learning face representation from predicting 10,000 classes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[12] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[13] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments,” *Univ. Massachusetts, Amherst, Tech. Rep. 07-49*, 2008.

[14] I. Goodfellow *et al.*, “Generative adversarial nets,” *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 2672–2680, 2014.

[15] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *CoRR*, vol. 1511.06434, 2016.

[16] L. Tran, X. Yin, and X. Liu, “Disentangled representation learning GAN for pose-invariant face recognition,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern*

Recognition, 2017, pp. 1283–1292.

[17] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, “3d-aided dual-agent gans for unconstrained face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, 2018.

[18] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4295–4304.

[19] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3D solution,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 146–155.

[20] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, “Synthesizing normalized faces from facial identity features,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3386–3395.

[21] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, “Regressing robust and discriminative 3D morphable models with a very deep neural network,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1493–1502.

[22] R. Huang, S. Zhang, T. Li, and R. He, “Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2458–2467.

[23] J. Zhao *et al.*, “Towards Pose Invariant Face Recognition in the Wild,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2207–2216.

[24] Y. Qian, W. Deng, and J. Hu, “Unsupervised face normalization with extreme pose and expression in the wild,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9843–9850.

[25] Z. Zhang *et al.*, “Semi-Supervised Face Frontalization in the Wild,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 909–922, 2020.

[26] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, “Towards High Fidelity Face Frontalization in the Wild,” *Int. J. Comput. Vis.*, pp. 1–20, 2019.

[27] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, “UV-GAN: Adversarial Facial UV Map Completion for Pose-Invariant Face Recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7093–7102.

[28] C. Peng, X. Gao, N. Wang, and J. Li, “Graphical representation for heterogeneous face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, 2016.

[29] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, “Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks,” *Int. J. Comput. Vis.*, vol. 127, no. 6–7, pp. 845–862, 2019.

[30] A. Litvin, K. Nasrollahi, S. Escalera, C. Ozcinar, T. B. Moeslund, and G. Anbarjafari, “A novel deep network architecture for reconstructing RGB facial images from thermal for face recognition,” *Multimed. Tools Appl.*, vol. 78, no. 18, pp. 25259–25271, 2019.

[31] A.-C. Guei and M. Akhloufi, “Deep learning enhancement of infrared face images using

generative adversarial networks," *Appl. Opt.*, vol. 57, no. 18, pp. D98–D107, 2018.

[32] B. Cao, N. Wang, J. Li, and X. Gao, "Data augmentation-based joint learning for heterogeneous face recognition," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 6, pp. 1731–1743, 2018.

[33] H. B. Bae, T. Jeon, Y. Lee, S. Jang, and S. Lee, "Non-Visual to Visual Translation for Cross-Domain Face Recognition," *IEEE Access*, vol. 8, pp. 50452–50464, 2020.

[34] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, 2019.

[35] A. Seal, D. Bhattacharjee, M. Nasipuri, C. Gonzalo-Martin, and E. Menasalvas, "Fusion of visible and thermal images using a directed search method for face recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 04, p. 1756005, 2017.

[36] M. Kanmani and V. Narasimhan, "Optimal fusion aided face recognition from visible and thermal face images," *Multimed. Tools Appl.*, pp. 1–25, 2020.

[37] W. Hu, H. Hu, and X. Lu, "Heterogeneous Face Recognition Based on Multiple Deep Networks with Scatter Loss and Diversity Combination," *IEEE Access*, vol. 7, pp. 75305–75317, 2019.

[38] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein cnn: Learning invariant features for nir-vis face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1761–1773, 2018.

[39] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-celeb-1M: A dataset and benchmark for large-scale face recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9907 LNCS, pp. 87–102.

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. 1411.7923, 2014.

[41] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873–4882.

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

[43] L. L. Chambino, J. S. Silva, and A. Bernardino, "Multispectral Facial Recognition: A Review," *IEEE Access*, vol. 8, pp. 207871–207883, 2020.

[44] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 348–353.

[45] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[46] S. Wang *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimed.*, vol. 12, no. 7, pp. 682–691, 2010.

[47] M. K. Bhowmik, P. Saha, A. Singha, D. Bhattacharjee, and P. Dutta, "Enhancement of robustness of face recognition system through reduced gaussianity in Log-ICA," *Expert Syst.*

Appl., vol. 116, pp. 96–107, 2019.

[48] R. B. Martin, M. Sluch, K. M. Kafka, R. Ice, and B. E. Lemoff, “Active-SWIR signatures for long-range night/day human detection and identification,” in *Active and Passive Signatures IV*, 2013, vol. 8734, p. 87340J.

[49] F. Nicolo and N. A. Schmid, “Long range cross-spectral face recognition: matching SWIR against visible light images,” *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1717–1726, 2012.

[50] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, “Face recognition by fusing thermal infrared and visible imagery,” *Image Vis. Comput.*, vol. 24, no. 7, pp. 727–742, 2006.

[51] R. Singh, M. Vatsa, and A. Noore, “Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition,” *Pattern Recognit.*, vol. 41, no. 3, pp. 880–893, 2008.

[52] S. Rajeev, K. M. Shreyas Kamath, Q. Wan, K. Panetta, and S. S. Agaian, “Illumination invariant NIR face recognition using directional visibility,” *IS T Int. Symp. Electron. Imaging Sci. Technol.*, no. 11, 2019.

[53] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[54] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, “Biometrics Recognition Using Deep Learning: A Survey,” *CoRR*, vol. 1912.00271, 2019.

[55] S. Minaee, P. Luo, Z. Lin, and K. Bowyer, “Going Deeper Into Face Detection: A Survey,” *CoRR*, vol. 2103.14983, 2021.

[56] W. Liu *et al.*, “SSD: Single shot multibox detector,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9905, pp. 21–37.

[57] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3FD: Single Shot Scale-Invariant Face Detector,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 192–201.

[58] B. Gary, “The OpenCV Library,” *Dr. Dobb’s J. Softw. Tools*, vol. 25, no. 2236121, pp. 120–123, 2008.

[59] J. Li *et al.*, “DSFD: Dual shot face detector,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5055–5064.

[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[61] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A face detection benchmark,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.

[62] Y. Wu and Q. Ji, “Facial Landmark Detection: A Literature Survey,” *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 115–142, 2019.

[63] D. E. King, “Dlib-ml: A Machine Learning Toolkit,” *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[64] A. Bulat and G. Tzimiropoulos, “How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks),” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1021–1030.

[65] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[66] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 Faces In-The-Wild Challenge: database and results,” *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.

[67] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9912 LNCS, pp. 483–499, 2016.

[68] Y. Wei, M. Liu, H. Wang, R. Zhu, G. Hu, and W. Zuo, “Learning Flow-Based Feature Warping for Face Frontalization with Illumination Inconsistent Supervision,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12357 LNCS, pp. 558–574, 2020.

[69] X. Wu, R. He, Z. Sun, and T. Tan, “A light cnn for deep face representation with noisy labels,” *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, 2018.

[70] L. Zhang, J. Liu, M. Luo, X. Chang, Q. Zheng, and A. G. Hauptmann, “Scheduled sampling for one-shot learning via matching network,” *Pattern Recognit.*, vol. 96, p. 106962, 2019.

[71] S. N. Tumpa and M. L. Gavrilova, “Score and rank level fusion algorithms for social behavioral biometrics,” *IEEE Access*, vol. 8, pp. 157663–157675, 2020.

[72] N. Srinivas, K. Veeramachaneni, and L. A. Osadciw, “Fusing correlated data from multiple classifiers for improved biometric verification,” in *12th International Conference on Information Fusion*, 2009, pp. 1504–1511.

[73] IEEE OTCBVS WS Series Bench; DOE University Research Program in Robotics under grant DOE-DE-FG02-86NE37968; DOD/TACOM/NAC/ARC Program under grant R01-1344-18; FAA/NSSA grant R01-1344-48/49; Office of Naval Research under grant #N000143010022., “Dataset 02: IRIS Thermal/Visible Face Databases,” 2005. http://vcipl-okstate.org/pbvs/bench/ (accessed Mar. 27, 2021).

[74] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” in *30th International Conference on Machine Learning*, 2013, no. PART 3, pp. 2176–2184.

[75] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[76] S. N. Borade, R. R. Deshmukh, and P. Shrishrimal, “Effect of Distance Measures on the Performance of Face Recognition Using Principal Component Analysis,” *Adv. Intell. Syst. Comput.*, vol. 384, pp. 569–577, 2016.

[77] C. Liu, “Discriminant analysis and similarity measure,” *Pattern Recognit.*, vol. 47, no. 1, pp. 359–367, 2014.