TÉCNICO LISBOA

ACADEMIA MILITAR
MILITARY ACADEMY

Reconhecimento Facial Através de Imagens Multiespectrais

Luis Carlos Lopes Chambino

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Orientadores: Professor Alexandre José Malheiro Bernardino

Professor José Silvestre Serra da Silva

Júri

Presidente: Professor João Fernando Cardoso Silva Sequeira

Orientador: Professor José Silvestre Serra da Silva

Vogais: Professor Paulo Luís Serras Lobato Correia

Tenente-Coronel TM (Eng) Henrique Martins dos Santos Cunha

Janeiro 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 dissertação de mestrado representa um desafio ultrapassado através de inúmeras horas de reflexão, estudo, investigação, redação e correções. Apesar de o trabalho ser individual, todo este esforço não teria o resultado pretendido sem o apoio direto e/ou indireto proporcionado por várias pessoas e instituições, e, como tal, este capítulo pretende mostrar o meu reconhecimento pessoal e gratidão a todos sem exceção, não por ordem de relevância, mas pelas palavras manifestadas.

Aos meus orientadores, Professor Alexandre Bernardino e Professor José Silvestre Silva agradeço por todo o tempo despendido para aconselhamento e orientação, e pela compreensão, rigor e espírito crítico demonstrados durante esta etapa. Para mim, isto é o que um orientador deve ser. Obrigado por acreditarem em mim.

Aos meus amigos e camaradas da Academia Militar, ao Tiago Fernandes, Ricardo Oliveira, Gilberto Gomes, Eduardo Osório, Rui Rita, Tiago Ramos, Eduardo Filipe, Manuel Passos, Joni Bernardo, Filipe Silva, Artur Machado e Sérgio Moreira, pela amizade, viagens e o inestimável apoio ao longo dos anos. Foxtrot todo o dia.

À minha família, em especial aos meus pais e irmão, que sempre me encorajaram e apoiaram e que me deram condições para estudar e perseguir o futuro que gostaria de ter, ser Oficial do Exército Português. A vós devo tudo aquilo que sou e que alguma vez consiga alcançar.

Mas palavra maior de gratidão vai para a Martinha. Podia ter feito este caminho sozinho, mas foi muito melhor fazê-lo acompanhado contigo... A tua alegria, motivação, companheirismo e carinho manifestados diariamente desde o dia em que te conheci fizeram, fazem, e farão a diferença em todas as etapas da minha vida.

A todos o meu profundo e sincero agradecimento!

Luis Carlos Lopes Chambino

Resumo

Esta dissertação de mestrado tem como objetivo o desenvolvimento e estudo de um sistema de reconhecimento facial multiespectral. O reconhecimento facial é um método de identificação ou autenticação da identidade de indivíduos através dos seus rostos.

Atualmente, os sistemas de reconhecimento facial que utilizam imagens multiespectrais obtêm melhores resultados, em comparação com aqueles que utilizem apenas imagens faciais da banda espectral do visível. Neste trabalho, é proposta uma arquitetura que utiliza múltiplas redes neurais convolucionais profundas e imagens multiespectrais para efetuar o reconhecimento facial. É realizado um estudo com o objetivo de avaliar o desempenho da adaptação de diversas camadas da rede neural base. Adicionalmente, foi realizado um segundo estudo para avaliar o desempenho das Máquinas Vetoriais de Suporte (SVM) e dos classificadores k-Nearest Neighbor para classificar os conjuntos de características multidimensionais obtidas através da arquitetura proposta.

Resultados experimentais nas bases de dados multiespectrais Tufts e CASIA NIR-VIS 2.0 indicam um desempenho competitivo no reconhecimento facial obtendo uma pontuação de *rank-1* de 99,7% e 99,8%, respetivamente.

É também proposto para este trabalho, um detetor de falsificação multiespectral. Este, utiliza imagens multiespectrais (na banda espectral do VIS, SWIR e LWIR) com o objetivo de identificar possíveis ataques de falsificação nas imagens faciais. Resultados experimentais comprovam a eficiência do detetor de falsificação multiespectral proposto, quando comparado com os detetores de pele YCbCr e HSV. Foram obtidas as seguintes taxas de deteção de falsificação de 13%, para YCbCr e HSV, e de 83% para o detetor de falsificação multiespectral proposto.

Palavras Chave: reconhecimento facial, imagens multiespectrais, infravermelho, detetor de falsificação.

Abstract

This master thesis aims at the development and study of a multispectral facial recognition system. Facial recognition is a method of identifying or authenticating the identity of individuals through their faces. Nowadays, facial recognition systems that use multispectral images obtain better results compared to those that use only facial images present in the visible spectral band.

In this work, an architecture that uses multiple deep convolutional neural networks and multispectral images to perform facial recognition is proposed. A study is carried out with the objective of evaluating the performance of the adaptation of several layers of the base neural network. Additionally, a second study was conducted to evaluate the performance of Vector Support Machines (SVM) and k-Nearest Neighbor classifiers to classify the sets of multidimensional characteristics obtained through the proposed architecture.

Experimental results in the Tufts and CASIA NIR-VIS 2.0 multispectral databases indicate a competitive performance in facial recognition obtaining a rank-1 score of 99.7% and 99.8%, respectively.

A multispectral presentation attack detector is also proposed for this work. This uses multispectral images (in VIS, SWIR and LWIR spectral band) in order to identify possible presentation attacks in facial images. Experimental results prove the efficiency of the proposed multispectral presentation attack detector when compared to YCbCr and HSV skin detectors. The following presentation attack detection rates of 13 % were obtained for YCbCr and HSV, and 83 % for the proposed multispectral presentation attack detector.

Keywords: facial recognition, multispectral images, infrared, presentation attack detector.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Siglas e Acrónimos

AAMA	Aquartelamento da Academia Militar Amadora
AM	Academia Militar
AUC	*Area Under the Curve*
CASIA	*China Academy of Sciences Institute of Automation*
CE	*Cross-Entropy*
CFC	*Adversarial Cross-spectral Face Completion*
CI	Corretamente Identificadas
CIA	Agência Central de Inteligência
CMC	*Cumulative Match Characteristic*
CNN	*Convolutional Neural Network*
DCNN	*Deep Convolutional Neural Network*
DSU	*Domain Specific Units*
EUA	Estados Unidos da América
FAIR	*Facebook Artificial Intelligence Research*
FAR	*False Acceptance Rate*
FIVE	Fusão das Imagens do Visível e do Infravermelho
FLOPS	*Floating Point Operations Per Second*
FUSIMIL	Fusão de Imagem Militar
GAN	*Generative Adversarial Network*
G-HFR	*Graphical Representation for Heterogeneous Face Recognition*
GPU	*Graphics Processing Unit*
HOG	*Histogram of Oriented Gradients*
HOGOM	*Histograms of Gabor Ordinal Measures*
ICA	*Independent Component Analysis*
IDICN	*Intraspectrum Discrimination and Interspectrum Correlation Analysis Deep Network*
IE	Identificações Efetuadas
kNN	*k-Nearest Neighbors*
LBP	*Local Binary Patterns*
Log-ICA	*Logarithmic-Independent Component Anlysis*
LWIR	*Long Wavelength Infrared*

MCA *Mutual Component Analysis*

MDNDC *Multiple Deep Network with Scatter Loss and Diversity Combination*

MFM *Max-Feature Map*

MWIR *Mid Wavelength Infrared*

NIR *Near Infrared*

RBF *Radial Basis Function*

ReLU *Rectified Linear Unit*

RLReLU *Randomized Leaky Rectified Linear Unit*

ROC *Receiver Operating Characteristic*

SCV *Stratified Cross-Validation*

SGR-DA *Sparse Graphical Representation based Discriminant Analysis*

SVM *Support Vector Machine*

SWIR *Short Wavelength Infrared*

UCL Última Camada Ligada

USTC-NVIE *University of Science and Technology of China-Natural Visible and Infrared facial Expression*

VIS visível

WCNN *Wasserstein Convolutional Neural Network*

W-kNN *Weighted k-Nearest Neighbors*

1 Introdução

O reconhecimento facial é uma capacidade dos seres humanos, exercida através da sua visão, fundamental para o relacionamento social. Com a constante evolução e facilidade de acesso de recursos computacionais, o interesse em incorporar as capacidades humanas em computadores tem aumentado constantemente, e conduzido a novas investigações nos fundamentos e aplicações industriais do reconhecimento facial. A necessidade crescente de uma forma confiável para reconhecer ou verificar a identidade de uma pessoa tem despertado uma pesquisa intensa no campo da biométrica [1].

1.1 Enquadramento e Motivação

Comparado com os vários tipos de traços biométricos que existem, como, por exemplo, a íris, a impressão digital, a assinatura das veias e o reconhecimento de voz, o reconhecimento facial tem a vantagem de permitir capturar com maior facilidade as características de uma pessoa. Adicionalmente, sistemas de reconhecimento faciais são um método cuja aplicação não é invasiva [1] [2].

Existem dois modos principais de aquisição de imagens em sistemas de reconhecimento facial: num ambiente controlado, onde uma pessoa coopera na aquisição de imagens, e num ambiente não controlado, vulgarmente conhecido em inglês por *in the wild*, onde uma pessoa não coopera ou não tem conhecimento que está a ser observada.

Nos dias de hoje, muitos dos sistemas biométricos de reconhecimento facial funcionam na banda espectral do visível (VIS). Os sistemas que utilizam apenas a banda espectral do VIS possuem diversos obstáculos, tais como oclusões, variação de poses, cooperação da pessoa e, o mais problemático, alterações na luminosidade. Como resultado, é necessário complementar estes sistemas de reconhecimento facial, quer com a utilização de outros sensores biométricos (e.g., impressão digital ou íris) ou outras bandas espectrais, a fim de minimizar estes problemas [3].

A utilização do espectro electromagnético infravermelho, nomeadamente as bandas espectrais *Near Infrared* (NIR), *Short Wavelength Infrared* (SWIR), *Mid Wavelength Infrared* (MWIR) e *Long Wavelength Infrared* (LWIR), têm sido utilizadas com sucesso em sistemas de reconhecimento facial, como complemento do espectro visível [1] [3]. A estes sistemas, que utilizam mais do que uma banda espectral, denominam-se de multiespectrais. A Tabela 1 indica as bandas espectrais mais utilizadas e aplicadas em reconhecimento facial.

Tabela 1: Intervalos espectrais utilizados em reconhecimento facial [4].

Nome da Banda Espectral	**Comprimento de Onda (μm)**
Visível	0,38 - 0,75
Near Infrared (NIR)	0,75 - 1,40
Short Wavelength Infrared (SWIR)	1,40 - 3,00
Mid Wavelength Infrared (MWIR)	3,00 - 8,00
Long Wavelength Infrared (LWIR)	8,00 - 15,00

A utilização do espectro infravermelho em sistemas de reconhecimento facial possui diversas vantagens quando comparada com o espectro visível. O infravermelho é imperceptível ao olho humano e, ao mesmo tempo, menos sensível às diferenças de luminosidade. Por exemplo, as câmaras nocturnas utilizadas na vigilância por vídeo utilizam *LEDs* com emissão no espectro infravermelho para iluminar o local, e realizar vigilância nocturna sem que as pessoas tenham conhecimento [5].

Dado que as bandas espectrais NIR e SWIR estão próximas da banda espectral do visível, é possível adaptar os métodos de aprendizagem automática treinados com imagens do espectro visível às bandas NIR e SWIR. As bandas espectrais MWIR e LWIR (também conhecida por térmica) permitem a utilização de sistemas de reconhecimento facial à noite, quando a luminosidade é muito baixa ou mesmo nula.

Grande parte dos sistemas atuais de reconhecimento facial estão vulneráveis a ataques de falsificação. Ao ser criado um sistema de reconhecimento facial onde não são tidos em consideração estes possíveis ataques, uma simples fotografia de uma pessoa seria o suficiente para o enganar. Sistemas que não são capazes de se defenderem destes ataques não são adequados para serem usados em situações sem supervisão, ou seja, em ambientes onde o reconhecimento facial é feito automaticamente [6] [7].

Os sistemas de reconhecimento facial multiespectral, em comparação com os sistemas de reconhecimento facial, que apenas utilizem a banda espectral do visível, podem ser utilizados como um método para adicionar uma camada de segurança extra, com o intuito de reconhecer uma pessoa com maior precisão no acesso a um local de alta segurança, a fim de garantir o acesso apenas a pessoas autorizadas. Estes locais podem ser hospitais, escolas, laboratórios e edifícios militares [3].

Através do desenvolvimento de melhores sistemas de reconhecimento facial, é possível garantir um controlo de acesso mais fiável e mais robusto, protegendo a propriedade e aumentando a segurança das pessoas.

1.2 Objetivos

Este trabalho tem como objetivo principal o desenvolvimento de um sistema de reconhecimento facial que utilize imagens multiespectrais. Este sistema é desenvolvido tendo em vista uma possível implementação em controlos de acesso a arrecadações de materiais de guerra, áreas de acesso limitado, acesso a paióis, entre outros usos no Exército Português.

1.3 Estrutura da Dissertação

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

- **Capítulo 1 - Introdução:** neste capitulo é descrita a motivação do trabalho apresentado nesta dissertação de mestrado, os objetivos e a estrutura da dissertação;
- **Capítulo 2 - Reconhecimento Facial:** neste capítulo são explanados conceitos importantes, como o funcionamento de um sistema de reconhecimento facial e a diferenciação de verificação e identificação em reconhecimento facial.
- **Capítulo 3 - Estado da Arte:** neste capítulo é feito um estudo do estado da arte sobre os métodos de reconhecimento facial multiespectral 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 módulos propostos na metodologia. Cada experiência é 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.

1.4 Contribuições Científicas

O presente trabalho resultou em três artigos científicos, o primeiro aceite para publicação na revista internacional cientifica com fator de impacto *IEEE Access*, denominada,

> L. Chambino, J. S. Silva e A. Bernardino, "Multispectral Facial Recognition: a Review," IEEE Access, vol. 8, pp. 207871-207883, 2020.

o segundo aceite para publicação na 26ª Conferência Anual Portuguesa de Reconhecimento de Padrões (RECPAD), realizada por via telemática a 30 de outubro de 2020,

> L. Chambino, J. S. Silva e A. Bernardino, "Multispectral Images Applied to Face Recognition", Portuguese Conference on Pattern Recognition (RECPAD2020), Universidade de Évora, 2020, pp. 15-16.

e o terceiro submetido para aceitação na jornadas das engenharias da Academia Militar, a realizar a 10 de fevereiro de 2021 no Aquartelamento da Academia Militar Amadora (AAMA),

> L. Chambino, J. S. Silva e A. Bernardino, "Reconhecimento Facial Através de Aprendizagem Profunda em Imagens Multiespectrais", Jornadas das Engenharias da Academia Militar, 2020, pg. 8.

2 Reconhecimento Facial

Biometria é a ciência de analisar características físicas ou comportamentais específicas a cada indivíduo com o intuito de autenticar ou identificar a sua identidade [1]. Este capítulo tem como intuito explanar de forma sucinta e clara o funcionamento do sistema de reconhecimento facial e onde este pode ser aplicado, fornecendo ao leitor uma base importante para os próximos capítulos.

Hoje em dia é possível observar um crescimento de aplicações que utilizam sistemas de reconhecimento facial, seja para uso colectivo, como em empresas, ou para uso pessoal, como nos telemóveis. Todo o sistema de reconhecimento facial tem como objetivo principal identificar ou verificar a identidade de uma pessoa através da imagem de entrada, encontrando a identidade correspondente na base de dados. A principal dificuldade é garantir que o processo seja realizado de forma expedita em tempo real, algo que não é possível de concretizar em todos os sistemas de reconhecimento facial, criando a necessidade contínua de implementar melhores sistemas de reconhecimento facial.

2.1 Sistemas Automáticos de Reconhecimento Facial

Existem diversos sistemas de reconhecimento facial, de uma forma geral, compostos pelas seguintes fases, ilustrado na Figura 1:

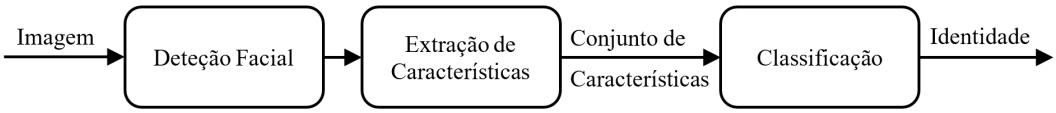


Figura 1: Esquema geral de um sistema de reconhecimento facial.

1. Entrada: esta é a primeira fase do sistema, a entrada pode ser uma ou várias imagens;

2. Deteção Facial: após a aquisição da imagem é necessário verificar a existência de pessoas na imagem, através de algoritmos especializados em detetar faces humanas;

3. Extração de Características: nesta fase, existem vários métodos, mas, de um modo geral, esses métodos extraem as características inerentes a cada pessoa e as analisam para obter um conjunto, ou vetor, de características que corresponda a uma identidade apenas. Estas características podem ser desde: sinais, distância entre olhos, pontos específicos na face humana transversais a todas as faces ou um conjunto de caraterísticas obtidos por uma rede neural profunda (este conceito irá ser abordado posteriormente);

4. Classificação: através das características obtidas da imagem na fase anterior, é efetuada a classificação da identidade da pessoa na imagem. O resultado final é uma previsão da identidade da imagem fornecida na entrada.

2.2 Aplicações

Os sistemas de reconhecimento facial podem ser aplicados em duas situações: para efetuar a autenticação da identidade fornecida ou para identificar a identidade de uma pessoa. Para os dois cenários é necessário existir um registo prévio da pessoa a identificar/autenticar. Na Figura 2 está ilustrado um possível método de registo de uma nova pessoa na base de dados. De forma a registar uma nova identidade na base de dados é necessário fornecer um número de identificação pessoal e uma (ou várias) imagem pessoal. Nas imagens fornecidas é efetuado o reconhecimento facial, não com o intuito de identificar a identidade da pessoa, mas sim com o intuito de agrupar as características extraídas. Este agrupamento de características irá representar agora a identidade da pessoa registada. Após o agrupamento é possível autenticar ou identificar a nova identidade registada.

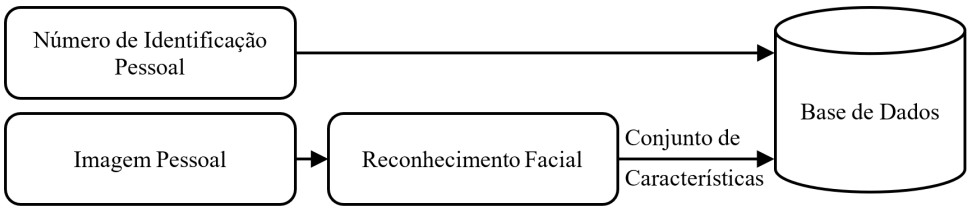


Figura 2: Situação de registo num sistema de reconhecimento facial.

2.2.1 Autenticação

No caso de autenticação de identidade, o utilizador terá que fornecer uma imagem (obtida no momento) e um número de identificação pessoal (idêntico ao fornecido na fase de autenticação) ou uma palavra chave, previamente definida.

Após a aquisição dos elementos necessário, a imagem é fornecida ao sistema de reconhecimento facial para extrair o vetor de características correspondente à identidade da imagem. Através do número de identificação fornecido, é extraído da base de dados o agrupamento de caraterísticas correspondente. Por fim, o sistema efetua uma comparação entre o vetor e agrupamento de características (i.e., o primeiro obtido da imagem e o segundo guardado na base de dados). Esta comparação é efetuada através da distância do vetor ao agrupamento de características, se esta distância for superior ao limite da distância previamente definido, implica que, a pessoa que se autentica não é a pessoa por ele indicada, sendo assim impedido o acesso.

Na Figura 3 está explanado o sistema de reconhecimento facial no caso de autenticação.

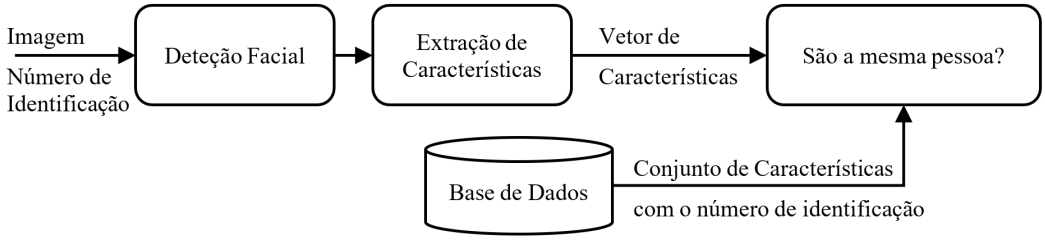


Figura 3: Situação de autenticação num sistema de reconhecimento facial.

2.2.2 Identificação

No segundo caso possível, o sistema de reconhecimento facial é aplicado para efetuar identificação, e necessita que o utilizador esteja registado na base de dados. No entanto, no momento de identificação o utilizador não necessita de comprovar a sua identidade, esta é comprovada apenas através da imagem obtida no momento de identificação (i.e., momento inicial).

A identificação é considerada um problema de fechado (em inglês, *closed-set*) quando o utilizador a identificar já existe na base de dados; é considerada um problema aberto (em inglês, *open-set*) se não se souber se o utilizador existe, ou não, na base de dados.

Na Figura 4 está explanado o sistema de reconhecimento facial para o caso de identificação.

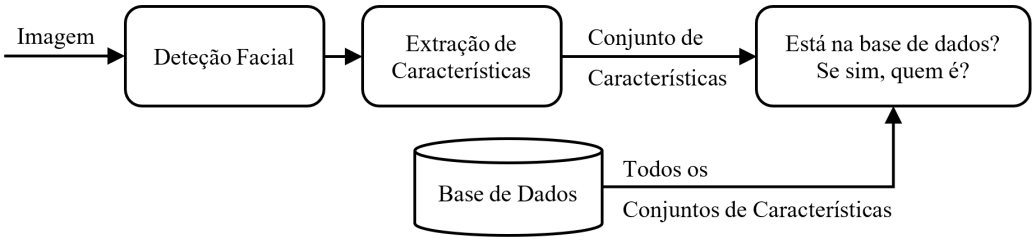


Figura 4: Situação de identificação num sistema de reconhecimento facial.

Na imagem adquirida inicialmente é efetuado reconhecimento facial com o objetivo de extrair o vetor de características. Obtido o vetor de características, é agora possível efetuar uma comparação com todos os conjuntos de características presentes na base de dados, a fim de identificar a identidade da pessoa na imagem.

A comparação, mais uma vez, é efetuada através da distância do vetor ao agrupamento de características. No caso particular do problema aberto, se essa distância for inferior ao limite previamente definido (i.e., o vetor está afastado de todo os conjuntos de características), a pessoa a identificar não se encontra presente na base de dados (i.e., identificação inconclusiva). Se a distância for superior ao limite da distância entre vetores, então, a pessoa identificada é a que tem a menor distância entre o vetor de características e o agrupamento de características. Esta última fase, é idêntica nos dois problemas, aberto e fechado.

3 Estado da Arte

Neste capítulo é apresentado uma revisão sistemática do reconhecimento facial multiespectral, elucidando assim o leitor dos métodos utilizados.

Para efetuar a revisão sistemática foi utilizada a base de dados *Web of Science*[1]. Foram considerados artigos publicados de janeiro de 2000 até junho de 2020 em revistas internacionais com fator de impacto (i.e., trabalhos publicados em conferencias não foram considerados). Nos campos de pesquisa avançada na base de dados da *Web of Science* foram utilizados os seguintes parâmetros de busca:

((TS=multispectral OR TS=spectral OR TS=thermal OR TS=SWIR) AND (TS="facial recognition" OR TS="facial detection" OR TS="face recognition" OR TS="face detection") NOT (TS=chemical) NOT (TS=vein) NOT (TS=emotion) NOT (TS=expression) NOT (TS=pedestrian) NOT (TS=palmprint) NOT (TS=eye) NOT (TS=alcohol) NOT (TS=blood))

O produto desta pesquisa foram 283 artigos publicados em 132 revistas internacionais científicas com fator de impacto. No entanto, para o presente estudo foi necessário efetuar uma segunda seleção mais refinada. Como tal, foram considerados apenas os artigos que efetuavam reconhecimento facial ou deteção facial com duas ou mais bandas espectrais (e.g., VIS-NIR, VIS-LWIR, VIS-NIR-LWIR, NIR-LWIR, entre outras possíveis combinações), reduzindo assim o número de artigos para 47. Estes artigos foram considerados os mais relevantes para esta dissertação de mestrado.

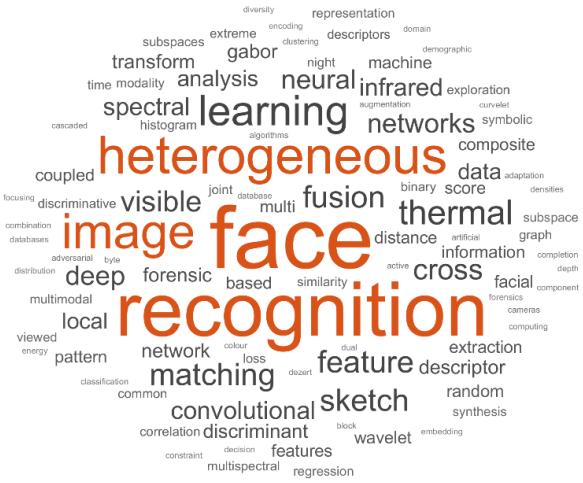


Figura 5: Nuvem de palavras das palavras-chave utilizadas nos artigos mais relevantes.

As palavras-chave escritas nos artigos ajudam o leitor a obter uma visão geral dos tópicos abordados por cada artigo. Todas as palavras-chave com mais de três caracteres foram compiladas e utilizadas para produzir uma nuvem de palavras, conforme ilustrado na Figura 5. É possível observar que o reconhecimento facial heterogéneo é mais estudado do que o reconhecimento facial homogéneo (conceitos

[1] A *Web of Science* é uma plataforma online que fornece acesso a várias bases de dados que providenciam citações de várias disciplinas académicas diferentes.

abordados posteriormente). Alem disso, as imagens térmicas (*thermal*) ou imagens LWIR são frequentemente utilizadas nos artigos mais relevantes.

3.1 Base de Dados

Nesta secção serão descritas as bases de dados mais relevantes nos artigos estudados. Numa primeira fase é efetuada uma análise da frequência de utilização das bases de dados ao longo dos artigos. De seguida, é efetuada uma análise das bases de dados públicas e das suas características, de forma a realçar as suas diferenças e semelhanças.

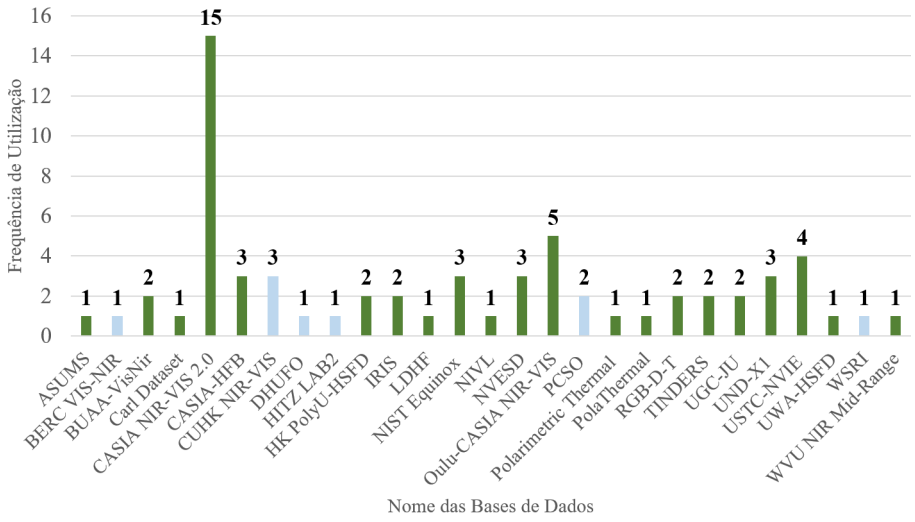


Figura 6: Frequência de utilização das bases de dados utilizadas nos artigos estudados.

As bases de dados públicas (barras a verde escuro na Figura 6) são utilizadas mais frequentemente, pois permitem a comparação de diferentes métodos facilitando ao investigador a escolha da base de dados mais adequada para o seu trabalho. Bases de dados privadas (barras a azul claro na Figura 6), são apenas utilizadas pelos criadores das mesmas, e consequentemente os métodos utilizados foram desenvolvidos pelos autores, dificultando assim uma comparação com diferentes métodos.

Através da Figura 6 é possível observar que as bases de dados mais utilizadas são a *China Academy of Sciences Institute of Automation* (CASIA) NIR-VIS 2.0 [8] (utilizada 15 vezes nos 47 artigos estudados), a Oulu-CASIA NIR-VIS [9] (utilizada 5 vezes) e a *University of Science and Technology of China-Natural Visible and Infrared facial Expression* (USTC-NVIE) [10] (utilizada 4 vezes).

A base de dados CASIA NIR-VIS 2.0 [8] destaca-se de outras bases de dados pelas seguintes razões: (i) a base de dados possui protocolos previamente definidos (i.e., quais as imagens a utilizar na fase de teste e de treino), sendo mais simples e fiável a comparação entre diferentes métodos, e (ii) composta por duas base de dados, a primeira com imagens originais, e a segunda, com as mesmas imagens faciais numa resolução de 128x128 *pixels*, agora com deteção facial e alinhamento facial efetuada.

Foi elaborada a Tabela 2 com informação relativa às bases de dados públicas que se encontram disponíveis para fins académicos. As bases de dados são classificadas pelo seu nome, ano de criação, bandas espectrais utilizadas, número de pessoas presentes, número de pessoas por imagem, e o número

total de imagens presentes. A informação relevante que não se enquadra em nenhum critério anterior foi adicionado em comentário. De realçar que, durante a construção desta tabela não foram considerados outros tipos de imagens, nomeadamente esboços [11] [12] [13] ou imagens com informação da profundidade [14] [15].

Tabela 2: Caraterísticas das bases de dados públicas mais relevantes.

Nome	Citado em	Ano	Banda Espectral	Número de Pessoas	Pessoas por Imagem	Número de Imagens	Melhor Resultado	Comentários
ASUMS [16]	[16]	2011	VIS, LWIR	96	6	576	97,9 % [16]	Com 4 variações de luminosidade.
BUAA-VisNir [17]	[18] [19]	2012	VIS, NIR	150	162	24 300	97,4 % [18]	Com 9 expressões faciais.
Carl Dataset [20]	[21]	2013	VIS, NIR, LWIR	41	180	7 380	75,6 % [21]	Com 3 variações de luminosidade e 5 expressões faciais diferentes.
CASIA NIR-VIS 2.0 [8]	[11] [13] [19] [18] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32]	2013	VIS, NIR	725	24	17 580	99,4 % [32]	-
CASIA-HFB [33]	[24] [26] [34]	2009	VIS, NIR	100	16	1 616	95,2 % [24]	Com variações nas expressões faciais.
HK PolyU-HSFD [35]	[36] [37]	2010	VIS, NIR	25	900	22 500	99,8 % [36]	Com 2 variações de luminosidade, 2 variações de pose e 2 expressões faciais diferentes.
IRIS [38]	[39] [40]	-	VIS, LWIR	30	141	4 228	96,0 % [40]	Com 5 variações de luminosidade e 3 expressões faciais diferentes.
LDHF [41]	[42]	2014	VIS, NIR	100	8	800	78,0 % [42]	Com 2 variações de luminosidade e 4 distâncias diferentes pessoa-câmara.
NIST Equinox [43]	[37] [44] [45]	2007	VIS, SWIR, MWIR, LWIR	95	-	-	99,6 %[44]	Com 3 variações de luminosidade e 3 expressões faciais diferentes.
NIVL [46]	[11]	2012	VIS, NIR	574	43	24 605	94,5 % [11]	-
NVESD [47]	[48] [49] [50]	2013	VIS, MWIR, LWIR	50	-	-	82,3 % [48]	-
Oulu-CASIA NIR-VIS [9]	[18] [19] [27] [30] [31]	2009	VIS, NIR	80	36	2 880	99,9 % [19]	Com 3 variações de luminosidade e 6 expressões faciais diferentes.
Polarimetric Thermal [3]	[3]	2019	VIS, LWIR	111	-	-	98,0 % [3]	Com 2 variações de luminosidade.
PolaThermal [51]	[11]	2016	VIS, LWIR	60	vídeo	vídeo	76,3 % [11]	Com várias expressões faciais diferentes e 3 distâncias diferentes entre pessoa-câmara.
RGB-D-T [14]	[14] [15]	2016	VIS, LWIR	51	900	45 900	86,9 % [15]	-
TINDERS [52]	[53] [54]	2009	VIS, NIR, SWIR	48	26	1 255	97,8 % [54]	Com 2 expressões faciais diferentes.
UGC-JU [55]	[56] [57]	2015	VIS, LWIR	84	39	6 552	99,2 % [56]	Com 22 poses diferentes (uma com óculos e 4 com oclusões) e 7 expressões faciais diferentes.
UND-X1 [58]	[44] [48] [49]	2004	VIS, LWIR	82	56	4 584	99,1 % [44]	-
USTC-NVIE [10]	[23] [28] [39] [59]	2010	VIS, LWIR	215	162	34 830	97,4 % [39]	Com 3 variações de luminosidade, 9 variações de pose e 3 expressões faciais diferentes.
UWA-HSFD [60]	[36]	2013	VIS, NIR	70	57	3 960	99,8 % [36]	-
WVU NIR Mid-Range [61]	[61]	2015	VIS, NIR, SWIR, LWIR	103	5 videos	515 videos	56,0 % [61]	Com 2 variações de luminosidade.

Na Tabela 2, é possível observar que grande parte das bases de dados são antigas, com uma idade média de 8 anos. Outro detalhe relevante extraído da Tabela 2, o número médio de pessoas presente numa base de dados multiespectral é de 138, número bastante inferior quando comparado com bases de dados com imagens na banda espectral do visível. A título de exemplo, a base de dados MS-Celeb-1M

[62] (composta apenas por imagens na banda espectral do visível) possui 79 099 pessoas.

De acordo com *Masi* [63], uma base de dados que contenha um número elevado de imagens faciais de diferentes pessoas é vantajoso para treinar uma rede neural profunda, pois permite cobrir a grande variedade na aparência humana. *Masi* também concluiu que uma base de dados com várias imagens da mesma pessoa, com variações de luminosidade e de pose, permite uma melhor aprendizagem da rede neural profunda. Como tal, uma base de dados que seja composta por várias imagens da mesma pessoa (i.e., número de imagens por pessoa elevado) tem a vantagem de generalizar melhor o modelo de uma rede neural através de o retreino dessa rede neural. Esta generalização apenas é possível se a rede neural tiver sido treinada inicialmente com uma base de dados com várias imagens de pessoas (i.e., número de imagens por pessoas reduzido).

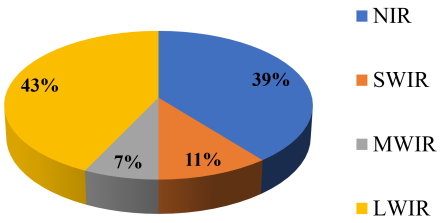


Figura 7: Distribuição das bandas espectrais nas bases de dados.

A Figura 7 ilustra a distribuição das bandas espectrais (NIR, SWIR, MWIR e LWIR) existente nas bases de dados multiespectrais públicas. Através da Figura 7 é possível observar que as bases de dados que contém imagens faciais nas bandas espectrais SWIR e MWIR são muito reduzidas, 11 % e 7 % respetivamente. Este facto deve-se ao preço elevado das câmaras SWIR e MWIR, quando comparadas com câmaras NIR ou LWIR.

3.2 Métricas Utilizadas

É necessário utilizar medidas de desempenho padronizadas de forma a avaliar o desempenho de um dado método e comparar com outros. Em reconhecimento facial multiespectral os métodos mais utilizados são: *rank-1*, taxa de falsa aceitação (FAR, em inglês *False Acceptance Rate*) e o tempo computacional utilizado pelo algoritmo [36].

A correspondência de identidades (i.e., atribuir a uma imagem facial uma identidade) é medida como a percentagem de tentativas de identificação para as quais a previsão da imagem facial é retornada nos N primeiros resultados classificados. O *rank-1* refere-se à percentagem de identidades previstas que retornam a sua primeira correspondência como correta (i.e., previu a identidade da pessoa corretamente). Esta é calculada através da divisão do número total de imagens corretamente identificadas (CI) como o número total de identificações efetuadas (IE):

$$\text{Rank-1} \ \ (\%) = \frac{CI}{IE} * 100 \tag{1}$$

Rank-N é uma extensão do *rank-1*, em que, ao invés de se verificar se a imagem mais provável é a mais correta, é verificado se a imagem correta esta entre as N imagens mais prováveis.

No caso de identificação num problema fechado, descrito no capítulo anterior, a métrica mais utilizada é a curva de características de correspondência cumulativa (CMC, em inglês *Cumulative Match Characteristic*), em que são traçados a taxa de identificação no eixo das ordenadas e a *rank-N* no eixo das abcissas. Para a comparação de métodos com base nestas curvas, os valores mais utilizados para N são 5 e 10 [64].

Para identificação em problema aberto, a métrica mais utilizada é a curva de característica de operação do recetor (ROC, em inglês *Receiver Operating Characteristic*), onde a taxa de verificação é representada em função da FAR. A partir da curva de ROC é possível obter a área sob a curva de ROC (AUC, em inglês *Area Under the Curve*), que corresponde à área por baixo da curva (ou seja, o integral) de ROC [65].

A taxa de verificação, ou taxa de verdadeiros positivos, mede a proporção dos atuais positivos que são corretamente classificados.

$$\text{Taxa de Verificação } (\%) = \frac{TP}{TP + FN} * 100 \tag{2}$$

onde, TP é o número de verdadeiros positivos e FN é o número de falsos negativos no conjunto de dados.

A taxa de falsa aceitação (FAR), ou taxa de falsos positivos, é uma estimativa empírica da probabilidade (i.e., a percentagem de vezes) que um sistema classifica incorretamente uma amostra biométrica pertencente à identidade reivindicada (i.e., ao impostor) quando a amostra pertence na realidade a um sujeito diferente (i.e., a pessoa correta).

Em sistemas de controlo de acesso, a FAR quantifica a probabilidade de um sistema biométrico (e.g., sistema de reconhecimento facial) a dar acesso a um utilizador não autorizado:

$$FAR\,(\%) = \frac{FP}{FP + TN} * 100 \tag{3}$$

onde, FP é o número de autorizações indevidas, e TN é o número de verdadeiros positivos. Por exemplo, um sistema que contenha uma taxa de FAR de 1% implica que em 100 classificações consideradas corretas, 99 foram corretamente autorizadas e 1 foi incorretamente autorizada. A personificação de identidades é uma das quebras de segurança mais importante na área de segurança biométrica, pois fornece autorização indevida a utilizadores que não a deveriam possuir [1].

Um sistema com taxas de FAR reduzidas é considerado mais seguro, prevenindo impostores de entrar. No entanto, valores de FAR são acompanhados por taxas de verificação mais reduzidas. Existe um compromisso entre a taxa de FAR e a taxa de verificação, sendo necessário afinar o algoritmo para cumprir os requisitos definidos para o sistema de reconhecimento facial. Os sistemas mais seguros têm associados taxas de FAR mais reduzidas. Os valores mais comuns para taxas de FAR são de 1% e 0,1% [1].

A avaliação de desempenho do sistema também é medida pelo tempo computacional utilizado pelo algoritmo. Quando vários algoritmos obtêm valores de classificação semelhantes, com taxas de FAR fixas, é utilizado o tempo necessário para identificar um número determinado de pessoas para demonstrar a superioridade do seu algoritmo [36].

3.3 Métodos

Nesta secção os artigos mais relevantes de reconhecimento facial multiespectral foram agrupados de acordo com o método utilizado. Na análise efetuada para cada método é fornecida uma pequena explicação, acompanhada pelos trabalhos mais relevantes, e recentes, em conjunto com um resumo do método abordado pelo autor, as bases de dados utilizadas, e os resultados obtidos.

Durante a análise de cada artigo foi possível verificar a existência de três abordagens distintas durante a implementação de métodos de reconhecimento facial: (i) multi-canal a multi-canal, (ii) multi-canal a mono-canal, (iii) e mono-canal a mono-canal, onde um canal pode ser uma banda espectral ou um intervalo espectral dentro de uma banda espectral.

A primeira abordagem utiliza os mesmos canais durante a fase de treino e teste. Usando esta abordagem, é possível utilizar toda a informação disponível de todos os canais, tendo como desvantagem os custos mais elevados da configuração.

A abordagem multi-canal a mono-canal utiliza todos os canais na fase de treino e, na fase de teste, utiliza apenas um canal. Esta abordagem é útil para reduzir os custos durante a implementação do sistema de reconhecimento facial, uma vez que só utilizamos uma única câmara. A abordagem é também conhecida como reconhecimento facial heterogéneo.

A última abordagem é a mais limitada, com as vantagens e limitações associadas ao canal utilizado. Neste caso, um único canal é utilizado na fase de formação e teste. A abordagem é também conhecida como reconhecimento facial homogéneo.

Através da análise sistemática foi possível destacar cinco métodos principais utilizados no reconhecimento facial multiespectral: reconhecimento por características, subespaço acoplado, síntese, fusão e redes neurais profundas. A Tabela 3 agrupa os 47 artigos estudados consoante o método utilizado pelos autores e a frequência de utilização de cada método nos artigos estudados. Através da Tabela 3 é possível observar que o método mais utilizado utiliza redes neurais profundas, que representam o estado da arte atual em reconhecimento de padrões e com resultados promissores nos trabalhos de reconhecimento facial multiespectral.

Tabela 3: Métodos mais utilizados.

Métodos	**Citado em**	**Percentagem**
Reconhecimento por Características	[23][28][42][53][54][66]	13 %
Subespaço Acoplado	[12][16][22][34][39][48][67][68][69][70]	21 %
Síntese	[3][15][19][50][59][71]	13 %
Fusão	[14][21][37][40][44][56][57][61][72][73]	21 %
Redes Neurais Profundas	[11][13][18][24][25][26][27][29][30][31][32][36][49][74][75]	32 %

Na Figura 8 está ilustrada a distribuição dos métodos por ano de publicação. É possível observar a predominância de artigos que utilizam o método de fusão de imagens em reconhecimento facial multiespectral à data de 2017. Desde então, o método de fusão foi superado pelo método de redes neurais

profundas, uma vez que este obtém resultados superiores.

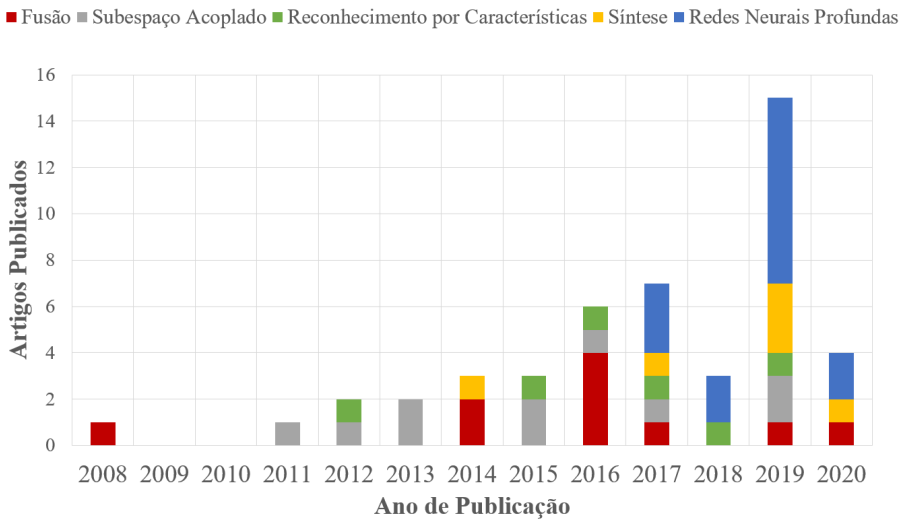


Figura 8: Distribuição dos métodos por ano de publicação.

Na Figura 9 está apresentado um gráfico *caixa de bigodes*[2] que resume o desempenho obtido nos artigos estudados para cada método. Deste modo, é possível efetuar uma comparação ao nível do desempenho de cada método. Por observação da Figura 9, a rede neural profunda obtém os melhores resultados, e como tal, justifica o aparecimento de novas redes neurais profundas e métodos dentro da área (também corroborado pela Figura 8).

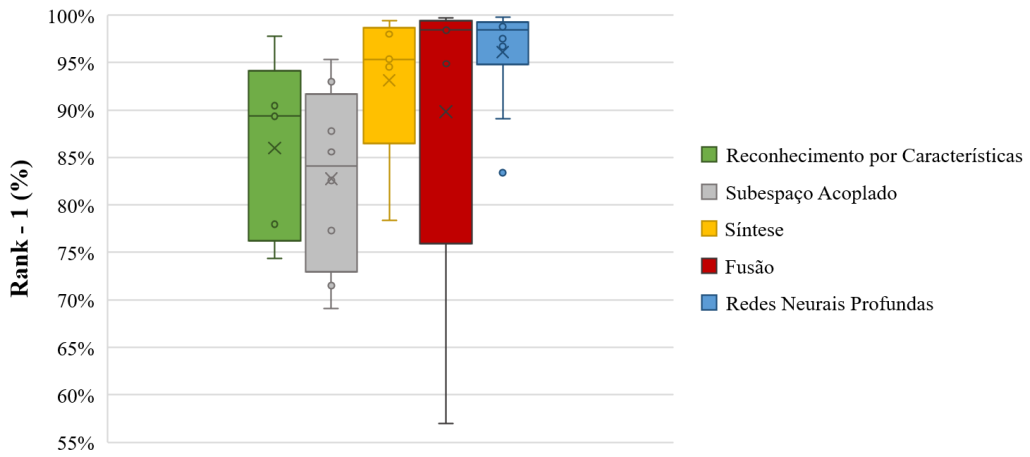


Figura 9: Desempenho obtido nos métodos descritos nos artigos estudados para cada método.

3.3.1 Reconhecimento por Características

Os métodos de representação de características procuram extrair as características que sejam mais invariantes à banda espectral utilizada. Através da extração de características faciais (e.g., contornos, cantos, olhos, boca, entre outros) é possível efetuar uma redução da informação fornecida pela imagem

[2] A caixa de bigodes (em inglês *boxplot*) é uma ferramenta gráfica utilizada para representar uma variação de dados observados de uma variável numérica por meio de quartis. É utilizada frequentemente para analisar e comparar a variação de uma variável entre diferentes grupos de dados.

inicial. Eliminando a informação irrelevante é possível reduzir o trabalho computacional desempenhado pelo classificador. Deste modo, é possível reduzir a discrepância existente entre as diferentes bandas espectrais utilizadas [22].

O método de reconhecimento de caraterísticas pode ser utilizado exclusivamente (em conjunto com um classificador), ou como base para outras metodologias de reconhecimento facial multiespectral (como será apresentado nas próximas subsecções) [22].

Uma das principais desvantagens ao utilizar este método é que alguns extratores de características, como por exemplo os padrões de binário local (LBP, em inglês *Local Binary Patterns*), ignoram a estrutura espacial do rosto (i.e., olhos, boca, nariz), sendo estes cruciais para ser possível obter bons desempenhos em sistemas de reconhecimento facial [23].

Shamia *et al.* [42] utilizou uma combinação entre um histograma de gradientes orientados (HOG, em inglês *Histogram of Oriented Gradients*) e LBP para extrair características faciais de imagens na banda espectral do NIR e efetuar reconhecimento facial para distâncias de 60, 100 e 150 metros. As características extraídas das duas imagens VIS e NIR foram comparadas utilizando a distância euclidiana. Foi utilizada a base de dados LDHF [41], que inclui imagens VIS e NIR capturadas a várias distâncias (1, 60, 100 e 150 metros). Para esta base de dados, foi obtida um *rank-1* de 72%, 78% e 32% para as distâncias de 60, 100 e 150 metros, respetivamente.

Peng *et al.* [23] desenvolveu uma representação gráfica baseada em reconhecimento facial heterogéneo (G-HFR, em inglês *Graphical Representation for Heterogeneous Face Recognition*) que utiliza redes de *Markov* (um gráfico não direcionado cujas ligações representam dependências probabilísticas simétricas [76]) para representar partes de imagens heterogéneas separadamente. A rede de *Markov* tem em consideração a compatibilidade espacial entre as partes vizinhas. As bases de dados CASIA NIR-VIS 2.0 [8] e USTC-NVIE [10] foram utilizadas para validar o método proposto, obtendo um *rank-50* de 83,3% e 95,4% para a primeira e segunda base de dados, respetivamente.

Num segundo trabalho de Peng *et al.* [28], foi proposto utilizar uma representação gráfica dispersa baseada numa análise discriminativa (SGR-DA, em inglês *Sparse Graphical Representation based Discriminant Analysis*) para representar imagens faciais em diferentes bandas espectrais. Os vetores dispersos adaptativos utilizados foram produzidos através de redes de *Markov*, foram bastante eficientes para reconhecimento facial heterogéneo. As bases de dados CASIA NIR-VIS 2.0 [8] e USTC-NVIE [10] foram utilizadas para validar o método proposto, obtendo um *rank-50* de 87,8% e 93,1% para a primeira e segunda base de dados, respetivamente.

3.3.2 Subespaço Acoplado

Os métodos que projetem as características de diferentes bandas espectrais num subespaço comum são conhecidos por métodos de subespaço acoplado, vulgarmente conhecido em inglês por *coupled subspace*. Este subespaço permite identificar a informação que é comum às diferentes bandas espectrais. Com esta abordagem, é possível reduzir a discrepância entre as imagens de diferentes bandas espectrais. No entanto, o poder discriminativo deste método é fortemente reduzido se a diferença entre as bandas espectrais for elevada, onde o melhor caso possível é para VIS-NIR e o pior caso para VIS-LWIR [22] [27].

Este método possui a desvantagem de, durante a projeção das imagens no subespaço, existir a possibilidade de haver sempre informação crucial a ser descartada, e como tal, diminuir o desempenho do sistema de reconhecimento facial multiespectral [23].

Jin *et al.* [22] apresentou um novo método para efetuar a extração de características. O método é denominado de aprendizagem de características discriminativas acopladas (em inglês *coupled discriminative feature learning*), e aplicado ao seu sistema de reconhecimento facial heterogéneo. Este método maximiza as variações inter-classe e minimiza as variações intra-classe. Foi utilizada a base de dados CASIA NIR-VIS 2.0 [8] para efetuar os testes e foram obtidos os seguintes resultados *rank-1*, e taxa de verificação com FAR a 1% e 0,1%: 71,5%, 55,1% e 67,7%, respetivamente.

Li *et al.* [12] propôs uma análise de componentes mútuas (MCA, em inglês *Mutual Component Analysis*) para estudar as características que são mútuas (comuns) às duas bandas espectrais de cada imagem, nomeadamente do VIS e do LWIR. Foi utilizada a base de dados CASIA NIR-VIS 2.0 [8] para efetuar os testes e obteve os seguintes resultados para *rank-1*, e taxa de verificação com FAR a 1% e 0,1%: 69,1%, 92,7% e 76,2%, respetivamente.

Bhowmik *et al.* [39] propôs uma variante para a análise de componentes independentes (ICA, em inglês *Independent Component Analysis*), através da aplicação de uma transformação logarítmica na ICA base, denominada de (Log-ICA). Foram propostas duas arquiteturas, Log-ICA I e Log-ICA II; a última apresentou melhores resultados. Foi concluído que ambas as arquiteturas propostas apresentam bons resultados em reconhecimento facial heterogéneo, e que podem ser aplicadas em reconhecimento de expressões faciais e reconhecimento facial de imagens com ruído. Foram utilizadas duas bases de dados VIS-LWIR, a IRIS [38] e USTC-NVIE [10]. Na primeira base de dados, foi obtido um *rank-1* de 88,2% e 90,5% para a Log-ICA I e Log-ICA II, respetivamente. Na segunda base de dados, foi obtido um *rank-1* de 96,0% e 97,4% para a Log-ICA I e Log-ICA II, respetivamente.

3.3.3 Síntese

Os métodos de síntese de imagem transformam uma imagem de uma banda espectral para outra banda espectral, permitindo assim comparar duas imagens mais facilmente.

Estes métodos permitem sintetizar uma imagem na banda espectral do visível utilizando como ponto de partida uma imagem de outra banda espectral (e.g., a banda espectral LWIR). A principal vantagem dos métodos de síntese de imagem é que, assim que a imagem LWIR é sintetizada numa imagem da banda espectral do visível, é possível aplicar métodos de reconhecimento facial preparados para imagens do visível [77].

Em contrapartida, o principal problema neste método é que a síntese de imagens é um processo complexo e, na maioria dos casos o desempenho do sistema de reconhecimento facial está em grande parte dependente da precisão da imagem sintetizada [23].

Osia *et al.* [50] emprega imagens na banda espectral do LWIR para produzir, através de síntese, imagens equivalentes na banda espectral do visível. A fim de demonstrar as vantagens do método proposto, foi efetuada extração de características através de LBP com o intuito de efetuar reconhecimento facial nas imagens sintetizadas. O método proposto foi testado na base de dados NVESD [47], que inclui ima-

gens nas bandas espectrais do VIS, MWIR e LWIR. Foi efetuado reconhecimento facial heterogéneo nas bandas espectrais VIS-MWIR, obtendo-se um *rank-1* de 75,3%. Foi também efetuado reconhecimento facial heterogéneo nas bandas espectrais VIS-LWIR, e obteve-se desta vez um *rank-1* de 81,4%.

Litvin *et al.* [15] propôs a utilização de uma rede neural convolucional para efetuar síntese de imagens na banda espectral do LWIR para o visível. O autor modificou a arquitetura da *FusionNet* [78] e o algoritmo de treino correspondente para diminuir o sobre ajuste. Tal foi possível através da introdução de: (i) *dropout*, (ii) *Randomized Leaky Rectified Linear Unit* (RLReLU) e de (iii) regularização ortogonal. O método foi testado para cada uma das três variações presentes na base de dados RGB-D-T [14]: pose, expressões, e variações de luminosidade; produzindo um *rank-1* de 86,9%, 97,5% e 99,2%, para cada respetiva variação.

He *et al.* [19] propôs um complemento facial multiespectral (CFC, em inglês *Adversarial Cross-spectral Face Completion*) que utiliza uma rede generativa antagónica (GAN, em inglês *Generative Adversarial Network*) com o objetivo de sintetizar imagens na banda espectral do visível através de imagens na banda espectral do NIR. Esta abordagem destaca-se de outros métodos, pois utiliza uma componente de repintura que sintetiza e preenche texturas de imagens na banda espectral do VIS através de texturas obtidas nas imagens da banda espectral do NIR. O método converte qualquer pose de imagens NIR numa imagem com pose frontal do VIS, resultando num par de imagens com texturas nas duas bandas espectrais. De seguida, as duas componentes obtidas (i.e., de repintura e de correção de pose) são ligadas e integradas numa *Deep Convolutional Neural Network* (DCNN). Por fim, no último passo é realizado reconhecimento facial através da imagem sintetizada utilizando a LightCNN [79]. A CFC foi testada em três base de dados: (i) CASIA NIR-VIS 2.0 [8], (ii) Oulu-CASIA NIR-VIS [9] e a (iii) BUAA-VisNir [17], obtendo os seguintes resultados para *rank-1*: 98,6%, 99,9% e 99,7%, respetivamente para cada base de dados. Utilizando as mesmas bases de dados, a CFC obteve as seguintes taxas de verificação com um FAR de 1% e 0,1%: (i) 99,2% e 97,3%, (ii) 98,1% e 90,7%, (iii) 98,7% e 97,88%, respetivamente para as três bases de dados.

3.3.4 Fusão

O desempenho global do sistema de reconhecimento facial multiespectral pode ser melhorado através da combinação de várias imagens numa única imagem, em função das imagens utilizadas. Os métodos mais relevantes de fusão de imagem aplicados em reconhecimento facial são: fusão de características e fusão de pontuações; podendo ser utilizados individualmente ou combinados.

A fusão de características combina as características de várias imagens, adquiridas durante a fase de extração de características, num vetor de características. Estas características incluem informação acerca de arestas, cantos, linhas e texturas, sendo calculadas e concatenadas num único vetor de características, para ser utilizado na deteção ou segmentação facial. O método de fusão de características também é empregue como método para reduzir a dimensão do vetor de características final, reduzindo assim a exigência computacional ao algoritmo [1] [80].

Fusão de pontuações melhora o desempenho global da classificação, combinando a saída de vários classificadores num único classificador. O método mais utilizado na fusão de pontuações é o voto por

maioria, em que a classificação obtida de cada classificador é tida em consideração para a votação. Esta votação consiste em descobrir que classificação ocorre com mais frequência, atribuindo-a ao classificador global. Outro método que pode ser utilizado na fusão de pontuações, é o peso adaptativo, ou dinâmico, onde é atribuído um peso dinâmico a cada classificador. Deste modo, classificadores que demonstrem um fraco desempenho irão ter atribuído um peso reduzido, e consequentemente, menor importância no classificação global [1] [80].

A utilização do método de fusão de imagens em sistemas de reconhecimento facial possui como principal vantagem a redução da taxa de erro e o custo de implementação através da aplicação de diversas câmaras de baixo custo, em vez de uma única câmara de custo elevado [81].

Seal *et al.* [57] aplicou um processo de fusão onde é calculada a soma ponderada da informação das imagens nas bandas espectrais do VIS e LWIR através de dois pesos. De forma a avaliar o método, inicialmente foram realizados dois reconhecimentos faciais independentes, o primeiro na imagem do VIS e o segundo na imagem do LWIR. Cada um produziu uma pontuação, que é igual à probabilidade de classificação correta para cada imagem. Numa segunda fase, o reconhecimento facial é efetuado a partir da imagem fundida, criada a partir do processo de fusão proposto onde os pesos são as pontuações previamente calculadas. Foi utilizada a base de dados UGC-JU [55] e produziu um *rank-1* de 98,4%.

Simón *et al.* [14] utilizou os extratores de características LBP, HOG e *Histograms of Gabor Ordinal Measures* (HOGOM) [82] para extrair as características de imagens VIS, LWIR e de profundidade. As características foram depois concatenadas num único vetor de características para treinar o algoritmo do k-vizinhos mais próximos ponderado (W-kNN, em inglês *Weighted k-Nearest Neighbors*). A ideia por trás do W-kNN é de dar mais peso aos pontos que estão próximos e menos peso aos pontos que estão mais distantes. Logo após, foi utilizado uma *Convolutional Neural Network* (CNN) para processar individualmente cada imagem original (VIS, LWIR e de profundidade). Por fim, o classificador final foi obtido pela fusão dos classificadores W-kNN e CNN com pesos diferentes. O autor também divulgou uma nova base de dados, a RGB-D-T, composta por imagens do visível, LWIR e de profundidade.

Kanmani *et al.* [40] propôs três métodos de fusão com o intuito de auxiliar no reconhecimento facial heterogéneo. Para o primeiro e o segundo métodos a imagem de entrada foi decomposta em coeficientes de alta e baixa frequência através da transformada *wavelet* discreta de duas árvores (em inglês *dual tree discrete wavelet transform*). De seguida, é utilizado uma técnica de otimização assente nos níveis de população [83] para encontrar os pesos ótimos para realizar a fusão das imagens VIS e LWIR. O terceiro método, aplica uma técnica de otimização, o *Self Tuning Particle Swarm Optimization*, deste modo é possível evitar a convergência prematura do *particle swarm*. Esta utiliza uma transformada de *curvelet* para efetuar uma decomposição da imagem, preservando as arestas ao longo das curvas. Para aprimorar a procura dos pesos ótimos, é utilizado o algoritmo de *brainstorm* [84]. Utilizando a base de dados IRIS [38] foi possível obter um *rank-1* de 94,2%, 94,5% e 96,0%, para os primeiro, segundo e terceiro algoritmo.

3.3.5 Redes Neurais Profundas

A rede neural artificial teve como inspiração o córtex visual do ser humano. Em comparação com os métodos clássicos, consegue obter resultados significativamente superiores. As redes neurais artificiais

podem ser constituídas por dezenas camadas, em que cada camada é treinada e responsável por detetar diferentes características numa dada imagem [85].

Atualmente, a rede neural artificial mais utilizada em reconhecimento facial é a rede neural convolucional profunda (DCNN, inglês *Deep Convolutional Neural Network*), que compreendem um número elevado de camadas, quando comparadas com as redes neurais tradicionais.

As redes DCNN são compostas por várias camadas de convolução, de ativação e de *pooling*. A camada de convolução aplica um conjunto de filtros de convolução nas imagens de entrada, em que cada filtro é sintonizado pelo processo de aprendizagem para certas características das imagens. A camada de ativação define o valor de saída tendo em consideração a entrada. Existem várias funções de ativação passiveis de serem utilizadas na camada de ativação, sendo as mais populares: a função *Rectified Linear Unit* (ReLU) e a função linear [86]. A camada de *pooling* simplifica a saída efetuando uma redução da resolução de modo não linear, reduzindo assim o número de parâmetros que a rede necessita de aprender. A repetição destas camadas permite identificar as características mais particulares e as mais únicas das imagens ao longo da rede neural [87].

O conjunto de características produzido por uma DCNN treinada para reconhecimento facial, vulgarmente conhecido em inglês por *embeddings*, é uma representação vetorial da identidade de uma pessoa, obtida através de uma imagem facial. Esta representação vetorial possui uma dimensão fixa, sendo necessário definir-la *a priori* durante a criação da arquitetura da DCNN. As dimensões mais utilizadas para o conjunto de características em reconhecimento facial são 128 e 256 [79] [88].

A aplicação de DCNN em sistemas de reconhecimento facial é relativamente simples. Inicialmente, uma imagem é introduzida na rede neural, sendo extraído um conjunto de características. Quando a mesma rede recebe outra imagem facial da mesma pessoa, a rede deve produzir um conjunto de características semelhantes, enquanto que o contrário deve acontecer quando é fornecido à entrada da rede uma imagem de uma pessoa diferente [87].

As atuais DCNN possuem como desvantagem o tempo de treino, dependente do desempenho da unidade de processamento gráfico (GPU, em inglês *Graphics Processing Unit*). Por vezes, as redes neurais não são apenas comparadas pelos resultados obtidos, mas também pelo tempo de treino e de classificação [36].

Hu *et al.* [27] desenvolveu múltiplas redes profundas com combinação de diversidade (MDNDC, em inglês *Multiple Deep Network with Scatter Loss and Diversity Combination*) para reduzir a variação intra-classe e aumentar a variação inter-classe. Foi utilizada a função de custo de dispersão (introduzida num outro trabalho do autor [30]), que permite reduzir a diferença entre imagens de bandas espectrais diferentes (i.e., VIS e NIR), preservando assim as informações da pessoa a ser identificada. São utilizadas diversas redes profundas para extrair as características. A combinação de diversidade é utilizada para ajustar adaptativamente os pesos de cada rede profunda. A MDNDC foi testada na CASIA NIR-VIS 2.0 [8] obtendo um *rank-1* de 98,9% e uma taxa de verificação de 99,6% e 97,6%, para um FAR de 1% e 0,1%, respetivamente. A MDNDC foi também testada na Oulu-CASIA NIR-VIS [9] obtendo um *rank-1* de 99,8% e uma taxa de verificação de 88,1% e 65,3%, para um FAR de 1% e 0,1%, respetivamente.

Peng *et al.* [29] propôs a utilização de uma estrutura de aprendizagem com um descritor local profundo

(em inglês, *deep local descriptor learning framework*) aplicado num sistema de reconhecimento facial heterogéneo, que é capaz de aprender informações discriminantes diretamente de imagens faciais. Foi proposta uma nova função de custo, a função de enumeração, para bandas espectrais diferentes de forma a eliminar a diferença entre diferentes bandas espectrais ao nível local, que é depois integrado numa CNN para extrair descritores locais profundos. O método foi testado na CASIA NIR-VIS 2.0 [8] e obteve um *rank-1* de 96,7%.

Pereira *et al.* [11] introduziu o conceito de unidade específica da banda espectral (DSU, do inglês *Domain Specific Units*). O autor enuncia que as características de alto nível das redes neurais convolucionais profundas codificam características faciais gerais que são independentes da banda espectral utilizada. Portanto, as características de baixo nível podem ser adaptadas para satisfazer uma dada banda espectral, permitindo a aprendizagem por transferência de conhecimento de uma rede pré-treinada. O autor utiliza um modelo DCNN previamente treinado em imagens faciais na banda espectral do visível, a Inception-ResNet v2 [89]. Foram utilizados dois métodos para retreinar a DCNN, a tripla rede neural e a rede neural siamesa. A DCNN foi testada em três base de dados: (i) CASIA NIR-VIS 2.0 [8], (ii) NIVL [46] e PolaThermal [51]. Foram obtidas as seguintes pontuações em *rank-1* para a tripla rede neural e a rede neural siamesa nestas bases de dados: (i) 96,3% e 90,1%, (ii) 94,5% e 92,2%, (iii) 76,3% e 50,9%, respetivamente.

He *et al.* [18] aplicou a distância de *Wasserstein* como função de custo no treino de uma CNN, a *Wasserstein Convolutional Neural Network* (WCNN). A distância de *Wasserstein* é a distância entre duas distribuições de probabilidade num dado espaço. O efeito de treinar uma rede com esta distância será o de reduzir a distância existente entre os conjuntos de características das imagens das bandas espectrais VIS e NIR. A WCNN foi testada em três bases de dados: (i) CASIA NIR-VIS 2.0 [8], (ii) Oulu-CASIA NIR-VIS [9] e BUAA-VisNir [17], alcançando as pontuações em *rank-1*: 98,7%, 98,0% e 97,4%, respetivamente para cada base de dados. Utilizando as mesmas bases de dados obtiveram as seguintes taxas de verificação para um FAR de 1% e 0,1%: (i) 99,5% e 98,4%, (ii) 81,5% e 54,6%, (iii) 96,0% e 91,9%, respetivamente para as três bases de dados.

Wu *et al.* [36] desenvolveu uma DCNN para reconhecimento facial multiespectral que explora a informação discriminante intra-espectral e a correlação da informação inter-espectral. O autor denomina a rede profunda de análise da correlação inter-espectral e discriminação intra-espectral (IDICN, em inglês *Intraspectrum Discrimination and Interspectrum Correlation Analysis Deep Network*) e efetua testes em duas bases de dados, a HK PolyU-HSFD [35] e a UWA-HSFD [60], obtendo um *rank-1* de 99,8% e 99,8%, respetivamente.

Bae *et al.* [32] introduziu dois módulos com o intuito de melhorar o reconhecimento facial heterogéneo, sendo a banda espectral da imagem final do visível. O primeiro módulo consiste num (i) pré-processamento em cadeia, para garantir que a gama de intensidades é semelhante entre a imagem a processar e a imagem alvo (i.e., a imagem final na banda espectral do visível); (ii) uma GAN cíclica (em inglês *CycleGAN*) para aprender o mapeamento entre a imagem de entrada (NIR) e a imagem final (VIS) através de um conjunto de dados de treino de pares de imagens, previamente preparadas e alinhadas. No segundo módulo, as imagens do conjunto de treino da base de dados e as suas imagens traduzidas são utilizadas para afinar

o modelo DCNN previamente treinado (a ResNet-101 [90] treinada com imagens faciais do visível da base de dados MS-Celeb-1M [62]) para obter uma representação vetorial com 512 dimensões. Durante a fase de testes foi utilizada a base de dados CASIA NIR-VIS [8]. Sem o módulo de pré-processamento, obteve-se um *rank-1* de 99,1% e uma taxa de verificação de 98,7% a um FAR de 0,1%. Com o módulo de pré-processamento, os autores obtiveram melhores resultados: um *rank-1* de 99,4% e uma taxa de verificação de 98,8% a um FAR de 0,1%.

4 Metodologia

Este capítulo versa sobre a metodologia adotada para a implementação de um sistema de reconhecimento facial multiespectral. Foram elaborados dois fluxogramas, o primeiro, apresentado na Figura 10, serve para dar uma visão simples, e, no entanto, elucidativa, ao leitor da metodologia utilizada, e o segundo, apresentado no Apêndice A, proporcionar ao leitor uma visão mais detalhada de cada módulo.

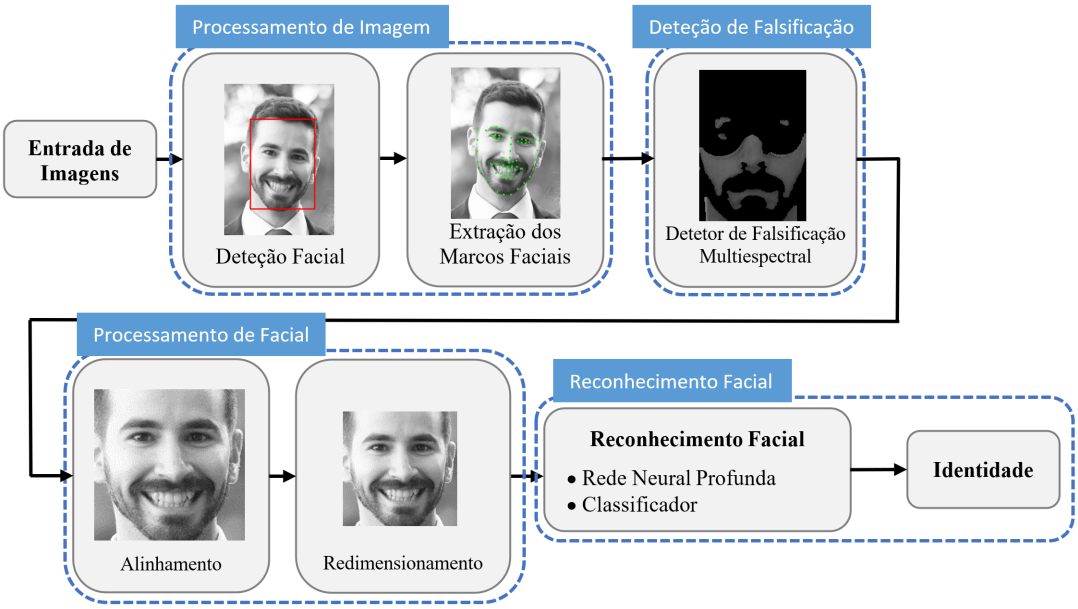


Figura 10: Fluxograma com a metodologia a adotar.

A construção da solução num problema de reconhecimento facial inicia-se pela aquisição de imagens. Nesta fase inicial do sistema são adquiridas as imagens multiespectrais (i.e., visível e infravermelho). A obtenção destas imagens pode ser realizada por vários equipamentos mono-espectrais ou por um equipamento de imagem capaz de obter imagens em vários intervalos espectrais. O único requisito nesta fase é de que todas as imagens sejam obtidas no mesmo instante, de forma a que possuam as mesmas condições de luminosidade e pose da pessoa.

O passo seguinte consiste em detetar a presença de faces humanas e extração dos marcos faciais (e.g. olhos, nariz, boca) na imagem adquirida. Este módulo, denominado de processamento de imagem, corresponde à Secção 4.1.

O sistema de reconhecimento facial proposto inclui um módulo para detetar e avisar potenciais ataques de falsificação (Secção 4.2), através da utilização de imagens multiespectrais, que emprega todas as bandas espectrais disponíveis para efetuar uma deteção de pele, prevenindo assim o sistema de reconhecimento facial de possíveis ataques.

No módulo seguinte (Secção 4.3) é realizado o processamento facial, sendo utilizados os marcos faciais, obtidos no primeiro módulo, para efetuar o alinhamento da face. Este módulo tem como principal objetivo a normalização da imagem a inserir na DCNN. Este processamento facial distingue-se do processamento de imagem na medida em que no processamento facial é efetuado o processamento apenas na face detetada

e não na imagem global, como é no caso do processamento de imagem.

O módulo de reconhecimento facial, apresentado na Secção 4.4, efetua uma extração do conjunto de características da pessoa a identificar através da rede proposta. De seguida, é efetuada a classificação da identidade através do conjunto de características calculadas pela DCNN a fim de obter a identificação da pessoa nas imagens de entrada.

4.1 Processamento de Imagem

Esta é a primeira fase do sistema de reconhecimento facial e tem como objetivo principal a deteção e extração dos marcos faciais das imagens multiespectrais. Uma incorreta obtenção na deteção facial e dos marcos faciais serão fatores limitativos no desempenho dos consequentes módulos.

Adquiridas as imagens é necessário, agora, efetuar uma conversão das imagens inicialmente obtidas num padrão de cores RGB (i.e., três canais com uma gama de valores de (0, 255) para cada canal), para um padrão de cores da gama do cinzento, que utiliza um canal. Esta conversão deve-se essencialmente a duas razões, a primeira imposta pelos algoritmos utilizados (e.g. extrator de marcos faciais apenas utiliza imagens a cinzento) e a segunda pelo facto de que convertendo as imagens para cinzento é possível reduzir a influência das variações de luminosidade na imagem [79].

Como método de deteção facial é utilizada uma rede neuronal profunda da OpenCV[3] [91]. Esta é baseada no detetor de múltiplas regiões de passagem simples (SSD, do inglês *single shot multibox detector*) desenvolvido por Liu *et al.* [92] e utiliza uma arquitetura semelhante à ResNet-10 [90]. O detetor divide a imagem inicial em diferentes retângulos com diferentes tamanhos e atribui diferentes pontuações a cada retângulo. Com estas pontuações o detetor ajusta os retângulos de modo a este se a ajustar à cara da pessoa. O modelo disponibilizado pela OpenCV foi treinado numa base de dados pública, no entanto não é indicada qual [91]. A deteção produz uma caixa delimitadora, indicando assim a localização da face humana detetada na imagem introduzida.

Os marcos faciais são utilizados para localizar e representar regiões salientes ou partes faciais da cara humana, como olhos, boca, nariz e a mandíbula humana. Após a obtenção dos marcos faciais é possível aplicá-los em diversas finalidades como por exemplo: no alinhamento facial, obtenção de uma estimativa da pose da cabeça, o piscar de olhos, ou até mesmo, detetar a sonolência de um condutor ao volante [93].

Para efetuar a extração dos marcos faciais foi utilizada uma rede neural disponibilizada para o efeito na biblioteca da Dlib[4] [94]. A rede neural da Dlib é baseada na implementação de Kazemi *et al.* [95] e foi treinada na base de dados iBUG 300-W [96]. Esta base de dados é composta por imagens na banda espectral do visível, onde cada imagem contém anotações precisas, detalhadas e consistentes de vários marcos faciais. O modelo possui como entrada uma imagem de níveis de cinzento e produz uma lista contendo a localização dos 68 marcos faciais. Esta localização corresponde às coordenadas (x, y) da

[3]OpenCV (Open Source Computer Vision Library) é uma biblioteca de software de código aberto com vista a agilizar o processo de implementação de ideias em visão computacional. Sendo este um produto de licença BSD (i.e., requer apenas o reconhecimento dos autores) é facilitado o uso e modificação do código base.

[4]Dlib é conjunto de ferramentas constituído principalmente por algoritmos de aprendizagem automática, com vista a resolver problemas do mundo real. Sendo este um produto de licença BSD (i.e., requer apenas o reconhecimento dos autores) é facilitado o uso e modificação do código base.

imagem fornecida na entrada.

Na Figura 11 está ilustrada a localização e numeração dos marcos faciais extraidos.

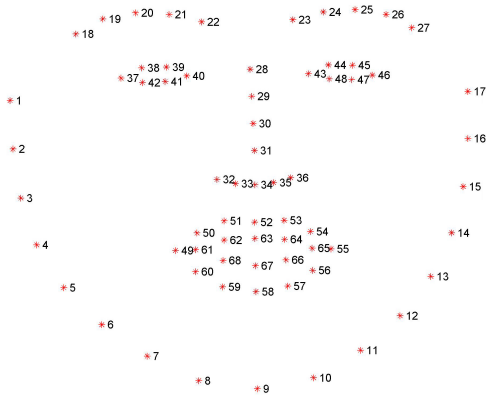


Figura 11: Localização e numeração dos 68 marcos faciais produzidos pela dlib.

Na Figura 12 está ilustrada a sequência de passos do módulo de processamento de imagens, simplificando a compreensão do módulo. O processamento de imagem será efetuado em todas as bandas espectrais adquiridas. Caso nalguma banda não se consiga obter o resultado proposto (i.e., efetuar deteção facial ou extração de marcos faciais) este irá socorrer-se de seguida de resultado proveniente da banda espectral do visível. Este é um mecanismo de segurança que apenas opera em caso de insucesso.

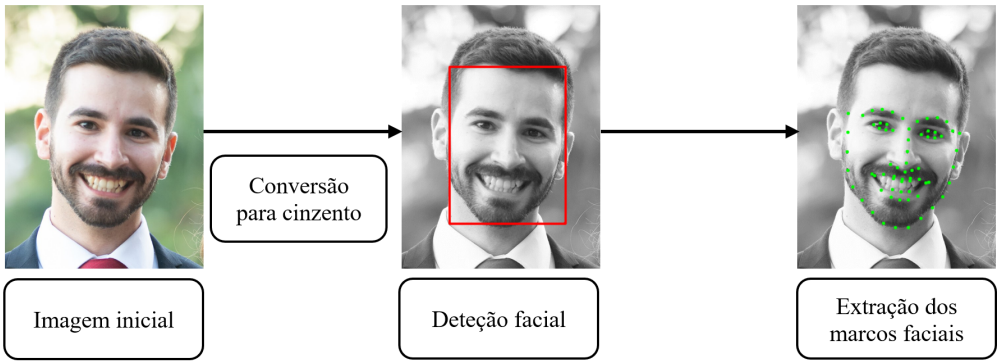


Figura 12: Esquema resumo do módulo de processamento de imagem.

4.2 Deteção de Falsificação

No módulo de deteção de falsificação é empregue um detetor de falsificação multiespectral, baseado no classificador proposto por Steiner *et al.* [6]. São utilizadas todas as bandas espectrais disponíveis para efetuar uma deteção de pele ou não pele.

Na banda espectral do visível os artefactos de ataque (e.g., máscaras) de falsificação são pensados de modo a que sejam difíceis de distinguir da pele. Deste modo, qualquer algoritmo de deteção de falsificação que utilize apenas imagens na banda do visível não consegue obter bons resultados. É, por isso, necessário complementar os detetores com outras bandas espectrais (e.g., NIR ou SWIR). Quanto maior o número de bandas espectrais usadas, maior a fidelidade de deteção de pele alcançada pelo detetor multiespectral de falsificação [97].

O detetor de falsificação multiespectral proposto efetua uma deteção de pele ao nível do *pixel*. É definido por $g_i(x, y)$ como *pixel* do canal i ($i = 1, 2, ..., n$) de n canais, onde n corresponde ao canal atribuído. A este corresponde um valor na escala do cinzento, de (0, 255), na localização (x, y).

Numa primeira etapa é efetuada a diferença normalizada, d[g_a, g_b], para todas as combinações possíveis, tendo em conta os canais disponíveis. A diferença normalizada é calculada da seguinte forma:

$$d[g_a, g_b] = \frac{g_a - g_b}{g_a + g_b} \tag{4}$$

com $1 \leq a \leq n$ e $a \leq b \leq n$. Portanto, para $n = 3$ (i.e., onde são utilizados 3 canais) é obtido o seguinte vetor de diferenças normalizadas $\overrightarrow{d}$, sendo este definido por:

$$\overrightarrow{d} = (d[g_1, g_2], d[g_1, g_3], d[g_2, g_3]) \tag{5}$$

para cada *pixel* (x, y). A diferença normalizada irá tomar valores de $-1 \leq d[g_a, g_b] \leq +1$.

Tendo obtido o vetor de diferenças normalizadas é possível agora empregar o detetor de pele. Esta seleção tem como objetivo efetuar uma deteção de pele ou não pele, descartando os *pixels* que são classificados como não sendo pele para cada diferença normalizada. Esta deteção prévia é efetuada através da aplicação de limites máximos e mínimos para cada diferença normalizada, sendo estes determinados empiricamente.

Após a deteção de pele é feita uma decisão entre as diferentes deteções para cada diferença normalizada. O resultado desta decisão é um mapa binário que contém os *pixels* que foram considerados como pele em todas as diferenças normalizadas. Esta máscara é um mapa binário, onde o "1" corresponde a pele e "0" corresponde a não pele.

Na segunda etapa é utilizado o mapa binário, produzido na etapa anterior, e os marcos faciais extraídos no módulo de processamento de imagem. Nesta é feita uma deteção de falsificação através da percentagem de marcos faciais que são considerados como pele. Esta percentagem é determinada pela divisão do total de marcos faciais que correspondem a coordenadas onde existe pele, pelo número total de marcos faciais utilizados, como é descrita na seguinte equação:

$$Pele(\%) = \frac{MarcosFaciais(1)}{MarcosFaciais(0) + MarcosFaciais(1)} * 100 \tag{6}$$

onde Marcos Faciais(0) corresponde ao número de marcos faciais que são equivalentes a não pele, e Marcos Faciais(1) corresponde ao número de marcos faciais que são equivalentes a pele. Se a percentagem de pele for inferior a um limite estipulado será emitido um aviso sobre um potencial ataque de falsificação. Se este limite for superado a imagem continuará para o módulo seguinte.

Está ilustrado na Figura 13 um esquema resumo que ilustra o módulo de deteção de falsificação.

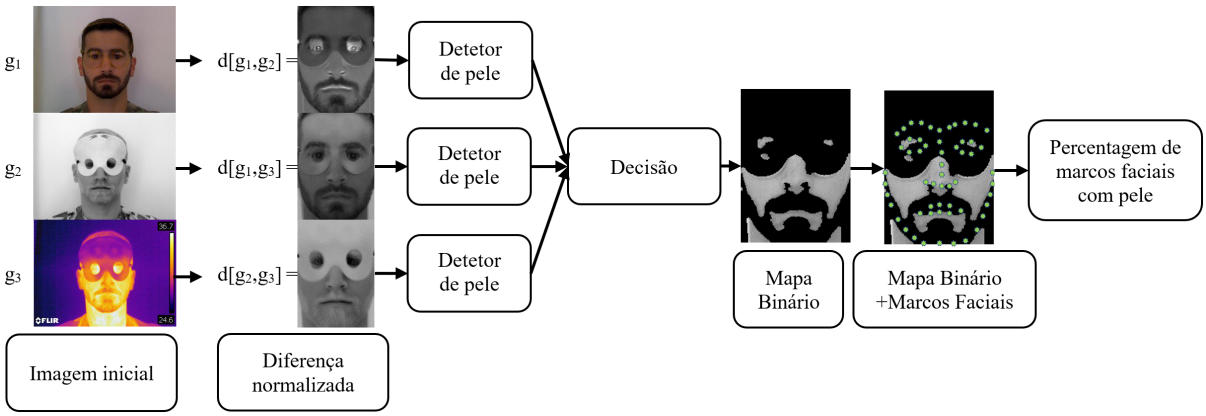


Figura 13: Esquema resumo do módulo de deteção de falsificação.

O detetor de falsificação multiespectral possui requisitos mínimos de utilização, sendo estes os seguintes: (i) as bandas espectrais e canais são os mesmos utilizados na fase de desenvolvimento do módulo, para que não ocorram erros na diferença normalizada, (ii) a gama de valores a utilizar necessita de ser definida previamente, senão não é possível detetar pele corretamente e (iii) são precisos de pelo menos duas bandas espectrais diferentes para ser possível operar corretamente, uma vez que é efetuado uma diferença normalizada entre imagens de diferentes canais.

4.3 Processamento Facial

Este módulo tem como finalidade o alinhamento facial e o redimensionamento da imagem de forma a normalizar a localização da face e a resolução da imagem introduzida para o módulo consequinte.

Inicialmente, é efetuado um alinhamento facial com o auxílio do retângulo delimitador da face e dos marcos faciais, obtidos no módulo de processamento de imagem. O alinhamento é efetuado através dos olhos da face a alinhar, o objetivo é rodar a face humana de forma a que os olhos fiquem numa linha horizontal, ou seja, garantindo que os olhos fiquem na mesma ordenada da imagem e que fiquem sempre na mesma localização.

Para o alinhamento são necessários apenas os marcos faciais dos olhos. Através da Figura 11 é possível constatar que os marcos faciais para o olho esquerdo são os marcos números 37 a 42, e para o olho direito são os marcos com os números 43 a 48. O centro de cada olho, ou centroide, é calculado através da média das coordenadas dos marcos faciais de cada olho.

Obtido o centro de cada olho é necessário calcular o ângulo de rotação entre os olhos. Para tal é calculada a diferença entre as coordenadas dos olhos e de seguida é utilizado o arco de tangente para obter-se o ângulo entre os olhos e o plano horizontal. É aplicada uma rotação da imagem utilizando o ângulo de rotação, calculado anteriormente. Esta rotação é aplicada no ponto central dos olhos, esta é a coordenada média equidistante entre os dois olhos. De seguida é efetuada uma transformação afim na imagem, tendo em consideração a localização desejada para os olhos.

Após o alinhamento facial, a imagem é sujeita a um recorte e redimensionamento. O recorte é feito na imagem que tinha sido previamente alinhada. Por fim, é efetuado o redimensionamento da imagem. A dimensão desta deve ser escolhida tendo em consideração o valor de entrada da DCNN a utilizar. No

nosso caso as imagens são redimensionadas para um tamanho de 144x144 *pixels*.

Está ilustrado na Figura 14 um esquema resumo que ilustra o módulo de processamento facial.

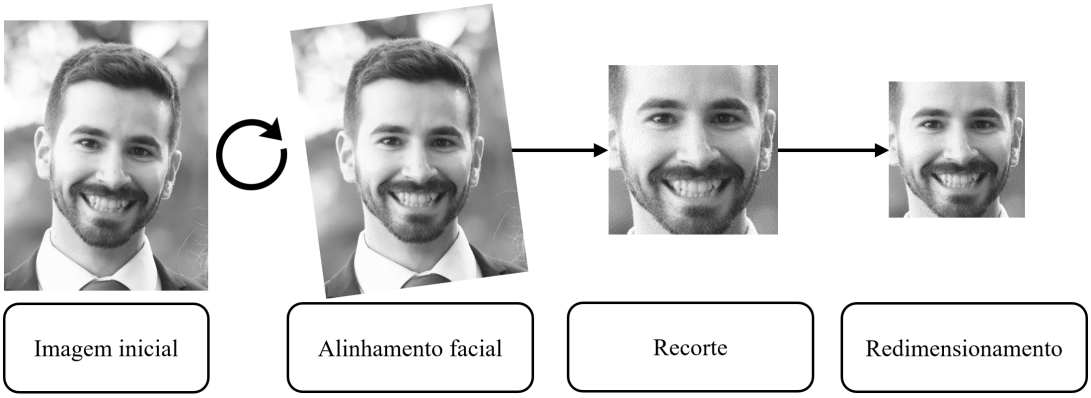


Figura 14: Esquema resumo do módulo de processamento facial.

4.4 Reconhecimento Facial

Neste módulo é proposto uma rede neuronal, através desta é efetuado uma extração do conjunto de características representativo das imagens faciais introduzidas. Seguidamente, é utilizado um classificador para efetuar a classificação dos conjunto de características extraídos afim de se obter a identidade da pessoa presente nas imagens.

4.4.1 Extração de Características

Este módulo tem como finalidade efetuar uma extração do conjunto de características representativo da pessoa a identificar através da DCNN.

De forma extrair o conjunto de caraterísticas é proposta uma rede neuronal com uma arquitectura inovadora, ilustrada na Figura 15. Através desta é possível utilizar vários canais, alocando cada canal para uma banda espectral, ou intervalo espectral (caso se esteja a utilizar vários intervalos espectrais na mesma banda).

A DCNN base utilizada por cada canal é a LightCNN [79]. Esta DCNN destaca-se de outras DCNN semelhantes pelo fato de empregar *Max-Feature Map* (MFM), uma extensão da função de ativação *maxout*, na sua arquitetura base. Através desta função de ativação, a *LightCNN* [79] obtém um número reduzido de parâmetros, como alternativa à ReLU. A rede tem como entrada imagens na gama do cinzento, de tamanho de 128x128 *pixels* e como saída um conjunto de características, representativo da identidade, da pessoa, de 256 dimensões. Esta rede possui também $12,6 \times 10^6$ parâmetros e exige cerca de $3,9 \times 10^9$ operações de ponto flutuante por segundo (FLOPS, do inglês *Floating Point Operations Per Second*).

Serão adaptadas camadas distintas da LightCNN [79] de forma a adaptar o modelo utilizado da LightCNN [79] a uma banda espectral diferente da do visível, ideia inspirada no trabalho de Pereira [11]. O canal atribuído à banda espectral do visível não será adaptado. A aprendizagem por transferência dos pesos de uma DCNN pré treinada para reconhecimento facial numa base de dados com um número elevado de imagens, é uma técnica muito utilizada para evitar o sobre ajuste, dado o número limitado

de imagens multiespectrais utilizadas na fase de treino [11]. O modelo da LightCNN [79] utilizado foi treinado em duas bases de dados, MS-Celeb-1M [62] e CASIA-WebFace [98], ambas constituídas por imagens faciais da banda espectral do visível.

Após a entrada das imagens na rede neuronal, é efetuada uma divisão de cada *pixel* por 255 para normalizar a game de valores para o intervalo entre 0 e 1.

Cada canal irá produzir um conjunto de características de 256-dimensões. Para garantir que o vetor final de características mantém a dimensão de 256 foi ampliada a arquitetura no final com uma camada. Esta camada final foi denominada de última camada ligada (UCL) e tem como entrada um conjunto de características com dimensão $N \times 256$, e como saída um vetor características de 256-dimensões.

A Figura 15 ilustra um caso genérico de utilização da rede proposta que emprega N canais. A verde estão indicada as camadas que vão ser adaptadas, e a azul as camadas que não são adaptadas. Da análise desta figura é possível observar que o *Canal 0*, atribuído à banda espectral do visível, não é adaptado, como já tinha sido mencionado.

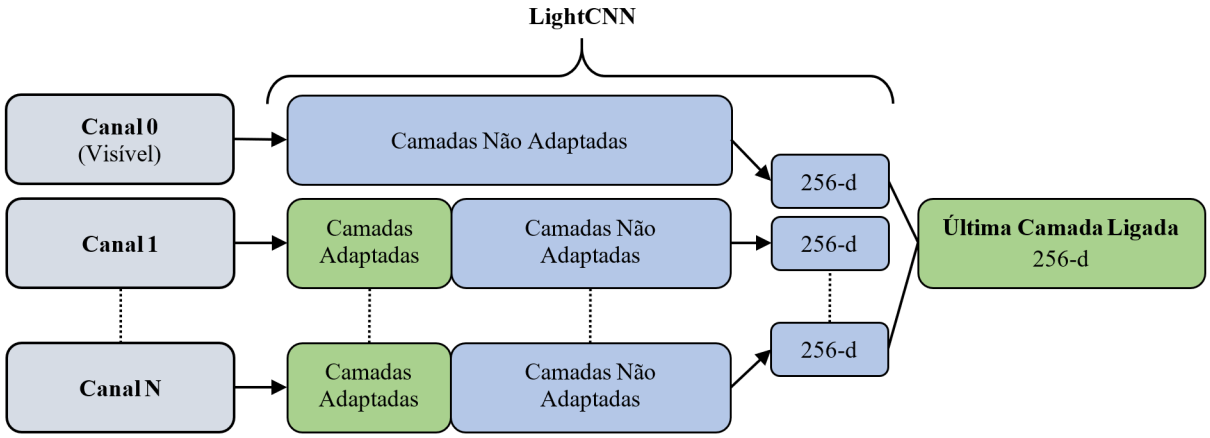


Figura 15: Esquema da arquitetura da DCNN proposta.

4.4.2 Classificador

Após obtenção do conjunto de características é necessário classificar-las de forma a obter a correspondente identidade da pessoa na imagem facial. São testados diversos classificadores afim de se averiguar qual o mais indicado a classificar o conjunto de características extraído pela rede proposta. São testados os classificadores mais utilizados, sendo estes, o *Support Vector Machine* (SVM) (com *kernel* linear ou RBF) e o *k-Nearest Neighbors* (kNN) [99].

O classificador máquina de vetores de suporte (SVM, do inglês *Support Vector Machine*) consiste numa metodologia de aprendizagem supervisionada, utilizada para a classificação estatística e análise de regressão que se baseia no conceito de planos que definem fronteiras de decisão.

A SVM constrói um hiperplano que separa os vectores de características que representam objetos de diferentes classes, e que maximiza a distância entre os pontos de classes diferentes. Baseia-se na ideia de que quanto maior for a distância entre as classes, menor será o erro da classificação, embora esta hipótese dependa dos dados utilizados e não se verifique exactamente. O número de hiperplanos pode variar, consoante o número de classes.

Em certos problemas de classificação, a distribuição dos dados não permite uma separação linear dos dados entre as classes. Uma possível solução para este problema consiste em mapear os dados num espaço com maior dimensionalidade, $\tilde{x}_i = \phi(x_i)$, onde x_i corresponde às características iniciais, e a função ϕ é tal que torna os pontos transformados, $\tilde{x}_i$, mais próprios, para ser separados por um hiperplano.

O algoritmo não necessita de saber explicitamente a função ϕ mas apenas o produto interno dos vectores transformados $\phi(x_i)^T\phi(x_j)$. Estes produtos internos podem ser calculados utilizando funções de *kernel* da forma $k(x_i, x_j) = \phi(x_i).\phi(x_j)$. As funções de *kernel* mais comuns são a linear (Equação 7) e a função de base radial (RBF, do inglês *Radial Basis Function*) (Equação 8).

$$\text{Kernel Linear: } k(x_i, x_j) = {x_i}^T x_j \tag{7}$$

$$\text{Kernel RBF: } k(x_i, x_j) = e^{-\gamma\|x_i + x_j\|^2} \tag{8}$$

O último classificador testado é o de k-vizinhos mais próximos (kNN, do inglês *k-Nearest Neighbors*). O algoritmo kNN é um dos classificadores mais simples é dos mais utilizados e apresenta bons resultados na resolução de problemas de classificação em sistemas de reconhecimento facial. Dado um vetor de características, este classificador atribui-lhe uma classe com base no cálculo da distância aos k vectores de características do conjunto de treino mais próximos, sendo escolhido aquele que obtiver a classe com maior frequência absoluta [99].

O desempenho deste classificador depende essencialmente do número de vizinhos a considerar (k) e a métrica de cálculo de distâncias escolhidas. Se k=1, então o vetor de características é atribuído à classe do vizinho mais próximo. As métricas de distâncias utilizadas habitualmente para este classificador são: euclidiana, *manhattan*, *chebyshev* [99].

Num sistema de reconhecimento facial a métrica mais utilizada é a distância euclidiana. Esta é definida pela distância entre dois pontos, P e Q num espaço-n [100]. Em coordenadas cartesianas, se P = (p_1, p_2, ..., p_n) e Q = (q_1, q_2, ..., q_n) forem dois pontos num espaço euclidiano então a distância entre os pontos P e Q pode ser calculada por,

$$d(P, Q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{9}$$

A Figura 16 ilustra um esquema resumo que exemplifica o módulo de reconhecimento facial, desde da entrada de imagens, para cada canal, até à identificação da identidade da pessoa presente nessas mesmas imagens.

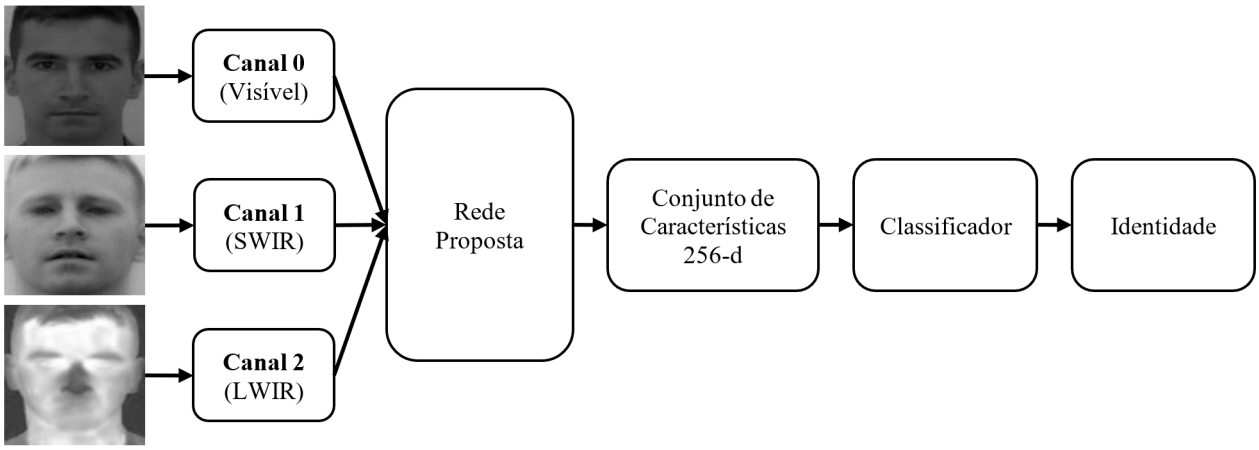


Figura 16: Esquema resumo do módulo de reconhecimento facial.

Para o esquema resumo, presente na Figura 16, foram empregue três canais, onde cada canal possui uma banda espectral distinta, a VIS, SWIR e LWIR, respetivamente. A rede proposta extrai o conjunto de características de 256-dimensões de três imagens. Após a extração, o classificador empregue vai utilizar esse mesmo conjunto para identificar a identidade da pessoa.

5 Resultados e Discussão

De forma a avaliar e comparar o desempenho da metodologia proposta com as técnicas do estado da arte foram realizados diversos testes. Neste capítulo serão enunciadas, apresentadas e discutidas as experiências efetuadas durante a realização da dissertação de mestrado.

Na Secção 5.1 são apresentadas as bases de dados multiespectrais utilizadas, assim como a razão da sua escolha. Na Secção 5.2, são descritas as experiências efetuadas para avaliar a robustez dos módulos de processamento de imagem e processamento facial. De seguida, na Secção 5.4, são descritas as experiências efetuadas para comparar os diversos detetores de falsificação utilizados. Na Secção 5.4 é explanado o processo de treino e avaliação da DCNN e a escolha das camadas adaptadas. É efetuado, adicionalmente, um estudo em relação ao classificador empregue, com a finalidade de escolher o classificador e os hiperparâmetros mais adequados para as bases de dados multiespectrais utilizadas.

Por último, na Secção 5.4.4 procede-se a uma avaliação do desempenho geral da metodologia proposta com outras metodologias do estado da arte, para cada uma das bases de dados multiespectrais utilizadas.

Como produto desta dissertação, o respetivo código para cada módulo está publicados em repositório púbico, na plataforma github[5] de forma a que qualquer um possa aceder e usar no seu projeto de reconhecimento facial multiespectral. No entanto, para se ter acesso às bases de dados multiespectrais utilizadas nesta dissertação de mestrado é necessário solicitar aos autores das mesmas.

5.1 Bases de Dados

De forma a avaliar corretamente os algoritmos empregues na metodologia proposta foram utilizados três bases de dados multiespectrais. Nesta secção serão descritas as três bases de dados, empregues durante os testes: a Tufts [101], a CASIA NIR-VIS 2.0 [8] e um conjunto de imagens adquiridas no AAMA.

Antes das bases de dados multiespectrais poderem ser utilizadas, é necessário proceder-se a uma limpeza e pré-processamento da mesma. A limpeza da base de dados tem em vista a exclusão de imagens inutilizáveis. Após a mesma, é efetuado um pré-processamento de imagens, sendo este composto pela deteção e por um alinhamento facial. Por último, procede-se ao redimensionamento das imagens presentes (já limpas) nas bases de dados multiespectrais.

5.1.1 CASIA NIR-VIS 2.0

A base de dados multiespectral CASIA NIR-VIS 2.0 [8] foi construída na *China Academy of Sciences Institute of Automation*, e as imagens faciais das pessoas que a constituem são predominantemente de etnia oriental. Esta base de dados é composta por duas bandas espectrais, VIS e NIR. Esta base de dados é composta por imagens da base de dados multiespectral CASIA-HFB [33], daí o nome 2.0.

Foi dada preferência a esta base de dados multiespectral, em relação a outras bases de dados, já que esta é empregue por um número elevado de investigadores e, por isso, uma boa forma de comparar

[5] `https://github.com/Cloroo/MultispectralFaceRecognitionSyS`.

diferentes metodologias.

Aquando a limpeza da base de dados, deparamo-nos apenas com pessoas cujas imagens não possuíam as duas bandas espectrais. De forma a resolver este problema, seis pessoas foram retiradas da base de dados multiespectral.

Após a limpeza da base de dados esta ficou com 17 489 imagens de 715 pessoas, removendo 23 imagens de 6 pessoas. Na Figura 17 estão ilustradas algumas imagens da base de dados multiespectral CASIA NIR-VIS 2.0 [8].

Figura 17: Ilustração de imagens da base de dados CASIA NIR-VIS 2.0 [8].

5.1.2 Tufts

A base de dados multiespectral Tufts [101] foi construída na Universidade de Tufts, nos Estados Unidos da América (EUA). Esta ao contrário da anterior, é caracterizada pela presença de imagens que representam uma maior diversidade étnica, reproduzindo uma diversidade que se aproxime com a realidade, e não sobre ajustar os dados a uma dada etnia. Esta base de dados é composta por três bandas espectrais: VIS, NIR e LWIR.

Antes da utilização da base de dados Tufts [101], foi necessário igualmente proceder a uma limpeza. Numa primeira fase, as imagens em falta e inutilizáveis (i.e., corruptas ou desfocadas) foram excluídas. Antes da limpeza da base de dados foram definidos os seguintes requisitos para a exclusão de imagens: imagens desfocadas, luminosidade reduzida ou nula, sombras e pessoas cujas imagens obtidas não possuam todas as bandas espectrais (e.g., VIS-NIR-LWIR). Na Figura 18 encontram-se exemplos de imagens excluídas da base de dados Tufts [101].

(a) Luminosidade reduzida. (b) Luminusidade nula. (c) Sombra. (d) Desfocado.

Figura 18: Exemplos ilustrativo de imagens excluídas da base de dados multiespectral Tufts [101].

Efetuada a limpeza, esta ficou com um total de 7 675 imagens de 109 pessoas, removendo assim 53 imagens e 4 pessoas. Na Figura 19 estão ilustradas algumas imagens da base de dados multiespectral Tufts [101].

Figura 19: Ilustração de imagens da base de dados multiespectral Tufts [101].

5.1.3 Academia Militar

De forma a validar e testar o módulo proposto para deteção de falsificação, foi necessário elaborar uma base de dados multiespectral, constituída por três bandas espectrais, VIS, SWIR e LWIR. A aquisição destas imagens foi efetuada através de equipamentos de captação de imagens disponibilizados pela Academia Militar (AM) no AAMA. As câmaras foram adquiridas durante a execução dos projetos Fusão de Imagem Militar (FUSIMIL) e Fusão das Imagens do Visível e do Infravermelho (FIVE), cujas especificações estão presentes na Tabela 4.

Tabela 4: Especificações das câmaras utilizada para a composição da base de dados, pertencentes à Academia Militar.

Nome do Equipamento	**Resolução [*pixels*]**	**Intervalo Espectral [nm]**
Flir T440BX Câmara de VIS e LWIR	320 x 240	(450-520), (515-600), (600-690) e (7 500-13 000)
WiDy 640V-S Câmara de SWIR	640 x 512	(900-1 700)

Para o estudo dos algoritmos empregues no módulo de deteção de falsificação foram utilizadas máscaras de diferentes materiais: plástico, papel, e de feltro. De realçar que a caraterística relevante das máscaras utilizadas não é o seu formato ou cor, mas sim o material de que são feitas.

Na Figura 20 estão presentes as máscaras utilizadas e também o material de que é feito.

(a) Máscara A (Plástico)

(b) Máscara B (Papel)

(c) Máscara C (Feltro)

Figura 20: Máscaras utilizadas durante os testes do módulo de deteção de falsificação.

A Figura 21 ilustra a disposição das câmaras multiespectrais e do utilizador. A distância das câmaras ao utilizador era de 1,5 metros, com o intuito de simbolizar a distância normal entre um utilizador e um sistema de reconhecimento facial.

Figura 21: Disposição das câmaras multiespectrais e da pessoa durante a aquisição de imagens.

Na Figura 22 estão ilustradas algumas imagens presentes na base de dados multiespectral elaborada durante a dissertação de mestrado.

Figura 22: Exemplo ilustrativo das imagens adquiridas presentes na base de dados multiespectral construída na Academia Militar.

5.2 Deteção e Alinhamento Facial

Com o intuito de confirmar a robustez, e consequentemente espelhar as limitações existentes nos módulos de processamento de imagem facial e da metodologia proposta, foram efetuados diversos testes. Estes, têm como objetivo refletir o impacto da pose do utilizador na deteção facial, na extração dos marcos faciais e, por último, no alinhamento facial.

Apesar do sistema de reconhecimento facial multiespectral proposto ser empregue num local onde seja expectável a colaboração do utilizador, foram testadas diversas poses com o intuito de ilustrar a falta e incorreta colaboração. Na realização dos testes foram utilizadas 17 poses diferentes: uma frontal, quatro horizontais, quatro verticais e oito poses diagonais. As combinações dos ângulos utilizadas foram respetivamente 0°, 45° e 90°, na horizontal e na vertical. Onde o ângulo de 0° corresponde à pose frontal do utilizador. Ângulos de 45° são consideradas como limite máximo sugerido pelos sistemas de reconhecimento facial. Os ângulos de 90° são consideradas casos extremos e difíceis, e como tal, é expectável obter resultados inferiores para estes ângulos. Na Figura 23 encontra-se ilustrado as poses utilizadas na aquisição de imagens.

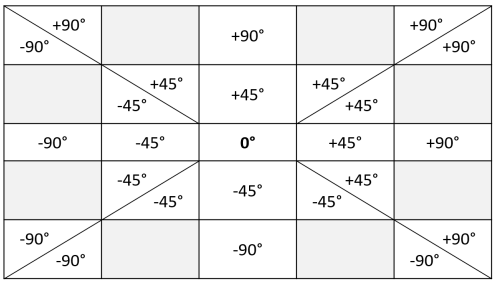


Figura 23: Ilustração das diversas combinações de pose.

Na Figura 24 encontram-se ilustradas as imagens adquiridas nas diferentes poses a uma distância de 1,50 metros. As imagens foram adquiridas através da câmara fotográfica de um telemóvel *Samsung S9+* com uma resolução de 4032x2268 *pixels*.

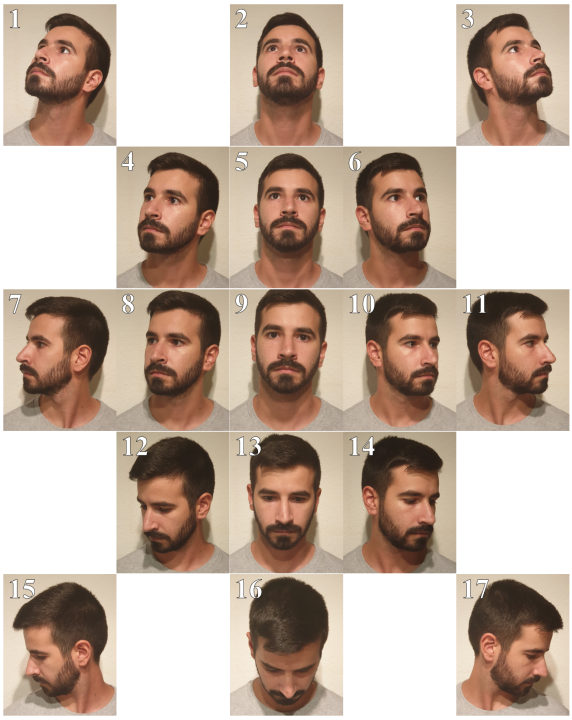


Figura 24: Ilustração das imagens utilizadas nos testes de deteção, extração de marcos faciais e alinhamento facial.

Na Figura 25 está ilustrada a sequência de testes a efetuar às 17 imagens adquiridas. No primeiro teste é aferida a capacidade do algoritmo de proceder à deteção facial nas diferentes poses. A deteção facial irá produzir um retângulo vermelho que irá conter a face da pessoa na imagem facial. Na localização deste retângulo serão extraídos os marcos faciais, representados a verde na imagem. Após a extração dos marcos faciais, é efetuado um alinhamento facial da face da pessoa detetada.

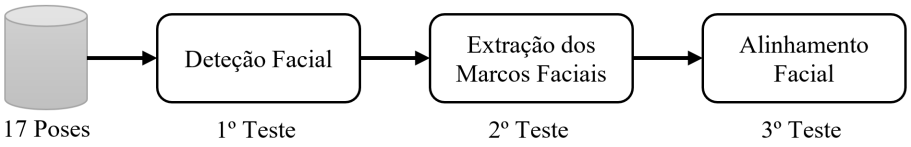


Figura 25: Sequência dos testes com 17 imagens em diferentes poses.

A capacidade do algoritmo de efetuar a deteção facial será, igualmente, testada em outras bandas espectrais. Para este estudo serão utilizadas as bases de dados multiespectrais Tufts [101] (VIS, NIR e LWIR) e CASIA NIR-VIS 2.0 [8] (VIS e NIR).

Para este conjunto de imagens, ilustrado na Figura 24, foi aplicado o algoritmo de deteção facial. A Figura 26 ilustra o resultado obtido pelo algoritmo de deteção facial, representando por um retângulo vermelho. Analisando as imagens resultantes da deteção facial, Figura 26, é possível observar que o algoritmo detetou com sucesso a totalidade do conjunto de imagens, independentemente da pose representada na imagem. Apesar da correta deteção da face pelo algoritmo, uma análise mais cuidadosa das imagens que ilustram as poses 1 e 3 permite constatar que este foi incapaz de enquadrar corretamente o retângulo (assinalado a vermelho na figura) da deteção facial na face. A incorreta marcação do retângulo facial pode ser um fator limitativo para os próximos módulos.

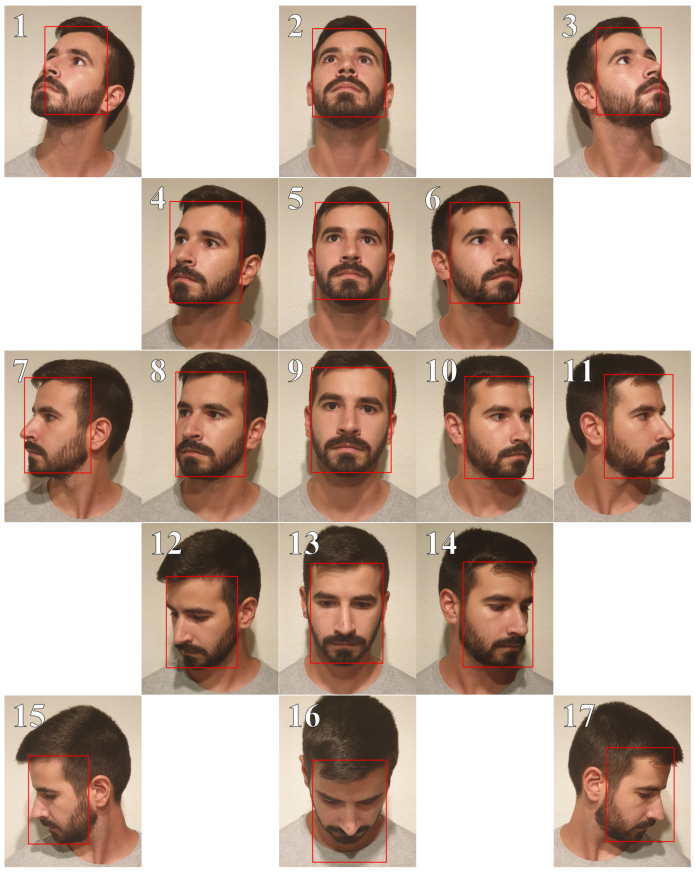


Figura 26: Localização da deteção facial nas imagens iniciais.

Através da informação da localização da face (i.e., em coordenadas (x,y)), obtida pelo algoritmo de deteção facial, é realizado uma extração dos marcos faciais. No Apêndice B são disponibilizados os resultados obtidos após extração dos marcos faciais de forma a permitir uma melhor compreensão destes resultados. Ainda assim, na Figura 27 estão selecionadas algumas imagens representativas de uma correta e incorreta extração de marcos faciais. É possível afirmar que se obteve uma correta extração de marcos faciais quando estes são sobrepostos com os pontos de referência produzidos pela dlib (i.e., imagem padrão, Figura 11) e que estes não se encontram excessivamente distantes.

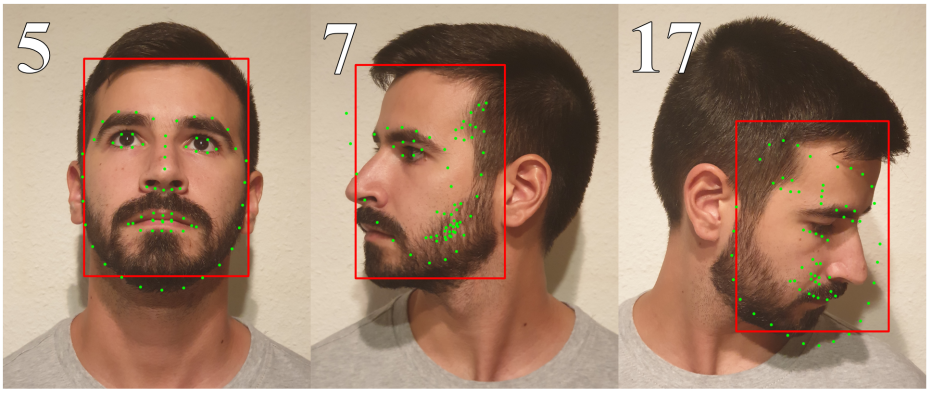


Figura 27: Localização dos marcos faciais para as poses número 5, 7 e 17, respetivamente.

Na pose número 5, da Figura 27, é possível observar que os marcos faciais foram extraídos corretamente, pois estão iguais à localização padrão. Contudo, para as poses número 7 e 17, todos os marcos faciais foram extraídos incorretamente. Como razões justificativas para este resultado temos que: houve uma incorreta localização da face, confirmado por outros testes aquando da fase de experimentação do algoritmo de extração de marcos faciais, e também pelo facto de o extrator de marcos faciais ter dificuldades em extrair os marcos ocultos (i.e., que não estejam visíveis na imagem). Como já supramencionado, o retângulo que contém a predição da face da pessoa (a vermelho na Figura 27), quando pequeno e incorretamente localizado, condiciona o algoritmo que efetua a extração dos marcos faciais.

Após análise do Apêndice B em detalhe, para as poses número 8 e 10 é possível observar que apesar de os marcos faciais do olho direito e esquerdo, respetivamente para cada imagem, tenham sido corretamente extraídos, o algoritmo de extração não consegue identificar corretamente os marcos faciais do restante olho, colocando a aproximação bastante próxima dos marcos corretamente extraídos. Do mesmo modo, é possível observar que, durante a extração dos marcos faciais dos olhos estes possuem como referência a esclera (i.e., elemento branco do olho humano), e que, quando esta não está visível na imagem os marcos faciais não são tão precisos.

Em conjunto com a informação da localização da face nas imagens e dos marcos faciais, é agora possível efetuar o alinhamento, recorte e redimensionamento das imagens. As imagens são alinhadas através dos marcos faciais dos olhos extraídos na fase anterior. Após o alinhamento facial é efetuado o recorte, e de seguida o redimensionado de imagem para uma resolução de 144x144 *pixels*. Na Figura 28 está ilustrado o produto final do alinhamento e redimensionamento de imagem.

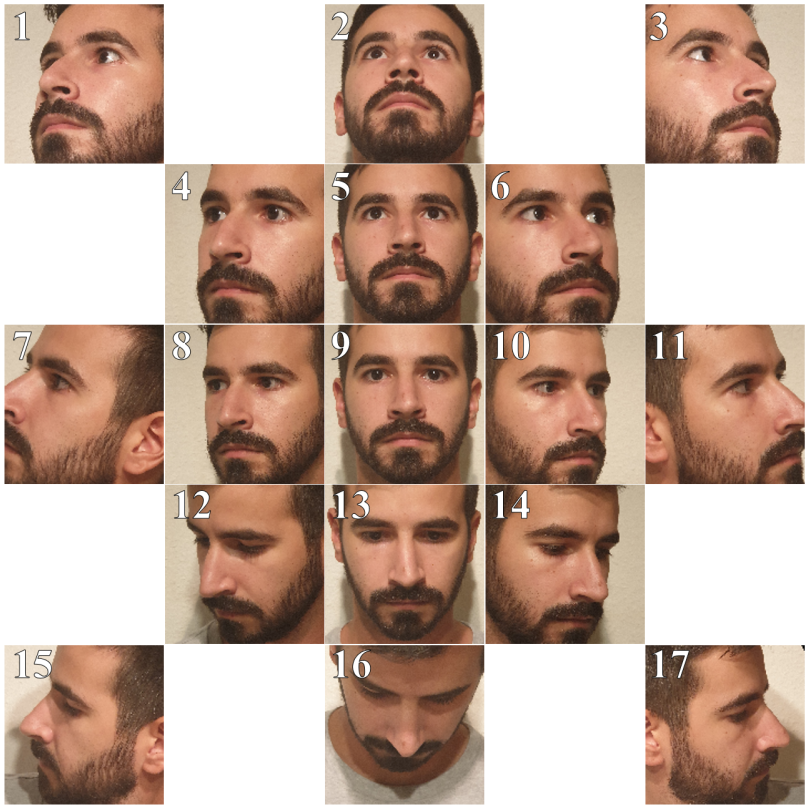


Figura 28: Ilustração do produto final do alinhamento facial e redimensionamento de imagem.

Após uma análise da Figura 28, é possível afirmar que nas 17 imagens faciais iniciais, foi possível efetuar um alinhamento facial correto, tendo em consideração as diferentes poses existentes nas imagens. É de salientar que, as imagens 7 e 11 foram as que obtiveram o pior alinhamento, excluindo grande parte da face, justificado pelo facto de, durante a extração dos marcos faciais a predição dos marcos faciais foram bastante diferentes do espectável, como já tinha sido supramencionado durante a discussão do teste anterior. Numa segunda análise à Figura 28, é possível afirmar que o alinhamento facial nas imagens é constante, isto é, a face encontra-se centrada no local esperado, independentemente da pose utilizada.

Efetuados os três estudos que ilustram as limitações dos algoritmos empregues é possível retirar as seguintes conclusões: (i) apesar da incorreta colocação do retângulo de deteção em algumas imagens faciais o algoritmo de deteção facial obtém os resultados desejáveis, independentemente da pose do utilizador (desde que esta esteja compreendida entre 0° e 90°) , (ii) apesar de alguns dos marcos faciais extraídos estarem indevidamente marcados, estes não vão inviabilizar o produto final do alinhamento facial.

Assim se concluí possível concluir que o principal entrave ao módulo de processamento de imagem, e consequentemente aos módulos em diante, será o algoritmo de deteção facial, na medida em que uma deteção facial incorreta irá condicionar os restantes módulos.

5.2.1 Bases de Dados Multiespectrais

Como tinha sido descrito previamente na Secção 5.2, será efetuado um teste nas bases de dados com os algoritmos de deteção e alinhamento facial utilizados na metodologia proposta.

É importante realçar que o modelo da rede neural utilizada no algoritmo que efetua a deteção facial foi treinada unicamente com imagens faciais na banda espectral do visível. Nas imagens faciais, de todas as bandas espectrais utilizadas, onde não tenha sido possível efetuar uma deteção facial automática foi necessário efetuar uma deteção facial manual individual. As imagens detetadas e alinhadas neste estudo serão utilizadas posteriormente no Secção 5.4.

Na Tabela 5 encontra-se disponível a relação de imagens corretamente detetadas e não detetadas para cada banda espectral da base de dados Tufts [101].

Tabela 5: Relação de imagens corretamente detetadas e não detetadas para cada banda espectral, para a base de dados Tufts [101].

Banda Espectral	**Detetadas**	**Não Detetadas**	**Total**	**Percentagem**
VIS	3 203	93	3 296	97,2 %
NIR	2 949	459	3 408	86,5 %
LWIR	775	196	971	79,8 %
Total	**6 927**	**748**	**7 675**	**90,3 %**

Numa primeira análise à Tabela 5, é possível constatar que o detetor facial obteve melhores resultados para a banda espectral VIS. Como supramencionado, o modelo da rede neural foi treinado com imagens no visível, e, como tal, é espectável que para bandas espectrais diferentes do visível o número de imagens

faciais detetadas seja inferior.

De uma análise mais detalhada da informação da Tabela 5 é possível verificar que, a banda espectral NIR obtém significativamente melhores resultados (86,5%) que a banda espectral LWIR (79,8%). Este resultado é justificado pela maior proximidade da banda espectral NIR ao visível, quando comparado com a banda espectral LWIR.

Na Tabela 6 encontra-se disponível a relação de imagens corretamente detetadas e não detetadas para cada banda espectral da base de dados CASIA NIR-VIS 2.0 [8].

Tabela 6: Relação de imagens corretamente detetadas e não detetadas para cada banda espectral, para a base de dados CASIA NIR-VIS 2.0 [8].

Banda Espectral	**Detetadas**	**Não Detetadas**	**Total**	**Percentagem**
VIS	5 013	56	5 069	99,0 %
NIR	11 438	955	12 393	92,3 %
Total	**16 451**	**1 011**	**17 462**	**94,2 %**

A Tabela 6 corrobora a informação já indicada na Tabela 5 para as bandas espectrais VIS e NIR. Para esta base de dados multiespectral foram detetadas um número superior de imagens. Isto deve-se ao facto de as imagens faciais multiespectrais presentes serem de resolução inferior. A base de dados multiespectral CASIA NIR-VIS 2.0 [8] possui imagens na resolução de 640x480 *pixels*. Comparativamente, a base de dados multiespectral Tufts [101] possui imagens na resolução de 3280x2464 *pixels*.

Após a deteção facial todas as imagens foram redimensionadas e alinhadas para uma resolução de 144x144 *pixels*. Pelo fato de ser utilizado um alargamento de dados, a imagem facial é redimensionada para uma resolução de 144x144 *pixels* em vez dos 128x128 *pixels* esperados, isto deve-se ao requisito mínimo da arquitetura proposta. Este alargamento de dados (abordado posteriormente na Secção 5.4) irá depois redimensionar as imagens para uma resolução de 128x128 *pixels*. No momento da análise do produto final do alinhamento facial, é possível afirmar que as imagens obtidas ficaram corretamente alinhadas.

5.3 Deteção de Falsificação

O próximo módulo a ser estudado é o de deteção de falsificação, módulo este já fundamentado no capítulo da metodologia. Os testes efetuados a este módulo têm como finalidade comprovar a vantagem da utilização de imagens multiespectrais em metodologias de deteção de ataques de falsificação.

O detetor de falsificação multiespectral proposto aplica uma filtragem de pele na imagem que emprega todas as bandas espectrais disponíveis, de forma a classificar pele ou não pele na imagem. De forma a comprovar a eficiência do detetor de pele multiespectral proposto, é efetuada uma comparação com outros dois detetores de pele, o detetor de pele YCbCr e o HSV. De seguida, é feita uma comparação entre o que foi detetado como sendo pele e os marcos faciais, extraídos no módulo de processamento de imagem. Se o número de marcos faciais que foram considerados pele for inferior a 75% então foi detetado um possível ataque de falsificação.

5.3.1 Detetores de Pele YCbCr e HSV

No detetor de pele YCbCr, é utilizado a gama de cores YCbCr (Y= Luminância, Cb= Crominância Azul, Cr=Crominância Vermelha) permitindo determinar na imagem representada o que é pele. Neste detetor são utilizados exclusivamente imagens na banda espectral do visível, na gama de cores RGB. De forma a aplicar o detetor, é necessário converter primeiro a imagem para a gama de cores YCbCr. Esta conversão, RGB → YCbCr, é efetuada através de uma função disponibilizada pela *OpenCV*.

Da mesma forma, para o classificador de pele HSV é utilizado a gama de cores HSV (H= Matiz, S= Saturação, V=Valor) de forma a determinar na imagem o que é pele. À semelhança do detetor anterior, este também utiliza apenas imagens na banda espectral do visível. Para a aplicação do detetor de pele, é necessário a conversão prévia da imagem para a gama de cores HSV. Esta conversão, RGB → HSV, é efetuada também através de uma função disponibilizada pela *OpenCV*.

A gama de valores utilizada para detetar o que é pele foi obtida através de outros trabalhos. Para o detetor de pele YCbCr foram considerados como pele a gama de valores (35, 138, 70) < (Y, Cb, Cr) < (255, 178, 133) [102] [103]. Do mesmo modo para o detetor de pele HSV, foram considerados como pele a gama de valores (0, 71, 50) < (H, S, V) < (20, 173, 255) [103].

Foram efetuados diversos testes utilizando as máscaras presentes na Figura 20. Na análise dos resultados será efetuada uma discussão mais detalhada para a máscara C (de feltro), e uma discussão superficial para as restantes máscaras. Foi dado seguimento a esta abordagem pelo facto de os resultados obtidos para as restantes máscaras serem semelhantes à máscara de feltro.

Foram aplicados os detetores de pele YCbCR e HSV com o intuito de se obter apenas da imagem introduzida originalmente o que era considerado pele. Na Figura 29 está ilustrada a imagem inicial utilizada e os produtos dos detetores YCbCr e HSV, Figura 29b e 29c, respetivamente. Nestas estão assinalado a preto o que foi considerado como não sendo pele.

(a) Imagem Inicial (b) Classificador YCbCr (c) Classificador HSV

Figura 29: Resultado final após aplicação dos detetores de pele YCbCr e HSV.

Numa primeira análise à Figura 29, é possível concluir que o detetor de pele HSV foi aquele que descartou maior quantidade de informação da imagem. Em contrapartida, o detetor de pele YCbCr considerou, incorretamente, como pele diversas zonas, como por exemplo na periferia das orelhas, e em áreas de barba próximas da pele.

Como é possível observar na Figura 29, ambos os detetores de pele não foram capazes de efetuar uma distinção correta entre o que é verdadeiramente pele e máscara, classificando assim incorretamente

a máscara como pele. Os resultados obtidos para este ataque, são consistentes com aqueles obtidos para outras tipologias de ataques utilizados (i.e., outras máscaras).

5.3.2 Detetor de Pele Multiespectral

Para o detetor de pele multiespectral, é necessário proceder à diferença normalizada nas imagens. Para tal, é necessário que estas estejam devidamente alinhadas. Como as imagens são adquiridas com diferentes resoluções é necessário proceder ao seu alinhamento prévio antes de se efetuar a diferença normalizada. O método utilizado para o efeito, neste módulo, consiste em encontrar a resolução média do retângulo da deteção facial e redimensionar as três imagens para esta resolução. Para que este alinhamento possa funcionar corretamente, é necessário que as imagens tenham sido adquiridas num curto espaço tempo e com um enquadramento semelhante (i.e., tamanho e pose da face tem de ser semelhante nas duas imagens no momento de aquisição).

A escolha da gama de valores para classificar como pele ou não pele, foi feita empiricamente, sendo estes definidos para cada diferença normalizada. Os valores para pele estão compreendidos entre (76, 51, 65) $<$ ($d[g_1, g_2]$, $d[g_1, g_3]$, $d[g_2, g_3]$) $<$ (131, 140, 127). Esta gama de valores foi ajustada de forma a ter uma aplicação geral e não apenas para casos particulares (i.e., numa secção de imagens específicas), como é o caso dos detetores de pele anteriores. Este ajuste de valores foi possível através de diversas experimentações, onde foram tiradas imagens em diversas sessões e com exposição a diferentes iluminações.

Na Figura 30 está ilustrado a diferença normalizada entre as imagens na banda espectral do VIS com o SWIR ($d[g_1, g_2]$), a diferença normalizada entre as imagens na banda espectral do VIS e do LWIR ($d[g_1, g_3]$) assim como a diferença normalizada entre as imagens na banda espectral do SWIR com o LWIR ($d[g_2, g_3]$), respetivamente.

Figura 30: Ilustração da diferença normalizada nas diferentes imagens $d[g_1, g_2]$, $d[g_1, g_3]$ e $d[g_2, g_3]$.

Após uma análise detalhada da Figura 30, é possível observar que o alinhamento de imagem foi efetuado com sucesso. Um incorreto alinhamento teria sido identificado pela existência de "sombras" significativas em zonas especificas onde são facilmente visíveis. Para este caso concreto, estas zonas especificas seriam no pescoço e nos olhos.

Obtido um bom alinhamento das imagens é possível, agora, aplicar o detetor de pele multiespectral na imagem de cada diferença normalizada. Após a deteção de pele é feita uma decisão entre as diferentes deteções. O resultado desta decisão é um mapa binário que contém os *pixels* que foram considerados como

pele em todas as diferenças normalizadas. Esta máscara é um mapa binário, onde o "1" corresponde a pele e "0" corresponde a não pele. Na Figura 31 está ilustrado o antes, imagem à esquerda, e o mapa binário após a aplicação do detetor de pele multiespectral, imagem à direita. À semelhança de outros detetores de pele utilizados, está assinalado a preto o que foi classificado como não sendo pele.

Figura 31: Ilustração do resultado final após aplicação do detetor de pele multiespectral proposto.

Numa primeira análise ao detetor de pele multiespectral, é possível observar que o detetor classificou corretamente o que era pele e o que não era, não classificando a máscara como sendo pele (ao contrário dos detetores de pele YCbCr e HSV).

5.3.3 Comparação entre Detetores de Falsificação

Através dos marcos faciais, obtidos no módulo anterior, e as deteções de pele é possível verificar se existe um ataque de falsificação nas imagens utilizadas. Para tal, é feita uma comparação entre os marcos faciais e o mapa binário. Esta comparação é através da percentagem de marcos faciais que são considerados como pele. Se a percentagem de marcos faciais considerados pele for inferior a 75% estamos perante um ataque de falsificação. Na comparação entre os detetores de pele, é classificado como uma deteção de falsificação correta quando a imagem utilizada possui uma máscara e este consegue detetar corretamente o ataque de falsificação.

Após análise dos resultados experimentais, para os três detetores de pele, é possível afirmar que os detetores de pele que utilizam a banda espectral do visível tem dificuldades em diferenciar a pele real (i.e., pele humana) de pele falsificada, proveniente das máscaras utilizadas.

Quando utilizados estes classificadores de pele, da banda espectral do visível, como método de deteção de falsificação, a taxa de deteção de falsificação (i.e., quando este consegue detetar corretamente um ataque de falsificação) foi apenas de 13%. Resultado justificado pelo facto de os classificadores utilizados não serem capazes de efetuar uma correta diferenciação da pele humana da máscara.

Quando comparados com o classificador que utiliza as três bandas espectrais, este obtém melhores resultados. A taxa de deteção de falsificação foi agora de 83%. Através destes resultados experimentais é também possível concluir que a excessiva luminosidade nas imagens da banda espectral visível influência, pela negativa, o resultado final do classificador.

5.4 Reconhecimento Facial

Neste capítulo são apresentados os resultados dos diversos testes efetuados relativos à rede de extracção de características de identidade e ao classificador final de identidade.. Numa primeira fase, são efetuados testes com o intuito de descobrir qual o conjunto de camadas da *LightCNN* [79] que devem ser adaptadas. Conhecido o melhor conjunto de camadas a adaptar, são então extraídos os conjuntos de características de cada imagem facial, produzindo assim um código interno exclusivo à pessoa presente na imagem facial.

Extraídos os conjuntos de características, é necessário agora classificá-los com o objetivo de se obter a identidade dessa pessoa. Foram escolhidos três classificadores para a tarefa: SVM-Linear, SVM-RBF e kNN. O próximo teste a efetuar será o de determinar quais os hiperparâmetros mais adequados para cada classificador, sendo cada um específico para cada base de dados multiespectral utilizada.

O terceiro teste efetuado será realizado com os valores dos hiperparâmetros mais adequados. Cada classificador é treinado com estes valores. É efetuada uma comparação entre os classificadores através de uma curva CMC e da pontuação em *rank-1* obtido para um conjunto de teste. Deste modo, é possível realizar uma comparação justa entre os classificadores, obtendo assim o melhor resultado (para cada base de dados multiespectral). O classificador que se revelar mais adequado, será utilizado para comparar a metodologia proposta com trabalhos de outros autores.

Antes de se iniciarem os testes ao módulo de reconhecimento facial, é necessário proceder à divisão das bases de dados em diferentes conjuntos, nomeadamente, treino, validação e teste. A percentagem de imagens de teste foi de 20% e de 80% para o conjunto de treino. De forma a obter os hiperparâmetros corretos, sem se sobre ajustar aos dados, foi dividido o conjunto de dados de treino em dois, um para o treino, com 64% das imagens do conjunto original, e outro para a validação, com 16% das imagens do conjunto original.

Como o número de imagens por pessoa não é igual nas bases de dados utilizadas, foi imperativo efetuar uma divisão estratificada da base de dados. A divisão estratificada procura garantir uma representação equitativa de cada pessoa em cada subconjunto, como ilustrado na Figura 32 pelas identidades A, B e C. Se porventura, o número de imagens por pessoa fosse igual nas bases de dados, tal divisão não seria necessária, uma vez que cada classe teria o seu número de imagens já equitativamente distribuídas por cada conjunto de dados.

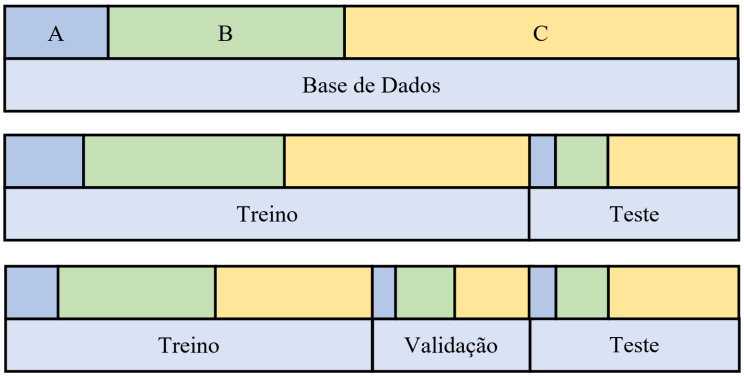


Figura 32: Exemplo ilustrativo da divisão estratificada de uma base de dados com três identidades (A, B e C) nos conjuntos de dados de treino, validação e teste. Na figura superior está ilustrado as identidades existentes na base de dados, onde o tamanho da barra corresponde à percentagem de identidades. Nas subsequentes figuras é possível observar a distribuição das identidades nos diferentes conjuntos.

5.4.1 Treino e Avaliação da Rede

Neste capítulo serão abordados as funções e os parâmetros utilizados durante o treino e avaliação da DCNN.

A DCNN foi implementada e treinada em *Pytorch*[6] na linguagem computacional *Python*. De forma a utilizar as imagens das bases de dados multiespectrais na DCNN foi necessário agrupar as imagens em diretorias segundo a classe (i.e., identidade da pessoa) a que pertencem, para o efeito foi desenvolvido código que permitisse criar uma estrutura de dados.

O tamanho do lote de imagens, vulgarmente conhecido em inglês por *batch size*, foi selecionado de forma a que os números de imagens por lote fosse o maior possível, evitando que a GPU ficasse sem memória durante a fase de treino. Foi preciso garantir, no entanto, que o número escolhido para o lote de imagens seja um expoente de 2[7], como é sugerido por [104] e [105]. Para a base de dados Tufts [101] e CASIA NIR-VIS 2.0 [8] foi utilizado um lote de imagens de 16 e 32, respetivamente.

Para efeitos de treino da DCNN foi selecionado o algoritmo de otimização Adam (em inglês *adaptative moment estimation*). Este, foi introduzido por Kingma *et al.* [106] e ao contrário do gradiente estocástico (outro algoritmo de otimização utilizado no treino de DCNN) o valor da taxa de aprendizagem varia ao longo do treino. Não obstante, é necessário escolher um valor inicial para a taxa de aprendizagem, os valores sugeridos por Kingma e pela *Pytorch* são de 0,001. Através de experimentação empírica foram testados outros valores para a taxa de aprendizagem, no entanto, o valor para o qual se obteve melhor resultado foi o valor sugerido pelos autores, isto é, um valor de aprendizagem de 0,001.

O número de épocas escolhido para treinar a DCNN foi definido através de estudos empíricos, diferindo consoante a base de dados multiespectral utilizada. Para a base de dados Tufts [101], o número de épocas a treinar foi de 10 épocas. Para a segunda base de dados, a CASIA NIR-VIS 2.0 [8], foram definidas

[6] *Pytorch* é uma biblioteca de aprendizagem automática de código aberto baseada na biblioteca *Torch*, utilizada em sistemas de visão computacional e de processamento de voz. A Pytorch foi desenvolvida pelo laboratório de pesquisa de inteligência artificial do *Facebook* (FAIR, em inglês *Facebook Artificial Intelligence Research*).

[7] Esta limitação deve-se ao alinhamento dos processadores virtuais nos processadores físicos da GPU.

50 épocas para o treino da DCNN. Para um número superior de épocas era evidente o sobre ajuste do modelo ao conjunto de dados de treino.

Foi utilizada a função de custo de entropia cruzada[8] (CE, em inglês *Cross-Entropy*) durante o treino da DCNN.

Na Última Camada Ligada (UCL) para transformar o conjunto de características proveniente de todos os canais num conjunto de características de 256-dimensões foi empregue a função de ativação linear. Esta é definida por,

$$y = xA^T + b \tag{10}$$

sendo que x corresponde ao conjunto de características de entrada, o A corresponde aos pesos da DCNN e b à polarização, vulgarmente conhecido em inglês por *bias*.

Foi utilizado diluição de camadas, vulgarmente conhecida em inglês por *dropout*, durante o treino da DCNN. Esta técnica de regularização foi empregue na UCL com o objetivo de reduzir o sobre ajuste da DCNN ao conjunto de dados de treino. O valor escolhido para p, probabilidade de colocar a zero a informação de elementos de uma camada, foi de 50%. O valor de p foi escolhido tendo em consideração o valor recomendado e experimentações previamente efetuadas na DCNN [107].

A Tabela 7 resume os valores utilizados em cada parâmetro para bases de dados distintas durante o treino da DCNN.

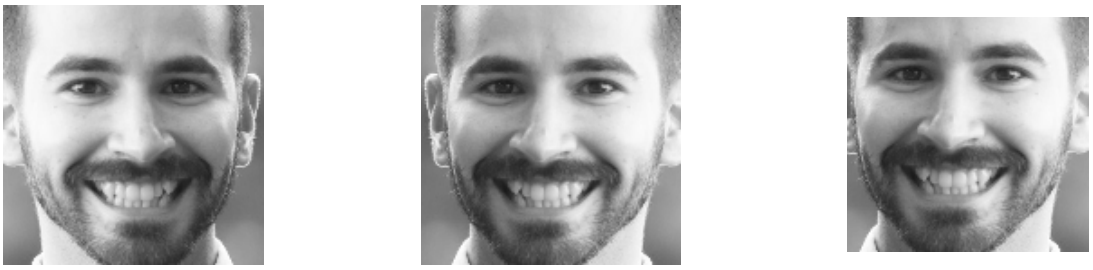
Tabela 7: Resumo dos valores utilizados em cada parâmetro para cada base de dados estudada para o treino da DCNN.

Parâmetro	**Tufts**	**CASIA NIR-VIS 2.0**
Tamanho do Lote	16	32
Algoritmo de Otimização	Adam	Adam
Taxa de Aprendizagem	0,001	0,001
Número de Épocas	10	50
Função de Custo	Entropia Cruzada	Entropia Cruzada

Foram empregues as seguintes técnicas de alargamento de dados pela seguinte ordem: espelhagem aleatória da imagem na horizontal e recorte aleatório das mesmas. De realçar que os métodos de aumento de imagem no treino não são os mesmos aplicados na fase de validação. Na fase de validação apenas é aplicado um recorte centrado da imagem para que esta fique com o tamanho adequado para a entrada da rede neural (128x128 *pixels*). Na Figura 33 estão ilustrados exemplos das técnicas utilizadas.

Durante o treino da rede era gravado em ficheiro, por época, o modelo da DCNN treinado até à época descrita. O modelo que obtivesse o menor custo no conjunto de validação era gravado no final como melhor modelo. Através das gravações do modelo por época é possível averiguar o desempenho do

[8]Note que a função de custo *Cross-Entropy* (CE) disponível na biblioteca da *Pytorch* não segue a definição convencional desta. Esta função de custo (da *Pytorch*) combina a função de custo *negative log likelihood* com a função *SoftMax* logarítmica, vulgarmente conhecida em inglês por *Log SoftMax*, para calcular a probabilidade logarítmica da DCNN.



(a) Imagem inicial. (b) Espelhagem aleatória na horizontal. (c) Recorte aleatório.

Figura 33: Imagens ilustrativas das técnicas de alargamento de dados utilizados.

modelo em diversas partes do treino. Estas gravações tinham adicionalmente como finalidade criar um ponto de controlo do modelo, caso a DCNN se deparasse com algum erro durante o treino.

5.4.2 Camadas Adaptadas

Como já supramencionado, aqui será efetuado um estudo às camadas da arquitetura *LightCNN* a adaptar, de forma a obter o melhor desempenho em *rank-1*.

A experiência iniciará com o treino apenas da UCL. Esta camada não existia previamente na arquitetura inicial da DCNN, tendo sido implementada na nossa metodologia para que no produto final da DCNN permanecesse um vetor de características de 256-dimensões. De seguida, as camadas iniciais serão adaptadas (inclusive a UCL) até que todas as camadas da DCNN sejam adaptadas. Em todas as experiências os pesos foram inicializados a partir do modelo inicial[9] da LightCNN [79].

A nomenclatura dos conjuntos adaptados segue a nomenclatura inicial dos autores da LightCNN [79]. A *LightCNN* é composta por 29 camadas. Nestas destacam-se 9 conjuntos de camadas: a primeira camada convolucional em conjunto com a primeira *Max-Feature Map* (MFM), denominada de *Conv1*, 4 conjuntos denominadas de *Group*, que constituem as camadas entre as camadas de *pooling*, e as restantes 4 camadas denominadas de *Block*, constituída por um bloco de camadas convolucionais no início de cada *Group*. As notações utilizadas na combinação das camadas adaptadas são as seguintes:

- **UCL:** Apenas a Última Camada Ligada (UCL) é adaptada;
- **Conv1-UCL ({1-1}+UCL):** A primeira camada convolucional é adaptada em conjunto com a primeira MFM e a UCL é adaptada;
- **Conv1-Block1-UCL ({1-2}+UCL):** Bloco de redes neurais residuais é adaptado em conjunto com as camadas anteriores;
- **Conv1-Block1-Group1-UCL ({1-3}+UCL):** Adapta o primeiro grupo da camada seguinte em conjunto com as camadas anteriores;
- **Conv1-N-UCL ({1-N}+UCL):** Adapta as camadas de 1 a N em conjunto com a UCL;

[9] O modelo utilizado está disponível, para *Pytorch*, no seguinte repositório: `https://github.com/AlfredXiangWu/LightCNN` .

- **Todas as Camadas:** Todas as camadas da LightCNN [79] em conjunto com a UCL são adaptadas.

O número de épocas definido para o treino da DCNN neste teste foi de 10 e 50 para as bases de dados Tufts [101] e CASIA NIR-VIS 2.0 [8], respetivamente. Após cada treino da DCNN foram extraídos os conjuntos de características de 256-dimensões das imagens de cada base de dados multiespectral. Para avaliar cada modelo foi utilizado o classificador SVM-Linear, com hiperparâmetro $C = 100$, para classificar os conjuntos de características extraídos.

Os resultados obtidos com a adaptação das diferentes camadas para cada base de dados multiespectral estão ilustrados nas Tabelas 8 e 9. Para cada tabela é possível observar o *rank-1* médio, em conjunto com o desvio padrão, obtido para cada conjunto de camadas adaptado, e também em qual *rank* é obtida uma taxa de identificação de 100%.

Tabela 8: Desempenho da DCNN quando diferentes combinações de camadas são adaptadas para a base de dados multiespectral Tufts.

Camada Adaptada	***Rank-1* Médio (Treino)**	***Rank-1* Médio (Teste)**	**Taxa de Identificação = 100%**
UCL	99,7% ± 0,1	99,7% ± 0,1	*Rank-6*
{1-1} + UCL	99,4% ± 0,1	99,5% ± 0,2	*Rank-4*
{1-2} + UCL	99,6% ± 0,1	99,6% ± 0,1	*Rank-3*
{1-3} + UCL	**99,8% ± 0,1**	**99,7% ± 0,1**	***Rank-2***
{1-4} + UCL	99,8% ± 0,1	99,7% ± 0,1	*Rank-5*
{1-5} + UCL	99,5% ± 0,1	99,5% ± 0,1	*Rank-5*
{1-6} + UCL	93,7% ± 0,6	93,3% ± 1,0	>10
{1-7} + UCL	95,0% ± 0,5	95,1% ± 1,0	>10
{1-8} + UCL	90,7% ± 0,6	90,7% ± 0,6	>10
{1-9} + UCL	91,3% ± 1,5	90,7% ± 0,6	>10
Todas as camadas	40,3% ± 7,5	39,1% ± 5,8	>10

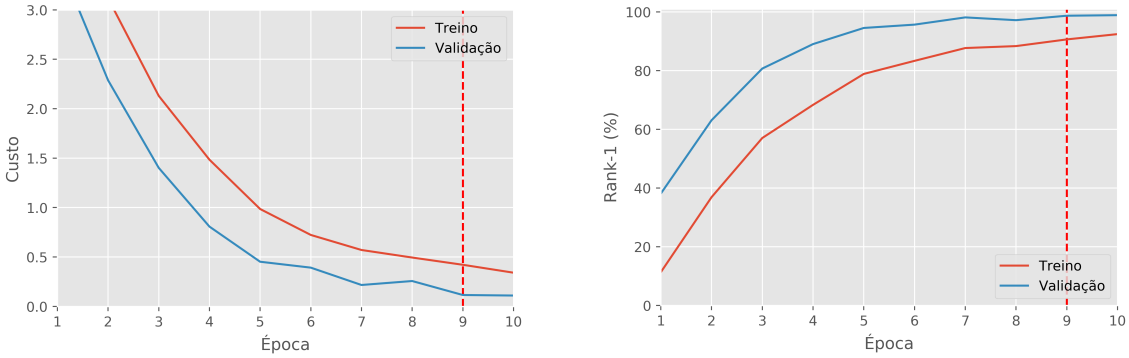
Tabela 9: Desempenho da DCNN quando diferentes combinações de camadas são adaptadas para a base de dados multiespectral CASIA NIR-VIS 2.0.

Camada Adaptada	***Rank-1* Médio (Treino)**	***Rank-1* Médio (Teste)**	**Taxa de Identificação = 100%**
UCL	99,7% ± 0,1	99,7% ± 0,1	>10
{1-1} + UCL	99,6% ± 0,1	99,6% ± 0,2	*Rank-10*
{1-2} + UCL	99,7% ± 0,1	99,7% ± 0,2	*Rank-9*
{1-3} + UCL	**99,8% ± 0,1**	**99,7% ± 0,1**	***Rank-9***
{1-4} + UCL	99,6% ± 0,1	99,6% ± 0,1	>10
{1-5} + UCL	99,5% ± 0,1	99,5% ± 0,1	>10
{1-6} + UCL	93,9% ± 1,0	93,9% ± 1,0	>10
{1-7} + UCL	94,2% ± 1,1	95,2% ± 0,8	>10
{1-8} + UCL	91,1% ± 0,8	91,0% ± 0,9	>10
{1-9} + UCL	90,7% ± 1,6	90,7% ± 1,6	>10
Todas as camadas	24,1% ± 11,5	32,1% ± 15,3	>10

Após análise das Tabelas 8 e 9 é possível concluir que o desempenho da rede obtém resultados favoráveis quando são adaptadas as camadas iniciais, obtendo-se os melhores resultados para as camadas {1-2} e {1-3}. Da camada {1-6} em diante, é notório a queda no desempenho da rede neural. Esta queda pode ser atribuída ao número elevado de parâmetros a adaptar, provocando assim um sobre ajuste da rede neural à base de dados. Esta observação é ainda transversal às bases de dados multiespectrais estudadas.

A seleção das camadas a adaptar foi efetuada através de uma escolha empírica. Os critérios de seleção foram os seguintes: a taxa de reconhecimento obtidas na fase de treino e de teste e o menor *rank* para uma taxa de reconhecimento de 100%.

Na Figura 34 está ilustrado o processo de treino do modelo ({1-3} + UCL), o melhor modelo para a base de dados multiespectral Tufts [101], com o gráfico do custo e do *rank-1* por época.



(a) Custo

(b) *Rank-1*

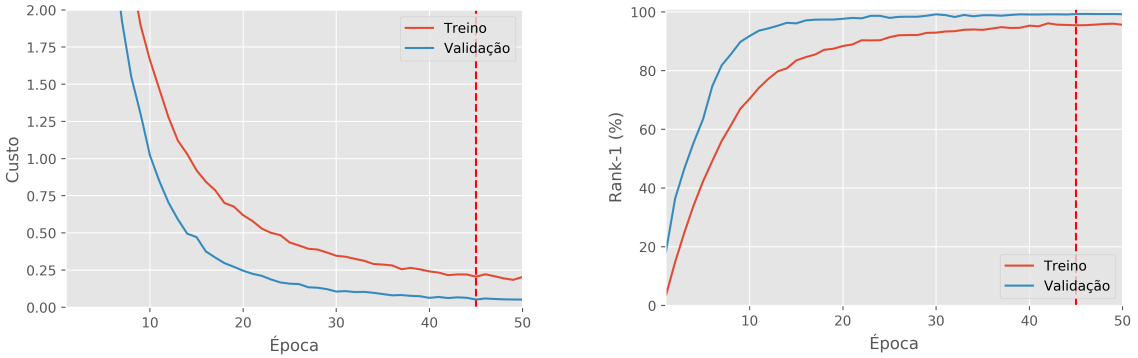
Figura 34: Processo de treino do modelo ({1-3} + UCL) para a base de dados Tufts [101].

Seria expectável que o custo durante o treino da DCNN fosse inferior para o conjunto de treino quando comparado com o conjunto de validação, no entanto, tal não acontece, como é aparente na Figura 34a. O mesmo seria expectável para o *rank-1*, Figura 34b.

A justificação para tal prende-se com o facto de se ter empregue uma diluição na UCL durante o treino. Esta técnica de regularização, irá fazer com que a fase de treino "perca" informação (i.e., seja zerada) na UCL. Desta forma, o conjunto de treino irá permanecer em constante desvantagem quando comparado com o conjunto de validação, onde não é empregue diluição da UCL.

Um outro fator que irá provocar um aumento no valor do custo durante o treino da DCNN, e consequentemente, um *rank-1* menor, é a utilização de alargamento de dados apenas para o conjunto de treino. O alargamento de dados empregues na arquitetura da DCNN são, previamente enunciados no Capítulo 4: a espelhagem aleatória na horizontal e recorte aleatório da imagem. Para o conjunto de validação é efetuado apenas o recorte de imagem, mas para este conjunto de dados, este é sempre efetuado no centro da imagem.

Na Figura 35 está ilustrado o processo de treino do modelo ({1-3} + UCL), o melhor modelo para a base de dados multiespectral CASIA NIR-VIS 2.0 [8], com o gráfico do custo e do *rank-1* por época. Assinalado a vermelho está ilustrado a melhor época, a época utilizada nos estudos consequentes.



(a) Custo

(b) *Rank-1*

Figura 35: Processo de treino do modelo ({1-3} + UCL) para a base de dados CASIA NIR-VIS 2.0 [8].

Numa primeira análise da Figura 35a, é possível constatar que até à época 15 o custo durante o treino da DCNN decresce de forma rápida. Este continuará a decrescer, mas agora de um modo mais suave, até à época 50.

Tal como já tinha sido realçado aquando o treino da DCNN com a base de dados Tufts [101], o custo durante o treino da DCNN com o conjunto de validação é inferior ao de treino, justificado pela utilização de alargamento de dados no conjunto de treino, e da utilização de diluição na UCL.

5.4.3 Estudo dos Hiperparâmetros

Efetuada a extração dos conjuntos de características de 256-dimensões, através da DCNN, é necessário agora proceder à sua classificação. Como já descrito no Capítulo 4, serão utilizados os classificadores kNN e SVM (linear e RBF). Para tal, foi necessário efetuar um estudo para cada um dos classificadores com

o intuito de identificar os melhores valores para cada hiperparâmetro.

De forma a proceder a uma correta escolha nos hiperparâmetros a utilizar, foi efetuada uma validação cruzada estratificada (SCV, em inglês *Stratified Cross-Validation*). A utilização de SCV permite efetuar uma escolha mais correta dos hiperparâmetros para bases de dados não balanceadas (i.e., número de imagens por pessoa não é constante na base de dados), conforme descrito por [108] e [109]. Durante a SCV o conjunto de dados de treino e de validação foram unificados. No entanto, durante a fase de treino do classificador (com os melhores hiperparâmetros já determinados) foi utilizado apenas o conjunto de treino (i.e., sem o conjunto de validação).

A utilização da SCV é limitado pela pessoa que contém no conjunto de treino e validação o menor número de imagens. Como tal, o número máximo de vezes que se pode fazer SCV é de 5 vezes para a base de dados multiespectral Tufts [101] e 4 vezes para a base de dados multiespectral CASIA NIR-VIS 2.0 [8].

Na Figura 36 encontra-se ilustrado um exemplo de SCV. Neste conjunto de dados, é possível constatar a presença de três identidades distintas (A, B e C), cada uma com mais imagens que a anterior. Neste exemplo é efetuado SCV três vezes, em que em cada validação é utilizado uma parte equitativa de cada identidade. Após a validação é possível obter a pontuação *rank-1* média e o desvio padrão desta.

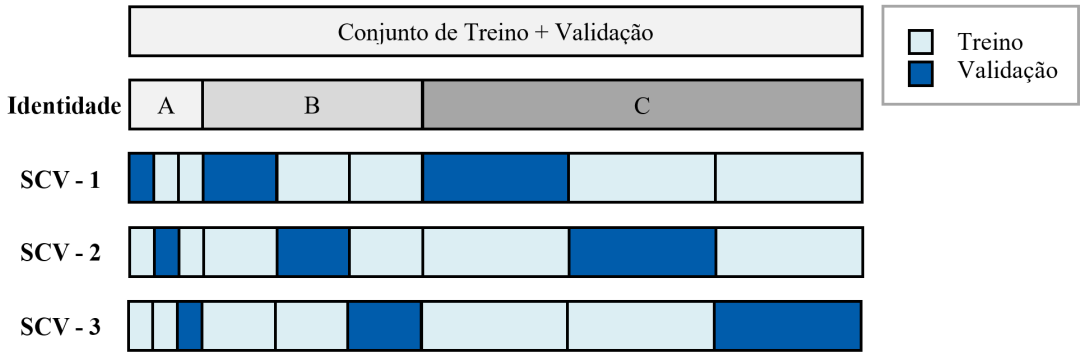


Figura 36: Exemplo ilustrativo de SCV efetuado três vezes para um conjunto de dados com três identidades.

O hiperparâmetro afinado para o algoritmo SVM-Linear foi o parâmetro de regularização (C). Este hiperparâmetro indica o grau de importância dada às classificações incorretas. Para valores altos de C, o algoritmo irá atribuir uma margem menor para o hiperplano. Inversamente, um menor valor para C a margem do hiperplano será superior, mesmo que este classifique incorretamente mais pontos.

De forma a encontrar os valores mais adequados, foram testados diversos valores para o hiperparâmetro C, o alcance de valores estudados foram de 10^{-10} a 10^{+5}, espaçados por uma década logarítmica. De realçar que, para determinar estes valores foi necessário proceder-se a um estudo prévio de forma a determinar os alcances de valores mais adequados para serem estudados.

Para o algoritmo SVM-RBF foram afinados os seguintes hiperparâmetros: o parâmetro de regularização (C) e o coeficiente de *kernel* (γ). O hiperparâmetro coeficiente de *kernel* tem como objetivo definir a influência de um ponto, no conjunto de dados, em relação a outros. Para valores baixos de γ são considerados os pontos bastante afastados da margem do hiperplano. O contrário acontece para valores

elevados de γ, em que neste caso apenas são considerados os pontos mais próximos do plano de decisão do hiperplano.

Com o intuito de encontrar os valores para os hiperparâmetros mais adequados, foram testadas várias combinações diferentes utilizando diversos valores para C e para γ. Para C foram utilizados valores de 10^{-4} a 10^{+7}, espaçados por uma década logarítmica. E para γ, foram utilizados valores de 10^{-10} a 10^{+2}, espaçados por uma década logarítmica. De realçar que, para determinar estes valores foi necessário fazer um estudo prévio para determinar os valores mais adequados para serem estudados.

Por último, para o algoritmo de kNN, o hiperparâmetro a afinar foi o número de vizinhos próximos (k). Este hiperparâmetro influência o número de pontos a considerar aquando a classificação. Para o hiperparâmetro k foram estudados os valores de 1 a 25.

Na Tabela 10 está um resumo dos hiperparâmetros e a gama de valores correspondente estudada para cada classificador.

Tabela 10: Hiperparâmetros e gama de valores estudada para cada classificador.

Classificador	**Hiperparâmetros**		
	Parâmetro de Regularização	**Coeficiente de *Kernel***	**Número de Vizinhos**
SVM-Linear	$10^{-10} \leq C \leq 10^{+5}$	-	-
SVM-RBF	$10^{-4} \leq C \leq 10^{+7}$	$10^{-10} \leq \gamma \leq 10^{+2}$	-
kNN	-	-	$1 \leq k \leq 25$

5.4.3.1 SVM-Linear

Os resultados obtidos no estudo dos hiperparâmetros para a SVM-Linear, na base de dados Tufts [101], encontram-se disponíveis na Tabela 15, do Apêndice C, e na Figura 37. Nesta tabela está identificado a negrito o melhor resultado obtido neste estudo.

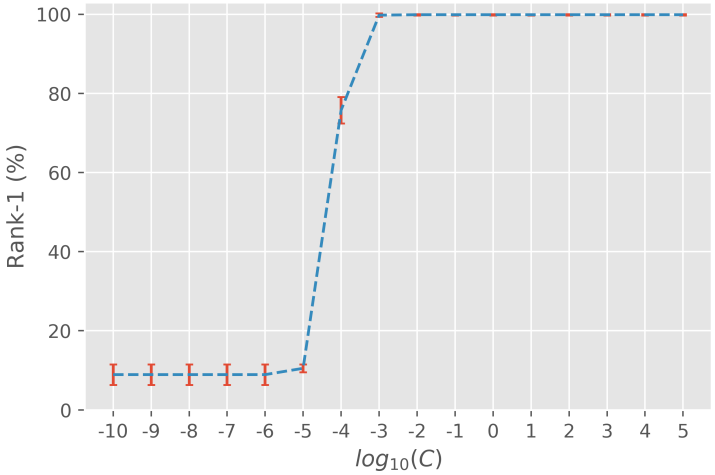


Figura 37: Afinamento do hiperparâmetro C para SVM-Linear para a base de dados Tufts [101].

Numa primeira análise em conjunto com a Tabela 15 e a Figura 37 é possível observar que para valores de $C > 10^{-2}$ a pontuação de *rank-1* média é constante, obtendo-se um valor de 99,89%, com um desvio padrão de 0,09%. Numa segunda análise, os valores mais inferiores de *rank-1* são para $C < 10^{-5}$, nestes valores o *rank-1* nunca foi superior a 11% acompanhado por desvios padrões de 1,29%.

Após este estudo, é possível concluir que qualquer valor de C, compreendido entre 10^{-2} e 10^{+5}, é valido para ser utilizado no classificador SVM-Linear. Assim, foi treinado um classificador SVM com *kernel* linear com o hiperparâmetro C de 10^{-2}. Este modelo será utilizado para ser avaliado com o conjunto de teste, podendo assim, posteriormente, ser comparado com o melhor modelo de cada classificador, para a base de dados Tufts [101].

Na base de dados CASIA NIR-VIS 2.0 [8], os resultados obtidos para o estudo dos hiperparâmetros para a SVM-Linear encontram-se disponíveis na Tabela 16 e na Figura 38, onde está assinalado a negrito o melhor resultado obtido para este estudo.

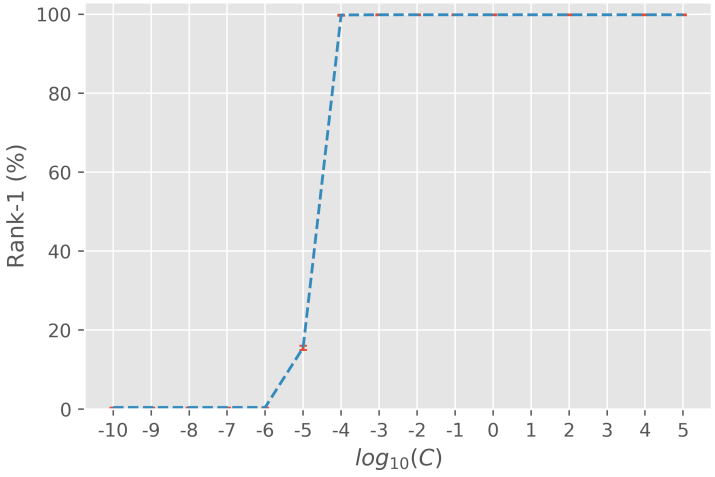


Figura 38: Afinamento do hiperparâmetro C para SVM-Linear para a base de dados CASIA NIR-VIS 2.0 [8].

Numa primeira análise em conjunto com a Tabela 16 e Figura 38 é possível observar que para valores de $C > 10^{-3}$ o *rank-1* médio é constante em 99,86%, com um desvio padrão de 0,06%. Numa segunda análise, os valores inferiores de *rank-1* são para $C < 10^{-5}$, nestes valores o *rank-1* nunca foi superior a 0,40% acompanhado por desvios padrões de 0%.

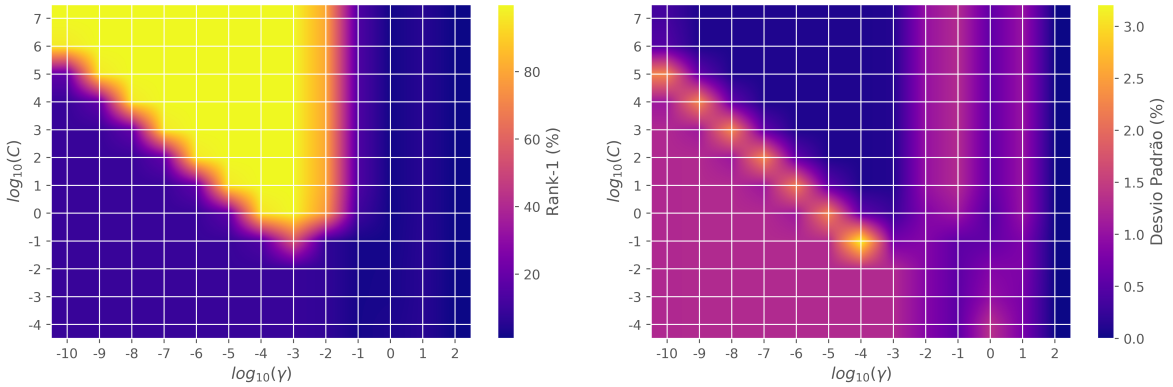
Após este estudo, é possível concluir que qualquer valor de C, compreendido entre 10^{-3} e 10^{+5}, é adequado para ser utilizado no classificador SVM-Linear. Assim, foi treinado um classificador SVM com *kernel* linear com o hiperparâmetro C de 10^{-3}. Este modelo será utilizado para ser avaliado com o conjunto de teste, podendo assim, posteriormente, ser comparado com o melhor modelo de cada classificador, para a base de dados CASIA NIR-VIS 2.0 [8].

5.4.3.2 SVM-RBF

Os hiperparâmetros afinados para o classificador SVM-RBF foram os seguintes: o parâmetro de

regularização (C) e o coeficiente de *kernel* (γ). Para C foram utilizados valores de 10^{-4} a 10^{+7}, espaçados por uma década logarítmica. E para γ, foram empregues valores de 10^{-10} a 10^{+2}, espaçados por uma década logarítmica.

A Figura 39 ilustra o *rank-1* médio (Figura 39a) e o desvio padrão correspondente (Figura 39b) para o estudo dos hiperparâmetros efetuado para a SVM-RBF.



(a) *Rank-1* Médio. (b) Desvio Padrão

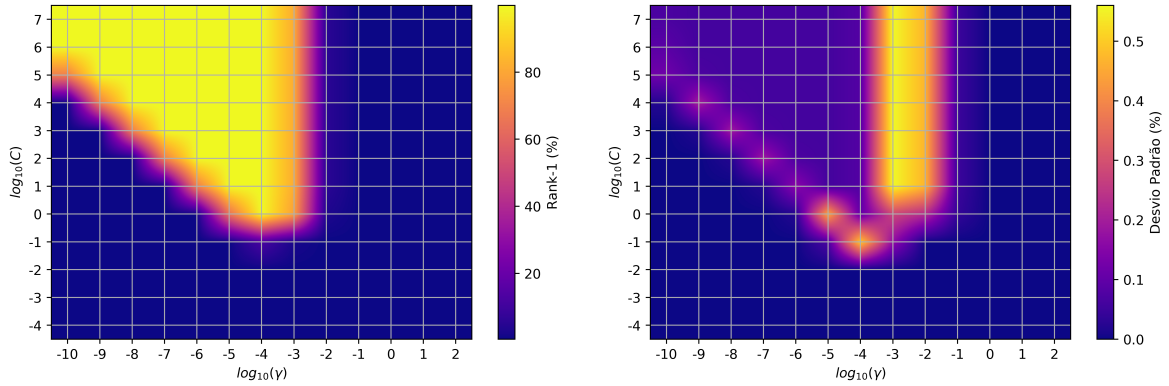
Figura 39: Afinamento do hiperparâmetro C e γ para o classificador SVM-RBF para a base de dados Tufts [101].

Numa primeira análise, verifica-se que para valores de γ superiores a 10^{-1} o *rank-1* diminui significativamente, independentemente do valor atribuído ao hiperparâmetro C.

Numa segunda análise, é possível constatar que o melhor resultado obtido foi um *rank-1* médio de 99.89% com desvio padrão de 0,09%, com os seguintes hiperparâmetros: $C = 10^{+1}$ e $\gamma = 10^{-4}$. Em comparação, o pior resultado obtido neste estudo foi um *rank-1* médio de 0,94%, com um desvio padrão aproximadamente de 0%, para a gama de valores de $\gamma = 10^{+2}$ independentemente do hiperparâmetro C.

Após este estudo, é possível concluir que para qualquer valor de C e γ compreendido na zona amarela da Figura 39a pode ser utilizado no classificador SVM-RBF. Assim, foi treinado um classificador SVM com *kernel* RBF com o hiperparâmetro C de 10^{+1} e o hiperparâmetro γ de 10^{-4}. Este modelo será utilizado para ser avaliado com o conjunto de teste, podendo assim comparar-se com o melhor modelo de cada classificador.

Na Figura 40 está ilustrado o *rank-1* médio (Figura 40a) e o desvio padrão correspondente (Figura 40b) para o estudo dos hiperparâmetros efetuado para a SVM-RBF, para a base de dados CASIA NIR-VIS 2.0 [8].



(a) *Rank-1* Médio.

(b) Desvio Padrão

Figura 40: Afinamento do hiperparâmetro C e γ para o classificador SVM-RBF para a base de dados CASIA NIR-VIS 2.0 [8].

Numa primeira análise, verifica-se que para valores de γ superiores a 10^{-2} o *rank-1* diminui significativamente, independentemente do valor atribuído ao hiperparâmetro C.

Numa segunda análise, foi possível constatar que o melhor resultado obtido foi um *rank-1* médio de 99,86% com desvio padrão de 0,06%, com os seguintes hiperparâmetros: $C = 10^{+1}$ e $\gamma = 10^{-5}$. Em comparação, o pior resultado foi um *rank-1* médio de 0,40%, com um desvio padrão aproximadamente de 0%, para a gama de valores de $C = 10^{-3}$ independentemente do hiperparâmetro γ.

Após este estudo, é possível concluir que para qualquer valor de C e γ compreendido na zona amarela da Figura 40a pode ser utilizado no classificador SVM-RBF. Como tal, foi treinado um classificador SVM com kernel RBF com o hiperparâmetro C de 10^{+1} e o hiperparâmetro γ de 10^{-5}. Este modelo será utilizado para ser avaliado com o conjunto de teste.

Confrontando os resultados obtidos nesta base de dados com a base de dados Tufts [101], é possível sublinhar uma grande diferença. Os resultados para este classificador, são mais precisos, isto é, o desvio padrão na base de dados CASIA NIR-VIS 2.0 [8] é significativamente menor, 0,56%, do que para a Tufts [101], 3,5%.

5.4.3.3 kNN

Para o último classificador, o kNN, o hiperparâmetro a afinar foi o número de vizinhos próximos (k). Este hiperparâmetro influencia o número de pontos a considerar aquando a classificação. Para o hiperparâmetro k foram estudados a gama de valores de 1 a 25.

Na Tabela 17, do Apêndice C, e Figura 41 estão ilustrados os resultados obtidos quando empregues diferentes números de vizinhos pelo classificador na base de dados Tufts [101], assinalado a negrito o melhor resultado obtido.

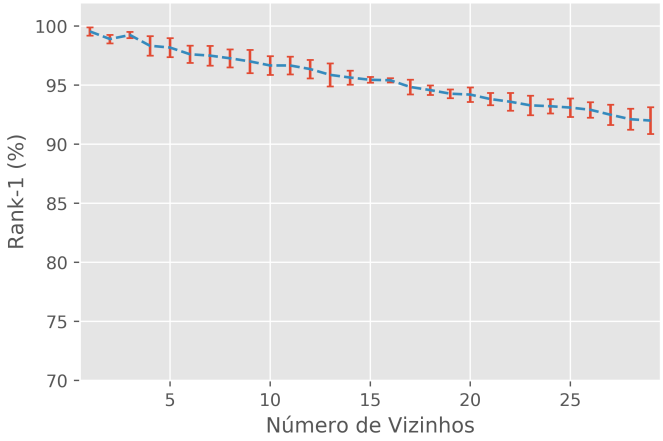


Figura 41: Afinamento do hiperparâmetro k para a kNN para a base de dados Tufts [101].

Analisando a Tabela 17 e Figura 41 é possível observar que o melhor resultado obtido é para $k = 1$, em que foi obtido um *rank-1* médio de 99,54%, com um desvio padrão de 0,35%. Numa segunda análise, é possível constatar que à medida que o número de vizinhos aumenta o respetivo *rank-1* médio decresce, aumentado o desvio padrão. Como tal, foi treinado um classificador kNN com hiperparâmetro k de 1. Este modelo será utilizado para ser avaliado com o conjunto de teste.

Na Tabela 18, do Apêndice C, e Figura 42 estão ilustrados os resultados obtidos quando utilizados diferentes números de vizinhos pelo classificador na base de dados CASIA NIR-VIS 2.0 [8], assinalado a negrito o melhor resultado obtido.

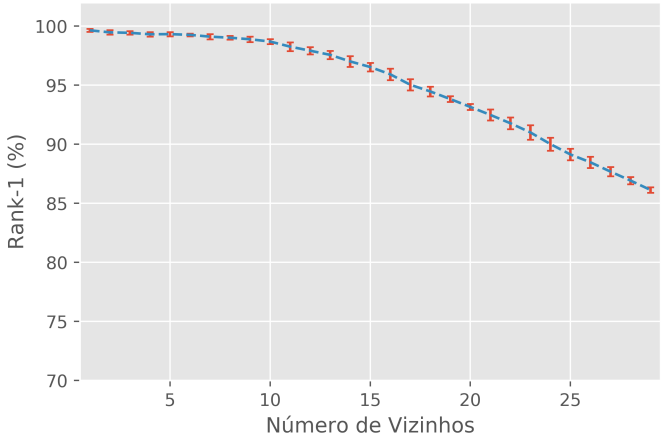


Figura 42: Afinamento do hiperparâmetro k para a kNN para a base de dados CASIA NIR-VIS 2.0 [8].

Analisando a Tabela 18 e Figura 42 é possível observar que o melhor resultado obtido é para $k = 1$, neste foi obtido um *rank-1* médio de 99,63% com um desvio padrão de 0,25%. Numa segunda análise é possível constatar que à medida que o número de vizinhos aumenta o respetivo *rank-1* médio diminui, aumentado o desvio padrão.

Obtido o melhor resultado foi treinado um classificador kNN com hiperparâmetro k de 1. Este modelo será utilizado para ser avaliado com o conjunto de teste, podendo assim comparar com outros

classificadores.

Quando comparado com os resultados obtidos para a base de dados Tufts [101], é notório que para valores de k compreendidos entre 1 e 5, a diferença no *rank-1* médio e o desvio padrão não é tão acentuada.

5.4.4 Comparação com o Estado da Arte

Obtidos os valores mais adequados para cada hiperparâmetro é efetuado um estudo de forma a determinar qual o melhor classificador para cada base de dados multiespectral. O desempenho de cada classificador será avaliado no conjunto de dados de teste, sendo que o melhor classificador será aquele que obtiver um *rank-1* superior. Em caso de empate, o fator de desempate utilizado será através da curva CMC, onde será dada primazia ao classificador que obtiver uma taxa de identificação de 100% primeiro.

Com a escolha do classificador, é efetuada uma comparação com as restantes metodologias identificadas no estado da arte.

5.4.4.1 Base de Dados Tufts

Na Tabela 11 encontra-se disponível um resumo dos melhores hiperparâmetros para cada classificador, e também o respetivo *rank-1*, obtido para o conjunto de teste na base de dados Tufts [101]. O classificador que obteve o melhor resultado está assinalado a negrito.

Tabela 11: Melhores hiperparâmetros obtidos para cada classificador na base de dados Tufts [101].

Classificador	**Hiperparâmetros**			***Rank-1***
	C	γ	k	
({1-3} + UCL)) + SVM-Linear	**0,01**	**-**	**-**	**99,7 %**
({1-3} + UCL)) + SVM-RBF	10	10^{-4}	-	99,2 %
({1-3} + UCL)) + kNN	-	-	1	99,2 %

Analisando a Tabela 11, é possível concluir que o classificador que obtém os melhores resultados para o conjunto de teste é o classificador SVM-Linear, obtendo um *rank-1* de 99,7%.

Utilizando os modelos dos classificadores empregues anteriormente, foi agora efetuado um estudo para diferentes classificações (i.e., *rank-N*), permitindo analisar a fidelidade da classificação de cada classificador. Na Figura 43 está ilustrado a curva CMC para os classificadores SVM-Linear, SVM-RBF e kNN para as primeiras dez classificações (i.e., até *rank-10*) para a base de dados Tufts [101].

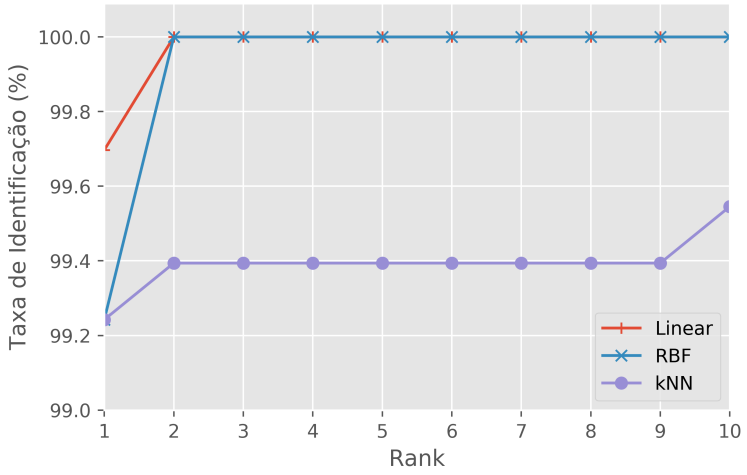


Figura 43: Curva CMC para os diferentes classificadores, para a base de dados Tufts [101].

Uma análise comparativa entre os três classificadores através da Tabela 11 sugere que o classificador SVM-Linear é aquele que obtém os melhores resultados, com um *rank-1* de 99,7%, classificando incorretamente duas imagens. Em contra partida, os classificadores SVM-RBF e kNN obtiveram o mesmo resultado, 99,2%, para o mesmo conjunto.

No entanto, quando a Tabela 11 é visualizada em conjunto com a Figura 43, é possível constatar que para o *rank-2*, independentemente do *kernel* utilizado no classificador SVM, este obtém uma taxa de identificação de 100%. Isto é, a totalidade das imagens faciais do conjunto de teste foram corretamente identificadas, dentro das duas com maior probabilidade. Comparativamente, o classificador kNN obtém uma taxa de identificação de 100% para *rank-102*.

Através da análise anterior, é escolhido o classificador SVM-Linear para classificar o conjunto de características de 256-dimensões produzidos pela DCNN.

Na Tabela 12 encontram-se explanados os resultados obtidos da metodologia proposta com outras metodologias da literatura. De realçar que a metodologia proposta utiliza a DCNN *LightCNN* como base, adaptando as camadas ({1-3} + UCL), para produzir um conjunto de características de 256-dimensões, que serão classificadas depois pelo classificador SVM-Linear.

Tabela 12: Resultados obtidos na metodologia proposta quando comparados com o estado da arte para a base de dados Tufts [101].

Método	**Rank-1**
TR-GAN [110]	88,7 %
Circular HOG [111]	94,5 %
Metodologia Proposta[10]	**99,7 %**

Como é possível observar na Tabela 12, através da metodologia proposta é possível obter um resultado

[10]De realçar que a base de dados multiespectral utilizada para o treino da DCNN foi limpa por nós. Outros autores não especificam se efetuaram limpeza ou não na base de dados.

bastante competitivo quando comparado com outras metodologias. Quando contabilizadas as 26 imagens excluídas [11] (na limpeza da base de dados) é obtido uma pontuação de *rank-1* de 95,9%. No entanto, este resultado continua a ser superior ao estado da arte para esta base de dados. De realçar que, sendo a base de dados recente, disponibilizada ao público para investigação em 2020[12], o número de autores que a utilizam é reduzida.

5.4.4.2 Base de Dados CASIA NIR-VIS 2.0

A Tabela 13 contém um resumo dos melhores hiperparâmetros para cada classificador, assim como o respetivo *rank-1*, obtido para o conjunto de teste na base de dados CASIA NIR-VIS 2.0 [8]. O classificador que obteve o melhor resultado está assinalado a negrito.

Tabela 13: Melhores hiperparâmetros obtidos para cada classificador para a base de dados CASIA NIR-VIS 2.0 [8].

Classificador	**Hiperparâmetros**			***Rank-1***
	C	γ	k	
({1-3} + UCL)) + SVM-Linear	**0,001**	**-**	**-**	**99,8 %**
({1-3} + UCL)) + SVM-RBF	10	10^{-5}	-	99,8 %
({1-3} + UCL)) + kNN	-	-	1	99,7 %

A partir dos modelos dos classificadores empregues anteriormente, foi agora efetuado um estudo para diferentes classificações (i.e., *rank-N*). Deste modo é possível observar a fidelidade dos resultados de cada classificador. Na Figura 44 está ilustrado a curva CMC para os classificadores SVM-Linear, SVM-RBF e kNN para as primeiras dez classificações (i.e., até *rank-10*) para a base de dados CASIA NIR-VIS 2.0 [8].

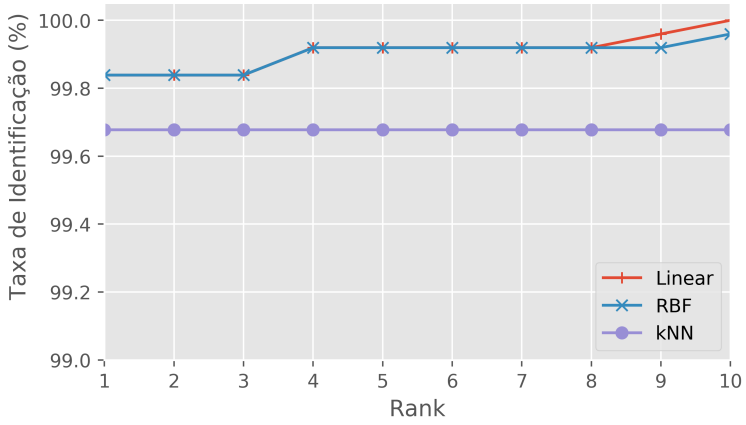


Figura 44: Curva CMC para os diferentes classificadores, para a base de dados CASIA NIR-VIS 2.0 [8].

[11] Note-se que foram excluídas 53 imagens faciais da base de dados multiespectral Tufts. No entanto, 27 imagens faciais foram excluídas porque 2 pessoas não tinham imagens em todas as bandas espectrais.

[12] A base de dados foi finalizada no ano de 2018 e disponibilizada ao público desde então. No entanto, apenas foi publicada no ano de 2020.

Através de uma análise comparativa entre os três classificadores constantes da Tabela 13, é possível concluir que os dois classificadores SVM obtém um resultado semelhante, um *rank-1* de 99,8%. No entanto, através de uma análise posterior à Figura 44, é possível constatar que do *rank-8* em diante o classificador SVM-Linear obtém uma taxa de identificação superior, quando comparado com o SVM-RBF. Os classificadores SVM obtém uma taxa de identificação de 100% para o *rank-10* e *rank-12*, para o *kernel* linear e RBF. Em contrapartida, o classificador kNN obtém uma taxa de identificação de 100% para o *rank-673*.

Na Tabela 13 encontram-se explanados os resultados obtidos na metodologia proposta com outras metodologias do estado da arte, comparadas através de um *rank-1* obtido. De realçar que a metodologia proposta utiliza a DCNN *LightCNN* como base, adaptando as camadas ({1-3} + UCL), para produzir um conjunto de características de 256-dimensões, que serão classificados pelo classificador SVM-Linear. De notar que, a tabela encontra-se enumerada pelo ano de publicação, e não pela classificação obtida.

Tabela 14: Resultados obtidos através da metodologia proposta quando comparados com o estado da arte para a base de dados CASIA NIR-VIS 2.0 [8].

Método	***Rank-1***	**Ano de Publicação**
CDFL [22]	71,5 %	2015
MCA [12]	69,1 %	2016
MTC-ELM [24]	89,1 %	2017
CEFDA [68]	85,6 %	2017
Oh *et al.* [25]	97,5 %	2017
LightCNN [79]	96,7 %	2018
MDNDC [27]	98,9 %	2019
Peng *et al.* [29]	96,7 %	2019
DSU [11]	96,3 %	2019
WCNN [18]	98,7 %	2019
DDFLJM [30]	98,8 %	2019
Peng *et al.* [31]	98,7 %	2019
CFC [19]	98,6 %	2019
CycleGAN [32]	99,4 %	2020
Metodologia Proposta	**99,8 %**	**2020**

Numa primeira análise da Tabela 14, é possível observar que a metodologia proposta obtém resultados superiores em *rank-1* quando comparado com o restante estado da arte. O trabalho mais atual, entre o estado da arte, que utiliza a base de dados CASIA NIR-VIS 2.0 [8] é o de Bae *et al.* [32] obteve uma pontuação em *rank-1* de 99,4%, inferior ao resultado obtido pela metodologia proposta, de 99,8%.

Numa segunda análise da Tabela 14, é possível observar que a arquitetura DCNN utilizada na metodologia proposta, a *LightCNN* [79], obteve um *rank-1* de 96,7%. Através da metodologia proposta foi possível alcançar um *rank-1* superior em 3,1%, um aumento significativo.

6 Conclusão

Atualmente, os sistemas de reconhecimento facial multiespectral continuam a ser complexos e exigentes, dados os inúmeros fatores a considerar no momento da deteção facial, extração dos marcos faciais e do reconhecimento facial.

Através do estudo sistemático do estado da arte, foi possível concluir que o método de reconhecimento facial mais utilizado, e aquele que obtém os melhores resultados, é o de redes neurais profundas. Dos artigos estudados, 36% utilizam redes neurais profundas como método de reconhecimento facial multiespectral. De salientar que, desde 2019, houve um reaparecimento dos métodos de síntese de imagens devido ao uso de redes neurais, principalmente, as GAN. Igualmente, foi possível concluir que a métrica mais utilizada para comparar métodos em diferentes bases de dados multiespectrais é a pontuação em *rank-1*.

O principal problema dos atuais sistemas de reconhecimento facial multiespectral é a falta de disponibilidade de bases de dados multiespectrais. Do estudo do estado da arte também se verificou que a base de dados pública mais utilizada é a CASIA NIR-VIS 2.0 [8]. Quando comparadas com as bases de dados com imagens na banda espectral do visível, as bases de dados multiespectrais públicas atuais são mais pequenas (em relação ao número total de imagens), o que pode levar a um sobre ajuste dos classificadores treinados nestas bases de dados. As bases de dados multiespectrais possuem diversas limitações, como, por exemplo: o número reduzido de imagens; o fato de não existir nenhuma base de dados com imagens faciais da mesma pessoa em diferentes bandas espectrais (e.g., VIS, NIR, SWIR e LWIR); a falta de variações de pose, luminosidade e variações da distância (entre a câmara e a pessoa).

Os atuais métodos de reconhecimento facial multiespectral obtém melhor desempenho quando comparados com os sistemas de reconhecimento facial que utilizam apenas imagens da banda espectral visível [50]. No entanto, o uso de redes neurais profundas como método para realizar o reconhecimento facial multiespectral ainda é limitado devido ao número reduzido de imagens (e pessoas) nas bases de dados multiespectrais atuais. Não obstante, as redes neurais profundas são os métodos mais utilizados para realizar o reconhecimento facial multiespectral, podendo produzir resultados bastante promissores. Quando aplicados na base de dados multiespectral CASIA NIR-VIS 2.0 [8] o melhor resultado obtido foi de uma pontuação em *rank-1* de 99,4%, considerando apenas os artigos estudados.

Nesta dissertação foi proposto um sistema de reconhecimento facial multiespectral. Este sistema tira proveito de imagens multiespectrais com a finalidade de obter melhores resultados. O sistema é composto por quatro módulos: processamento de imagem, deteção de falsificação, processamento facial e reconhecimento facial.

Para o módulo de deteção de falsificação, é proposto um detetor de falsificação multiespectral. Este tira proveito de todas as imagens multiespectrais disponíveis no sistema de reconhecimento facial. O detetor de falsificação multiespectral realiza uma deteção da pele humana presente na imagem facial. Através da pele detetada é efetuada uma comparação com os marcos faciais, de modo a averiguar a existência de um ataque de falsificação, ou seja possível dissimulação de identidade. Foram realizadas

diversas experiências com a finalidade de demonstrar a superioridade deste detetor com outros. Para esta validação, o detetor de falsificação proposto é comparado com outros detetores de pele, nomeadamente YCbCr e HSV. Ambos os detetores utilizados na comparação apenas utilizam imagens na banda espectral do visível. Os resultados experimentais comprovam a eficiência do detetor de falsificação multiespectral proposto (que utiliza imagens na banda espectral do VIS, SWIR e LWIR), quando comparado com os do tipo YCbCr e HSV, conseguindo efetuar uma correta identificação do que é pele mesmo quando são utilizadas máscaras. Um sistema de segurança que utilize os detetores de pele, aplicados na fase de comparação, não teria detetado os utilizadores com máscaras. No entanto, o detetor de falsificação multiespectral proposto consegue, com sucesso, uma correta avaliação, validando assim o conceito do classificador proposto.

Neste trabalho, é proposta uma nova rede de extração de características para o reconhecimento facial utilizando imagens multiespectrais. Esta possui diversos canais, em que a cada um é atribuído uma banda espectral ou intervalo espectral. Cada canal utiliza a rede neural convolucional profunda, a LightCNN [79], de forma a extrair conjuntos de características de 256-dimensões. São adaptados conjuntos de camadas da LightCNN [79] de cada canal, com a finalidade de adaptar cada um a uma banda espectral. Neste processo excetua-se o canal que irá receber imagens na banda espectral do visível. De forma a manter um conjunto de características de 256-dimensões como saída da arquitetura, foi implementada uma última camada ligada. Esta camada final tem como finalidade efetuar uma transformação linear na totalidade dos conjuntos de características de 256-dimensões em um conjunto de características de 256-dimensões singular. Diversas camadas da LightCNN foram adaptadas com a finalidade de averiguar quais as que apresentam melhores resultados. Através da experimentação é possível afirmar que as melhores camadas a adaptar, independentemente da base de dados multiespectrais utilizada, são as camadas iniciais, nomeadamente ({1-3} + UCL).

Foram também efetuados estudos extensivos nos classificadores utilizados de forma a classificar os conjuntos de características extraídos pela arquitetura. Num primeiro estudo, foram averiguados quais os valores mais adequados a serem utilizados pelos hiperparâmetros de cada classificador. Os hiperparâmetros estudados foram: parâmetro de regularização, coeficiente de *kernel* e o número de vizinhos. Dos resultados experimentais conclui-se que é necessário realizar um estudo prévio para cada base de dados multiespectral utilizada. Esta observação não se adequa para o classificador kNN, uma vez que o melhor número de vizinhos foi sempre de 1, independentemente da base de dados multiespectral. Após o estudo dos hiperparâmetros para cada classificador, foi treinado cada classificador com os hiperparâmetros mais adequados. Com os resultados experimentais, é possível concluir que o classificador que melhor obtem o melhor resultado, através dos conjuntos de características de 256-d, é o classificador SVM com o *kernel* linear. Este resultado foi obtido para as duas bases de dados multiespectrais. Estudos extensivos nas bases de dados multiespectrais ilustram a superioridade da metodologia proposta. Foram obtidas as pontuações em *rank-1* de 99,7% e 99,8% para as bases de dados multiespectrais Tufts [101] e CASIA NIR-VIS 2.0 [8]. Em comparação com outras metodologias identificadas no estado da arte, as melhores pontuações em *rank-1* para estas bases de dados foi de 94,5% e 99,8%, respetivamente.

O reconhecimento facial multiespectral, sendo uma tecnologia recente, ainda tem muito espaço para

evoluir e melhorar. Os principais objetivos dos sistemas de reconhecimento facial multiespectral continuam a ser a segurança e a vigilância, especialmente em locais críticos, como aeroportos ou em áreas militares classificadas, nomeadamente paióis ou arrecadações de material de guerra.

6.1 Trabalho Futuro

Após o termino de um trabalho de investigação, há sempre caminhos que não foram trilhados assim como ideias que surgem no decorrer do mesmo. Neste projeto, em que as áreas do saber abrangentes vão desde a aprendizagem automática, até ao reconhecimento facial, as hipóteses de trabalho futuro são diversas. Assim destacam-se algumas que são consideradas exequíveis a médio/longo prazo:

Aperfeiçoamento dos algoritmos de extração de marcos faciais. Atualmente, os métodos existentes possuem lacunas quando são utilizadas diferentes poses nas imagens faciais, não conseguindo prever a correta localização de marcos faciais ocultos pela face humana. Uma possível abordagem a este problema seria a utilização de bases de dados, na banda espectral do visível, que dispõem de imagens com marcos faciais em partes ocultas da face, de forma a adaptar os modelos previamente treinados. Um exemplo de base de dados é a Menpo [112]. Através desta abordagem, é expectável que o algoritmo consiga extrair os marcos faciais independentemente da pose do utilizador.

Apesar do detetor de falsificação multiespectral ter sido validado como prova de conceito, é necessário comparar com máscaras de outros materiais, nomeadamente silicone. Esta tipologia de máscara é de ampla utilização por serviços de informação. Um caso mediático exemplificativo da utilização desta tipologia de máscaras foi numa reunião na casa branca entre a diretora do centro de disfarces da Agência Central de Inteligência (CIA) com o presidente dos EUA da altura, George Bush, esta conseguiu ocultar a sua identidade durante a reunião. O Presidente apenas teve conhecimento no momento em que a diretora retirou a máscara[13].

Através da dissertação de mestrado é observável a lacuna existente na literatura de bases de dados multiespectrais com diversas imagens em bandas espectrais diferentes. Inicialmente estava proposto a criação de uma base de dados multiespectral, através das câmaras disponibilizadas pela Academia Militar. No entanto, devido à pandemia não foi possível este ano. Como tal, é proposto a construção de uma base de dados multiespectral. A elaboração de uma base de dados desta tipologia, seria bastante benéfico para a comunidade científica, e também para a Academia Militar.

Dada a atual situação pandémica, seria interessante efetuar um estudo sobre o desempenho da utilização de máscaras de proteção individual em sistemas de reconhecimento facial. E, consequentemente, a possibilidade de adaptar uma rede neural convolucional profunda a reconhecer, com sucesso, a identidade de uma pessoa.

[13] `https://www.wsj.com/articles/the-cias-former-chief-of-disguise-drops-her-mask-11576168327`.

Referências

[1] A. Jain, A. Ross, and K. Nandakumar, *Introduction to Biometrics*. Springer, 2011.

[2] R. Munir and R. Khan, "An Extensive Review on Spectral Imaging in Biometric Systems: Challenges and Advancements," *Journal of Visual Communication and Image Representation*, vol. 65, no. 1, p. 14–26, 2019.

[3] W. Zhang, X. Zhao, J. Morvan, and L. Chen, "Improving Shadow Suppression for Illumination Robust Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 611–624, 2019.

[4] A. D'Amico, C. Natale, F. Castro, S. Iarossi, A. Catini, and E. Martinelli, "Volatile Compounds Detection by IR Acousto-Optic Detectors," in *Unexploded Ordnance Detection and Mitigation*. Springer Netherlands, 2009, pp. 21–59.

[5] S. Hu, N. Short, B. Riggan, M. Chasse, and M. Sarfraz, "Heterogeneous Face Recognition: Recent Advances in Infrared-to-Visible Matching," in *International Conference on Automatic Face Gesture Recognition*. Washington DC, USA: IEEE, 2017, pp. 883–890.

[6] H. Steiner, A. Kolb, and N. Jung, "Reliable Face Anti-Spoofing using Multispectral SWIR Imaging," in *International Conference on Biometrics*. Halmstad, Sweden: IEEE, 2016, pp. 1–8.

[7] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network," *IEEE Transactions on Information Forensics and Security*, vol. 15, no. 1, pp. 42–55, 2020.

[8] S. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 Face Database," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Portland, United States of America: IEEE, 2013, p. 348–353.

[9] G. Zhao, X. Huang, M. Taini, S. Li, and M. Pietikainen, "Facial Expression Recognition From Near-Infrared Videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[10] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference," *Image and Vision Computing*, vol. 12, no. 7, p. 682–691, 2010.

[11] T. D. Pereira, A. Anjos, and S. Marcel, "Heterogeneous Face Recognition Using Domain Specific Units," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1803–1816, 2019.

[12] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Mutual Component Analysis for Heterogeneous Face Recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, pp. 1–23, 2016.

[13] W. Hu and H. Hu, "Fine Tuning Dual Streams Deep Network with Multi-Scale Pyramid Decision for Heterogeneous Face Recognition," *Neural Processing Letters*, vol. 50, no. 2, pp. 1465–1483, 2019.

[14] M. Simón, C. Corneanu, K. Nasrollahi, O. Nikisins, S. Escalera, Y. Sun, H. Li, Z. Sun, T. Moeslund, and M. Greitans, "Improved RGB-D-T based Face Recognition," *IET Biometrics*, vol. 5, no. 4, pp. 297–303, 2016.

[15] A. Litvin, K. Nasrollahi, S. Escalera, C. Ozcinar, T. Moeslund, and G. Anbarjafari, "A Novel Deep Network Architecture for Reconstructing RGB Facial Images From Thermal for Face Recognition," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 25 259–25 271, 2019.

[16] Y. Zheng, "Orientation-based Face Recognition Using Multispectral Imagery and Score Fusion," *Optical Engineering*, vol. 50, no. 11, 2011.

[17] D. Huang, J. Sun, and Y. Wang, "The BUAA-VisNir Face Database Instructions," IRIP-TR- 12-FR-001, BEIHANG UNIVERSITY, Beijing, China, Tech. Rep., 2012.

[18] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1761–1773, 2019.

[19] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial Cross-Spectral Face Completion for NIR-VIS Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1025–1037, 2019.

[20] V. Espinosa-Duro, M. Faundez-Zanuy, and J. Mekyska, "A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums," *Cognitive Computation*, vol. 5, no. 1, pp. 119–135, 2013.

[21] C. Chen and A. Ross, "Matching Thermal to Visible Face Images Using Hidden Factor Analysis in a Cascaded Subspace Learning Framework," *Pattern Recognition Letters*, vol. 72, pp. 25–32, 2016.

[22] Y. Jin, J. Lu, , and Q. Ruan, "Coupled Discriminative Feature Learning for Heterogeneous Face Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, 2015.

[23] C. Peng, X. Gao, N. Wang, , and J. Li, "Graphical Representation for Heterogeneous Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 301–312, 2017.

[24] Y. Jin, J. Li, C. Lang, and Q. Ruan, "Multi-task Clustering ELM for VIS-NIR Cross-Modal Feature Learning," *Multidimensional Systems and Signal Processing*, vol. 28, no. 3, pp. 905–920, 2017.

[25] B. Oh, K. Oh, A. Teoh, Z. Lin, and K. Toh, "A Gabor-based Network for Heterogeneous Face Recognition," *Neurocomputing*, vol. 261, pp. 253–265, 2017.

[26] A. Guei and M. Akhloufi, "Deep Learning Enhancement of Infrared Face Images Using Generative Adversarial Networks," *Applied Optics*, vol. 57, no. 18, pp. D98–D107, 2018.

[27] 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. 75 305–75 317, 2019.

[28] C. Peng, X. Gao, N. Wang, and J.Li, "Sparse Graphical Representation Based Discriminant Analysis for Heterogeneous Face Recognition," *Signal Processing*, vol. 156, pp. 46–61, 2019.

[29] C. Peng, N. Wang, J. Li, and X. Gao, "DLFace: Deep Local Descriptor for Cross-Modality Face Recognition," *Pattern Recognition*, vol. 90, pp. 161–171, 2019.

[30] W. Hu and H. Hu, "Discriminant Deep Feature Learning Based on Joint Supervision Loss and Multi-layer Feature Fusion For Heterogeneous Face Recognition," *Computer Vision and Image Understanding*, vol. 184, pp. 9–21, 2019.

[31] C. Peng, N. Wang, J. Li, and X. Gao, "Re-Ranking High-Dimensional Deep Local Representation for NIR-VIS Face Recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4553–4565, 2019.

[32] H. Bae, T. Jeon, Y. Lee, S. Jang, and S. Lee, "Non-Visual to Visual Translation for Cross-Domain Face Recognition," *IEEE Access*, vol. 8, no. 7, pp. 50 452–50 464, 2020.

[33] S. Li, Z. Lei, and M. Ao, "The HFB Face Database for Heterogeneous Face Biometrics research," in *Computer Vision and Pattern Recognition Workshops*. Miami, USA: IEEE, 2009, pp. 1–8.

[34] Z. Lei, S. Liao, A. Jain, and S. Li, "Coupled Discriminant Analysis for Heterogeneous Face Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1707–1716, 2012.

[35] W. Di, L. Zhang, D. Zhang, and Q. Pan, "A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference," *IEEE Transactions on Systems Man and Cybernetics Part A - Systems and Humans*, vol. 40, no. 6, pp. 1354–1361, 2010.

[36] F. Wu, X. Jing, X. Dong, R. Hu, D. Yue, and L. Wang, "Intraspectrum Discrimination and Interspectrum Correlation Analysis Deep Network for Multispectral Face Recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 1009–1022, 2020.

[37] H. Zhao and S. Sun, "Sparse Tensor Embedding Based Multispectral Face Recognition," *Neurocomputing*, vol. 133, pp. 427–436, 2014.

[38] S. Li, D. Yi, Z. Lei, and S. Liao, "Fusion of Visual, Thermal, and Range as a Solution to Illumination and Pose Restrictions in Face Recognition," in *International Carnahan Conference on Security Technology*. Albuquerque, United States of America: IEEE, 2004, pp. 325–330.

[39] M. Bhowmik, P. Saha, A. Singha, D. Bhattacharjee, and P. Dutta, "Enhancement of Robustness of Face Recognition System Through Reduced Gaussianity in Log-ICA," *Expert Systems with Applications*, vol. 116, pp. 96–107, 2019.

[40] M. Kanmani and V. Narasimhan, "Optimal Fusion Aided Face Recognition from Visible and Thermal Face Images," *Multimedia Tools and Applications*, pp. 1–25, 2020.

[41] D. Kang, H. Han, A. Jain, and S. Lee, "Nighttime Face Recognition at Large Standoff: Cross-Distance and Cross-Spectral Matching," *IEEE Transactions on Systems Man and Cybernetics Part A - Systems and Humans*, vol. 47, no. 12, pp. 3750–3766, 2014.

[42] D. Shamia and D. Chandy, "Intelligent System for Cross-Spectral and Cross-Distance Face Matching," *Computers Electrical Engineering*, vol. 71, no. 2, pp. 915–924, 2018.

[43] G. Bebis, A. Gyaourova, S. Singh, and I. Pavlidis, "Face Recognition by Fusing Thermal Infrared and Visible Imagery," *Image and Vision Computing*, vol. 24, no. 7, pp. 727–742, 2006.

[44] 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 Recognition*, vol. 41, no. 3, pp. 880–893, 2008.

[45] S. Sun, H. Zhao, and B. Jin, "Robust Tensor Preserving Projection for Multispectral Face Recognition," *Mathematical Problems in Engineering*, no. 597245, 2013.

[46] J. Bernhard, J. Barr, K. Bowyer, and P. Flynn, "Near-IR to Visible Light Face Matching: Effectiveness of Pre-Processing Options for Commercial Matchers," in *International Conference on Biometrics Theory.* Arlington, United States of America: IEEE, 2015, pp. 1–8.

[47] K. Byrd, "Preview of the Newly Acquired NVESD-ARL Multimodal Face Database," *SPIE DSS*, vol. 8734, p. 8734–8734, 2013.

[48] S. Hu, J. Choi, A. Chan, and W. Schwartz, "Thermal-to-Visible Face Recognition Using Partial Least Squares," *Journal of the Optical Society of America A-Optics Image Science and Vision*, vol. 32, no. 3, pp. 431–442, 2015.

[49] M. Sarfraz and R. Stiefelhagen, "Deep Perceptual Mapping for Cross-Modal Face Recognition," *International Journal of Computer Vision*, vol. 122, no. 3, pp. 426–438, 2017.

[50] N. Osia and T. Bourlai, "Bridging the Spectral Gap Using Image Synthesis: a Study on Matching Visible to Passive Infrared Face Images," *Machine Vision and Applications*, vol. 28, no. 5-6, pp. 649–663, 2017.

[51] S. Hu, N. Short, B. Riggan, C. Gordon, K. Gurton, M. Thielke, P. Gurram, and A. Chan, "A Polarimetric Thermal Database for Face Recognition Research," in *Computer Vision and Pattern Recognition Workshops.* Las Vegas, United States of America: IEEE, 2016, p. 187–194.

[52] R. Martin, M. Sluch, K. Kafka, R. Ice, and B. Lemoff, "Active-SWIR Signatures for Long-Range Night/Day Human Detection and Identification," *Active and Passive Signatures IV*, vol. 8734, no. 4, p. 121–129, 2013.

[53] Z. Cao, N. Schmid, and T. Bourlai, "Composite Multilobe Descriptors for Cross-spectral Recognition of Fulland Partial Face," *Optical Engineering*, vol. 55, no. 8, pp. 83–107, 2016.

[54] F. Nicolo and N. Schmid, "Long Range Cross-Spectral Face Recognition: Matching SWIR Against Visible Light Images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1717–1726, 2012.

[55] A. Seal, D. Bhattacharjee, M. Nasipuri, and D. Basu, "Preview of the Newly Acquired NVESD-ARL Multimodal Face Database," *Multimedia Tools and Applications*, vol. 74, no. 9, p. 2913–2937, 2015.

[56] A. Seal, D. Bhattacharjee, and M. Nasipuri, "Human Face Recognition Using Random Forest Based Fusion of a-trous Wavelet Transform Coefficients From Thermal and Visible Images," *Aeu-International Journal of Electronics and Communications*, vol. 70, no. 8, pp. 1041–1049, 2016.

[57] 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," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 4, p. 1756005, 2017.

[58] X. Chen, P. Flynn, and K. Bowye, "Visible -light and Infrared Face Recognition," in *ACM Workshop on Multimodal User Authentication.* Springer Netherlands, 2003, pp. 48–55.

[59] B. Cao, N. Wang, J. Li, and X. Gao, "Data Augmentation-based Joint Learning for Heterogeneous Face Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1731–1743, 2018.

[60] M. Uzair, A. Mahmood, and A. Mian, "Hyperspectral Face Recognition using 3D-DCT and Partial Least Squares," in *British Machine Vision Conference*, vol. 1. Bristol, United Kingdom: BMVA, 2013.

[61] T. Bourlai, N. Mavridis, and N. Narang, "On Designing Practical Long Range Near Infrared-based Face Recognition Systems," *Image and Vision Computing*, vol. 52, pp. 25–41, 2016.

[62] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," in *European Conference on Computer Vision.* Amsterdam, The Netherlands: Springer International Publishing, 2016, pp. 87–102.

[63] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep Face Recognition: A Survey," in *Conference on Graphics, Patterns and Images*, vol. 1. Parana, Brazil: IEEE, 2018, pp. 471–478.

[64] Y. Zhao, J. Lin, Q. Xuan, , and X. Xi, "HPILN: a Feature Learning Framework for Cross-Modality Person Re-Identification," *IET Image Processing*, vol. 13, no. 14, pp. 2897–2904, 2019.

[65] M. Song, X. Shang, and C. Chang, "3-D Receiver Operating Characteristic Analysis for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–23, 2020.

[66] N. Short, S. Hu, P. Gurram, K. Gurton, and A. Chan, "Long Range Cross-Spectral Face Recognition: Matching SWIR Against Visible Light Images," *IEEE Transactions on Information Forensics and Security*, vol. 40, no. 6, pp. 882–885, 2015.

[67] B. Klare and A. Jain, "Heterogeneous Face Recognition Using Kernel Prototype Similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, 2013.

[68] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous Face Recognition: A Common Encoding Feature Discriminant Approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2079–2089, 2017.

[69] S. Angadi and S. Hatture, "Face Recognition Through Symbolic Modeling of Face Graphs and Texture," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 12, p. 1956008, 2019.

[70] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Li, "Regularized Discriminative Spectral Regression Method for Heterogeneous Face Matching," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 353–362, 2013.

[71] K. Gurton, A. Yuffa, and G. Videen, "Enhanced Facial Recognition for Thermal Imagery Using Polarimetric Imaging," *Optics Letters*, vol. 39, no. 13, pp. 3857–3859, 2014.

[72] Z. Xie, S. Zhang, X. Yu, and G. Liu, "Infrared and Visible Face Fusion Recognition Based on Extended Sparse Representation Classification and Local Binary Patterns for the Single Sample Problem," *Journal of Optical Technology*, vol. 86, no. 7, pp. 408–413, 2019.

[73] J. Gui, Z. Sun, J. Cheng, S. Ji, and X. Wu, "How to Estimate the Regularization Parameter for Spectral Regression Discriminant Analysis and its Kernel Version?" *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 211–223, 2014.

[74] K. Siwek and S. Osowski, "Deep Neural Networks and Classical Approach to Face Recognition - Comparative Analysis," *Przeglad Elektrotechniczny*, vol. 94, no. 4, pp. 1–4, 2018.

[75] J. Chmielinska and J. Jakubowski, "Face Recognition Based on Deep Learning Techniques and Image Fusion," *Przeglad Elektrotechniczny*, vol. 95, no. 11, pp. 150–154, 2019.

[76] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, 1998.

[77] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Heterogeneous Face Recognition Using Kernel Prototype Similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1410–1422, 2012.

[78] T. Quan, D. Hildebrand, and W. Jeong, "FusionNet: A Deep Fully Residual Convolutional Neural Network for Image Segmentation in Connectomics," 2016.

[79] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[80] N. Bento, J. Silva, and J. Bioucas-Dias, "Detection of Camouflaged People," *International Journal of Sensor Networks and Data Communications*, vol. 5, no. 3, pp. 143–148, 2016.

[81] M. Liggins, D. Hall, and J. Llinas, *Handbook of Multisensor Data Fusion: Theory and Practice*. CRC Press, 2017.

[82] Z. Chai, Z. Sun, H. Mendez-Vazquez, R. He, and T. Tan, "A Comprehensive Database for Benchmarking Imaging Systems," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 1, p. 14–26, 2014.

[83] M. Omran, A. Engelbrecht, and A. Salman, "Particle Swarm Optimization Method for Image Clustering," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 3, pp. 297–321, 2005.

[84] G. F. C. S. G. Hermosilla, F. Gallardo, "Face Recognition Based on Deep Learning Techniques and Image Fusion," *Sensors*, vol. 15, no. 8, p. 17944–17962, 2015.

[85] M. Krišto and M. Ivasic-Kos, "An Overview of Thermal Face Recognition Methods," in *International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia, 2018, pp. 1098–1103.

[86] P. Ramachandran, B. Zoph, and Q. Le, "Searching for Activation Functions," in *International Conference on Learning Representations*. Vancouver, Canada: OpenReview.net, 2018.

[87] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep Learning Applications in Medical Image Analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[88] F. Schroff, D. Kalenichenk, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*. Boston, United States of America: IEEE, 2015, pp. 815–823.

[89] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, Inception-ResNet and the Impact of Residual Connections on Learning," in *AAAI Conference on Artificial Intelligence*. San Francisco, United States of America: AAAI Press, 2017, p. 4278–4284.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, United States of America: IEEE, 2016, pp. 770–778.

[91] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[92] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "SSD: Single Shot MultiBox Detector," in *European Conference on Computer Vision*, vol. 9905. Amsterdam, The Netherlands: Springer International Publishing, 2016.

[93] D. Acharya, Z. Huang, D. Paudel, and L. Van Gool, "Covariance Pooling for Facial Expression Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Salt Lake City, United States of America: IEEE, 2018, pp. 480–4807.

[94] D. King, "Dlib-ml: A Machine Learning Research," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[95] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment With an Ensemble of Regression Trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, United States of America, 2014, pp. 1867–1874.

[96] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces In-The-Wild Challenge: Database and Results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[97] H. Steiner, S. Sebastian, K. Andreas, and J. Norbert, "Design of an Active Multispectral SWIR Camera System for Skin Detection and Face Verification," *Journal of Sensors*, vol. 2016, pp. 1–16, 2016.

[98] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning Face Representation from Scratch," 2014.

[99] X. Fu, J. Lu, X. Zhang, X. Yang, and I. Unwala, "Intelligent In-Vehicle Safety and Security Monitoring System with Face Recognition," in *IEEE International Conference on Computational Science and IEEE International Conference on Embedded and Ubiquitous Computing Engineering*. New York, United States of America: IEEE, 2019, pp. 225–229.

[100] L. Hu, M. Huang, S. Ke, and C. Tsai, "The Distance Function Effect on k-Nearest Neighbor Classification for Medical Datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, 2016.

[101] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A Comprehensive Database for Benchmarking Imaging Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, 2020.

[102] F. Chelali, N. Cherabit, and A. Djeradi, "Face Recognition System Using Skin Detection in RGB and YCbCr Color Space," in *World Symposium on Web Applications and Networking*. Sousse, Tunísia: IEEE, 2015, pp. 1–7.

[103] S. Kolkur, D. Kalbande, P. Shimpi, C. Bapat, and J. Jatakia, "Human Skin Detection Using RGB, HSV and YCbCr Color Models," in *International Conference on Communication and Signal Processing*. Lonere, India: Atlantis Press, 2016, pp. 324–332.

[104] D. Mishkin, N. Sergievskiy, and J. Matas, "Systematic Evaluation of Convolution Neural Network Advances on the Imagenet," *Computer Vision and Image Understanding*, vol. 161, no. C, pp. 11–19, 2017.

[105] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[106] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*. San Diego, USA: Yoshua Bengio and Yann LeCun, 2015.

[107] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, p. 1929–1958, 2014.

[108] G. Forman and M. Scholz, "Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, p. 49–57, 2010.

[109] I. Tsamardinos, A. Rakhshani, and V. Lagani, "Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization," *International Journal on Artificial Intelligence Tools*, vol. 24, no. 5, p. 1540023, 2015.

[110] L. Kezebou, V. Oludare, K. Panetta, and S. Agaian, "TR-GAN: Thermal to RGB Face Synthesis with Generative Adversarial Network for Cross-Modal Face Recognition," in *Mobile Multimedia/Image Processing*, vol. 11399. SPIE, 2020.

[111] S. Rajeev, K. Shreyas, Q. Wan, K. Panetta, and S. Agaian, "Illumination Invariant NIR Face Recognition Using Directional Visibility," *Electronic Imaging, Image Processing: Algorithms and Systems*, pp. 273–1–273–7, 2019.

[112] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The Menpo Facial Landmark Localisation Challenge: A Step Towards the Solution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 2116–2125.

A Fluxograma Detalhado da Metodologia

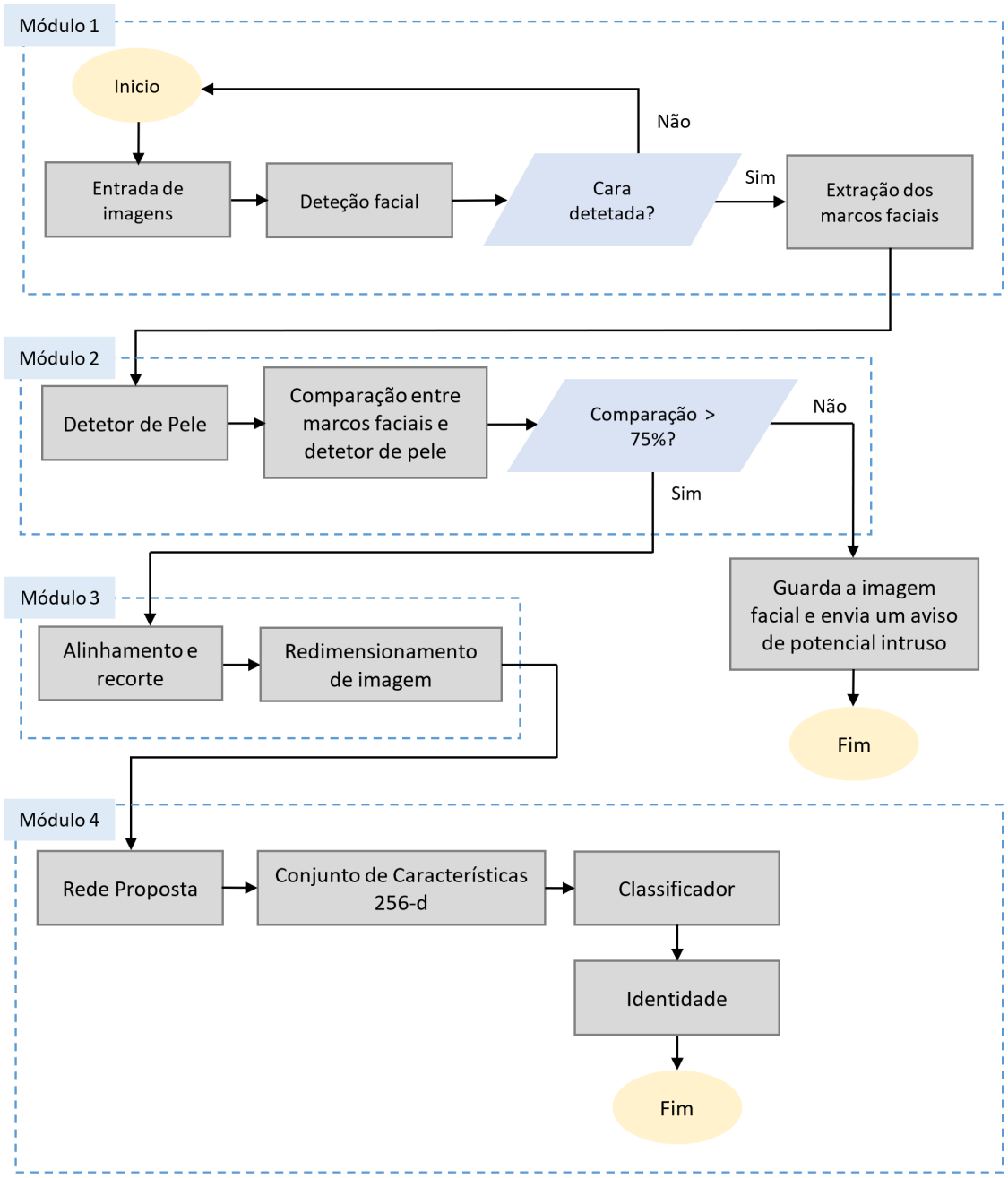


Figura 45: Fluxograma da metodologia adoptada, versão detalhada.

B Extração dos Marcos Faciais

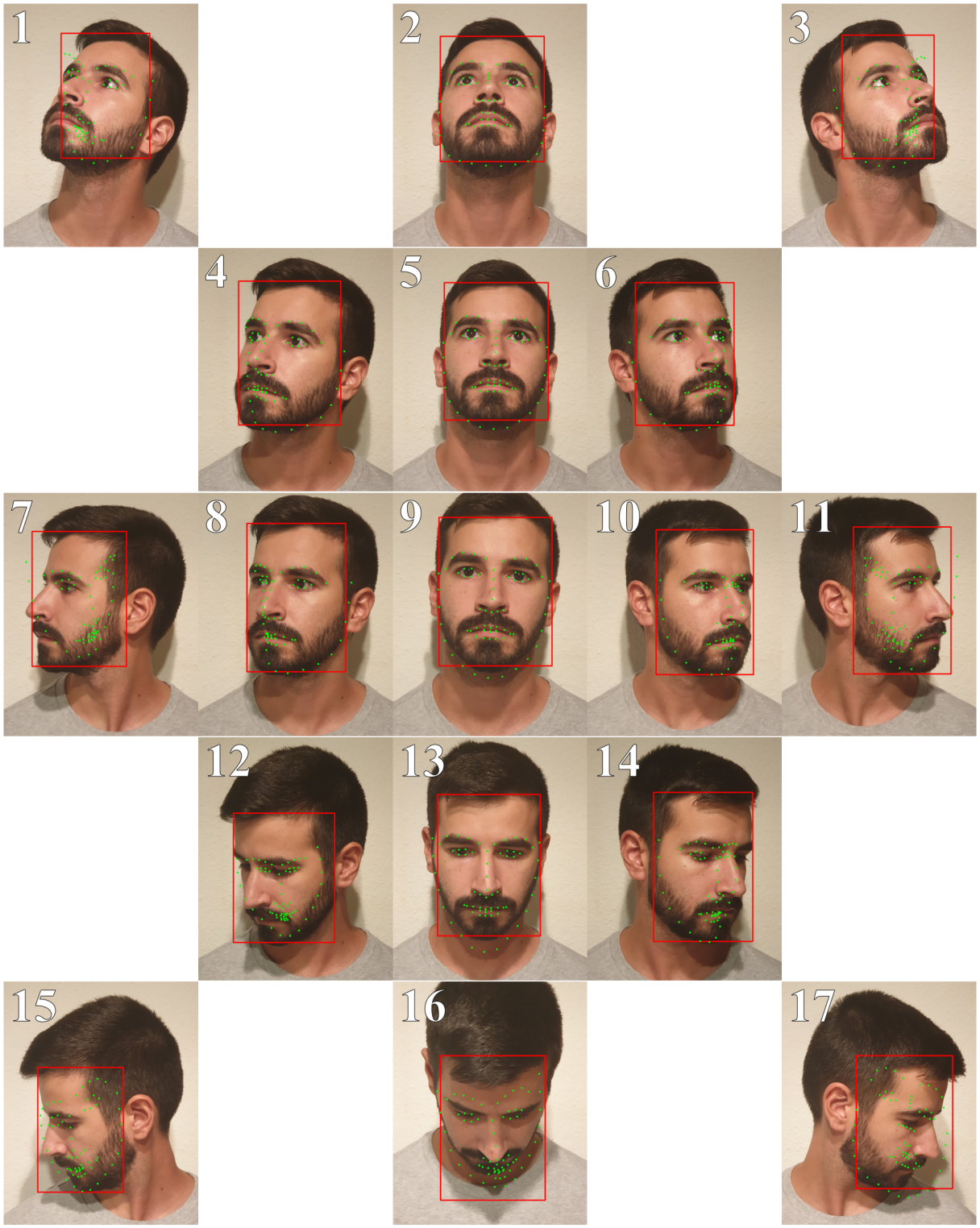


Figura 46: Ilustração dos marcos faciais extraídos das imagens iniciais, após deteção facial.

C Resultados Numéricos para o Estudo dos Hiperparâmetros

SVM-Linear

Resultados numéricos do afinamento do hiperparâmetro C para o classificador SVM-Linear nas bases de dados multiespectrais Tufts [101] e CASIA NIR-VIS 2.0 [8].

Tabela 15: Afinamento do hiperparâmetro C para o classificador SVM-Linear, para a base de dados Tufts [101].

C	***Rank-1*** **(Valor Médio)**	***Rank-1*** **(Desvio Padrão)**
10^{-10}	8,91 %	1,29 %
10^{-9}	8,91 %	1,29 %
10^{-8}	8,91 %	1,29 %
10^{-7}	8,91 %	1,29 %
10^{-6}	8,91 %	1,29 %
10^{-5}	10,50 %	0,48 %
10^{-4}	75,76 %	1,67 %
10^{-3}	99,77 %	0,22 %
10^{-2}	**99,89 %**	**0,09 %**
10^{-1}	99,89 %	0,09 %
10^{+0}	99,89 %	0,09 %
10^{+1}	99,89 %	0,09 %
10^{+2}	99,89 %	0,09 %
10^{+3}	99,89 %	0,09 %
10^{+4}	99,89 %	0,09 %
10^{+5}	99,89 %	0,09 %

Tabela 16: Afinamento do hiperparâmetro C para o classificador SVM-Linear, para a base de dados CASIA NIR-VIS 2.0 [8].

C	**Rank-1 (Valor Médio)**	**Rank-1 (Desvio Padrão)**
10^{-10}	0,40 %	0,00 %
10^{-9}	0,40 %	0,00 %
10^{-8}	0,40 %	0,00 %
10^{-7}	0,40 %	0,00 %
10^{-6}	0,40 %	0,00 %
10^{-5}	15,53 %	0,28 %
10^{-4}	99,80 %	0,07 %
10^{-3}	**99,86 %**	**0,06 %**
10^{-2}	99,86 %	0,06 %
10^{-1}	99,86 %	0,06 %
10^{+0}	99,86 %	0,06 %
10^{+1}	99,86 %	0,06 %
10^{+2}	99,86 %	0,06 %
10^{+3}	99,86 %	0,06 %
10^{+4}	99.86 %	0.06 %
10^{+5}	99.86 %	0.06 %

kNN

Resultados numéricos do afinamento do hiperparâmetro k para o classificador kNN nas bases de dados multiespectrais Tufts [101] e CASIA NIR-VIS 2.0 [8].

Tabela 17: Afinamento do hiperparâmetro k para o classificador kNN, para a base de dados Tufts [101].

Número de Vizinhos	***Rank-1*** **(Valor Médio)**	***Rank-1*** **(Desvio Padrão)**
1	**99,54 %**	**0,35 %**
2	98,90 %	0,37 %
3	99,24 %	0,27 %
4	98,33 %	0,83 %
5	98,18 %	0,80 %
6	97,60 %	0,73 %
7	97,50 %	0,84 %
8	97,00 %	0,76 %
9	96,66 %	0,99 %
10	96,66 %	0,80 %

Tabela 18: Afinamento do hiperparâmetro k para o classificador kNN, para a base de dados CASIA NIR-VIS 2.0 [8].

Número de Vizinhos	***Rank-1*** **(Valor Médio)**	***Rank-1*** **(Desvio Padrão)**
1	**99,63 %**	**0,25 %**
2	99,47 %	0,38 %
3	99,43 %	0,29 %
4	99,30 %	0,39 %
5	99,31 %	0,36 %
6	99,25 %	0,24 %
7	99,10 %	0,45 %
8	99,01 %	0,33 %
9	98,88 %	0,46 %
10	98,69 %	0,43 %

D Estudo para Futura Implementação na Academia Militar

Neste apêndice é elaborado um estudo com o propósito de implementar um sistema de reconhecimento facial multiespectral na Academia Militar. Este estudo é pertinente dado o número substancial de locais sensíveis, nomeadamente arrecadações de material de guerra, e também, de material sensível.

Durante o estudo este estudo foi necessário escolher equipamentos que sejam confiáveis a longo prazo (i.e., tenham um prazo de vida relativamente longos) e que o material/equipamento justifique o preço monetário do mesmo.

Os equipamentos considerados para este estudo foram relativamente às câmaras, servidores e algoritmos (i.e., código) utilizados pelo sistema de reconhecimento facial multiespectral. Equipamento de reduzido custo não foram considerados para este estudo monetário, como por exemplo, (i) cabos para interligar os diversos equipamentos, e (ii) caixas para armazenar e proteger os diversos equipamentos.

Foi dado escolha a compra de uma única câmara multiespectral, nas bandas espectrais do VIS, SWIR e LWIR. No entanto, a inexistência de uma câmara multiespectral no mercado com as três bandas impossibilitou a escolha de uma.

Após um estudo do mercado de câmaras multiespectrais, constata-se a existência de câmaras multiespectrais nas bandas espectrais VIS e LWIR, em conjunto (i.e., no mesmo equipamento). Contudo, monetariamente, é preferível adquirir câmaras com bandas espectrais distintas (i.e., uma câmara para cada banda espectral). Como tal, foi dado primazia a câmaras de banda espectral única.

As câmaras multiespectrais que se melhor adequam ao estudo, tendo em consideração aos requisitos monetários e específicos, são: (i) Wyze Cam Pan, banda espectral do VISível, (ii) FLIR Tau SWIR, banda espectral do SWIR, e por último, (iii) FLIR Boson, banda espectral do LWIR. Dado o elevado preço das câmaras multiespectrais SWIR e LWIR, foi dado primazia a empresas com renome no mercado, como é o caso da empresa FLIR.

Com principal de armazenar a informação da base de dados (i.e., imagens em conjunto com a identificação da pessoa) e executar o sistema de reconhecimento facial multiespectral é necessário escolher um servidor. Dado a elevada oferta no mercado, a escolha de servidores é relativamente simplificada. Do servidor escolhido destacam-se duas características principais: os 8 GB de RAM, e os 2 TB de armazenamento interno.

Apesar de o código desenvolvido por nós utilizar algoritmos e bibliotecas desenvolvidos por outros, estes possuem uma licença BSD. Uma licença BSD requer que para o sistema, se for utilizado para fins comerciais, seja colocado um aVISo de isenção de responsabilidade.

Na Tabela 19 encontra-se disponível o equipamento necessário para cada módulo e o seu respetivo preço.

Tabela 19: Preçário do material necessário para a concretização de um sistema de reconhecimento facial multiespectral.

Módulo	Equipamento	Preço
Câmaras	VIS - Wyze Cam Pan	35,0 €
	SWIR – FLIR Tau SWIR	12 660,0 €
	LWIR – FLIR Boson	2 500,0 €
Servidor	Fujitsu Primergy	609,9 €
Processamento de Imagem	Código em Python	* 0,0 €
	Algoritmo de Deteção Facial	** 0,0 €
	Algoritmo de Extração de Marcos Faciais	* 0,0 €
Deteção de Falsificação	Código em Python	* 0,0 €
Processamento Facial	Código em Python	* ** 0,0 €
Reconhecimento Facial	Código em Python	* ** 0,0 €
	Custo Total	**15 804,9 €**

O sistema de reconhecimento facial multiespectral proposto é passível de ser empregue em locais onde seja necessário efetuar reconhecimento facial com fidelidade/desempenho elevado. Proferindo assim uma elevada camada de proteção ao local implementado. Os locais possíveis de implementação podem ser desde paióis, arrecadações de material de guerra os gabinetes ou salas com material sensível.

* Código elaborado por minha pessoa.

** Algumas bibliotecas (não elaboradas pela minha pessoa) utilizadas no código são de licença BSD. Como tal, é no sentido legal é possível utilizar o código para fins comerciais. Sendo necessário, no entanto, colocar um aviso de isenção de responsabilidade.