Centro Federal de Educação Tecnológica de Minas Gerais
Programa de Pós-graduação em Modelagem Matemática e Computacional

Descrição de Características Locais para Correspondência entre Imagens do Espectro Visível e Infravermelho

Cristiano Fraga Guimarães Nunes

Orientador: Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Agosto de 2023

Cristiano Fraga Guimarães Nunes

Descrição de Características Locais para Correspondência entre Imagens do Espectro Visível e Infravermelho

Tese apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional.

Linha de pesquisa: Métodos Matemáticos Aplicados.

Orientador: Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Centro Federal de Educação Tecnológica de Minas Gerais
Programa de Pós-graduação em Modelagem Matemática e Computacional
Belo Horizonte
Agosto de 2023

Nunes, Cristiano Fraga Guimarães
N972d Descrição de características locais para correspondência entre imagens do espectro visível e infravermelho / Cristiano Fraga Guimarães Nunes. – 2023.
137 f.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Flávio Luis Cardeal Pádua.
Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Imagem multiespectral – Teses. 2. Reconhecimento de padrões – Teses. 3. Imagens infravermelhas – Teses. 4. Processamento de imagem – Teses. I. Pádua, Flávio Luis Cardeal. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título.

CDD 006.37

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6ª 1592 / Cefet/MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"DESCRIÇÃO DE CARACTERÍSTICAS LOCAIS PARA CORRESPONDÊNCIA ENTRE IMAGENS DO ESPECTRO VISÍVEL E INFRAVERMELHO".

Tese de Doutorado apresentada por **Cristiano Fraga Guimarães Nunes**, em 31 de agosto de 2023, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Documento assinado digitalmente
FLAVIO LUIS CARDEAL PADUA
Data: 31/08/2023 14:25:25-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Flávio Luis Cardeal Pádua (Orientador)
Centro Federal de Educação Tecnológica de Minas Gerais

Documento assinado digitalmente
DARLAN NUNES DE BRITO
Data: 31/08/2023 16:44:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Darlan Nunes Brito
Universidade Federal de Ouro Preto

Documento assinado digitalmente
CELSO LUIZ DE SOUZA
Data: 31/08/2023 17:06:36-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Celso Luiz de Souza
Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais

Documento assinado digitalmente
ELIZABETH FIALHO WANNER
Data: 02/09/2023 06:22:28-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Elizabeth Fialho Wanner
Centro Federal de Educação Tecnológica de Minas Gerais

Documento assinado digitalmente
ADRIANO CESAR MACHADO PEREIRA
Data: 31/08/2023 15:38:58-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Adriano César Machado Pereira
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação Stricto Sensu em Modelagem Matemática e Computacional

Agradecimentos

Ao refletir sobre esta importante etapa da minha jornada, é essencial expressar minha profunda gratidão a todos que, de forma direta ou indireta, contribuíram para essa realização.

Em primeiro lugar, agradeço a Deus e a Nossa Senhora da Conceição Aparecida, por me proporcionarem saúde, força e sabedoria ao longo deste percurso. Minha fé me manteve forte nos momentos mais desafiadores.

À minha família, meus pais e irmãs, cuja presença constante e amor incondicional foram essenciais para minha formação. Vocês foram minha âncora, mesmo nos momentos em que minha presença se fez mais rara.

Ao meu estimado orientador, professor Flávio Cardeal, minha eterna gratidão por toda a dedicação, paciência e confiança depositadas em mim. Não foi somente seu conhecimento acadêmico que fez a diferença, mas também sua humanidade e capacidade de compreender e orientar-me nos momentos em que mais necessitava.

À minha querida amiga Bruna, minha gratidão é imensa. Sua ajuda incansável na revisão deste texto foi crucial, e sua paciência e dedicação não têm preço. Muito obrigado pelo auxílio nessa fase tão importante.

Estendo também meus agradecimentos aos colegas do Grupo de Pesquisas Interdisciplinares em Informação Multimídia do CEFET-MG e a todos os amigos que fiz durante o doutorado. Nossa convivência enriqueceu este percurso e reforçou o sentimento de coletividade.

Por fim, e não menos importante, agradeço ao CEFET-MG, instituição que não apenas se destaca por seu ensino público, gratuito e de qualidade, mas também foi vital em minha formação.

Caminhante, não há caminho. O caminho se faz ao caminhar.

(Antônio Machado, poeta Espanhol)

Resumo

Neste trabalho são propostos dois métodos descritores, intitulados como MF-Net (*Multispectral Features Network*) e SORIF (*Scale-Orientation Robust Infrared Features*), projetados para realizar a descrição de características locais no problema da correspondência entre imagens provenientes do espectro visível (VIS) e do espectro infravermelho (IV). As imagens do espectro infravermelho são fundamentais na área de Visão Computacional, uma vez que possibilitam a extração de informações adicionais ao espectro visível. Neste cenário, um dos desafios mais relevantes relacionados à descrição de características locais reside na falta de uma relação previsível entre os valores de intensidade dos pixels das imagens capturadas das diferentes faixas do espectro eletromagnético. O primeiro método, MF-Net, emprega uma abordagem baseada em aprendizagem profunda, inspirada no recente sucesso de abordagens baseadas em aprendizado de máquina em várias aplicações. Este método incorpora uma nova camada de aprendizagem profunda para lidar com relações não-monotônicas nos valores de intensidade das imagens. Comparativamente a outros métodos baseados em aprendizagem profunda, o método descritor MF-Net demonstrou superior eficácia em termos de precisão e maior eficiência em tempo de processamento para treinamento da rede. O segundo método descritor deste trabalho, SORIF, foi desenvolvido para suprir uma lacuna na literatura sobre a descrição de características locais, em que os métodos descritores baseados em modelagem manual apresentavam limitações ao lidar com variações geométricas em imagens VIS/IV. Uma estratégia adotada por este método consistiu na introdução de uma etapa para calcular a orientação de parte da imagem a ser descrita. Adicionalmente, provou ser eficaz em certas variações geométricas, como na rotação das imagens, o que é extremamente útil em aplicações práticas. Ambos os métodos foram avaliados em um processo de correspondência de características locais de imagens, por meio das medidas de desempenho de precisão e revocação. Os experimentos foram conduzidos por meio da utilização de bases de dados amplamente reconhecidas na literatura e disponíveis publicamente, que incluem pares de imagens obtidas do espectro visível e do espectro infravermelho. Em suma, os progressos possibilitados pelos métodos descritores apresentados neste trabalho representam avanços na descrição de características locais em imagens VIS/IV para aplicação no problema da correspondência entre imagens.

Palavras-chave: Descritor de características locais. Correspondência entre imagens. Imagens infravermelho.

Abstract

This work proposes two descriptor methods, titled MF-Net (Multispectral Features Network) and SORIF (Scale-Orientation Robust Infrared Features), designed for local feature description in images from the visible spectrum (VIS) and the infrared spectrum (IV). Infrared spectrum images are fundamental in Computer Vision, as they allow the extraction of additional information to the visible spectrum. In this scenario, one of the most significant challenges related to describing local features lies in the need for a predictable relationship between the intensity values of the pixels from images captured from different electromagnetic spectrum bands. The first method, MF-Net, employs an approach based on deep learning inspired by the recent success of machine learning-based approaches in many applications. This method incorporates a new deep learning layer to deal with non-monotonic relationships in image intensity values. Compared to other deep learning-based methods, the MF-Net descriptor method has shown superior accuracy and greater processing time efficiency for network training. The second descriptor method of this work, SORIF, was developed to address a gap in the literature on the description of local features, where handcrafted descriptor methods had limitations in dealing with geometric variations in VIS/IV images. A strategy adopted by this method involved introducing a step to calculate the orientation of part of the image. As a result, the method outperformed other state-of-the-art algorithms. Additionally, it proved effective concerning the geometric variations of images, such as rotation variations between images, which is extremely useful in practical applications. Both methods were evaluated by matching local features using precision and recall performance measures. The experiments were conducted using databases widely recognized in the literature and publicly available, which include pairs of images obtained from the visible and infrared spectrums. In summary, the advancements promoted by the descriptor methods presented in this work represent advances in describing local features in VIS/IV images for application in the image matching problem.

Keywords: Local feature descriptor. Feature matching. Infrared images.

Lista de Figuras

Figura 1 – Principais técnicas aplicadas ao problema da correspondência entre imagens. 16
Figura 2 – Exemplo de características locais (cantos) extraídas de uma imagem. . . . 18
Figura 3 – Variação nos valores de intensidade dos pixels ao longo de um segmento de reta. 19
Figura 4 – Escopo do trabalho no problema da correspondência entre imagens. . . . 20
Figura 5 – Exemplos de aplicações em que é importante a utilização das imagens VIS/IV. 21
Figura 6 – Exemplo de cenário considerado neste trabalho. 24
Figura 7 – Exemplos de características locais em uma imagem. 27
Figura 8 – Etapas de um processo convencional de correspondência entre imagens. . 29
Figura 9 – Descritor de uma característica local. 30
Figura 10 – Processo para a correspondência de dois descritores. 31
Figura 11 – Erro entre o ponto-chave detectado e sua projeção. 33
Figura 12 – Exemplo de aplicação das medidas de desempenho de revocação e precisão. 34
Figura 13 – Espectro eletromagnético. 35
Figura 14 – Exemplo de composição de uma imagem multiespectral. 36
Figura 15 – Exemplos de imagens do espectro infravermelho para um mesmo objeto (face humana). 36
Figura 16 – Exemplos de imagens do espectro infravermelho para uma mesma cena. . 37
Figura 17 – Valores de intensidade dos pixels em uma imagem multiespectral. 38
Figura 18 – Analogia entre um neurônio e um neurônio artificial. 39
Figura 19 – Detalhes de um neurônio artificial. 40
Figura 20 – Funções de ativação comumente utilizadas em redes neurais. 41
Figura 21 – Arquitetura de rede neural de múltiplas camadas. 42
Figura 22 – Arquitetura típica de uma CNN. 43
Figura 23 – Camada de convolução com dois filtros. 44
Figura 24 – Exemplo de funcionamento da camada de agrupamento. 45
Figura 25 – Camada totalmente conectada. 46
Figura 26 – Estrutura utilizada para o treinamento da arquitetura DeepDesc. 51
Figura 27 – Estrutura utilizada para o treinamento das arquiteturas PN-Net e TFeat. . 52
Figura 28 – Estrutura de descrição da arquitetura L2-Net. 53
Figura 29 – Estrutura utilizada para o treinamento da arquitetura Q-Net. 58
Figura 30 – Objetivo do método MF-Net: calcular um descritor a partir de uma janela recortada. 65
Figura 31 – Metodologia para construção e avaliação da arquitetura MF-Net. 65
Figura 32 – Exemplo de funcionamento da camada de mapeamento. 66

Figura 33 – Combinação das imagens na camada de mapeamento. 68
Figura 34 – Exemplo das etapas realizadas na camada de mapeamento. 69
Figura 35 – Diferentes arquiteturas avaliadas. 69
Figura 36 – Detalhes da arquitetura construída neste trabalho. 70
Figura 37 – Estrutura utilizada para o treinamento da rede. 71
Figura 38 – Metodologia utilizada para avaliação dos métodos descritores baseados em aprendizagem profunda. 74
Figura 39 – Exemplos de pares de imagens da base de dados Potsdam. 76
Figura 40 – Exemplos de pares de imagens da base de dados EPFL. 77
Figura 41 – Curvas de precisão por revocação para diferentes arquiteturas na base de dados Potsdam. 79
Figura 42 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam. 80
Figura 43 – Curvas de precisão por revocação para diferentes arquiteturas na base de dados EPFL. 80
Figura 44 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL. 81
Figura 45 – Curvas de precisão por revocação para a base de dados Potsdam. 82
Figura 46 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam. 83
Figura 47 – Curvas de precisão por revocação para a base de dados EPFL. 83
Figura 48 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL. 84
Figura 49 – Valores médios para o tempo de processamento em cada par de imagens. 85
Figura 50 – Valores médios para o tempo de treinamento em cada conjunto de janelas. 85
Figura 51 – Resultados qualitativos para um par de images da base Potsdam. 86
Figura 52 – Resultados qualitativos para um par de images da base EPFL. 87
Figura 53 – Extração da sub-região e cálculo do histograma. 93
Figura 54 – Montagem do descritor SORIF. 94
Figura 55 – Metodologia utilizada para avaliação dos métodos descritores baseados em modelagem manual. 96
Figura 56 – Exemplos de pares de imagens das bases de dados utilizadas nos experimentos. 98
Figura 57 – Curvas de precisão por revocação para a base de dados Potsdam. 99
Figura 58 – Curvas de precisão por revocação para a base de dados EPFL. 99
Figura 59 – Curvas de precisão por revocação para a base de dados CVC. 100
Figura 60 – Curvas de precisão por revocação para a base de dados CVC2. 101
Figura 61 – Curvas de precisão por ângulo de rotação da imagem. 101
Figura 62 – Curvas de precisão por valores de escala. 102
Figura 63 – Valores médios para o tempo de processamento em cada par de imagens. 103
Figura 64 – Curvas de precisão por revocação para a base de dados Potsdam. 106
Figura 65 – Diferenças entre os filtros Gabor e Log-Gabor em relação à frequência de resposta. 126

Figura 66 – Exemplo de imagens de resposta aos filtros Log-Gabor. 127
Figura 67 – Diferenças na extração de bordas entre uma imagem do espectro visível e uma imagem do espectro infravermelho para uma mesma cena. 128
Figura 68 – Pirâmide de escala montada pelo algoritmo SIFT. 129
Figura 69 – Detecção de máximos e mínimos locais do algoritmo SIFT. 130
Figura 70 – Etapas para a construção do descritor SIFT. 131
Figura 71 – Exemplo de funcionamento do detector de cantos FAST. 132
Figura 72 – Construção do descritor EOH. 133
Figura 73 – Cálculo do descritor TFeat. 134
Figura 74 – Mapeamento de pontos entre dois planos. 135

Lista de Tabelas

Tabela 1 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam. 136
Tabela 2 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL. 136
Tabela 3 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam. 136
Tabela 4 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL. 137

Lista de Quadros

Quadro 1 – Principais trabalhos apresentados sobre métodos descritores baseados em aprendizagem profunda (*deep learning*). 63
Quadro 2 – Principais trabalhos apresentados sobre métodos descritores baseados em modelagem manual (*handcrafted*). 63
Quadro 3 – Camadas da arquitetura TFeat. 134
Quadro 4 – Parâmetros do algoritmo SIFT. 138
Quadro 5 – Parâmetros do algoritmo FAST. 138
Quadro 6 – Parâmetros dos algoritmos MFD e LGHD. 138
Quadro 7 – Implementações utilizadas nos experimentos deste trabalho. 139

Lista de Abreviaturas e Siglas

BRIEF	*Binary Robust Independent Elementary Features*
BRISK	*Binary Robust Invariant Scalable Keypoints*
CNN	*Convolutional Neural Network*
DOAP	*Descriptors Optimized for Average Precision*
EHD	*Edge Histogram Descriptor*
EOH	*Edge Oriented Histogram*
FAST	*Features from Accelerated Segment Test*
FFT	*Fast Fourier Transform*
FIR	*Far Infrared*
FREAK	*Fast Retina Keypoint*
GFTT	*Good Features to Track*
IV	Infravermelho
LGHD	*Log-Gabor Histogram Descriptor*
LWIR	*Long-Wave Infrared*
MF-Net	*Multispectral Features Network*
MFD	*Multispectral Feature Descriptor*
MPEG	*Moving Picture Experts Group*
MWIR	*Mid-Wave Infrared*
NIR	*Near-Infrared*
NNDR	*Nearest Neighbor Distance Ratio*
ORB	*Oriented FAST and Rotated BRIEF*
RANSAC	*RANdom SAmple Consensus*
ReLU	*Rectified Linear Unit*
RIFT	*Radiation-variation Insensitive Feature Transform*
SAR	*Synthetic Aperture Radar*
SIFT	*Scale Invariant Feature Transform*
SORIF	*Scale-Orientation Robust Infrared Features*
SURF	*Speeded Up Robust Features*
SWIR	*Short-Wave Infrared*
VIS	Visível
VIS/IV	Visível/Infravermelho

Sumário

1 – Introdução **16**
1.1 Motivação 20
1.2 Definição do problema de pesquisa 23
1.3 Objetivos 25
1.4 Contribuições 25
1.5 Organização do texto 26

2 – Fundamentação Teórica **27**
2.1 Características locais 27
2.2 Correspondência de características locais 28
2.3 Medidas de desempenho para métodos descritores 32
2.4 Imagens do espectro infravermelho 35
2.5 Redes neurais artificiais 38
2.6 Redes neurais convolucionais 41
2.6.1 Camada de convolução 43
2.6.2 Camada de ativação 45
2.6.3 Camada de agrupamento 45
2.6.4 Camada totalmente conectada 46

3 – Trabalhos Relacionados **47**
3.1 Correspondência entre imagens 47
3.2 Métodos descritores para imagens obtidas do espectro visível 49
3.3 Métodos descritores para imagens VIS/IV 55
3.4 Discussão sobre a literatura 60

4 – *Multispectral Features Network* (MF-Net) **64**
4.1 Camada de mapeamento 66
4.2 Arquitetura da rede 69
4.3 Treinamento da rede 71
4.4 Metodologia de avaliação 73
4.4.1 Bases de dados 76
4.4.2 Procedimento estatístico 77
4.5 Resultados experimentais 78
4.5.1 Análise exploratória sobre as diferentes arquiteturas 78
4.5.2 Avaliação de desempenho da arquitetura MF-Net 82
4.6 Discussão 86

5 – *Scale-Orientation Robust Infrared Features* (SORIF) **90**
5.1 Método proposto 90
5.1.1 Cálculo da orientação 91
5.1.2 Cálculo do descritor 92
5.2 Metodologia de avaliação 95
5.2.1 Bases de dados 97
5.3 Resultados experimentais 98
5.4 Discussão 103

6 – Discussão Geral **106**

7 – Conclusão **109**

Referências **112**

Apêndices **124**

APÊNDICE A – Filtros Log-Gabor **125**

APÊNDICE B – Métodos para descrição de características locais **129**
B.1 Algoritmo SIFT 129
B.2 Algoritmo FAST 132
B.3 Algoritmo EOH 132
B.4 Algoritmo TFeat 133

APÊNDICE C – Homografia 2D **135**

APÊNDICE D – Tabelas de resultados do método MF-Net **136**

APÊNDICE E – Parâmetros utilizados nos experimentos **138**

APÊNDICE F – Implementações utilizadas **139**

Capítulo 1

Introdução

O problema da correspondência entre imagens (do inglês, *image matching*) é uma área de pesquisa de grande importância no âmbito da Visão Computacional e Processamento de Imagens Digitais. Esse problema visa encontrar conteúdos ou estruturas equivalentes, como padrões ou algum ponto de interesse, entre duas ou mais imagens de uma mesma cena, ou um mesmo objeto. A partir da obtenção da correspondência, diversas aplicações podem ser realizadas, tais como: registro de imagens[1], rastreamento de objetos, reconhecimento facial, recuperação de imagens, dentre outras (MA et al., 2020; JIN et al., 2020; BELLAVIA; COLOMBO, 2020).

Alguns autores costumam classificar as técnicas desenvolvidas para a solução do problema da correspondência entre imagens em dois grupos principais, quais sejam: (i) técnicas baseadas em área e (ii) técnicas baseadas em características, conforme ilustradas na Figura 1. Nas técnicas baseadas em área a correspondência é realizada pelo uso direto dos valores de intensidade dos pixels[2] das imagens, pelo uso de critérios de similaridade como a correlação entre duas regiões. Nas técnicas baseadas em características, os elementos a serem comparados consistem em conjuntos numéricos, fornecidos por características ou informações locais contidas nas imagens (LIU et al., 2021; MA et al., 2020; LI; HU; AI, 2020).

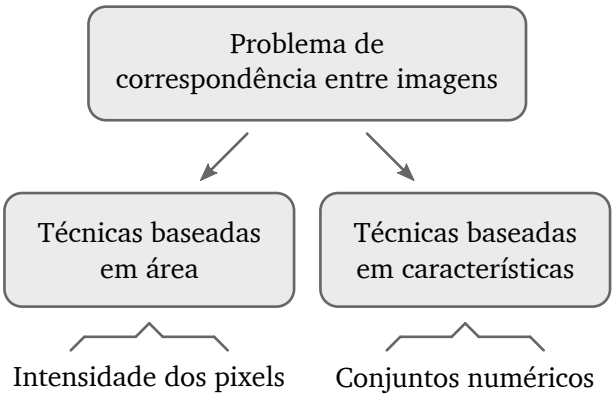


Figura 1 – Principais técnicas aplicadas ao problema da correspondência entre imagens.

[1] O registro de imagens é o processo de alinhamento espacial de duas ou mais imagens de uma mesma cena, obtidas de diferentes pontos de vista, diferentes sensores ou instantes de tempo (MA et al., 2020; ZITOVÁ; FLUSSER, 2003).

[2] O termo *pixel* é de origem inglesa, formada pela aglutinação das palavras *picture* e *element*. Em uma imagem digital, ele representa o menor elemento que pode ser endereçado (LYON, 2006). Apesar de alguns autores traduzirem este termo como píxel (e seu plural, píxeis), neste trabalho optou-se por utilizar o termo em inglês (com seu plural correspondente *pixels*).

As técnicas baseadas em área visam o alinhamento espacial entre as imagens, e por este motivo, são adequadas para aplicações de registro de imagens (KUPPALA; BANDA; BARIGE, 2020). Nessas aplicações, o uso da medida de informação mútua[3], como critério de similaridade, tem sido importante historicamente, principalmente tratando-se de imagens médicas, em que as imagens possuem poucos detalhes de texturas (JIANG et al., 2021; ZHUANG et al., 2016). Essas técnicas, no entanto, são sensíveis às transformações geométricas e deformações locais contidas nas imagens (LIU et al., 2021; MA et al., 2020).

Por outro lado, as técnicas baseadas em características locais possuem vantagens práticas. Dentre tais vantagens, possuem baixo custo computacional devido à consideração de poucos dados, boa exatidão e baixa sensibilidade às variações geométricas da cena. Comparadas às técnicas baseadas em área, as técnicas baseadas em características são adequadas para a realização da correspondência entre imagens obtidas do espectro visível e do espectro infravermelho com grandes diferenças entre si (LIU et al., 2021; MA et al., 2020; LI; HU; AI, 2020).

Também, as técnicas baseadas em características são mais flexíveis, por visarem a correspondência de conteúdos entre imagens, e não necessariamente o alinhamento espacial destas. Essa flexibilidade possibilita uma aplicação mais ampla, como no caso de duas imagens em que há a necessidade de realizar a correspondência de um mesmo objeto, em cenas diferentes, em que não é possível realizar o alinhamento espacial destas imagens. Nesse sentido, tem-se observado a crescente utilização das técnicas baseadas em características, tornando-se um objeto de estudo interessante para muitos pesquisadores (LIU et al., 2021; MA et al., 2020).

No âmbito das técnicas baseadas em características, uma característica local é definida como um padrão de imagem que difere de sua vizinhança imediata, como: pontos, cantos, bordas ou algum padrão específico, conforme exemplificado na Figura 2. Nesse cenário, o desenvolvimento de métodos para realizar a descrição de características que oferecem melhor discriminação para realizar correspondências entre imagens é de interesse significativo para a área de Visão Computacional (JOSHI; PATEL, 2020; SCHONBERGER et al., 2017). Nesse sentido, a literatura indica que as aplicações das características locais, no contexto da correspondência, ainda não foram suficientemente exploradas (MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; JIN et al., 2020).

A tarefa de descrição de características locais, no entanto, ainda é um desafio devido a alguns fatores, como: diferenças de iluminação, obstruções, mudanças no ângulo de visão, presença de ruídos ou borrões nas imagens. Ambiguidades na cena também se tornam um fator desafiador, pois diferentes regiões podem ter aparências similares, quando observadas sob pontos de vista distintos, dificultando assim sua discriminação (NUNES; PÁDUA, 2017). Alguns exemplos práticos desses desafios incluem a obstrução causada por objetos em

[3] No âmbito da teoria da informação, a informação mútua é uma medida que representa a dependência estatística entre duas variáveis aleatórias ou a quantidade de informações que uma variável contém sobre a outra (MAES et al., 1997; PLUIM; MAINTZ; VIERGEVER, 2003).

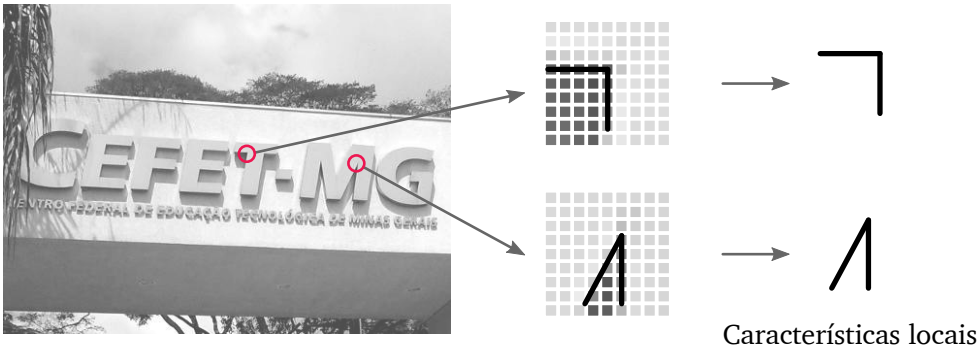


Figura 2 – Exemplo de características locais (cantos) extraídas de uma imagem.

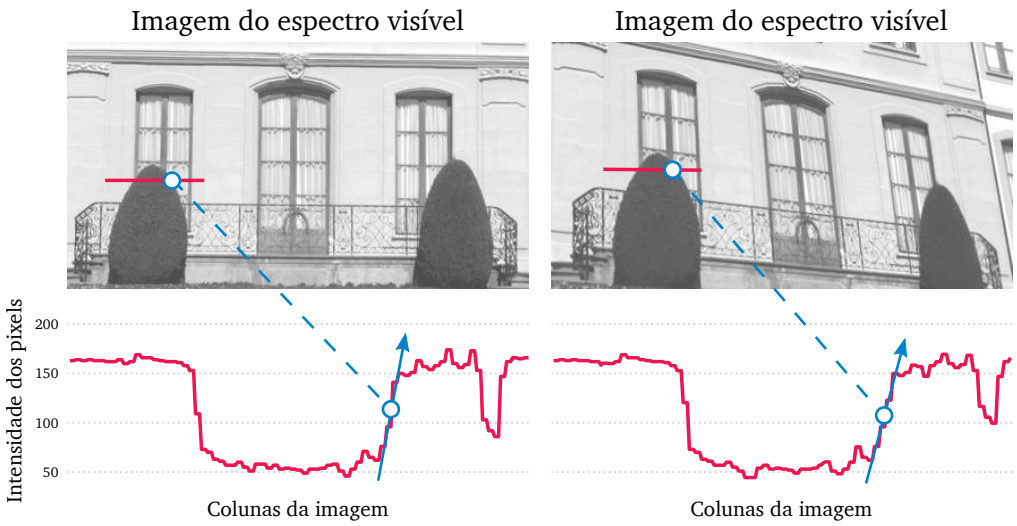
imagens de cenas externas, anatomias diferentes entre imagens médicas de um mesmo paciente, dentre outros (JIANG et al., 2021; MA et al., 2020).

No contexto das imagens obtidas do espectro visível e do espectro infravermelho (imagens VIS/IV), a descrição de características possui um desafio ainda maior, pois, a relação entre os valores de intensidade dos pixels de diferentes imagens pode ser não-monotônica[4] (BARUCH; KELLER, 2021; FAN et al., 2019; MA; MA; YU, 2019). Isso ocorre devido às diferenças na reflexão da luz para diferentes faixas do espectro eletromagnético. Nesse sentido, a correspondência entre imagens consiste em um problema desafiador devido à falta de relações implícitas ou explícitas entre as regiões correspondentes, sendo ainda um problema não resolvido (JIANG et al., 2021; MA et al., 2020).

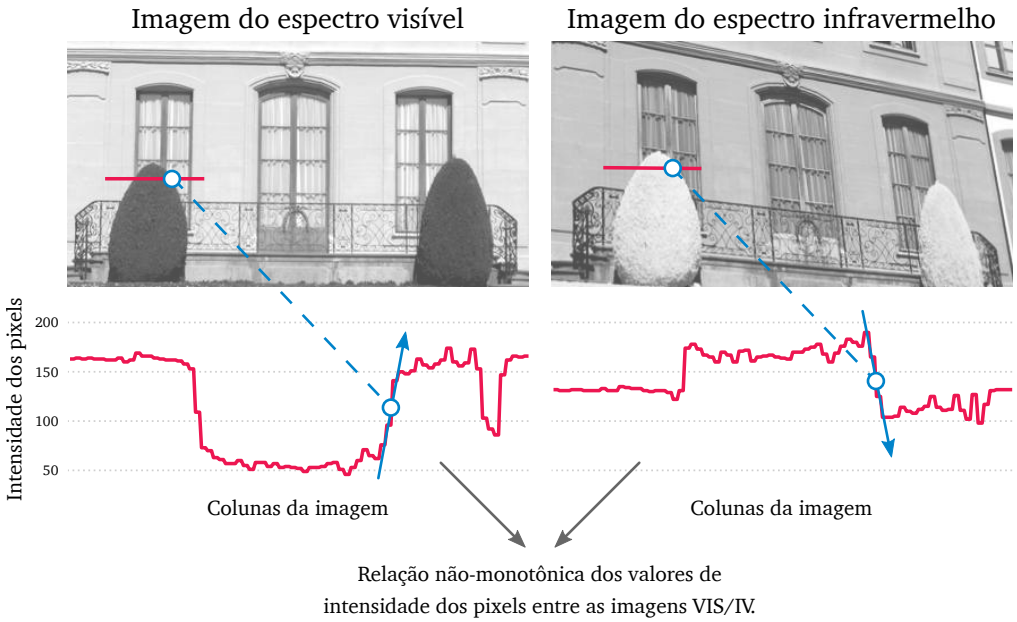
Para ilustrar a diferença entre os valores de intensidade dos pixels das imagens obtidas de diferentes faixas do espectro eletromagnético, a Figura 3 exibe os valores dessas intensidades ao longo de um segmento de reta (em vermelho), representando regiões correspondentes. Nessa figura, há um par de imagens obtidas de uma mesma faixa de frequência e outro par de faixas distintas. Observa-se, por meio das taxas de variação (setas no gráfico) que, a relação entre os valores de intensidade das imagens capturadas de distintas faixas do espectro eletromagnético pode afetar a imagem e dificultar a descrição dessas regiões. Apesar dessas diferenças, as aparências das formas dos objetos, representadas por suas bordas, tendem a permanecer preservadas em ambas as imagens (JIANG et al., 2021; WU; ZHU, 2021).

Há uma vasta literatura que propõe métodos para a descrição de características locais. No entanto, tais métodos foram fundamentalmente projetados considerando-se as características de imagens obtidas no espectro visível. Embora essas técnicas funcionem bem em imagens deste espectro, elas frequentemente não possuem resultados satisfatórios quando utilizadas em imagens VIS/IV, pois grande parte desses métodos se baseiam no cálculo de gradiente ou na informação de intensidade dos pixels (DELLINGER et al., 2015;

[4] Na matemática, uma função é dita monotônica quando ela preserva a relação de ordem, ou seja: $x \leq y \longleftrightarrow f(x) \leq f(y)$ ou $x \leq y \longleftrightarrow f(x) \leq f(y)$. Para uma função monotônica crescente, por exemplo, se o valor da variável dependente aumenta, o valor da função também aumenta. Uma função não-monotônica, portanto, não preserva essa relação de ordem (ZHANG et al., 2015).



(a) Par de imagens de mesmo espectro e diferentes pontos de vista.



(b) Par de imagens de diferentes espectros e diferentes pontos de vista.

Figura 3 – Variação nos valores de intensidade dos pixels ao longo de um segmento de reta.

Fonte: Adaptado de Nunes e Pádua (2017).

SALEEM; SABLATNIG, 2014). Também, comparado à quantidade de trabalhos que propõem soluções para o espectro visível, há poucos trabalhos na literatura que abordam imagens obtidas de diferentes faixas do espectro eletromagnético (JOSHI; PATEL, 2020). Com base nesse contexto e na necessidade de técnicas específicas para a solução desse problema, este trabalho propõe desenvolver métodos descritores de características locais para serem aplicados em problemas de correspondência entre imagens VIS/IV. A Figura 4 exibe um diagrama de blocos que ilustra o escopo deste trabalho, contendo o problema geral, bem como o contexto de aplicação.

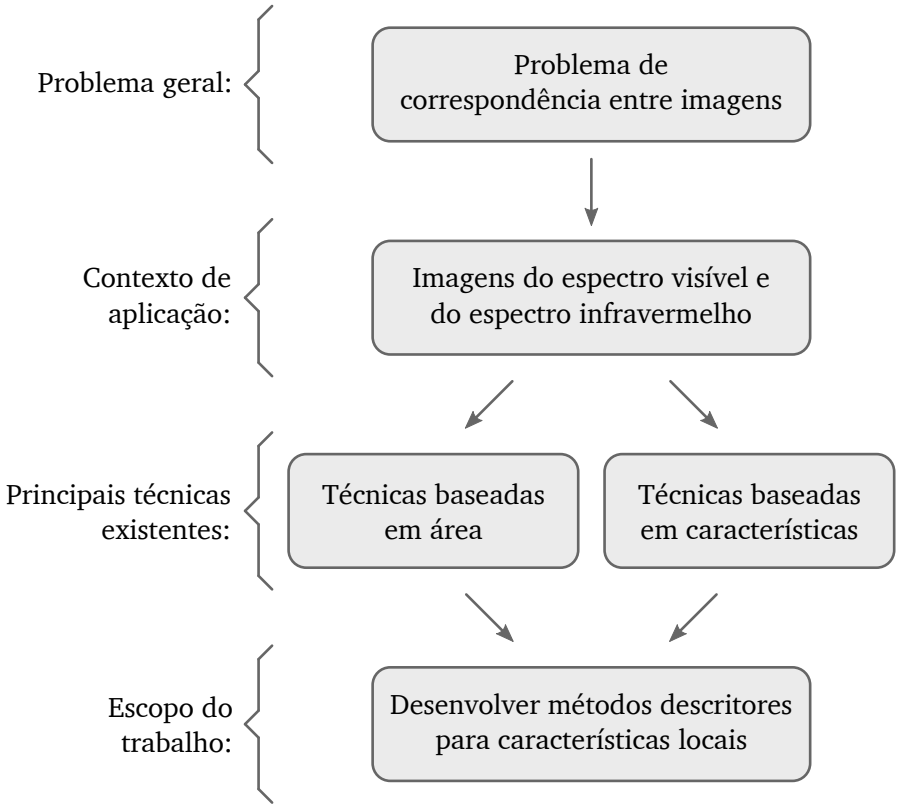


Figura 4 – Escopo do trabalho no problema da correspondência entre imagens.

Neste trabalho, o escopo de abrangência se limita a propor métodos para a descrição de características locais, que se trata de uma tarefa preliminar e crucial para o problema da correspondência entre imagens (MA et al., 2020; CHEN et al., 2017; SALEEM et al., 2017). Outros desafios referentes ao problema da correspondência, como na extração de características ou do alinhamento geométrico das imagens, não são contemplados aqui.

1.1 Motivação

As imagens obtidas do espectro infravermelho desempenham um papel fundamental na área de Visão Computacional, ao complementarem o espectro visível com informações adicionais sobre uma cena ou objeto (JOSHI; PATEL, 2020; SALEEM; BAIS, 2020). Essa capacidade amplia as análises em várias aplicações, como sistemas de vigilância (LE et al., 2020), navegação guiada por imagem (XAUD; LEITE; FROM, 2019), registro de imagens (MENG et al., 2021) e reconhecimento facial (HE et al., 2020), entre outras (TEUTSCH; SAPPA; HAMMOUD, 2021), conforme os exemplos ilustrados na Figura 5.

O uso de imagens obtidas do espectro infravermelho é amplamente valorizado devido às propriedades únicas desse tipo de radiação eletromagnética. Em comparação com a luz visível, o infravermelho pode penetrar e refletir de maneira diferente em vários materiais, fornecendo informações que muitas vezes não são visíveis a olho nu. Essa diferença na

(a) Sistemas de vigilância.

(b) Navegação guiada por imagem.

(c) Registro de imagens.

(d) Reconhecimento facial.

Figura 5 – Exemplos de aplicações em que é importante a utilização das imagens VIS/IV.

reflexão é particularmente útil para identificar materiais, detectar variações de temperatura e obter imagens em condições de baixa luminosidade. Além disso, certos materiais podem refletir a luz visível, mas absorver a radiação IV, e vice-versa. Isso permite que sistemas de visão computacional diferenciem materiais ou detectem anomalias em superfícies que poderiam ser indetectáveis utilizando apenas o espectro de luz visível. Em suma, a capacidade do infravermelho de interagir diferentemente com diferentes materiais e em diferentes condições torna-o uma ferramenta inestimável na área de processamento de imagens e visão computacional (JOSHI; PATEL, 2020; SALEEM; BAIS, 2020).

Para que um conjunto de imagens capturadas de distintas faixas do espectro eletromagnético seja efetivamente utilizado, é necessário realizar o registro dessas imagens de tal forma que todas elas possam estar alinhadas em uma coordenada em comum (LI et al., 2015). O registro é comumente realizado por processos de correspondência entre imagens, no entanto, executar esse procedimento ainda apresenta grandes desafios, dado que os pontos nos quais se deseja mapear são, na maioria das vezes, localmente distintos (QIN et al., 2014). No contexto das imagens capturadas de diferentes faixas do espectro eletromagnético, essa tarefa possui um desafio ainda maior, posto que a aquisição de imagens de mesmas cenas por diferentes sensores pode resultar em variações de aparência significativas. Tais variações se manifestam de formas não-lineares e desconhecidas, como mapeamentos de valores de intensidade não-monotônicos e texturas não correspondentes, o que dificulta a extração e descrição de características (BARUCH; KELLER, 2021; JIANG et al., 2021).

Nas últimas décadas, métodos descritores de características locais foram propostos para diferentes aplicações em processos de correspondência entre imagens. Dentre esses métodos desenvolvidos, o SIFT (*Scale Invariant Feature Transform*) (LOWE, 2004) e o SURF (*Speeded-Up Robust Feature*) (BAY; TUYTELAARS; GOOL, 2006) são algoritmos tradicionais e amplamente utilizados (JOSHI; PATEL, 2020; FU et al., 2018a). Devido à robustez desses métodos em relação à variação de iluminação e variações geométricas de escala e rotação, eles obtêm uma eficácia satisfatória em muitas tarefas, no entanto, esses métodos geralmente são falhos em obter correspondências corretas quando aplicados em tarefas que utilizam imagens VIS/IV (JOSHI; PATEL, 2020; MA; MA; YU, 2019).

Na literatura, alguns pesquisadores têm adaptado métodos desenvolvidos exclusivamente para imagens obtidas do espectro visível, visando construir métodos descritores para lidar especialmente com as imagens VIS/IV (JIANG et al., 2021). Tais adaptações ainda não possuem uma eficácia satisfatória, principalmente quando a diferença entre as faixas do espectro eletromagnético em que as imagens foram capturadas é muito grande (LI et al., 2016). Cabe lembrar que, comparado à quantidade de trabalhos que propõem soluções para o espectro visível, há poucos trabalhos na literatura que abordam imagens obtidas de diferentes faixas do espectro eletromagnético (JOSHI; PATEL, 2020).

Dentre os métodos desenvolvidos, há aqueles baseados em aprendizado de máquina e aqueles baseados em modelagem manual (do inglês, *handcrafted*[5]) (JOSHI; PATEL, 2020; MA et al., 2020; CHEN; ROTTENSTEINER; HEIPKE, 2020). Em geral, os métodos propostos para descrição de características em imagens VIS/IV, não abordam os desafios referentes às questões geométricas (variações em ponto de vista, escala e rotação), como também às questões fotométricas (variações de iluminação da cena, exposição da imagem, borrões ou ruídos) (KIM et al., 2020; MIKOLAJCZYK et al., 2005). Além disso, para que esses métodos descritores obtenham um resultado adequado é necessário, muitas vezes, realizar uma etapa de pós-processamento para desconsiderar as correspondências incorretas (MA; MA; YU, 2019).

Em relação aos métodos baseados em modelagem manual, grande parte dos estudos na literatura baseiam-se nos algoritmos SIFT e SURF (WANG; SHI; CAO, 2019). Sobre esses métodos, a literatura mostra que eles consistem em técnicas de alto custo computacional devido à sua complexidade e difícil implementação (BARAJAS-GARCÍA; SOLORZA-CALDERÓN; GUTIÉRREZ-LÓPEZ, 2019; RUBLEE et al., 2011), o que motiva a construção de técnicas mais eficazes e simples. Em se tratando dos métodos baseados em aprendizagem de máquina, há poucos trabalhos na literatura que se dedicam à descrição de características em imagens VIS/IV (JOSHI; PATEL, 2020). A ausência de um conjunto consolidado de trabalhos que usam essa abordagem também motiva este trabalho no desenvolvimento de uma técnica nessa linha, para aprimorar as soluções existentes, dado que trabalhos na literatura indicam

[5] Um método baseado em modelagem manual refere-se àqueles projetados especificamente por um especialista, em vez de serem gerados automaticamente ou aprendidos por um sistema computacional (MA et al., 2020).

que esta abordagem é promissora e ainda não foi suficientemente explorada (CHEN; ROTTENSTEINER; HEIPKE, 2020; KUPPALA; BANDA; BARIGE, 2020; BELLAVIA; COLOMBO, 2020).

Situado no contexto das imagens VIS/IV e diante dos motivos apresentados, referente às lacunas existentes na literatura, este trabalho propõe desenvolver métodos descritores de características locais para serem aplicados em problemas de correspondência entre imagens VIS/IV. Nesse sentido, este trabalho explora tanto a abordagem baseada em modelagem manual quanto a abordagem baseada em aprendizado de máquina. A hipótese de pesquisa é de que ao utilizar funções e métodos adequados às características dessas imagens obtenham-se melhores resultados comparados a algoritmos do estado da arte.

1.2 Definição do problema de pesquisa

No contexto das imagens obtidas do espectro visível (VIS) e do espectro infravermelho (IV), o foco deste trabalho consiste em avançar no desenvolvimento de novos métodos práticos para descrever características locais no âmbito da correspondência entre imagens. É importante ressaltar que, apesar deste trabalho se situar nesse contexto, não é o escopo deste trabalho propor processos de correspondência entre imagens. O problema tratado neste trabalho pode ser expresso da seguinte forma:

> *Dada uma cena ou um objeto, retratados por um par*[6] *de imagens obtidas do espectro visível e do espectro infravermelho, recupere os descritores de características locais que possuam a menor distância entre si por um critério de similaridade, de modo que estes representem as mesmas regiões no mundo real.*

A Figura 6 ilustra o cenário considerado neste trabalho, em que os métodos descritores são aplicáveis. Mais especificamente, o cenário do problema de pesquisa assume as seguintes premissas:

- As imagens representam uma mesma cena ou um mesmo objeto, capturadas a partir de pontos de vista e/ou instantes de tempo diferentes;
- Possíveis distorções e/ou ruídos introduzidos durante a obtenção das imagens são desconsiderados;
- As imagens são compostas por pares, obtidos tanto do espectro visível quanto do espectro infravermelho;
- A resolução das imagens é desconhecida.

[6] O trabalho se concentra em pares de imagens, mas em casos de múltiplas imagens, a correspondência entre elas pode ser realizada utilizando uma estratégia par a par.

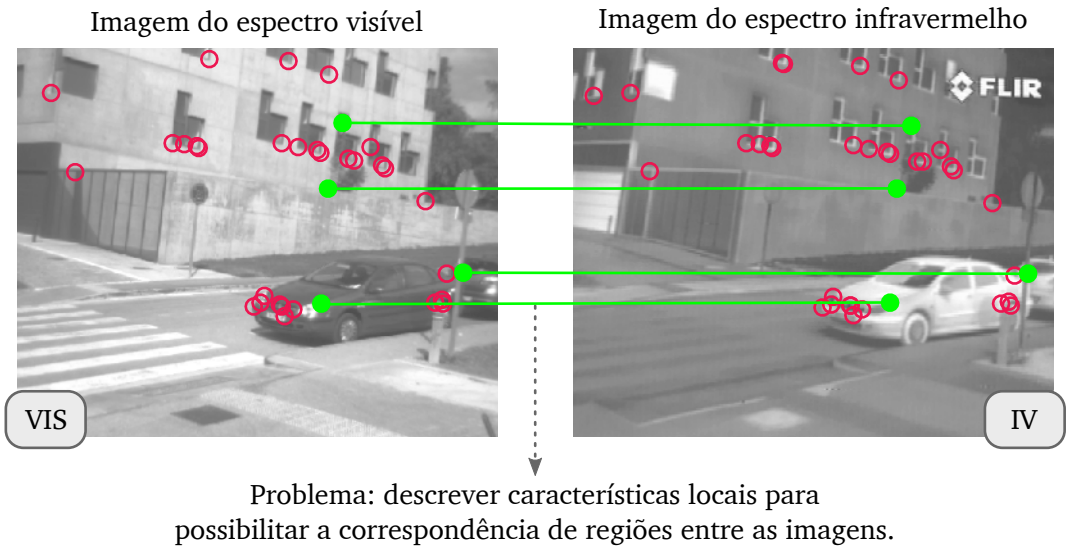


Figura 6 – Exemplo de cenário considerado neste trabalho.

Fonte: Adaptado de Saleem e Bais (2020).

Idealmente, a descrição das características deveria corresponder a partes com um significado representativo de um objeto (por exemplo, as bordas de uma imagem deveriam representar rodovias no mundo real). Na prática, no entanto, não seria possível obter uma interpretação de alto nível. Em vez disso, a descrição da característica local é realizada com base nos padrões de intensidade dos pixels da região local. Desse modo, uma solução ideal para descrição de características locais deste trabalho deve possuir as seguintes propriedades (JOSHI; PATEL, 2020; TUYTELAARS; MIKOLAJCZYK, 2008):

- *Localidade:* a obtenção das informações de descrição das características deve ser a mais local possível, de modo a reduzir a probabilidade de obstruções;
- *Distinção:* os padrões de intensidade, devem mostrar variações de modo que estas possam ser unicamente identificadas;
- *Eficiência:* preferencialmente, a descrição de características locais precisa ser rápida, de modo a possibilitar aplicações em tempo real e utilizar uma menor quantidade possível de recursos;
- *Robustez:* a descrição de características locais deve ser menos sensível em relação às diferenças nos valores de intensidade e pequenas deformações contidas na imagem;
- *Robustez a diferentes faixas do espectro eletromagnético:* no domínio das imagens VIS/IV, a descrição de características locais deve considerar a falta de relações implícitas ou explícitas das regiões correspondentes entre as imagens.

É evidente que a importância de algumas propriedades depende de uma aplicação específica. As propriedades de localidade e distinção, por exemplo, competem entre si, pois quanto menos informações estão disponíveis no padrão de intensidade subjacente, mais difícil se torna realizar a correspondência. Também, a distinção em excesso torna a descrição muito específica, a ponto de perder a capacidade de generalização e tornar a solução menos robusta. Nesse aspecto, este trabalho não se atém a avaliar propriedades específicas das soluções propostas, dado que algumas delas podem se correlacionar inversamente. Em vez disso, foram assumidas medidas de desempenho de avaliação que refletem essas propriedades de maneira geral. Das propriedades mencionadas anteriormente, a mais desejável seria a robustez a diferentes faixas do espectro eletromagnético, dado que essa propriedade é a que distingue os métodos que conseguem operar em distintas faixas do espectro eletromagnético.

1.3 Objetivos

O objetivo deste trabalho consistiu em desenvolver métodos descritores de características locais para serem utilizados em problemas de correspondência entre imagens obtidas do espectro visível (VIS) e do espectro infravermelho (IV). Mediante este trabalho, buscou-se também contribuir com pesquisas na área de Visão Computacional por meio da utilização das soluções aqui desenvolvidas ou mesmo para o desenvolvimento de novas soluções. Os objetivos específicos, envolveram:

- Melhorar a acurácia da descrição de características locais de métodos do estado da arte para encontrar correspondências entre imagens obtidas do espectro visível (VIS) e do espectro infravermelho (IV);
- Investigar a utilização de redes neurais convolucionais para descrever características locais. Neste tópico, este trabalho propôs um novo método projetado especialmente para imagens VIS/IV baseado em aprendizado de máquina;
- Desenvolver e apresentar um novo método, baseado em modelagem manual, para descrever características locais em imagens VIS/IV.

1.4 Contribuições

Do ponto de vista prático, este trabalho apresenta duas principais contribuições para a área de Processamento de Imagens Digitais, especialmente, para o campo de descrição de características locais:

- O desenvolvimento de um método, baseado em aprendizado de máquina, no qual se construiu uma arquitetura de rede neural convolucional para aprender descritores de características locais em imagens VIS/IV. Para a construção desta rede, este trabalho

apresenta uma nova camada, intitulada *camada de mapeamento*, para lidar com a relação não-monotônica dos valores de intensidade dos pixels entre as imagens;

- O desenvolvimento de um método, baseado em modelagem manual, com base na combinação de contribuições do estado da arte, no que se refere a métodos do espectro visível e métodos matemáticos adequados às características particulares das imagens VIS/IV.

Do ponto de vista teórico, essas contribuições fornecem conhecimentos intermediários a pesquisadores para a construção ou até mesmo para o aperfeiçoamento de novos algoritmos. Por isso, este trabalho contribui no âmbito científico ao apresentar novas soluções e abordagens para a descrição de características locais em imagens VIS/IV, mais especificamente, métodos descritores.

Vale salientar que este trabalho possui um caráter interdisciplinar e essa particularidade se manifesta ao se considerar que modelos e funções no campo da matemática são importantes para a construção das soluções para realizar a descrição de imagens VIS/IV. Além disso, como as técnicas desenvolvidas neste trabalho possuem aplicações diretas em outras áreas do conhecimento, este estudo possibilita a transferência das soluções para outros campos.

1.5 Organização do texto

Este trabalho está organizado em sete capítulos, incluindo o presente. No Capítulo 2 são apresentados alguns dos principais conceitos necessários que fundamentam o desenvolvimento deste trabalho. O Capítulo 3 apresenta uma revisão dos principais estudos relacionados ao tema, descrevendo seus resultados e suas contribuições. O Capítulo 4 apresenta a primeira contribuição deste trabalho, apresentando o método baseado em redes neurais convolucionais para descrever características locais em imagens VIS/IV. Este capítulo possui também a metodologia, os resultados e a discussão que envolve essa contribuição. Em seguida, no Capítulo 5, é apresentada a segunda contribuição desta tese, apresentando o desenvolvimento do método baseado em modelagem manual, contendo também a metodologia, os resultados e a discussão que envolve essa proposta. No Capítulo 6 é realizada uma discussão geral sobre as soluções propostas. Por fim, no Capítulo 7 são apresentadas as conclusões, bem como as perspectivas de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentadas algumas definições e técnicas centrais que fundamentam o desenvolvimento deste trabalho. A Seção 2.1 exibe o conceito das características locais e a explicação sobre como elas podem ser representadas por descritores. Ainda sobre esse conceito, a Seção 2.2 mostra as principais maneiras para realizar a correspondência de descritores, e a Seção 2.3 apresenta algumas das medidas de desempenho predominantemente utilizadas para a avaliação dessa correspondência, também utilizadas para a avaliação da eficácia de descritores. Na Seção 2.4 é apresentado o conceito de imagem multiespectral, dado que este termo possui correlação direta ao tema deste trabalho. Por fim, a Seção 2.5 e a Seção 2.6 apresentam as definições de redes neurais e redes neurais convolucionais, respectivamente, dado que esta última é empregada em uma das propostas deste trabalho.

2.1 Características locais

Uma característica local é definida como um padrão de imagem que difere de sua vizinhança imediata. Ela está associada a uma alteração de propriedades de uma região da imagem ou de várias propriedades simultaneamente. As propriedades da imagem comumente consideradas consistem em intensidade, cor ou textura. Conforme mencionado na introdução deste trabalho, uma característica local pode representar pontos, cantos, bordas ou algum padrão específico na imagem, conforme exemplificado na Figura 7 (KUPPALA; BANDA; BARIGE, 2020; TUYTELAARS; MIKOLAJCZYK, 2008; MIKOLAJCZYK; SCHMID, 2005).

De certa forma, uma característica local ideal seria um ponto com uma localização no espaço e sem extensão espacial (LIU; LI; PAN, 2019). Na prática, entretanto, as imagens são compostas em unidades espaciais (pixels), e para descrever as características locais, uma vizinhança local de pixels precisa ser analisada, dando feições locais a uma extensão espacial. Para algumas aplicações, como a calibração de câmera ou reconstrução 3D, esta extensão espacial não é relevante, e apenas a localização obtida no processo de detecção

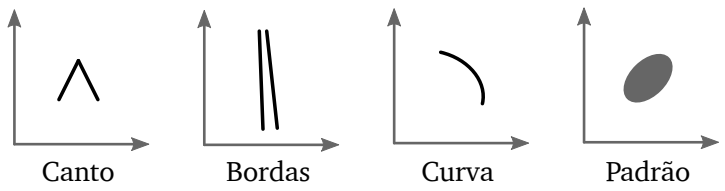


Figura 7 – Exemplos de características locais em uma imagem.

Fonte: Adaptado de Krig (2016).

da característica é utilizada. Normalmente nesses casos é utilizado o termo *ponto-chave* ou *ponto de interesse*, para denotar o centro da região onde se encontra uma determinada característica (TUYTELAARS; MIKOLAJCZYK, 2008).

Existem diversos cenários que motivam a utilização de características locais. As primeiras aplicações, para as quais as características locais foram úteis, exigiam que estas tivessem uma representação semântica no mundo real. Por exemplo, as bordas detectadas em imagens aéreas deveriam corresponder às rodovias, ou, a detecção de padrões locais (do inglês, *blobs*) eram usados para identificar impurezas em alguma tarefa de inspeção. Dessa maneira, o uso das características locais era limitado a uma interpretação semântica específica no contexto do problema (TUYTELAARS; MIKOLAJCZYK, 2008; JOSHI; PATEL, 2020).

As aplicações atuais, no entanto, possibilitaram a expansão do uso das características locais de tal forma que o que elas realmente representam não é relevante, desde que suas localizações possam ser determinadas com precisão e de forma estável. Nessas aplicações, as características locais fornecem um conjunto de pontos de ancoragem bem localizados e individualmente identificáveis. Esta é, por exemplo, a situação na maioria das aplicações de correspondência ou rastreamento e, especialmente, para calibração de câmeras ou reconstrução tridimensional. Outros cenários de aplicação incluem alinhamento e mosaico de imagens, classificação de cena, análise de textura, recuperação de imagens e reconhecimento de objetos (JOSHI; PATEL, 2020; MA et al., 2020).

A ideia básica em se utilizar características locais consiste em representar os sinais naturais sob uma dimensionalidade reduzida. Segundo Penev e Atick (1996), os sinais sensoriais dos animais possuem dimensões altas, como na retina humana, em que existem mais de seis milhões de células capazes de distinguir dezenas tonalidades de luz. A partir da atividade desse grande conjunto de receptores, o cérebro precisa descobrir onde e quais objetos existem no campo de visão, como também recuperar os atributos de cor e textura. Por esse motivo, o processamento sensorial deve reduzir a dimensionalidade do espaço de entrada. Com base nessa linha de raciocínio, todos os problemas de visão de alto nível se tornam mais tratáveis quando formulados em um espaço de baixa dimensão. Além disso, uma boa generalização depende de encontrar a representação correta de baixa dimensão.

2.2 Correspondência de características locais

As técnicas baseadas em características são comumente compostas por processos de correspondência de características locais. Esses processos podem ser divididos em três etapas principais, quais sejam: detecção, descrição e correspondência (MA et al., 2020; CHEN et al., 2017; SALEEM et al., 2017), como ilustrado na Figura 8. Essa figura também destaca o escopo de estudo deste trabalho, que consiste em desenvolver métodos descritores de características locais para correspondência entre imagens VIS/IV.

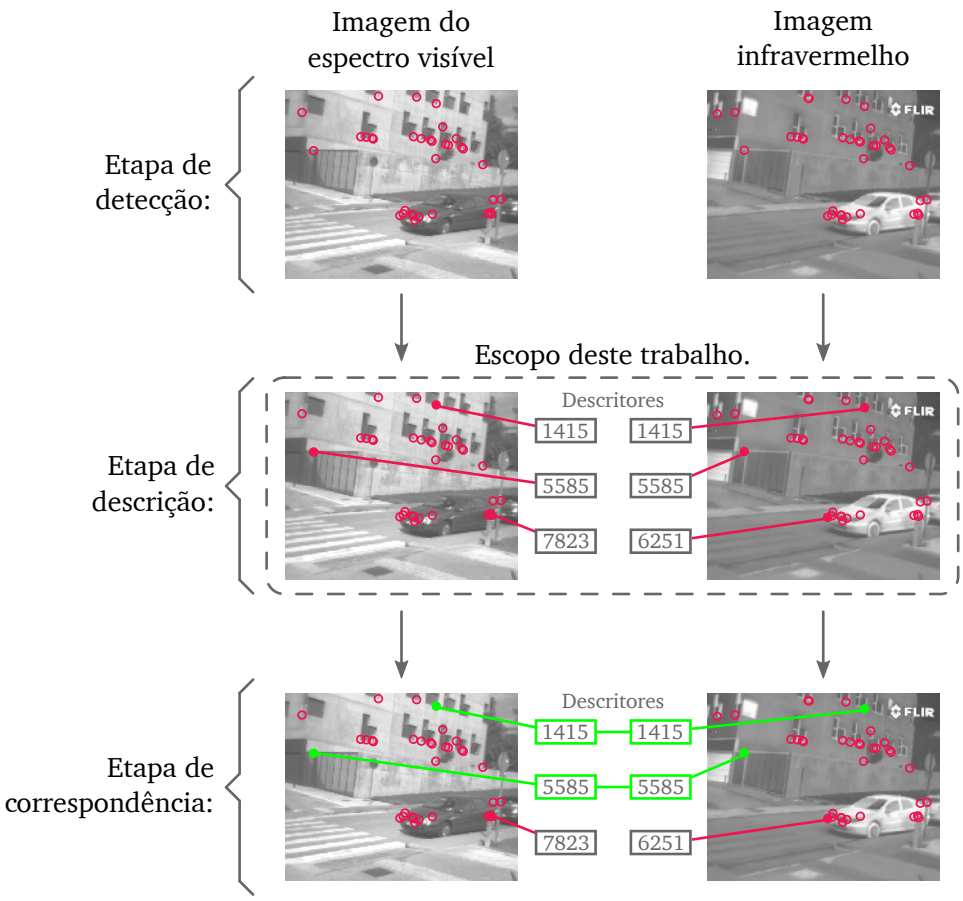


Figura 8 – Etapas de um processo convencional de correspondência entre imagens.

Na etapa de detecção, dada uma característica (por exemplo, um canto, uma borda ou um determinado padrão) deseja-se encontrar todas as regiões, pertencentes a uma imagem, onde se encontra essa característica. A etapa de descrição tem o objetivo de descrever a região onde está localizada uma característica no intuito de unicamente identificá-la e distingui-la das demais regiões, atribuindo uma espécie de assinatura a cada ponto-chave. Essa assinatura, denominada descritor, é representada por um vetor numérico, sendo utilizada para calcular a similaridade entre as regiões de outra imagem. Por fim, a etapa de correspondência tem o objetivo de corresponder cada descritor de uma imagem de referência com o descritor mais semelhante de outra imagem, dado um critério de similaridade (por exemplo, distância euclidiana) (JOSHI; PATEL, 2020).

Sobre as etapas do processo convencional de correspondência entre imagens, alguns autores ainda consideram uma etapa de pós-processamento para desconsiderar correspondências incorretas. Essa etapa comumente emprega o uso de restrições ou verificações geométricas globais sobre pontos detectados para remover as correspondências incorretas ao final do processo. Entre os métodos utilizados nesta etapa, se destaca o método RANSAC (*RANdom SAmple Consensus*) (FISCHLER; BOLLES, 1981) e algumas de suas variantes (JIANG et al., 2021; MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; FAN et al., 2019).

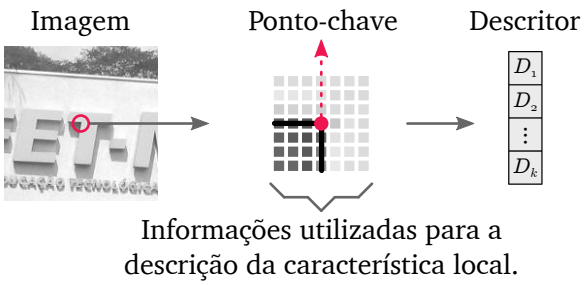


Figura 9 – Descritor de uma característica local.

A etapa de pós-processamento, por utilizar verificações geométricas, é útil em tarefas de registro de imagens, em que se deseja realizar o alinhamento espacial de duas ou mais imagens de uma mesma cena. Essa etapa auxilia a correspondência de padrões semelhantes que se repetem ao longo de uma imagem (por exemplo, imagens aéreas de uma floresta homogênea), por remover as correspondências incorretas entre esses padrões repetidos. No entanto, essa etapa não é eficaz para realizar correspondências entre pares de imagens de um mesmo objeto sob diferentes cenas, ao ser difícil estabelecer um mapeamento geométrico dos pontos entre essas imagens (KUPPALA; BANDA; BARIGE, 2020).

Na maioria das aplicações práticas, as características locais são descritas de modo que possam ser identificadas e correspondidas. Assim, algumas medidas são tomadas de uma região centrada de uma característica local e convertidas em descritor, conforme ilustrado na Figura 9. Em outras palavras, um descritor de característica D é usado para representar as informações locais em torno de um ponto de interesse de forma estável e discriminativa, sob a forma de uma sequência numérica de tamanho k, de modo que duas características correspondentes sejam o mais próximas possível (MA et al., 2020).

Para realizar a correspondência de características locais e determinar o quão similar é uma região da outra, é necessário definir um critério de similaridade conforme a natureza de seu descritor. Esse critério de similaridade é comumente associado à distância entre dois descritores (MA et al., 2020). Para descritores representados por uma sequência de números reais ($D \in \mathbb{R}^k$) é usualmente utilizada a distância euclidiana (norma L_2) (BELLAVIA; COLOMBO, 2020; CALONDER et al., 2010), já os descritores constituídos por uma sequência binária ($D \in \{0,1\}^k$) é empregada a distância de Hamming, devido à sua simplicidade e rapidez (BELLAVIA; COLOMBO, 2020; JOSHI; PATEL, 2020). A distância euclidiana e a distância de Hamming, para os descritores D_a e D_b de tamanho k, são dadas pela Equação (1) e Equação (2), respectivamente:

$$d_{Euclidiana}(D_a, D_b) = \|D_a - D_b\|_2 = \sqrt{\sum_{i=1}^{k} (D_{a_i} - D_{b_i})^2}, \tag{1}$$

$$d_{Hamming}(D_a, D_b) = \sum_{i=1}^{k} |D_{a_i} - D_{b_i}|. \tag{2}$$

Utilizando tais critérios de similaridade, é possível buscar por regiões similares entre duas ou mais imagens que possuem a menor distância, utilizando pares de descritores correspondentes. Os pares de descritores correspondentes representam os vizinhos mais próximos no espaço de características (LOWE, 2004). No entanto, a associação de dois descritores que possuem a menor distância pode não representar uma correspondência correta, devido tanto à ineficácia dos métodos de detecção e descrição quanto aos desafios impostos pelas imagens (MIKOLAJCZYK; SCHMID, 2005).

Na literatura, há algumas estratégias para reduzir a quantidade de correspondências incorretas. As duas estratégias de correspondência comumente utilizadas são: (i) correspondência baseada em limiar e (ii) razão da distância do vizinho mais próximo (NNDR, do inglês *Nearest Neighbor Distance Ratio*) (MOUATS et al., 2018; MIKOLAJCZYK; SCHMID, 2005). Na primeira abordagem, duas regiões são combinadas se a distância entre seus descritores estiverem abaixo de um certo limiar. Na segunda, o limiar é aplicado à relação de distância entre o primeiro e o segundo vizinho mais próximo (valor NNDR). A Figura 10 ilustra esse processo de correspondência de dois descritores. Nessa última abordagem, duas regiões são combinadas se satisfazerem a Equação (3):

$$\frac{\|D_a - D_b\|_2}{\|D_a - D_c\|_2} < \tau, \tag{3}$$

em que τ representa o valor NNDR, D_b é o primeiro e D_c é o segundo vizinho mais próximo do descritor D_a. Nesse caso, valores de τ próximos a 1 retornam um número maior de resultados, porém, com maior taxa de correspondências incorretas.

A utilização da primeira ou da segunda estratégia não influencia o resultado durante a comparação de métodos descritores. A utilização da abordagem via NNDR penaliza descritores semelhantes, dado que a distância do descritor mais próximo pode ser muito similar às distâncias de outros descritores (MOUATS et al., 2018).

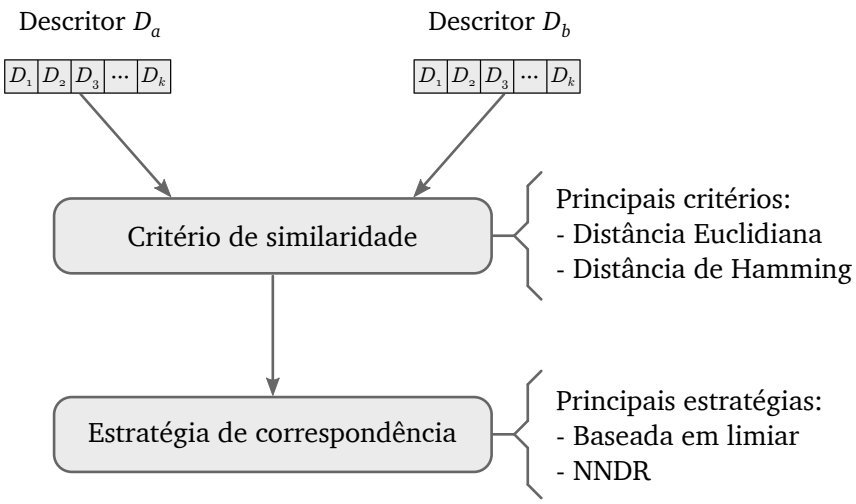


Figura 10 – Processo para a correspondência de dois descritores.

2.3 Medidas de desempenho para métodos descritores

A avaliação de métodos descritores é comumente realizada em processos de correspondência entre imagens. Nesse processo, é necessário definir previamente pares de regiões correspondentes que representam os mesmos pontos no mundo real, também chamados de correspondências reais[1]. Essa tarefa pode ser realizada manualmente, definindo pares de coordenadas nas imagens, ou utilizando um método detector de características locais (HE; LU; SCLAROFF, 2018; MOUATS et al., 2018; MIKOLAJCZYK; SCHMID, 2005).

Muitas das vezes pode ser trabalhoso definir manualmente um conjunto de coordenadas correspondentes entre vários pares de imagens. Por esse motivo, na literatura é usualmente utilizado um método detector para extrair as coordenadas dos pontos-chave de uma imagem, projetando-os em outra imagem (MA et al., 2020; AGUILERA; SAPPA; TOLEDO, 2015).

Para realizar essa projeção, é necessário conhecer a relação geométrica dos pontos-chave entre o par de imagens, definida por uma matriz de homografia[2] que mapeia os pontos de um plano para outro. Essa matriz é utilizada como um gabarito (*ground truth*) sendo fornecida em bases de dados que contêm pares de imagens de uma mesma cena, para avaliação e comparação de algoritmos diversos. Vale ressaltar que, o processo de obtenção dessas matrizes pode incluir algum erro, mas que, no entanto, esse erro é comumente desprezado para comparação de métodos detectores e descritores (MOUATS et al., 2018).

Após definir as correspondências reais, é possível verificar se as combinações entre os descritores (correspondências retornadas ao final do processo) estão corretas ou não. Para tanto, calcula-se o erro em pixels por meio da distância entre o ponto projetado e o ponto correspondido. Se essa distância for menor que um certo limiar, então considera-se que a correspondência retornada entre os descritores, associados a esses pontos, está correta (SALEEM; BAIS, 2020; HEINLY; DUNN; FRAHM, 2012).

A Figura 11 ilustra o processo de identificação de uma correspondência correta, em que o pixel P'_A consiste na projeção do pixel P_A na Imagem B, por meio da matriz de homografia H. O par de pixels P_A e P_B consiste em uma correspondência retornada, por meio da combinação de seus descritores associados. A distância euclidiana d entre o pixel P'_A e P_B consiste no erro em pixel entre a projeção e o ponto correspondido. Neste trabalho, assume-se que uma correspondência é correta se essa distância for menor que 5, assim como comumente utilizado em trabalhos da literatura (YU; ZHAO, 2020; FU et al., 2018a; LIU et al., 2018).

---

[1] Na literatura, é comum utilizar o termo, do inglês, *correspondence* para se referir às regiões correspondentes que representam um mesmo ponto no mundo real. Para denotar a correspondência de descritores (combinação de descritores) é utilizado o termo *matching*. Para tornar o texto mais claro e evitar possíveis confusões, neste trabalho optou-se por utilizar a expressão "correspondências reais" para se referir às regiões das imagens que representam um mesmo ponto no mundo real, e "correspondência retornada" para denotar a combinação de descritores retornada ao final de um processo de correspondência.

[2] O Apêndice C explica em mais detalhes a homografia entre planos.

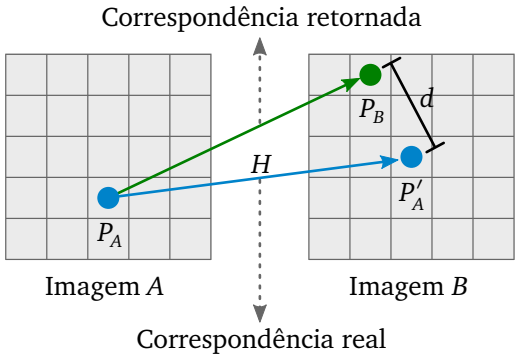


Figura 11 – Erro entre o ponto-chave detectado e sua projeção.

Entendido os conceitos de "correspondências reais", referente às projeções de pontos entre as imagens por meio da matriz de homografia, e o conceito de "correspondências retornadas", referente às combinações de descritores associados aos pontos detectados, é possível definir as medidas de desempenho de revocação e precisão para a avaliação dos métodos de descrição de características. Essas medidas de desempenho são utilizadas em diversos trabalhos que abordam o estudo e a avaliação de características locais, sendo estas as medidas de desempenho predominantes (JIANG et al., 2021; JOSHI; PATEL, 2020; MOUATS et al., 2018; SCHONBERGER et al., 2017).

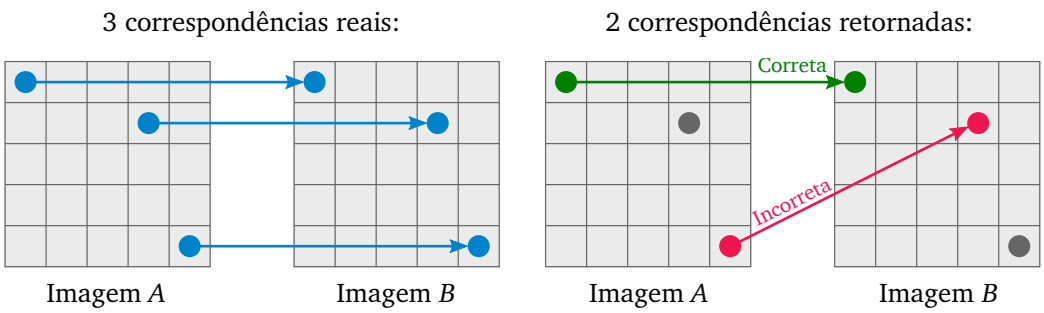
A revocação, definida pela Equação (4), representa a quantidade de correspondências corretas retornadas, em relação à quantidade de total de correspondências reais existentes em um par de imagens.

$$Revocação = \frac{|correspondências\ corretas\ retornadas|}{|total\ de\ correspondências\ reais|}. \tag{4}$$

Já a precisão é uma medida de desempenho que representa a quantidade de correspondências corretas retornadas em relação ao total de correspondências retornadas (corretas e incorretas), definida pela Equação (5):

$$Precisão = \frac{|correspondências\ corretas\ retornadas|}{|total\ de\ correspondências\ retornadas|}. \tag{5}$$

Ambas as medidas de desempenho dizem respeito à tarefa de descrição de características locais (JOSHI; PATEL, 2020; MIKOLAJCZYK; SCHMID, 2005). A medida de desempenho de revocação pode ser interpretada como uma medida de quantidade, já a precisão, uma medida de qualidade dessas correspondências retornadas (VULTUR, 2018). Seus valores podem variar entre 0 e 1, sendo desejado que se tenha, simultaneamente, maiores valores de revocação e precisão. A Figura 12 exemplifica o uso destas medidas de desempenho.



$$Revocação = \frac{|correspondências\ corretas\ retornadas|}{|total\ de\ correspondências\ reais|} = \frac{1}{3}$$

$$Precisão = \frac{|correspondências\ corretas\ retornadas|}{|total\ de\ correspondências\ retornadas|} = \frac{1}{2}$$

Figura 12 – Exemplo de aplicação das medidas de desempenho de revocação e precisão.

Em problemas práticos, é mais importante que as correspondências sejam precisas, indicando que as localizações das características foram determinadas de forma geometricamente estável, comparada à quantidade de correspondências obtidas (SCHONBERGER et al., 2017). Em outras palavras, o valor da medida de precisão é mais importante para grande parte dos problemas práticos.

As medidas de desempenho de revocação e precisão podem se contrapor entre si, pois um método que possui como resultado um valor alto de precisão pode possuir um valor baixo de revocação, e vice-versa. Nesse caso, usualmente utiliza-se uma medida que representa um balanço entre essas duas medidas de desempenho, intitulada medida F_1 (do inglês, F_1-*score*) (JIANG et al., 2021; WANG; ZHU; YUAN, 2013). Assim como a precisão e a revocação, seus valores podem variar entre 0 e 1, e quanto mais próximo de 1, maior é a precisão e a revocação. A medida F_1 é definida pela Equação (6):

$$F_1 = 2 \times \frac{Precisão \times Revocação}{Precisão + Revocação}. \tag{6}$$

É importante ressaltar que, apesar da avaliação de métodos descritores ser realizada em um processo de correspondência entre imagens (HE; LU; SCLAROFF, 2018; MIKOLAJCZYK; SCHMID, 2005), durante a avaliação de métodos descritores não se utilizam etapas de pós-processamento para desconsiderar correspondências incorretas, pelo uso de restrições ou verificações geométricas globais. Isso se deve ao fato de que essa etapa de pós-processamento beneficia o processo de correspondência de características locais. Nesse caso, essa etapa pode introduzir vieses ao avaliar os métodos descritores, pois, ela é uma etapa adicional a esse processo e as medidas de desempenho de avaliação mensuram o resultado ao final de todas as etapas.

2.4 Imagens do espectro infravermelho

Antes de caracterizar uma imagem do espectro infravermelho, é importante discorrer sobre o termo *imagem multiespectral*, cuja definição está relacionada com o contexto deste trabalho. Uma imagem multiespectral consiste em um conjunto de imagens de uma mesma cena, capturadas por diferentes faixas de comprimentos de onda do espectro eletromagnético, ilustrado na Figura 13. Elas carregam informações espaciais e espectrais sobre uma determinada cena ou objeto. A obtenção dessas imagens também pode ser realizada por diferentes fontes e/ou diferentes instantes de tempo (JOSHI; PATEL, 2020; PAOLETTI et al., 2018; PETROU; PETROU, 2010; ANUTA, 1970). A Figura 14 exemplifica uma imagem multiespectral formada por imagens obtidas do espectro visível e do espectro infravermelho.

Apesar de o conceito *imagens multiespectrais* ser bem definido em trabalhos que abordam o Sensoriamento Remoto, representando imagens únicas, obtidas de uma mesma cena e que contêm várias faixas alinhadas espacialmente, alguns trabalhos utilizam esse termo para denotar múltiplas imagens individuais. Essas imagens individuais são obtidas de diferentes faixas do espectro eletromagnético tanto de uma mesma cena quanto de um mesmo objeto (FAN et al., 2019; FANG et al., 2019; FU et al., 2018b).

Além disso, apesar desse conceito se referir a qualquer faixa do espectro eletromagnético, alguns trabalhos na literatura empregam esse termo para denotar imagens obtidas das faixas do espectro visível e infravermelho (imagens VIS/IV), devido às inúmeras aplicações práticas dessas faixas (LIU; LI; PAN, 2019; FAN et al., 2019).

A faixa do espectro eletromagnético referente ao infravermelho pode ser subdividida em diferentes faixas de comprimentos de onda (JOSHI; PATEL, 2020; DESTA; BUXTON; JANSEN, 2020; RICAURTE et al., 2014), quais sejam: infravermelho próximo (NIR, do inglês *Near-Infrared*), de ondas curtas (SWIR, do inglês *Short-Wave Infrared*), de ondas médias (MWIR, do inglês *Mid-Wave Infrared*), de ondas longas (LWIR, do inglês *Long-Wave Infrared*) e infravermelho distante (FIR, do inglês *Far Infrared*). Nesse sentido, uma imagem do espectro

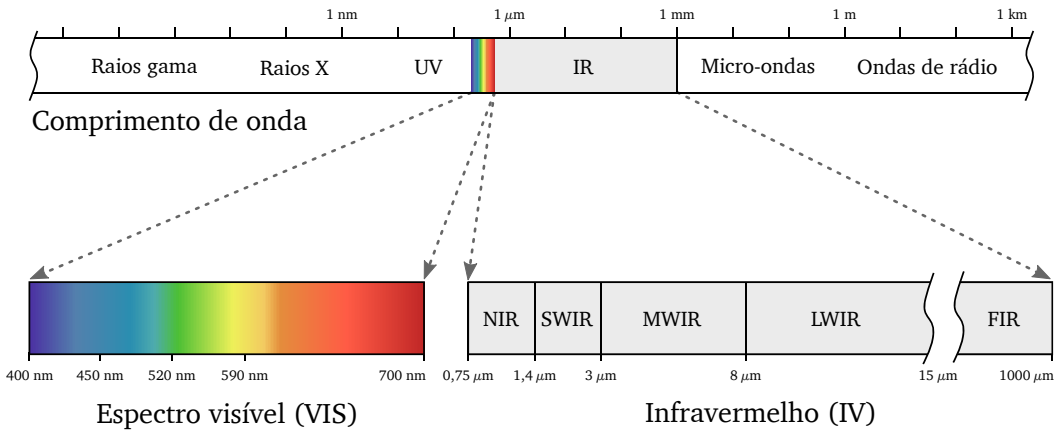


Figura 13 – Espectro eletromagnético.

Fonte: Adaptado de Joshi e Patel (2020).

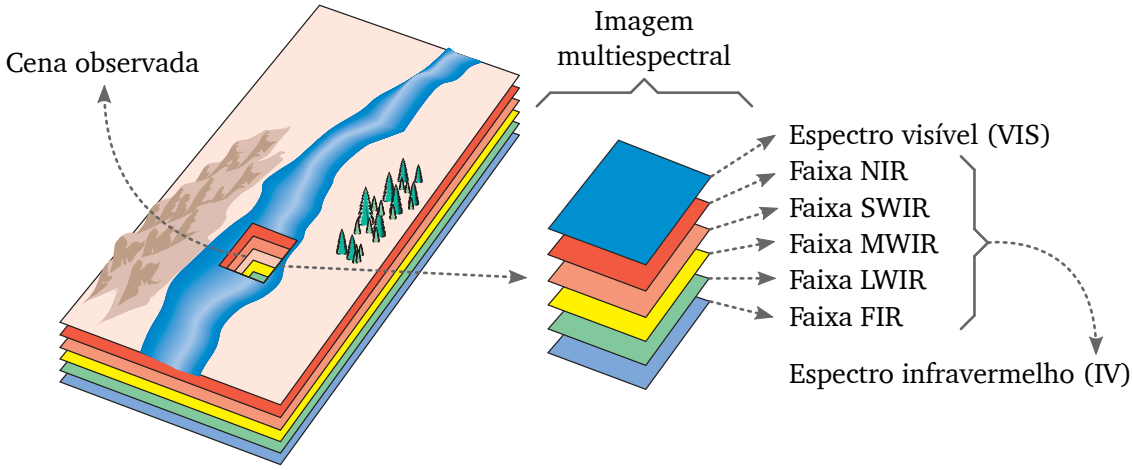


Figura 14 – Exemplo de composição de uma imagem multiespectral.

Fonte: Adaptado de Shaw e Burke (2003).

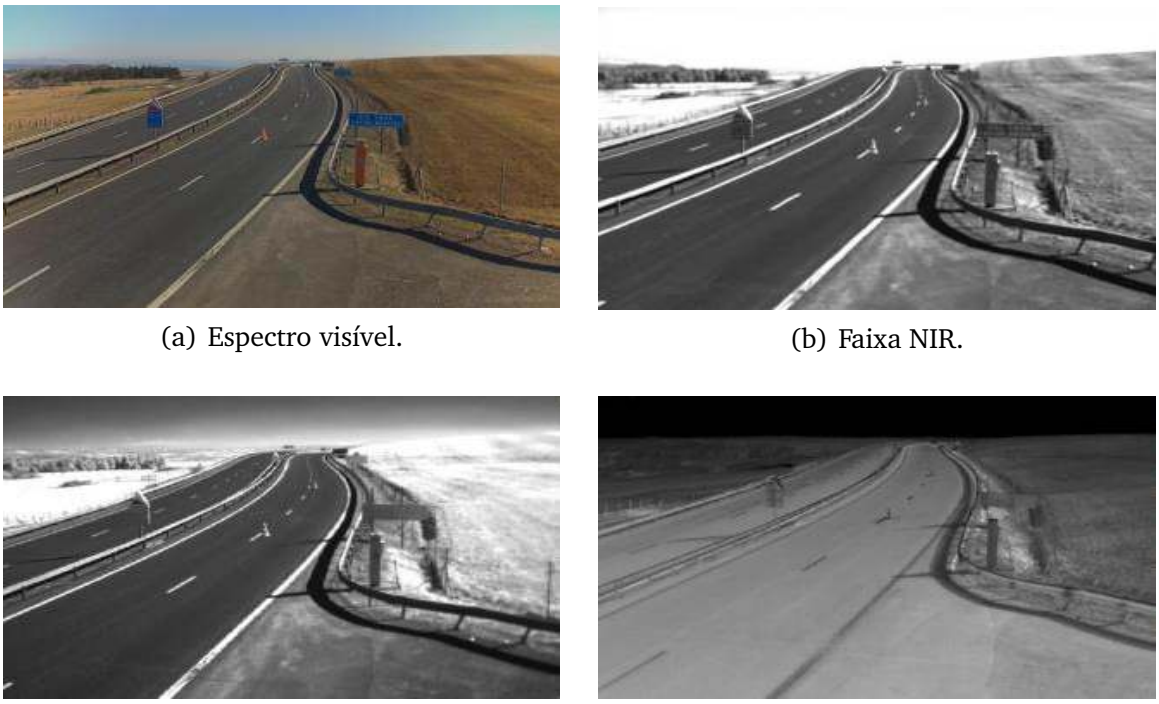
infravermelho seria uma imagem obtida por um sensor capaz de capturar alguma dessas faixas, podendo esta ser visualizada em níveis de cinza (TEUTSCH; SAPPA; HAMMOUD, 2021). As Figura 15 e Figura 16 ilustram exemplos de imagens do espectro infravermelho para um mesmo objeto e para uma mesma cena, respectivamente.

As imagens obtidas das subdivisões do espectro infravermelho possuem aplicações específicas. Por exemplo, as imagens da faixa NIR são geralmente usadas em aplicações de detecção e rastreamento de olhos, a faixa SWIR é adequada para uso em ambientes com forte neblina e a faixa MWIR é comumente usada em cenários militares para detectar temperaturas acima da temperatura corporal. Também, as imagens LWIR e FIR são empregadas em problemas de vigilância por vídeo, detecção de pedestres e de faces. A detecção e descrição de características locais na faixa espectral LWIR também é interessante para aplicações relacionadas a movimento de objetos, em que as condições de iluminação tendem a mudar mais rapidamente do que a temperatura (LE et al., 2020; RICAURTE et al., 2014).

(a) Faixa VIS. (b) Faixa NIR. (c) Faixa SWIR. (d) Faixa MWIR. (e) Faixa LWIR.

Figura 15 – Exemplos de imagens do espectro infravermelho para um mesmo objeto (face humana).

Fonte: Adaptado de Hu et al. (2017).



(a) Espectro visível.

(b) Faixa NIR.

(c) Faixa SWIR.

(d) Faixa LWIR.

Figura 16 – Exemplos de imagens do espectro infravermelho para uma mesma cena.

Fonte: Adaptado de Pinchon et al. (2018).

Para que as imagens do espectro infravermelho possam ser utilizadas nas aplicações práticas anteriormente mencionadas, é necessário que estas estejam alinhadas espacialmente, dado que essas imagens podem ser obtidas de diferentes fontes e/ou diferentes instantes de tempo. Nesse caso, o alinhamento dessas imagens é comumente realizado por um processo de registro de imagens, em que a detecção e descrição de características locais consistem em tarefas essenciais (JIANG et al., 2021; LI et al., 2015).

Em relação à extração e descrição de características nas imagens VIS/IV, assim como nas imagens multiespectrais, há um desafio devido à falta de uma relação bem definida entre os valores de intensidade de seus pixels. Para ilustrar essa questão, a Figura 17 exemplifica uma comparação entre esses valores para diferentes comprimentos de onda. Para o pixel P_A, a primeira faixa possui um valor de intensidade baixo, enquanto a segunda faixa possui um valor de intensidade alto. No entanto, esse mesmo padrão de intensidade se difere ao considerar o pixel P_B, devido às diferenças na reflexão da luz para diferentes faixas do espectro eletromagnético.

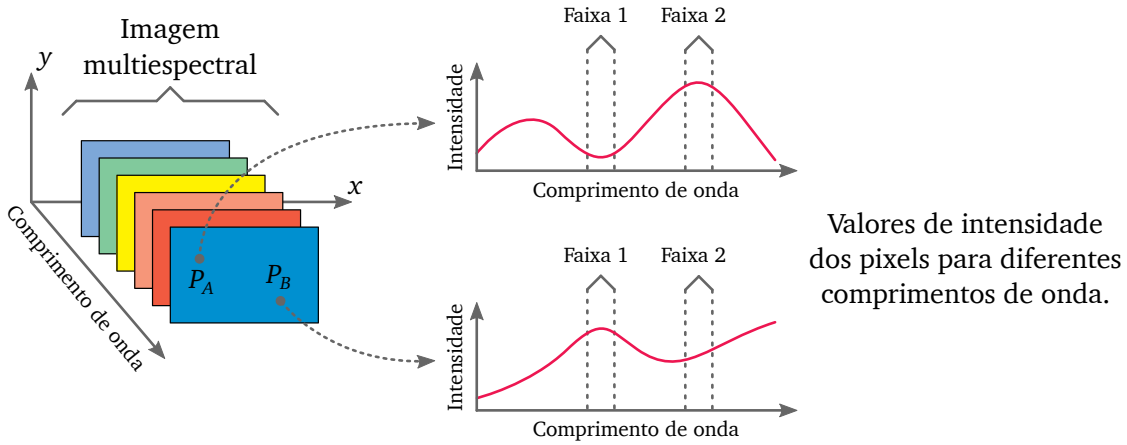


Figura 17 – Valores de intensidade dos pixels em uma imagem multiespectral.

Fonte: Adaptado de Joshi e Patel (2020).

2.5 Redes neurais artificiais

Um conceito relevante que precede a apresentação das soluções propostas neste trabalho é o das redes neurais artificiais. Para elucidar esse conceito, torna-se essencial introduzir o aprendizado de máquina, contexto no qual a rede neural artificial se insere.

O aprendizado de máquina (do inglês, *Machine Learning*) é uma subárea da Ciência da Computação e Inteligência Artificial, nas quais seus conhecimentos fornecem um conjunto de ferramentas estatísticas para estimar funções por meio da aprendizagem de dados. Ela aborda a questão de como construir algoritmos que melhoram automaticamente por meio da experiência. Um algoritmo de aprendizado de máquina é composto por um algoritmo de otimização, uma função de perda, um modelo e os dados para aprendizagem (JOSHI; PATEL, 2020; GOODFELLOW et al., 2016; JORDAN; MITCHELL, 2015).

Geralmente, os algoritmos de aprendizado de máquina podem ser categorizados como aprendizagem supervisionada ou aprendizagem não supervisionada. Na aprendizagem supervisionada, o programa recebe tanto a entrada quanto a saída desejada, por exemplo, imagens de objetos com rótulos correspondentes do que é representado. O objetivo desse aprendizado (também denominado treinamento) consiste em construir uma relação entre a entrada e a saída apresentada, podendo ser utilizada em problemas de classificação ou de regressão. Em contraste, na aprendizagem não supervisionada o objetivo do treinamento é encontrar similaridades nos dados fornecidos, podendo ser utilizado em problemas de agrupamento, por exemplo, para criar grupos a partir de dados semelhantes (JORDAN; MITCHELL, 2015; GOODFELLOW et al., 2016).

Um dos modelos de grande relevância na área de aprendizado de máquina é a rede neural artificial. Inspirada na neurociência, uma rede neural artificial consiste num modelo computacional composto por um sistema de várias unidades de processamento simples e associadas entre si. As unidades de processamento e suas conexões podem ser interpretadas desempenhando um papel análogo aos neurônios e suas conexões sinápticas em um sistema

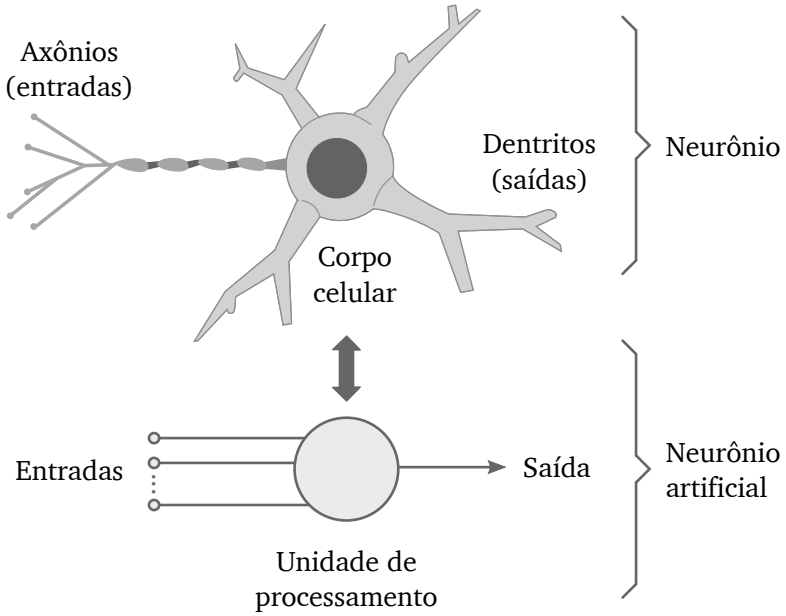


Figura 18 – Analogia entre um neurônio e um neurônio artificial.

Fonte: Adaptado de Haykin (2009).

nervoso, conforme ilustrado na Figura 18. A unidade de processamento das redes neurais artificiais pode ser chamada também de *neurônio artificial*, já que ela é inspirada no neurônio que também representa a unidade do processamento de informações numa rede neural biológica (GOODFELLOW et al., 2016; MCCULLOCH; PITTS, 1943).

O diagrama da Figura 19 mostra os detalhes de uma unidade de processamento que forma a base para o projeto das redes neurais artificiais. Nesse diagrama é possível identificar os elementos básicos do modelo de neurônio, quais sejam: (i) um conjunto de m "pesos" que representam as sinapses ou elos de conexão de um neurônio. Mais especificamente, um sinal x_j, na entrada j e conectado ao neurônio i, é multiplicado pelo peso w_{ij}; (ii) uma "Junção de soma" para somar os sinais de entrada, ponderados pelos respectivos pesos sinápticos do neurônio e (iii) uma "Função de ativação" $\Phi(\cdot)$ para limitar a amplitude da saída y_i do neurônio i.

O modelo neural da Figura 19 inclui um componente aplicado externamente, denotado por b_i. Este componente, conhecido como *bias* ou viés, adiciona uma constante à entrada da função de ativação, permitindo que a saída da função seja ajustada independentemente dos valores de entrada. Esse componente é um parâmetro crucial, por oferecer ao modelo neural a flexibilidade para se adaptar melhor aos dados. Em outras palavras, ajustando esse valor, o modelo pode alterar sua capacidade de ativação mesmo quando todas as entradas são zero. Sem ele, a capacidade do modelo de se ajustar e aprender com os dados seria limitada, uma vez que estaria restrito a responder apenas com base nas ponderações das entradas.

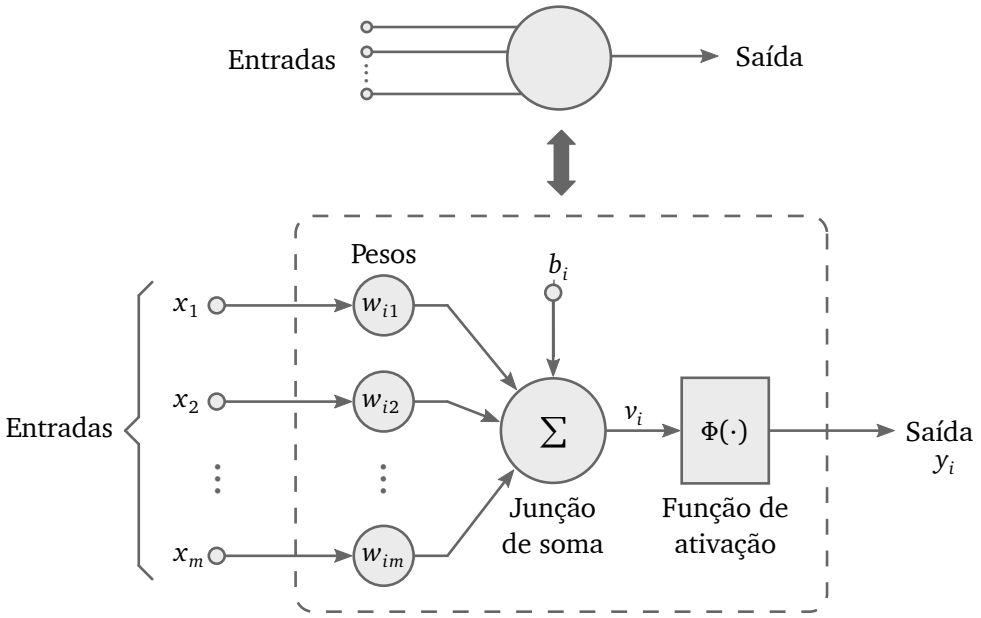


Figura 19 – Detalhes de um neurônio artificial.

Fonte: Adaptado de Haykin (2009).

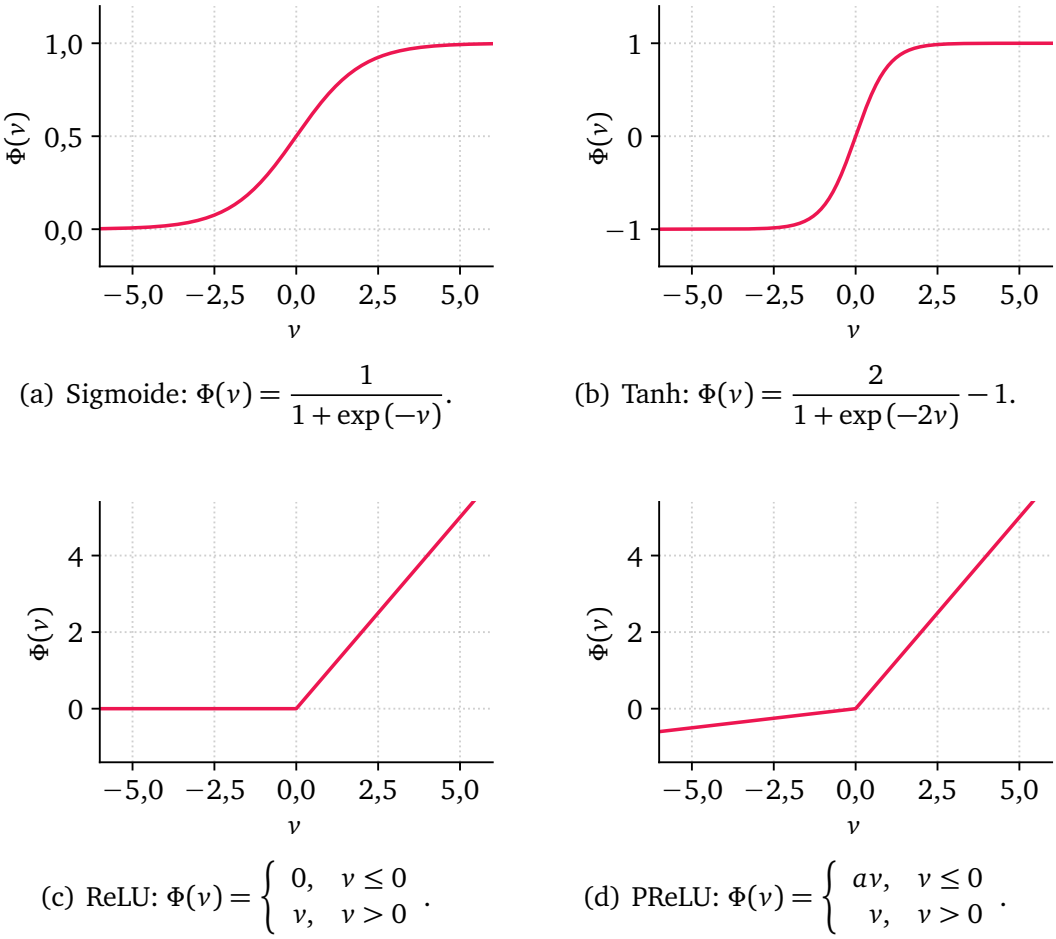
Em termos matemáticos, é possível descrever o neurônio representado na Figura 19 por meio da Equação (7) e da Equação (8) (HAYKIN, 2009; PAL; MITRA, 1992):

$$y_i = \Phi(v_i) \tag{7}$$

$$v_i = \sum_{j=1}^{m} w_{ij} x_j + b_i \tag{8}$$

Em um neurônio artificial, podem ser empregadas diversas funções de ativação $\Phi(v)$, que determinam sua saída com base em v. A função de ativação é inspirada na biologia, em que um neurônio é ativado somente se as entradas forem relevantes o suficiente. As funções mais comuns na literatura são: função sigmoide (também é conhecida como função logística), tangente hiperbólica (Tanh), ReLU (do inglês, *Rectified Linear Unit*), PReLU (do inglês, *Parametric Rectified Linear Unit*), dentre outras, conforme ilustradas na Figura 20 (APICELLA et al., 2021; KHAN et al., 2020).

Os neurônios de uma rede neural artificial podem ser associados de diversas maneiras. A organização dessa estrutura recebe o nome de arquitetura, e está intimamente relacionada com o algoritmo de aprendizagem utilizado para o treinamento (KHAN et al., 2020; HAYKIN, 2009). A Figura 21 exemplifica uma arquitetura de rede neural bastante comum, organizada em camadas. Arquiteturas organizadas dessa forma constituem uma classe de arquiteturas denominadas *Multilayer Neural Networks*. Nessa classe de arquiteturas, a entrada da rede recebe o nome de *camada de entrada*, as camadas seguintes são chamadas de *camada oculta*



(a) Sigmoide: $\Phi(v) = \dfrac{1}{1+\exp(-v)}$.

(b) Tanh: $\Phi(v) = \dfrac{2}{1+\exp(-2v)} - 1$.

(c) ReLU: $\Phi(v) = \begin{cases} 0, & v \leq 0 \\ v, & v > 0 \end{cases}$.

(d) PReLU: $\Phi(v) = \begin{cases} av, & v \leq 0 \\ v, & v > 0 \end{cases}$.

Figura 20 – Funções de ativação comumente utilizadas em redes neurais.

Fonte: Adaptado de Apicella et al. (2021).

e a última é intitulada *camada de saída*. Os dados nesse caso se propagam em um único sentido, que vai da entrada até a saída (LECUN; BENGIO; HINTON, 2015; PAL; MITRA, 1992).

A definição dos valores dos pesos de uma rede neural artificial é realizada por um processo de treinamento. Nesse processo, é necessário definir um algoritmo de otimização e uma função de perda para mensurar o erro atual da rede. O algoritmo de otimização ajusta os valores dos pesos para minimizar o resultado da função de perda (KHAN et al., 2020).

2.6 Redes neurais convolucionais

Uma Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*), também conhecida como *ConvNet*, consiste em uma categoria de rede neural artificial desenvolvida especialmente para reconhecer padrões em dados representados na forma de múltiplas matrizes. Vale ressaltar que na literatura há diversas outras redes neurais, inspiradas nas redes neurais convolucionais, dentre as quais se destacam as redes *Fully Convolutional*

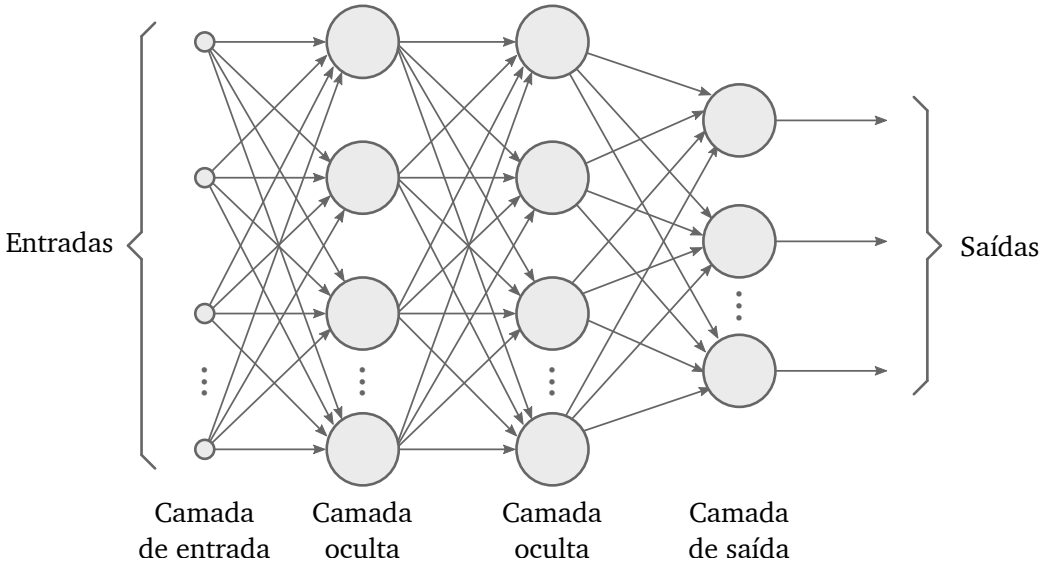


Figura 21 – Arquitetura de rede neural de múltiplas camadas.

Fonte: Adaptado de LeCun, Bengio e Hinton (2015).

Neural Network (FCN), *Region-based* CNN (R-CNN), bem como suas variantes *Faster* R-CNN e *Mask* R-CNN. Essas redes são comumente utilizadas para detecção e segmentação de objetos (KHAN et al., 2020; LECUN; BENGIO; HINTON, 2015).

As técnicas comuns de aprendizado de máquina eram ineficazes ao processar dados naturais (por exemplo, imagens e áudio). Por décadas, a construção de um sistema de aprendizado de máquina exigia uma engenharia cuidadosa e uma experiência de domínio sobre o problema para projetar um método que transformasse os dados brutos em uma representação adequada (LECUN; BENGIO; HINTON, 2015).

A necessidade em se trabalhar com sinais naturais em sua forma bruta motivou a construção das CNNs. Elas se baseiam na hipótese de que esses sinais são geralmente compostos por hierarquias de abstração. Nas imagens, por exemplo, os pixels podem representar arestas locais e então formar contornos. Tais contornos podem ser agrupados para formar objetos mais complexos (KHAN et al., 2020; LECUN; BENGIO; HINTON, 2015).

Portanto, uma CNN é estruturada pela composição de módulos capazes de transformar uma representação de nível mais baixo (começando com a entrada bruta dos dados) para uma de nível superior, de maior abstração. Assim, as CNNs possuem algumas características que as tornam adequadas para o processamento de sinais naturais, como: conexões locais para uma determinada região, uso de pesos compartilhados, subamostragem de sinais e a possibilidade de utilizar muitas camadas (LECUN; BENGIO; HINTON, 2015). Vale ressaltar que, uma rede neural que contém muitas camadas recebe o nome de rede *profunda*. Por esse motivo, alguns métodos de aprendizado de máquina são classificados como métodos de *aprendizagem profunda* (do inglês, *Deep Learning*), como nas redes neurais convolucionais (LITJENS et al., 2017).

Uma arquitetura típica de CNN é estruturada em pilhas de camadas (explicadas nas próximas subseções) e seguem um determinado padrão, conforme exemplificado na Figura 22. Os dados seguem um fluxo entre essas camadas, de tal forma que cada camada é responsável por aplicar uma determinada transformação. As primeiras camadas são responsáveis por detectar padrões de baixo nível, como cantos, bordas ou algum padrão mais simples. As camadas seguintes realizam uma associação dos padrões detectados para formar representações mais elaboradas, que correspondem a partes do objeto em questão (KHAN et al., 2020; LECUN; BENGIO; HINTON, 2015).

A seguir, são explicados o funcionamento e a utilidade das principais camadas utilizadas em CNNs, quais sejam: camada de convolução, camada de ativação, camada de agrupamento e camada totalmente conectada.

2.6.1 Camada de convolução

Essa camada é responsável por identificar padrões na imagem de entrada por meio da utilização de filtros. Esses filtros consistem em matrizes de pesos aprendidos durante

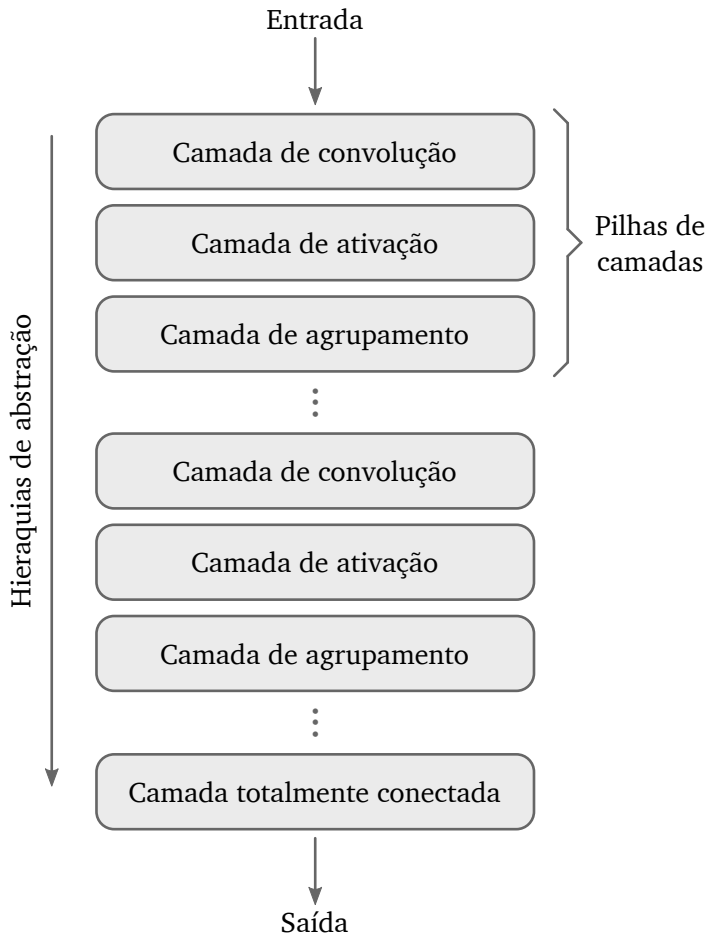


Figura 22 – Arquitetura típica de uma CNN.

Fonte: Adaptado de Patterson e Gibson (2017).

o processo de treinamento da rede. Os valores desses pesos são, geralmente, iniciados de forma aleatória.

A convolução é uma operação matemática realizada entre duas funções para produzir uma terceira. Ela é uma das operações mais importantes na área de processamento de sinais. A convolução discreta entre duas funções bidimensionais $F(x,y)$ e $G(x,y)$ é definida pela Equação (9) (GOODFELLOW et al., 2016; PAVEL; DAVI, 2013). A convolução é muito similar à correlação cruzada de dois sinais, utilizada para detecção de padrões e que já era utilizada nos primeiros trabalhos de registro de imagens (BAJCSY; KOVACIC, 1989).

$$C(x,y) = F(x,y) * G(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} F(i,j) \cdot G(x-i, y-j). \tag{9}$$

A camada de convolução em uma CNN, realiza a operação de convolução entre uma imagem de entrada $F(x,y)$ e um filtro $G(x,y)$. O resultado desse processo é conhecido como mapa de características. Esse mapa representa, graficamente, as respostas da imagem de entrada ao padrão definido no filtro, atuando dessa forma como um detector de padrões. Em outras palavras, o mapa de características informa qual região da imagem de entrada foi encontrado o padrão definido.

Uma camada de convolução pode conter n filtros, produzindo n imagens de saída para cada imagem de entrada. A Figura 23 ilustra um exemplo de camada de convolução, em que são aplicados dois filtros para uma mesma imagem de entrada.

Nessa camada, pode ser definido o tamanho do filtro, o tamanho do salto entre cada filtro (do inglês, *stride*) e um preenchimento (do inglês, *padding*) ao redor da entrada quando ela está sendo processada pelos filtros.

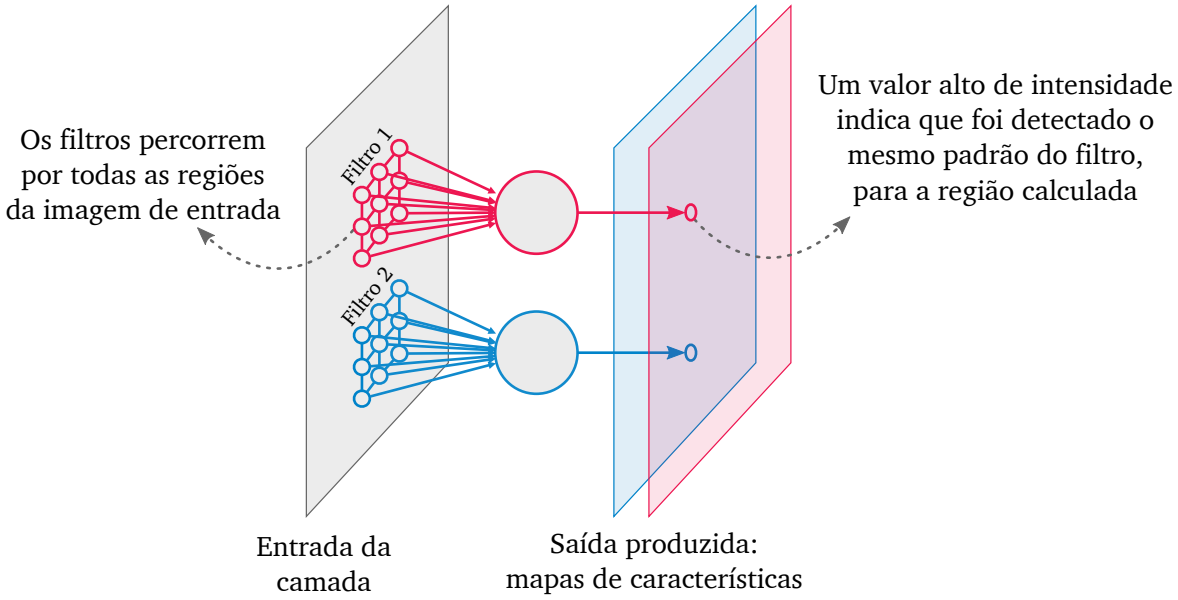


Figura 23 – Camada de convolução com dois filtros.

2.6.2 Camada de ativação

A camada de ativação tem por objetivo aplicar funções de ativação nos dados de entrada com o propósito de suprimir informações irrelevantes e possibilitar a generalização da rede (KHAN et al., 2020). Por esse motivo, essa camada é usualmente utilizada após uma camada de convolução, para filtrar informações contidas nos mapas de características.

Assim como em uma rede neural artificial em sua forma simples, a função de ativação é aplicada para cada valor de entrada e produz uma saída de mesmo tamanho. Nessa camada, portanto, para cada valor de pixel da entrada $C(x, y)$ aplica-se a função de ativação $\Phi(\cdot)$ para produzir a saída $Y(x, y)$, conforme definido pela Equação (10):

$$Y(x, y) = \Phi(C(x, y)). \tag{10}$$

Para essa função de ativação $\Phi(\cdot)$, podem ser utilizadas diferentes funções, como função sigmoide, tangente hiperbólica, ReLU e PReLU (conforme explicadas anteriormente) (APICELLA et al., 2021).

2.6.3 Camada de agrupamento

Outra categoria de camada importante para o funcionamento das CNNs é a camada de agrupamento (do inglês, *pooling layer*). Sua principal função consiste em realizar um agrupamento espacial dos dados de entrada, para reduzir a quantidade de informações e filtrar somente as partes mais relevantes de uma determinada região da imagem. Em outras palavras, essa camada é responsável por realizar uma subamostragem da entrada, produzindo uma imagem de menor tamanho na saída, conforme ilustrado na Figura 24. Essa

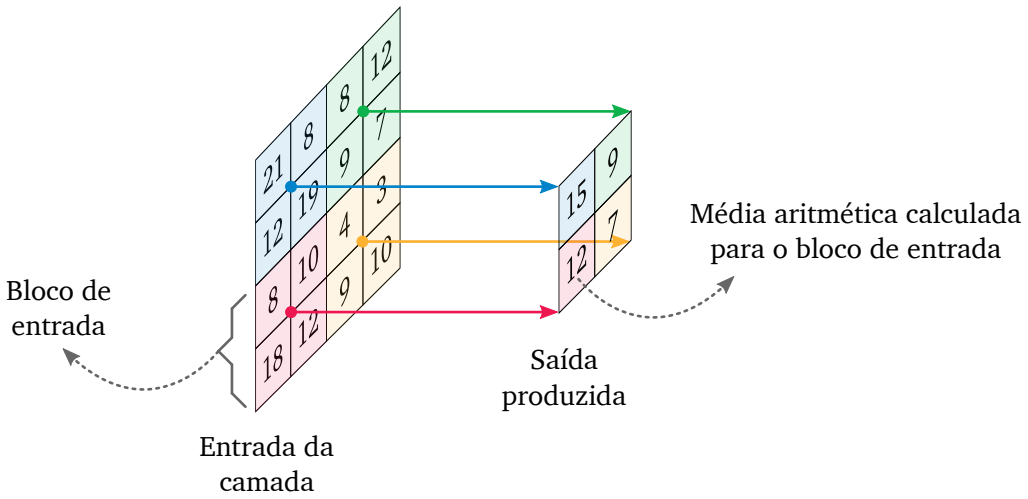


Figura 24 – Exemplo de funcionamento da camada de agrupamento.

Fonte: Adaptado de Buduma (2017).

camada afere robustez à rede, por eliminar informações irrelevantes ou ruídos e também maior velocidade, por reduzir os dados.

Cada camada de agrupamento recebe n imagens de entrada e fornece n imagens de saída, porém, com tamanho reduzido. O agrupamento dos dados na imagem é realizado em blocos, em que cada bloco pode ser sumarizado no valor máximo ou na média aritmética de seus valores. Nessa camada, pode ser definido o tamanho do bloco, o tamanho do salto entre cada bloco e um preenchimento ao redor da entrada quando ela está sendo processada pelos blocos.

2.6.4 Camada totalmente conectada

Uma camada totalmente conectada (do inglês, *fully connected layer*) consiste em uma camada de neurônios artificiais (como apresentado na Figura 19), em que suas entradas são interligadas em todos os valores da matriz de entrada, conforme ilustrado na Figura 25.

Essa camada é comumente utilizada no final da rede para possibilitar aplicações de classificação, de tal forma que o tamanho de sua saída representa a quantidade de classes possíveis de objetos (PATTERSON; GIBSON, 2017). Essa camada produz como saída uma matriz unidimensional de tamanho n, para qualquer dimensão de entrada. Algumas arquiteturas utilizam mais de uma camada desse tipo. O conjunto de saídas dessa camada pode ser utilizado também como um vetor de características.

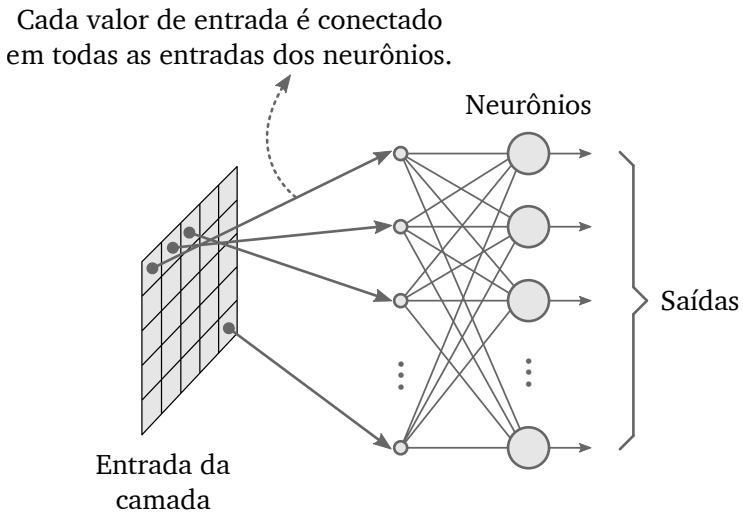


Figura 25 – Camada totalmente conectada.

Fonte: Adaptado de Buduma (2017).

Capítulo 3

Trabalhos Relacionados

Neste capítulo é apresentada uma revisão bibliográfica dos trabalhos mais relevantes na literatura, relativos ao problema de pesquisa deste trabalho. No intuito de organizar este capítulo, os assuntos foram separados em seções.

Inicialmente, na Seção 3.1 são apresentados alguns dos principais trabalhos relacionados ao problema da correspondência entre imagens. Em relação a esse problema, há diversos estudos que propõem soluções e processos para a correspondência entre imagens. Foram selecionados os estudos mais adequados e relevantes que realizaram uma revisão bibliográfica sobre o tema, especialmente aqueles que se alinham com os objetivos deste trabalho.

Em seguida, na Seção 3.2 são apresentados os trabalhos que propõem a construção ou a avaliação de métodos descritores de características locais para imagens do espectro visível. Na Seção 3.3, são abordados os estudos que sugerem a construção de métodos descritores para imagens oriundas do espectro visível e do infravermelho, assim como outros trabalhos correlatos. Cabe ressaltar a existência de um grande conjunto de estudos que abordam esses assuntos e que, portanto, é apresentado e discutido apenas um subconjunto composto por trabalhos cujas soluções e metodologias inspiram ou se assemelham àquelas apresentadas neste trabalho. Por fim, na Seção 3.4 é realizada uma discussão sobre o estado da arte, sumarizando os principais pontos abordados neste capítulo, os desafios ainda em aberto e as relações sobre a metodologia aqui desenvolvida.

3.1 Correspondência entre imagens

Conforme enunciado na introdução deste trabalho, o problema da correspondência entre imagens (do inglês, *image matching*) é uma área de pesquisa de grande importância no âmbito da Visão Computacional e Processamento de Imagens Digitais. Para este problema, a literatura apresenta diversas soluções, tais como aquelas baseadas em área e aquelas baseadas em características. Apesar da diversidade entre as técnicas para correspondência entre imagens, tem-se notado um aumento na adoção de técnicas com abordagem baseada em características, tornando-se um objeto de estudo interessante para muitos pesquisadores (LIU et al., 2021; MA et al., 2020; LI; HU; AI, 2020; BELLAVIA; COLOMBO, 2020; LENG et al., 2019).

O problema da correspondência entre imagens está fortemente relacionado a trabalhos sobre o registro de imagens, em que o objetivo consiste em realizar alinhamento espacial de duas ou mais imagens de uma mesma cena, obtidas de diferentes pontos de vista, diferentes sensores ou instantes de tempo (MA et al., 2020; SAXENA; SINGH, 2014; ZITOVÁ; FLUSSER, 2003). Alguns autores consideram ainda que o registro de imagens pode ser tratado como um problema da correspondência entre imagens (LENG et al., 2019; MA et al., 2020).

Nesse sentido, o trabalho de Zitová e Flusser (2003) apresenta uma revisão bibliográfica sobre as soluções existentes para a realização do registro de imagens. No estudo, as soluções existentes foram categorizadas em técnicas baseadas em área e técnicas baseadas em características. Segundo os autores, as técnicas baseadas em área têm sido comumente utilizadas para o registro de imagens médicas, especialmente por meio da utilização da informação mútua, como critério de similaridade. No entanto, os autores enfatizaram que essas técnicas são sensíveis às transformações geométricas e deformações locais contidas nas imagens. Como conclusão, os autores indicam que as técnicas baseadas em área apresentam-se mais apropriadas para circunstâncias em que ocorrem variações na iluminação da cena ou quando as imagens provêm de sensores distintos.

No trabalho de Kuppala, Banda e Barige (2020), complementar ao estudo apresentado por Zitová e Flusser (2003), foi apresentada uma nova revisão bibliográfica sobre correspondência entre imagens e registro de imagens, incluindo métodos baseados em CNN. Nesse estudo os autores apresentam que técnicas baseadas em área são mais adequadas para imagens com poucos detalhes significativos, sendo frequentemente empregadas em imagens médicas. Em contrapartida, as técnicas baseadas em características são mais adequadas em aplicações gerais, por se tratarem de métodos mais robustos às transformações geométricas. Segundo a revisão realizada neste estudo, as técnicas baseadas em características locais para realizar a correspondência entre imagens são compostas por três etapas principais, quais sejam: detecção, descrição e correspondência de características locais. Também, em relação à utilização de métodos baseados em CNN, os autores indicam que o custo computacional ainda é alto, mas que a utilização de tais métodos é promissora e ainda precisa ser explorada.

Em outro estudo, apresentado por Chen, Rottensteiner e Heipke (2020) foi apresentada uma revisão bibliográfica sobre métodos que utilizam a abordagem baseada em características para o problema da correspondência. Nesse trabalho os autores discutiram a utilização de métodos baseados em CNN e métodos baseados em modelagem manual (do inglês, *handcrafted*). Como conclusão, os autores indicam que os métodos baseados em CNN são promissores, mas que ainda há muito o que ser explorado. Nesse sentido, o estudo indica a importância de se integrar o conhecimento sobre o domínio do problema com as CNNs para atuar como um guia durante o aprendizado das redes.

Similarmente ao estudo apresentado por Chen, Rottensteiner e Heipke (2020), no estudo de Ma et al. (2020) é realizada uma revisão bibliográfica sobre soluções existentes para o problema da correspondência entre imagens. Os autores corroboram que técnicas

baseadas em características têm ganhado popularidade devido à sua flexibilidade e robustez, tornando-as adequadas para uma vasta gama de aplicações, além do registro de imagens. Esse estudo indica que os métodos baseados em área são sensíveis às variações de iluminação da cena, úteis apenas em casos em que há poucas deformações geométricas entre as imagens. Os métodos baseados em CNN são promissores, mas assim como apontado no estudo de Chen, Rottensteiner e Heipke (2020), ainda há a necessidade de se explorar mais tal abordagem, comparada aos métodos baseados em modelagem manual.

3.2 Métodos descritores para imagens obtidas do espectro visível

No âmbito das soluções para o problema da correspondência entre imagens, que utilizam técnicas baseadas em características, as tarefas de detecção e descrição de características locais consistem em tarefas fundamentais (LENG et al., 2019; CHEN; ROTTENSTEINER; HEIPKE, 2020; MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020).

Entre os algoritmos para a realização de tais tarefas se destaca o SIFT[1] (*Scale Invariant Feature Transform*), proposto por Lowe (2004). Este algoritmo é composto por um método detector e um método descritor de características locais, ambos robustos no que diz respeito às variações geométricas de escala e rotação. O algoritmo SURF (*Speeded Up Robust Features*), proposto pelos autores Bay, Tuytelaars e Gool (2006) e inspirado no método SIFT, também é um algoritmo relevante na literatura e apresenta uma eficiência superior. No entanto, trabalhos na literatura mostraram que o algoritmo SURF possui uma eficácia inferior ao SIFT (BARAJAS-GARCÍA; SOLORZA-CALDERÓN; GUTIÉRREZ-LÓPEZ, 2019).

Em relação à tarefa de detecção de características locais, no trabalho de Mikolajczyk et al. (2005) foi realizado um estudo comparativo sobre alguns métodos, incluindo o SIFT e SURF. Os autores apresentaram a medida de desempenho de repetibilidade para avaliar a eficácia da extração de características desses algoritmos. Em outro estudo relevante do mesmo grupo de pesquisadores, foi realizado um estudo comparativo sobre métodos descritores de características locais (MIKOLAJCZYK; SCHMID, 2005). Os autores avaliaram a robustez dos algoritmos no que diz respeito a variações de iluminação, diferentes pontos de vista e presença de distorções causadas pela compressão digital da imagem. Para a avaliação dos métodos, foram adotadas duas medidas de desempenho: revocação e precisão (conforme apresentadas na Seção 2.3 deste trabalho). No estudo, os autores mostraram ainda que os métodos construídos com base no algoritmo SIFT conseguem descrever características com eficácia, mesmo sob condições extremas. A avaliação dos métodos descritores no estudo foi realizada em um processo de correspondência entre imagens. A metodologia de avaliação, bem como as medidas de desempenho utilizadas, serviram como referência para avaliar a

[1] O funcionamento do algoritmo SIFT é explicado em detalhes no Apêndice B deste trabalho.

eficácia das soluções propostas neste trabalho.

Na literatura há também métodos descritores que têm por objetivo possuir um baixo custo computacional em detrimento da eficácia da descrição de características. Por exemplo, o método descritor BRIEF (*Binary Robust Independent Elementary Features*), proposto no trabalho de Calonder et al. (2010), segue essa linha de obter baixo custo computacional e rápido processamento. No entanto, a solução apresentada no estudo possui limitações no que diz respeito às mudanças de escala entre as imagens.

Inspirados no método descritor BRIEF, os métodos ORB (*Oriented FAST and Rotated BRIEF*) (RUBLEE et al., 2011), BRISK (*Binary Robust Invariant Scalable Keypoints*) (LEUTENEGGER; CHLI; SIEGWART, 2011) e FREAK (*Fast Retina Keypoint*) (ALAHI; ORTIZ; VANDERGHEYNST, 2012) foram projetados para manter um equilíbrio entre o custo computacional e a eficácia na descrição de características. Um dos fatores que contribuem para o baixo custo computacional dos métodos ORB, BRISK e FREAK se deve à utilização de descritores compostos por números binários para a representação das características, no lugar de descritores compostos por números reais, como utilizado nos métodos SIFT e SURF.

Todos esses métodos descritores consistem em alternativas eficientes, comparados aos algoritmos SIFT e SURF, podendo produzir resultados equivalentes em algumas aplicações (BRITO et al., 2016). Tais métodos descritores foram avaliados e estudados no trabalho de Heinly, Dunn e Frahm (2012), em que é realizada uma comparação entre diversos métodos detectores e descritores de características locais, para serem utilizados em processos de correspondência entre imagens, complementando o estudo de Mikolajczyk e Schmid (2005). Como conclusão, o estudo Heinly, Dunn e Frahm (2012) confirma a eficiência dos métodos que utilizam descritores compostos por números binários.

Com o advento do sucesso das Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Network*) em diversas aplicações na área de Visão Computacional, pesquisadores têm desenvolvido novas arquiteturas de CNN com o propósito de realizar a descrição de características locais de maneira similar aos métodos tradicionais SIFT e SURF (JOSHI; PATEL, 2020; USS et al., 2020). As CNNs e outras técnicas de aprendizagem profunda são empregadas em diversos níveis, como detecção de características, cálculos de similaridade ou até mesmo a predição de transformação entre imagens (KUPPALA; BANDA; BARIGE, 2020; JOSHI; PATEL, 2020). Vale ressaltar que essas redes têm recebido muita atenção nos últimos anos por seu sucesso em tarefas como classificação de cenas (DEDE; APTOULA; GENC, 2019), reconhecimento de objetos (REN et al., 2019), segmentação de imagens (NGO; CHA; HAN, 2020) e recuperação de imagens (QI et al., 2019).

Na linha das abordagens que utilizam CNNs para descrição de características locais, se destacam as arquiteturas DeepCompare e MatchNet, propostas nos trabalhos de Zagoruyko e Komodakis (2015) e Han et al. (2015), respectivamente. Ambos os trabalhos compartilham uma ideia similar de utilizar janelas retangulares de imagens para realizar a correspondência. Nesse caso, a comparação entre janelas é realizada pela própria arquitetura, e em contraste

aos métodos tradicionais (como o SIFT e o SURF, por exemplo), que comparam descritores por meio da distância euclidiana, esses trabalhos utilizaram uma camada específica para aprender uma medida personalizada. Ambos os estudos indicam ser promissora a utilização de abordagens baseadas em CNN para a aprendizagem e descrição de características. Uma crítica que pode ser feita sobre a metodologia utilizada nesses estudos, consiste na utilização de uma medida personalizada para comparar regiões, pois isso inviabiliza o uso da rede como uma substituição imediata dos métodos descritores tradicionais.

Como as arquiteturas DeepCompare e MatchNet, propostas nos trabalhos anteriores, dependem de medidas personalizadas para comparar regiões, os autores Simo-Serra et al. (2015) propuseram uma solução semelhante ao trabalho de Zagoruyko e Komodakis (2015), intitulada DeepDesc. Na solução proposta, em vez de usar uma medida personalizada para medir a similaridade entre duas janelas de imagens, é utilizada diretamente a distância euclidiana durante o treinamento. Sua estrutura de treinamento, ilustrada na Figura 26, consiste em uma estrutura siamesa com uma função de perda euclidiana que reforça pequenas distâncias entre as janelas de imagens semelhantes, além disso, cada CNN funciona como um descritor de características. Durante a etapa de teste da rede, é necessária apenas uma das duas CNNs para calcular o descritor de uma região da imagem. Os autores apresentaram que o descritor calculado dessa maneira pode ser usado como uma substituição imediata dos descritores de características tradicionais, como método descritor SIFT.

Inspirado no trabalho de Simo-Serra et al. (2015), os autores Balntas et al. (2016a) propõem uma rede de treinamento tripla, intitulada PN-Net, que utiliza um par de janelas de imagens correspondentes e uma janela não correspondente em cada amostra de treinamento, como ilustrada na Figura 27. A estrutura de treinamento empregada no estudo consiste em três cópias de uma mesma CNN com uma função de perda que utiliza três argumentos, inspirada também na distância euclidiana. Essencialmente, em cada etapa do treinamento, a rede seleciona um par de regiões similares e uma amostra não similar entre esses pares. Durante a etapa de teste da rede é necessária apenas uma das três CNNs para calcular o

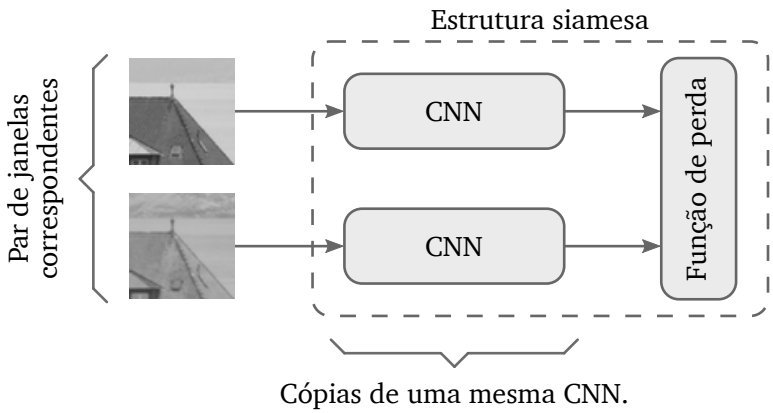


Figura 26 – Estrutura utilizada para o treinamento da arquitetura DeepDesc.

Fonte: Adaptado de Simo-Serra et al. (2015).

descritor de uma região da imagem. Os resultados do trabalho indicam uma melhora na velocidade geral de treinamento, em comparação com as técnicas anteriores baseadas em CNN. Em outro trabalho do mesmo grupo de pesquisa, em Balntas et al. (2016b), é proposta uma rede muito similar à rede PN-Net, denominada TFeat[2]. Entre essas soluções, de maneira geral, a principal diferença consiste nos parâmetros de treinamento. Nesse novo estudo, os autores também investigaram a utilização de diferentes funções de perda para redes triplas.

No intuito de comparar métodos descritores baseados em CNN e métodos baseados em modelagem manual (como os algoritmos SIFT e SURF), os autores Schonberger et al. (2017) avaliaram os métodos DeepDesc, TFeat, SIFT e algumas de suas variantes, por meio dos protocolos de avaliação definidos nos trabalhos de Mikolajczyk e Schmid (2005) e Heinly, Dunn e Frahm (2012). Em conclusão, os métodos baseados em modelagem manual apresentaram um desempenho igual ou superior àqueles baseados em CNN. Também, foi mostrado que os métodos descritores baseados em CNN apresentam alta variação em eficácia para diferentes bases de dados, sendo esta abordagem sensível a diferentes imagens e cenários.

Ainda no estudo de Schonberger et al. (2017), foi mostrado que a utilização de medidas de desempenho como o FPR95[3], comparada às medidas de desempenho de revocação e precisão, não é uma medida de desempenho adequada para avaliar métodos de descrição de características locais. Segundo os autores, valores altos de FPR95 não traduzem, necessariamente, altas taxas de correspondências corretas.

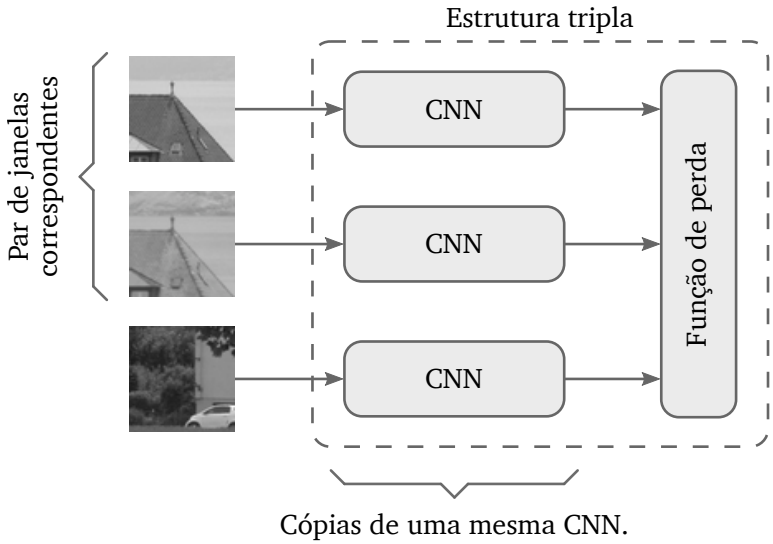


Figura 27 – Estrutura utilizada para o treinamento das arquiteturas PN-Net e TFeat.

Fonte: Adaptado de Balntas et al. (2016a).

[2] O funcionamento da arquitetura TFeat é explicado em detalhes no Apêndice B deste trabalho.
[3] A medida FPR95, ou Taxa de Falsos Positivos a 95% de Taxa de Verdadeiros Positivos, indica a proporção de identificações incorretas feitas por um sistema quando ele retorna 95% dos verdadeiros positivos.

No trabalho de Tian, Fan e Wu (2017) é proposta outra rede para descrever características locais que possibilita a utilização da distância euclidiana para correspondência de descritores, intitulada L2-Net. A solução proposta utiliza duas CNNs para descrever uma mesma janela de imagem, conforme a Figura 28. O descritor resultante é formado pela combinação dos dois vetores resultantes das saídas das CNNs. A estrutura de treinamento segue a estrutura similar à adotada na arquitetura DeepDesc, no entanto, a função de perda utilizada tem o objetivo de tornar os descritores de correspondências corretas o mais próximos possível no espaço euclidiano, por meio da distância dos vizinhos mais próximos. Também, no estudo é utilizada uma amostragem das correspondências, pois, segundo os autores, a quantidade de correspondências incorretas geralmente é maior, comparada às corretas. Nesse sentido, a amostragem utilizada realiza o balanceamento dessas quantidades. Em conclusão, a solução proposta se mostrou superior aos métodos descritores DeepDesc e PN-Net, em relação à precisão média. A arquitetura produzida, no entanto, é mais complexa, pois essa exige duas CNNs para descrever uma única janela, diferentemente das arquiteturas PN-Net, TFeat e DeepDesc, que necessitam apenas de uma CNN para uma janela de imagem.

Com base na utilização da arquitetura L2-Net, os autores He, Lu e Sclaroff (2018) desenvolveram uma medida personalizada para otimizar o descritor produzido e obter um maior valor médio de precisão. A abordagem desenvolvida, intitulada DOAP (do inglês, *Descriptors Optimized for Average Precision*) define uma função de perda em que as correspondências corretas devem ter um peso maior durante o treinamento em relação às correspondências incorretas. Neste estudo, os autores demonstraram que o método é superior em eficácia aos métodos SIFT, MatchNet e L2-Net em relação à precisão média. Assim como os métodos DeepCompare e MatchNet, uma crítica que pode ser feita sobre o método proposto consiste na utilização de uma medida personalizada para comparar regiões, o que inviabiliza seu uso em processos convencionais de correspondência entre imagens como uma substituição imediata dos métodos descritores tradicionais.

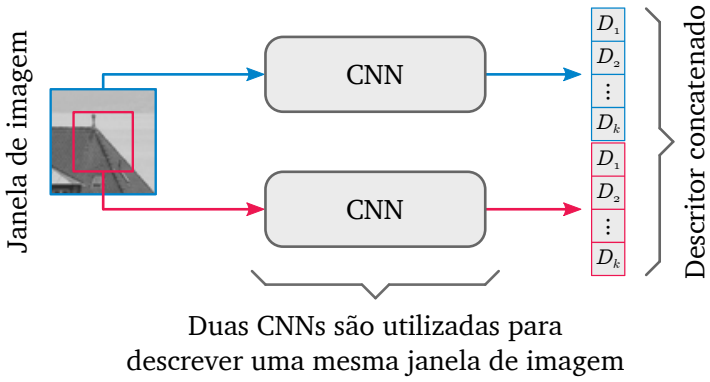


Figura 28 – Estrutura de descrição da arquitetura L2-Net.

Fonte: Adaptado de Tian, Fan e Wu (2017).

Outro método descritor baseado em CNN é apresentado no trabalho de Dong et al. (2019), denominado DescNet. No estudo é proposta uma arquitetura de CNN que contém apenas camadas de convolução e ativação para realizar a descrição de características locais. Assim como as arquiteturas MatchNet, PN-Net, TFeat a arquitetura proposta também permite a utilização de seus descritores em processos convencionais de correspondência entre imagens, pela distância euclidiana. A metodologia de treinamento da rede é inspirada no trabalho de Balntas et al. (2016b), em que é utilizado um par de janelas de imagens correspondentes e uma janela não correspondente em cada amostra de treinamento, como ilustrada na Figura 27. Neste estudo, os autores demonstraram que o método é superior aos métodos SIFT, MatchNet e TFeat em relação à precisão. Como crítica, o método descritor DescNet possui mais camadas de convolução que a arquitetura TFeat, de tal forma que o método proposto apresenta sete camadas em comparação à arquitetura TFeat que apresenta somente duas. Isso torna o modelo desenvolvido mais complexo e exige mais parâmetros de treinamento, o que torna seu processamento mais lento.

Ainda sobre os trabalhos que utilizam a abordagem via CNN para descrição de características locais, ainda há aqueles que propõem soluções que combinam a etapa de detecção e a etapa de descrição em apenas uma rede. Essas soluções são semelhantes àquelas que foram projetadas especialmente para a descrição de características locais, como apresentadas anteriormente. A principal diferença de tais soluções consiste na forma de treinamento e no desenho da estrutura da rede (MA et al., 2020). Dentre essas soluções se destacam os métodos LIFT (YI et al., 2016), SuperPoint (DETONE; MALISIEWICZ; RABINOVICH, 2018), RF-Net (SHEN et al., 2019) e D2-Net (DUSMANU et al., 2019). Essas soluções, no entanto, não permitem a maleabilidade de se utilizar métodos detectores específicos e suas implementações são, geralmente, mais complexas.

Em um recente estudo realizado por Bellavia e Colombo (2020) é realizada uma comparação do método tradicional SIFT com métodos baseados em CNN, como os métodos DeepDesc, L2-Net e DOAP. Outros métodos baseados em modelagem manual, como os métodos BRIEF, BRISK e FREAK também foram avaliados. Neste estudo também foi investigado possibilidades de se compactar o descritor SIFT para obter uma melhor eficiência sem comprometer sua eficácia. Como conclusão, os autores mostraram que o método descritor SIFT pode ter sua eficiência melhorada e que, apesar dos métodos baseados em CNN alcançarem melhores taxas de precisão, o algoritmo SIFT ainda mantém um balanço entre precisão e custo computacional. Nesse caso, os métodos baseados em modelagem manual possuem um custo computacional menor, comparado aos métodos baseados em CNN. Outro ponto importante indicado neste estudo se refere a considerar a orientação das características locais que, segundo os autores, possuem um papel importante durante o desenvolvimento de um descritor, pois esse atributo afere maior robustez ao método de descrição.

3.3 Métodos descritores para imagens VIS/IV

Na literatura há alguns métodos descritores de características locais propostos para serem robustos às variações de intensidade. Nessa linha, o trabalho de Chen et al. (2010), propõe a construção de um método descritor de características locais para o registro de imagens de retina humana, em que as imagens são obtidas do espectro infravermelho e do espectro de luz visível. O método proposto neste trabalho, intitulado PIIFD (*Partial Intensity Invariant Feature Descriptor*), consiste na modificação do algoritmo SIFT, em que os autores utilizam a orientação dos gradientes no lugar da direção dos gradientes da imagem para a construção do descritor (ver Seção B.1 para mais detalhes). Nesse sentido, a construção do descritor se baseou na hipótese de que o gradiente calculado entre imagens obtidas de diferentes faixas do espectro eletromagnético, para uma mesma região no mundo real, possuem mesma direção, mas sentidos opostos. Os autores mostraram que a solução proposta é superior ao algoritmo SIFT para a utilização em registro de imagens VIS/IV.

Outro trabalho também na linha das imagens VIS/IV, os autores Aguilera et al. (2012) apresentaram o método descritor EOH[4] (*Edge Oriented Histogram*). Esse método foi projetado sob a hipótese de que as bordas entre imagens VIS/IV, para uma mesma cena, permanecem preservadas. Nesse sentido, o método utiliza como uma etapa de pré-processamento a detecção de bordas por meio do algoritmo Canny (CANNY, 1987). Então, a partir das bordas das imagens, o método utiliza os passos do descritor EHD (*Edge Histogram Descriptor*) (MANJUNATH et al., 2001), em que filtros Sobel (SOBEL; FELDMAN, 1968) são utilizados para extrair informações dessas bordas e calcular um histograma que contém a orientação de cada uma delas, representando o descritor final. Para a avaliação do método proposto, o estudo utiliza um processo de correspondência entre imagens, composto pelo método detector SIFT e o descritor EOH. Segundo Qin et al. (2014), uma das críticas ao EOH consiste em sua precisão, principalmente porque o detector Canny necessita de parâmetros previamente definidos para um conjunto específico de imagens. No método EOH, portanto, a detecção das bordas das imagens podem variar, dificultando a definição de parâmetros adequados para imagens VIS/IV.

Similar ao método descritor EOH, os autores Mouats e Aouf (2013) apresentaram o PC-EHD. Como diferença ao EOH, o método descritor utiliza como uma etapa de pré-processamento a detecção de bordas por um banco de filtros Log-Gabor[5] , no lugar do algoritmo Canny. Então, a partir das bordas das imagens, o método utiliza os passos do descritor EHD (*Edge Histogram Descriptor*), da mesma forma utilizada no método descritor EOH. Os autores mostraram que o método proposto é superior ao método descritor EOH e SIFT, para realizar a descrição em imagens VIS/IV. Segundo os autores, a eficácia obtida pelo método é atribuída à utilização dos filtros Log-Gabor para a extração das bordas das imagens.

[4] O funcionamento do método descritor EOH é explicado em detalhes no Apêndice B deste trabalho.
[5] Os filtros Gabor e Log-Gabor são detalhados no Apêndice A deste trabalho.

Inspirado no método descritor SIFT, no trabalho de Saleem e Sablatnig (2014) é apresentado o método descritor NG-SIFT (*Normalized Gradients SIFT*). Esse método é similar ao algoritmo SIFT, no entanto, ele utiliza o cálculo de contraste local no lugar da magnitude dos gradientes do algoritmo (ver Seção B.1 para mais detalhes). Assim como no método EOH, esse método foi projetado sob a hipótese de que as bordas entre imagens VIS/IV, para uma mesma cena, permanecem preservadas. Para a comparação dos métodos, os autores utilizaram as medidas de desempenho de revocação e precisão. Em conclusão, os autores mostraram que o método proposto possui melhor precisão, em comparação ao algoritmo SIFT, quando aplicado em imagens VIS/IV. Apesar de o método NG-SIFT ter apresentado uma eficácia maior, comparado ao SIFT, os autores não aperfeiçoaram a eficiência do método em relação ao tempo de processamento do SIFT. Dessa forma, a descrição de características apresentada no estudo possui uma complexidade similar ao algoritmo SIFT.

Outro método descritor baseado no algoritmo SIFT, porém, para aplicações em correspondência de imagens SAR[6], é apresentado no trabalho de Dellinger et al. (2015). O algoritmo proposto no estudo, intitulado SAR-SIFT, apresenta um novo cálculo de gradiente adaptado para imagens SAR e que também trata ruídos em imagens desse tipo. Para eliminar correspondências incorretas, ao final do processo, os autores aplicaram o algoritmo RANSAC (*RANdom SAmple Consensus*) (FISCHLER; BOLLES, 1981) como uma etapa de pós-processamento. Posteriormente, o algoritmo SAR-SIFT foi aprimorado no trabalho de Wang et al. (2015), em que foi melhorada a distribuição espacial das características detectadas e adicionada a funcionalidade de controle no número de características. Apesar do algoritmo SAR-SIFT se apresentar eficaz ao lidar com ruídos, ele não considera relação não-monotônica dos valores de intensidade dos pixels entre as imagens SAR (CHEN et al., 2017).

Com o objetivo de construir um descritor de características robusto às variações de intensidade, os autores Aguilera, Sappa e Toledo (2015) apresentam o método LGHD (*Log-Gabor Histogram Descriptor*). A solução proposta no estudo possui uma estrutura similar aos métodos descritores EOH e PC-EHD para a realização da descrição de pontos, no entanto, diferentemente desses métodos que utilizam filtros Sobel para extração de informações de bordas, o LGHD emprega filtros Log-Gabor para a execução dessa tarefa. Para verificar a eficácia do método, os autores utilizaram bases de dados que contêm diversos tipos de imagens, como: mapas de profundidade, imagens obtidas sob diferentes níveis de iluminação e imagens VIS/IV. O método LGHD aplica 24 filtros Log-Gabor em 16 janelas em torno de um ponto-chave, fornecendo um descritor de tamanho 384. No estudo, para realizar a detecção das características foi empregado o detector FAST[7] (ROSTEN; DRUMMOND, 2006), por ser indicado como um método detector rápido e robusto às variações de intensidade. Apesar da eficácia do método proposto comparado aos algoritmos SIFT e SURF, o LGHD é sensível às variações geométricas de escala e rotação. Também, o fato do método LGHD resultar em um

[6] SAR (do inglês, *Synthetic Aperture Radar*) é um radar utilizado para produzir imagens de objetos. Uma imagem SAR é obtida por sinais recebidos da antena do radar que se desloca ao longo de uma trajetória para produzir uma imagem de alta resolução do objeto de interesse (RICHARDS, 2013).

descritor de tamanho 384 (três vezes maior que o descritor SIFT, que possui tamanho 128) pode tornar a etapa de correspondência de descritores mais lenta, comparada ao SIFT.

Inspirado no algoritmo LGHD, os autores Nunes e Pádua (2017) apresentaram o método descritor MFD (*Multispectral Feature Descriptor*). Diferentemente do método descritor LGHD, projetado para atender diversos tipos de imagens, o MFD foi construído especialmente para lidar com imagens VIS/IV. Para resolver o problema da eficiência do LGHD os autores definiram um conjunto de 10 filtros Log-Gabor com parâmetros ajustados para tratar exclusivamente da extração de textura das imagens, resultando em um descritor de tamanho 160. Os autores mostraram que o método proposto é superior aos métodos descritores SIFT, SURF, EOH e LGHD, em eficácia para realizar a correspondência de características entre imagens VIS/IV. Além disso, trabalhos na literatura mostraram que o MFD é superior a outros métodos, como: ORB, BRISK e NG-SIFT (SALEEM; BAIS, 2020; MA; MA; YU, 2019; FANG et al., 2019; SALEEM, 2019; FU et al., 2018b). Apesar do desempenho convincente indicado na literatura, o método MFD não considera variações geométricas de escala e rotação, assim como os métodos EOH e LGHD.

Na literatura há também estudos, como os trabalhos de Saleem et al. (2017) e Mouats et al. (2018), que avaliaram o comportamento de métodos detectores e descritores, projetados originalmente para lidar apenas com imagens obtidas do espectro visível, porém, no contexto das imagens VIS/IV. Nesses trabalhos foram avaliados os métodos detectores SIFT, SURF, ORB, BRISK, Harris (HARRIS; STEPHENS, 1988), FAST (ROSTEN; DRUMMOND, 2006) (*Features from Accelerated Segment Test*) e GFTT (*Good Features to Track*) (SHI; TOMASI, 1994). Dos principais descritores, foram avaliados os métodos SIFT, SURF, BRIEF, ORB, BRISK, FREAK, EOH e NG-SIFT. Os estudos utilizaram diversas bases de imagens de cenas internas e externas, obtidas de diferentes faixas do espectro eletromagnético como infravermelho próximo (NIR, do inglês *Near-Infrared*), de ondas longas (LWIR, do inglês *Long-Wave Infrared*) e infravermelho distante (FIR, do inglês *Far Infrared*).

Em ambos os trabalhos, os autores seguiram os mesmos protocolos de avaliação definidos em Mikolajczyk et al. (2005) e Mikolajczyk e Schmid (2005), por meio das medidas de desempenho de repetibilidade, revocação e precisão. Como conclusão, ambos os trabalhos mostraram que os métodos detectores de cantos Harris e FAST consistem em algoritmos mais eficazes no que diz respeito às diferenças nos valores de intensidade entre diferentes imagens, sendo estes mais adequados para a detecção de características locais em imagens VIS/IV. Em especial, o resultado de Mouats et al. (2018) mostrou que o método detector FAST foi, pelo menos, sete vezes mais rápido que os demais detectores. Em relação ao processo de descrição de características, os trabalhos concluem que os métodos que produzem descritores binários (ORB, BRIEF, BRISK e FREAK) possuem melhor eficácia no contexto avaliado, além de também possuírem um custo computacional baixo comparado aos outros métodos. O resultado de Mouats et al. (2018) também indicou que os métodos BRISK e FREAK são

[7] O funcionamento do método detector FAST é explicado em detalhes no Apêndice B deste trabalho.

os mais adequados no âmbito das imagens VIS/IV, principalmente quando há variações geométricas de escala e rotação.

Um dos primeiros estudos a investigar abordagens baseadas em CNN para descrição de características locais em imagens VIS/IV, foi o trabalho de Aguilera et al. (2016). No estudo, os autores avaliaram arquitetura DeepCompare em bases de dados compostas por imagens obtidas de diferentes faixas do espectro eletromagnético, como infravermelho próximo (NIR) e infravermelho de ondas longas (LWIR). Os autores compararam os resultados obtidos em relação aos métodos descritores SIFT, EOH e LGHD, por meio das medidas de desempenho de precisão e FPR95. Os resultados do estudo mostraram que, em alguns casos, o método DeepCompare obtém melhor precisão em relação a métodos baseados em modelagem manual, quando utilizado em imagens VIS/IV.

Ainda no contexto das abordagens que utilizam CNNs, os autores Aguilera et al. (2017) propõem uma arquitetura especialmente projetada para aprender características locais em imagens VIS/IV. A arquitetura, denominada Q-Net, é inspirada na arquitetura PN-Net, no entanto, os autores propõem uma solução composta por quatro cópias de uma mesma CNN, durante o processo de treinamento, conforme ilustrada na Figura 29. Para a avaliação, os autores utilizaram a medida de desempenho FPR95 e compararam as arquiteturas DeepDesc, PN-Net e os métodos EOH e LGHD. Os autores mostraram neste estudo que o método Q-Net possui uma acurácia de até 3% superior aos métodos avaliados. Os autores avaliaram também diferentes tamanhos de janelas de imagens, ao descrever características locais utilizando CNNs. Nesse sentido, os autores mostraram, por um experimento empírico,

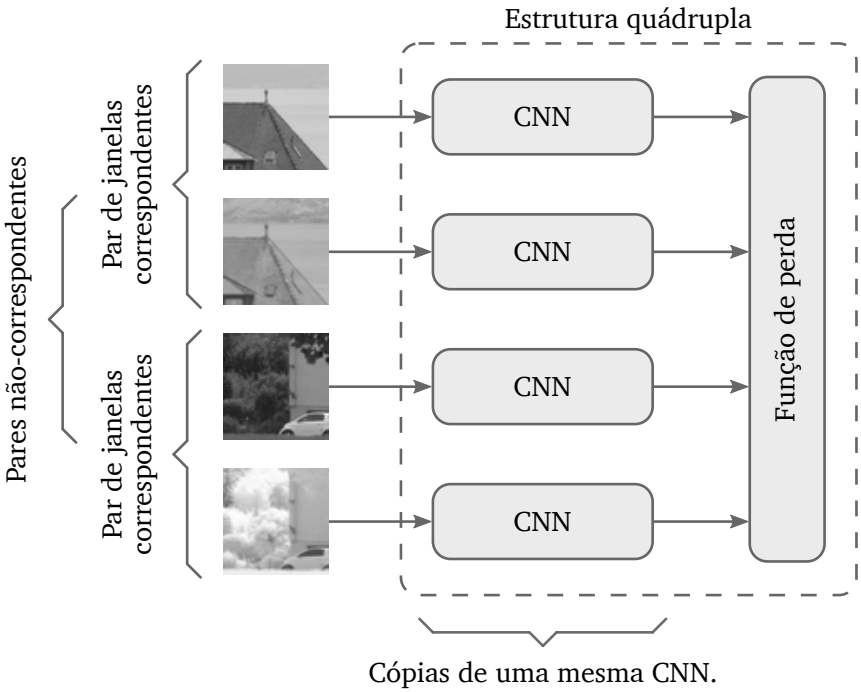


Figura 29 – Estrutura utilizada para o treinamento da arquitetura Q-Net.

Fonte: Adaptado de Aguilera et al. (2017).

que janelas de imagem de tamanho 64×64 produzem melhores resultados durante a descrição de características locais. Um ponto negativo do método apresentado no estudo está relacionado ao seu tempo de processamento, mais lento que a arquitetura PN-Net devido à sua complexidade. Outra crítica que se pode fazer sobre este estudo, está relacionada à medida de desempenho utilizada, tal qual, segundo Schonberger et al. (2017), a medida de desempenho FPR95 não é apropriada para o processo de avaliação de métodos descritores.

Inspirado no método descritor EHD, Fu et al. (2018b) apresentaram o método descritor para imagens VIS/IV, baseado em modelagem manual, denominado HODM (*Histogram of Directional Maps*). Esse método aplica filtros Sobel na região da característica local para extrair as bordas da imagem em vários ângulos, e, para cada imagem de resposta aos filtros, são extraídos informações de textura e de bordas da região (estrutura). Para cada uma dessas informações é calculado um histograma que contabiliza a maior resposta dos filtros para cada ângulo do filtro Sobel, por fim o descritor final é formado pela concatenação desses vetores. Os autores mostraram neste trabalho que o método HODM supera os métodos SIFT, SURF, NG-SIFT, PIIFD, EOH e MFD, em eficácia e eficiência. Entretanto, por sua construção, este método não é robusto às variações geométricas de escala e rotação.

Com base no método HODM, o mesmo grupo de pesquisadores modificaram seu método e propuseram o método HOMPC (*Histograms of Oriented Magnitude and Phase Congruency*), apresentado no estudo em Fu et al. (2018a). A estrutura do método proposto é a mesma definida pelo método HODM, com a diferença de que os filtros Sobel foram substituídos pelos filtros Log-Gabor, para extrair as bordas da imagem sob diversas maneiras. Os autores mostraram neste trabalho que o método HODM supera os métodos SIFT, SURF, NG-SIFT, PIIFD, EOH, PCEHD, LGHD e MFD, em eficácia. No estudo, o método foi avaliado sob diferentes variações geométricas de escala e rotação, e, os autores mostraram que o método é robusto às pequenas variações geométricas.

O trabalho de Li, Hu e Ai (2020) apresenta o método RIFT (*Radiation-variation Insensitive Feature Transform*) para descrever características locais. O método proposto neste trabalho, possui uma estrutura similar ao método LGHD, no entanto, ele aplica 6 filtros Log-Gabor em 36 janelas em torno de um ponto-chave, fornecendo um descritor de tamanho 216. Os autores mostraram neste trabalho que o método RIFT supera os métodos SIFT, SAR-SIFT e PIIFD, em eficácia para realizar a correspondência de características entre imagens multiespectrais, em geral, incluindo imagens VIS/IV. Segundo os autores, uma das limitações do método RIFT consiste em utilizá-lo em cenários em que há variações geométricas de escala e rotação. Apesar da eficácia nos resultados, uma crítica que pode ser feita ao estudo apresentando consiste na ausência da comparação do método proposto com métodos mais recentes, como MFD, LGHD e HODM e HOMPC.

Na literatura há também estudos recentes que propuseram processos de correspondência entre imagens VIS/IV inspirado nos métodos apresentados até aqui. Em Saleem e Bais (2020) é proposto processo de correspondência entre imagens VIS/IV com base no método

descritor NG-SIFT. Nos estudos de Wang et al. (2020) e Xu et al. (2021) foram desenvolvidas técnicas para correspondência com base nos contornos das imagens e inspirado no método descritor EOH. Em Li et al. (2021) é desenvolvido um processo de correspondência para imagens multiespectrais utilizando variações dos métodos descritores LGHD e MFD, envolvendo os filtros Log-Gabor. Nessa mesma linha, em Liu et al. (2021), foi apresentado um processo que envolve a utilização de uma modificação do método descritor MFD e inclui a etapa de verificação de restrições geométricas para realização das correspondências. Esses estudos, em geral, se baseiam na hipótese de que as estruturas e texturas entre imagens VIS/IV, para uma mesma cena, permanecem preservadas. Por esse motivo, tais estudos utilizam técnicas para extração de bordas e texturas para guiar a correspondência. Entretanto, comparado aos estudos anteriores, esses trabalhos não apresentaram uma contribuição significativa no que diz respeito ao processo de extração ou de descrição de características locais.

No trabalho de Wu e Zhu (2021) é proposto um método descritor que combina informações sobre o gradiente da região com informações de orientação de bordas, denominado HGEO (*Histogram of Maximum Gradient and Edge Orientation*). O método descritor consiste em um vetor formado por dois vetores concatenados: um contendo informações de gradientes e outro com informações de bordas de uma determinada região. As informações de gradientes são obtidas pela aplicação de filtros gaussianos na região contendo a característica local e produzem um histograma que contabiliza as variações de gradiente. Para a extração de informações de bordas, o método utiliza filtros Sobel na região e então calcula um histograma que representa o índice contendo as maiores respostas aos filtros (similarmente aos métodos EHD e EOH). Os autores mostraram no estudo que o método proposto foi superior ao método EOH. No entanto, o estudo não considerou os métodos mais recentes apresentados na literatura. Um problema referente a este método se refere ao cálculo de gradiente, pois, tanto as intensidades dos pixels quanto o cálculo do gradiente nessas imagens podem variar significativamente, não sendo, portanto, uma informação estável (LI et al., 2021).

3.4 Discussão sobre a literatura

Por meio dos trabalhos apresentados neste capítulo é possível observar que o uso de características locais para correspondência entre imagens é uma tarefa fundamental na visão computacional (KUPPALA; BANDA; BARIGE, 2020; CHEN; ROTTENSTEINER; HEIPKE, 2020; MA et al., 2020). Os trabalhos de revisão de literatura indicam que a utilização de técnicas baseadas em características é mais adequada para aplicações gerais de correspondência entre imagens, comparada às técnicas baseadas em área (ZITOVÁ; FLUSSER, 2003; MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020). Esse fato também motivou o presente trabalho a explorar a implementação de métodos para a solução do problema da correspondência entre imagens, que se enquadra no conjunto de técnicas baseadas em características.

Os estudos relacionados à extração de características locais em imagens VIS/IV são recentes e possuem questões ainda não resolvidas. Quando comparados aos métodos utilizados em imagens do espectro visível, fica evidente a necessidade de uma investigação mais aprofundada nesta área. Uma direção de pesquisa promissora consiste em personalizar métodos modernos para descrição de características locais para satisfazer diferentes requisitos em tarefas práticas (MA et al., 2020). Por meio da revisão bibliográfica fica evidenciado também que os métodos responsáveis pela descrição de características possuem grande importância em processos de correspondência entre imagens, considerada uma etapa crítica e com custo computacional relevante (CHEN; ROTTENSTEINER; HEIPKE, 2020; MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; LENG et al., 2019).

Os métodos descritores tradicionais e soluções baseadas em CNN, desenvolvidos especialmente para imagens obtidas do espectro visível, foram estudados e avaliados, como nos trabalhos de Heinly, Dunn e Frahm (2012), Schonberger et al. (2017), Bellavia e Colombo (2020) e Chen, Rottensteiner e Heipke (2020). Tais soluções foram também posteriormente avaliadas em imagens VIS/IV, como nos trabalhos de Saleem et al. (2017) e Mouats et al. (2018), e como a literatura mostra, o algoritmo SIFT ainda serve de inspiração para o desenvolvimento de novos métodos detectores e descritores ou na utilização em novas abordagens para correspondência de características (BELLAVIA; COLOMBO, 2020).

Em geral, os métodos baseados em modelagem manual (do inglês, *handcrafted*), como o método SIFT, ainda possuem uma eficácia e eficiência balanceadas, comparadas aos métodos baseados em CNN. Os métodos baseados em CNN podem ser desenvolvidos para se obter melhor eficácia tanto para detecção quanto para a descrição de características locais, porém ainda tais abordagens não foram suficientemente exploradas. Nesse sentido, uma questão importante consiste em como integrar o conhecimento de domínio especialista às abordagens baseadas em CNN (BELLAVIA; COLOMBO, 2020; CHEN; ROTTENSTEINER; HEIPKE, 2020; MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; JOSHI; PATEL, 2020).

No campo dos métodos desenvolvidos especialmente para lidar com imagens VIS/IV, o método descritor EOH, apresentado no estudo de Aguilera et al. (2012), serviu de inspiração para a construção de outros métodos, como os métodos PC-EHD, LGHD, MFD, HOMPC, RIFT e HGEO. A abordagem utilizada para a construção dos métodos PC-EHD, LGHD, MFD, RIFT e HOMPC, por meio das funções Log-Gabor, se mostra promissora, dado que essa abordagem visa solucionar o problema da relação não-monotônica dos valores de intensidade dos pixels entre as imagens VIS/IV. Também, esses métodos não são inspirados no algoritmo SIFT, considerado pela literatura um método que possui um custo computacional significativo e ineficiente no âmbito das imagens VIS/IV.

As soluções para descrição de características locais baseadas em CNNs para imagens VIS/IV possuem desafios a serem explorados. Nesse sentido, grande parte dos trabalhos na literatura não tratam a relação não-monotônica dos valores de intensidade dos pixels vizinhos das imagens VIS/IV de maneira direta, o que indica a importância de um estudo

nessa linha. Vale ressaltar também que apenas alguns métodos nessa linha permitem a utilização de seus descritores em processos convencionais de correspondência entre imagens por critérios de similaridades conhecidos como a distância euclidiana (norma L_2). Também, no contexto das imagens VIS/IV, os métodos baseados em modelagem manual são predominantes, comparados aos métodos baseados em CNN para descrição de características locais.

Os estudos apresentados neste capítulo também indicam que a avaliação dos métodos descritores é, habitualmente, realizada em processos de correspondência entre imagens, por etapas bem definidas (detecção, descrição e correspondência de características locais) (MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; MIKOLAJCZYK; SCHMID, 2005). Essa metodologia facilita a avaliação das soluções propostas, ao ser possível avaliar cada uma destas etapas separadamente. Adicionalmente, como a literatura mostra, as medidas de desempenho de revocação e precisão, apresentadas no trabalho de Mikolajczyk e Schmid (2005), ainda são comumente empregadas em grande parte dos estudos que abordam a construção ou a avaliação de métodos descritores de características locais (JOSHI; PATEL, 2020; MOUATS et al., 2018; SCHONBERGER et al., 2017). Tais medidas de desempenho foram também empregadas aqui, com o objetivo de avaliar as soluções propostas para descrição de características.

Como o presente trabalho tem o objetivo de apresentar métodos descritores de características locais em imagens VIS/IV, se faz necessário definir um método detector de características para avaliar a eficácia dos métodos propostos. Nesse sentido, os trabalhos de Mouats et al. (2018) e Saleem et al. (2017) sugerem a utilização dos métodos detectores FAST e Harris no âmbito das imagens VIS/IV, em especial o método detector FAST, por possuir uma velocidade de detecção superior.

Dentre os estudos apresentados neste capítulo, alguns como Fu et al. (2018a) e Ma, Ma e Yu (2019) exibem metodologias semelhantes à proposta neste trabalho referente ao processo de correspondência entre imagens. Nos estudos, os métodos descritores são avaliados de maneira separada, da mesma forma como será conduzida no presente trabalho. No que se refere às soluções para lidar com os desafios das imagens VIS/IV, as funções Log-Gabor apresentadas nos trabalhos de Aguilera, Sappa e Toledo (2015), Nunes e Pádua (2017) e Fu et al. (2018a), bem como abordagens via CNN, como empregadas nos trabalhos de Balntas et al. (2016b) e Aguilera et al. (2017), se mostram promissoras.

Neste capítulo, delimitou-se uma revisão da literatura sobre as abordagens e desafios relacionados ao problema de correspondência entre imagens. Além disso, foram discutidos estudos sobre métodos descritores para imagens obtidas tanto do espectro visível quanto do infravermelho. No Quadro 1 e no Quadro 2 são listados, em ordem cronológica, os principais estudos sobre métodos descritores de características locais apresentados aqui. As percepções derivadas dos trabalhos aqui mapeados estabeleceram uma base sólida para a construção dos métodos descritores propostos que serão apresentados nos capítulos seguintes.

Trabalho	**Método**	**Imagens**
Zagoruyko e Komodakis (2015)	DeepCompare	VIS
Han et al. (2015)	MatchNet	VIS
Simo-Serra et al. (2015)	DeepDesc	VIS
Balntas et al. (2016a)	PN-Net	VIS
Balntas et al. (2016b)	TFeat	VIS
Aguilera et al. (2017)	Q-Net	VIS/IV
Tian, Fan e Wu (2017)	L2-net	VIS
He, Lu e Sclaroff (2018)	DOAP	VIS
Dong et al. (2019)	DescNet	VIS

Quadro 1 – Principais trabalhos apresentados sobre métodos descritores baseados em aprendizagem profunda (*deep learning*).

Trabalho	**Método**	**Imagens**
Lowe (2004)	SIFT	VIS
Bay, Tuytelaars e Gool (2006)	SURF	VIS
Calonder et al. (2010)	BRIEF	VIS
Chen et al. (2010)	PIIFD	VIS/IV
Rublee et al. (2011)	ORB	VIS
Leutenegger, Chli e Siegwart (2011)	BRISK	VIS
Alahi, Ortiz e Vandergheynst (2012)	FREAK	VIS
Aguilera et al. (2012)	EOH	VIS/IV
Mouats e Aouf (2013)	PC-EHD	VIS/IV
Saleem e Sablatnig (2014)	NG-SIFT	VIS/IV
Dellinger et al. (2015)	SAR-SIFT	VIS/IV
Aguilera, Sappa e Toledo (2015)	LGHD	VIS/IV
Nunes e Pádua (2017)	MFD	VIS/IV
Fu et al. (2018b)	HODM	VIS/IV
Fu et al. (2018a)	HOMPC	VIS/IV
Li, Hu e Ai (2020)	RIFT	VIS/IV
Wu e Zhu (2021)	HGEO	VIS/IV

Quadro 2 – Principais trabalhos apresentados sobre métodos descritores baseados em modelagem manual (*handcrafted*).

Capítulo 4

Multispectral Features Network (MF-Net)

Conforme discutido na revisão bibliográfica deste trabalho, a utilização de Redes Neurais Convolucionais (CNNs) para descrição de características locais, apesar de promissora, ainda há muito o que ser explorado (CHEN; ROTTENSTEINER; HEIPKE, 2020; SCHONBERGER et al., 2017). No contexto das imagens VIS/IV, a utilização de CNNs para a construção de descritores é ainda desafiadora, posto que a relação entre os valores de intensidade dos pixels de diferentes imagens pode ser não-monotônica, conforme apresentado na introdução deste trabalho (BARUCH; KELLER, 2021; FAN et al., 2019; MA; MA; YU, 2019). Entretanto, essa abordagem tem obtido sucesso em diversas outras tarefas, incluindo a detecção de características, o que motiva explorar abordagens via CNN para descrever características locais, dada a lacuna a ser preenchida na literatura (JOSHI; PATEL, 2020; KUPPALA; BANDA; BARIGE, 2020).

No contexto das imagens VIS/IV, uma das limitações das CNNs se baseia no fato de que a distribuição espacial dos pixels na rede é tratada, de maneira geral, linearmente. Nesse caso, não há nenhuma camada em uma CNN típica capaz de lidar diretamente com relação não-monotônica dos valores de intensidade dos pixels vizinhos. De modo geral, a CNN aprende padrões em imagens por meio da operação de convolução, calculada sobre os valores de intensidade dos pixels das imagens (GOODFELLOW et al., 2016). No entanto, esses valores de intensidade não seguem um mesmo padrão quando se trata de imagens obtidas de diferentes faixas do espectro eletromagnético.

Nesse contexto, portanto, este trabalho propôs construir uma camada de CNN que possibilita que imagens obtidas de diferentes faixas do espectro eletromagnético tenham uma representação similar na rede após o mapeamento, para permitir o aprendizado de padrões que representam a mesma forma no mundo real. Em outras palavras, essa camada, intitulada neste trabalho *camada de mapeamento*, seria responsável por realizar o mapeamento entre o espectro visível e o espectro infravermelho. Nesse sentido, este trabalho propôs construir também a arquitetura de CNN que incorpora essa camada, para descrever características locais entre imagens VIS/IV.

A arquitetura da CNN proposta neste trabalho, intitulada *Multispectral Features Network* (MF-Net), tem o propósito de produzir um descritor a partir de uma imagem de entrada. Mais especificamente, o objetivo do método proposto consiste em calcular um descritor $D \in \mathbb{R}^k$ de uma janela de imagem $J \in \mathbb{R}^{n \times n}$, conforme ilustrado na Figura 30. O descritor D consiste na matriz unidimensional resultante da última camada da CNN e possui

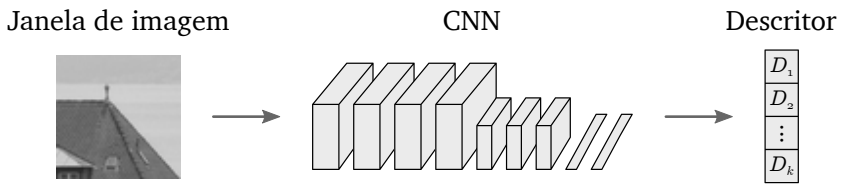


Figura 30 – Objetivo do método MF-Net: calcular um descritor a partir de uma janela recortada.

tamanho k. A janela de imagem J representa uma característica local de uma imagem obtida do espectro visível ou do espectro infravermelho.

As principais etapas para construção e avaliação do método proposto são ilustradas na Figura 31. Cada uma dessas etapas é explicada nas seções subsequentes deste capítulo. Na Seção 4.1 é detalhada a construção e funcionamento da camada de mapeamento. Essa camada foi incorporada em uma arquitetura de CNN sob diferentes maneiras (produzindo diferentes arquiteturas), com o objetivo de verificar o comportamento de sua inserção por uma análise exploratória, conforme apresentado na Seção 4.2. Na Seção 4.3 é explicado o processo utilizado no treinamento dessas redes e na Seção 4.4 são explicados os passos e o processo de avaliação utilizado nos experimentos.

Neste capítulo também são apresentados os resultados do método proposto. Os resultados foram divididos em duas principais seções. A Subseção 4.5.1 apresenta o resultado da análise exploratória das diferentes arquiteturas produzidas, e a Subseção 4.5.2 exibe a avaliação e comparação do método proposto com os demais métodos do estado da arte. Os resultados são discutidos na Seção 4.6 em que também são apresentados os principais pontos do método proposto, bem como suas limitações.

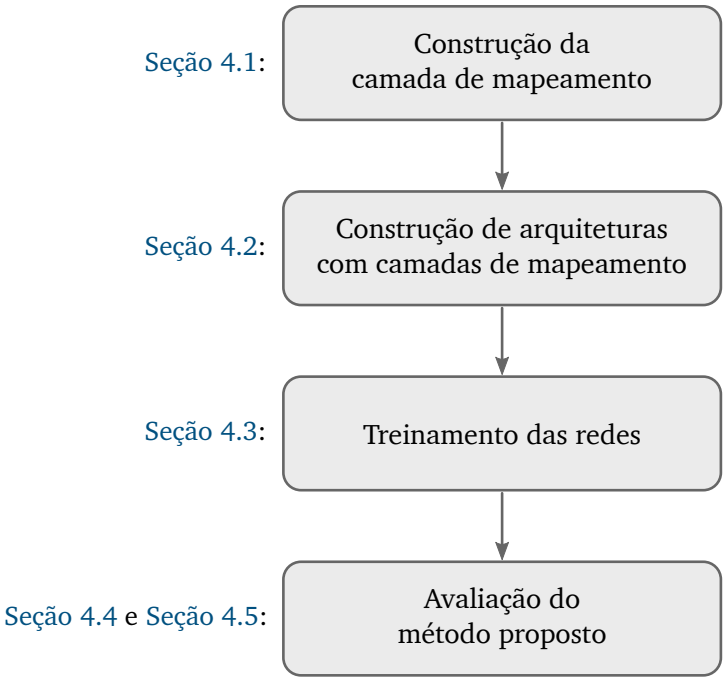


Figura 31 – Metodologia para construção e avaliação da arquitetura MF-Net.

4.1 Camada de mapeamento

Conforme mencionado no início deste capítulo, este trabalho apresenta uma nova camada de CNN, para possibilitar que imagens obtidas de diferentes faixas do espectro eletromagnético tenham uma representação similar na rede, conforme ilustrada na Figura 32. Essa camada, intitulada *camada de mapeamento*, é inspirada em trabalhos recentes da literatura, que utilizam informações de bordas das imagens VIS/IV para produzir descritores de características locais (FU et al., 2018a; LIU; LI; PAN, 2019).

Cabe lembrar aqui que, as aparências das formas dos objetos nas imagens VIS/IV tendem a permanecer preservadas (JIANG et al., 2021; WU; ZHU, 2021; FANG et al., 2019), e, o uso das informações de bordas é importante nesse contexto de descrição de características. Nesse contexto, dado que a camada de mapeamento foca na detecção de bordas nas imagens em sua saída, podemos inferir que há uma perda de informação quando essa camada é aplicada em regiões da imagem desprovidas de bordas. Essa questão, no entanto, é útil para a utilização em imagens VIS/IV, dado que regiões planas entre essas imagens podem apresentar níveis diferentes de intensidade.

Nesta camada, para a extração das informações de bordas, empregou-se um conjunto de filtros Log-Gabor[1] (FIELD, 1987), devido ao seu sucesso na literatura para detecção de bordas em imagens, principalmente no contexto das imagens VIS/IV. Também, esses filtros permitem utilizar diferentes parâmetros, como os parâmetros de escala e orientação, os quais os tornam bastante úteis para a análise de textura em imagens (YU; ZHAO, 2020; LIU et al., 2021; LI et al., 2021; WALIA; VERMA, 2016).

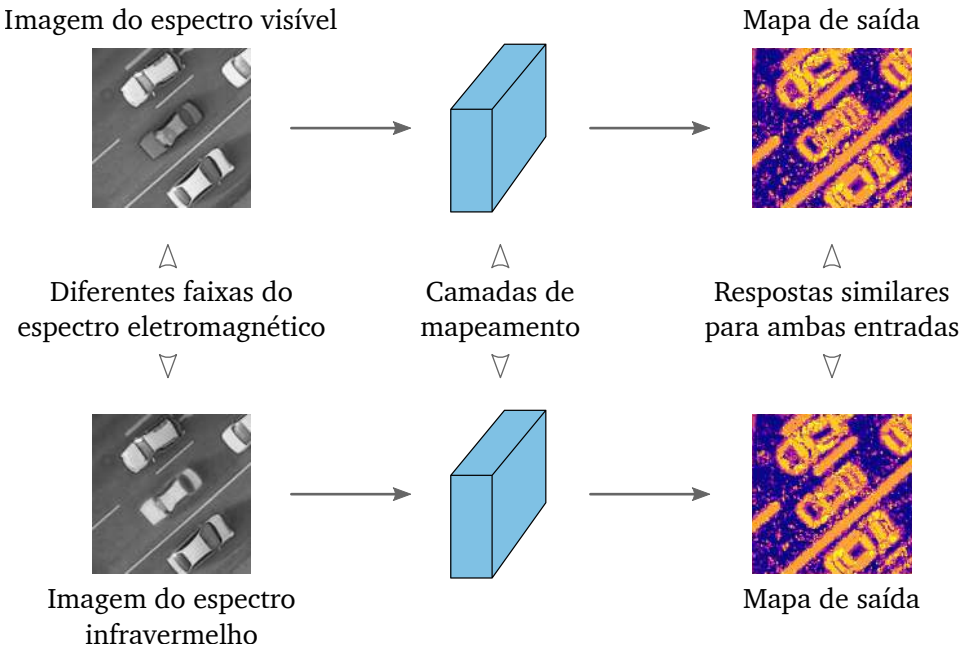


Figura 32 – Exemplo de funcionamento da camada de mapeamento.

[1] Para mais detalhes sobre os filtros Log-Gabor, veja o Apêndice A deste trabalho.

O filtro Log-Gabor (WALIA; VERMA, 2016), no domínio da frequência e utilizando coordenadas polares, é definido conforme a Equação (11):

$$F_{s,\omega}(r,\theta) = \exp\left(-\frac{\log\left(r/\lambda\kappa^{s}\right)^{2}}{2\sigma_r^2}\right).\exp\left(-\frac{(\theta-\omega)^2}{2}\right), \tag{11}$$

em que r e θ representam o raio e o ângulo do filtro em coordenadas polares, s e ω consiste na escala e no ângulo de orientação do filtro, σ_r representa o desvio padrão gaussiano, λ consiste no comprimento de onda da menor escala e κ representa uma constante multiplicativa de escala entre filtros sucessivos.

Para a utilização dos filtros, foram definidos os parâmetros $\lambda = 3{,}0$, $\kappa = 2{,}0$ e $\sigma_r = 0{,}65$, conforme sugeridos no trabalho de Walia e Verma (2016), por apresentarem melhores resultados em tarefas de descrição de imagens. Também, utilizou-se o mesmo conjunto de filtros definido pelo método descritor MFD (NUNES; PÁDUA, 2017), dado que os autores demonstraram resultados efetivos em imagens VIS/IV. Esse conjunto de filtros é formado pela associação de 5 valores de orientação (ω indo de 0 até 180 graus, igualmente divididos) por 2 valores de escala ($s \in \{1,2\}$, da Equação (11)), o que resulta em um conjunto de 10 filtros.

Embora os parâmetros da camada de mapeamento sejam pré-definidos, o fato de essa camada estar em uma arquitetura CNN, permite o aprendizado de todos os parâmetros por uma simples modificação em sua implementação. Assim, seria possível aprender os parâmetros da função Log-Gabor para produzir os filtros Log-Gabor adaptativamente. No entanto, objetivou-se produzir o menor modelo possível com uma menor quantidade de parâmetros, como também usar parâmetros que já foram validados na literatura, como os valores definidos em Walia e Verma (2016) e Nunes e Pádua (2017).

Inicialmente, na camada de mapeamento, são calculadas as imagens de resposta aos filtros Log-Gabor para a imagem de entrada I, sob diferentes níveis de escala s e orientação ω (10 filtros no total). Nesse caso, cada imagem de resposta $R_{s,\omega}$ é obtida conforme a Equação (12):

$$R_{s,\omega} = |\,\mathrm{IFFT}(\mathrm{FFT}(I) * F_{s,\omega})|, \tag{12}$$

em que "||" representa o módulo (valor absoluto), FFT representa a Transformada Rápida de Fourier e IFFT a sua função inversa. O filtro Log-Gabor é denotado por $F_{s,\omega}$ para a escala s e orientação ω (conforme a Equação (11)).

Note na Equação (12) que a utilização dos filtros Log-Gabor é feita no domínio da frequência, dado que a construção desses filtros nesse domínio é mais simples. Então, para extrair as bordas de uma imagem I, esta imagem é transformada do domínio espacial para o domínio da frequência, por meio da Transformada Rápida de Fourier (FFT). Em seguida, é realizada a convolução da imagem I com o filtro $F_{s,\omega}$. Por fim, a imagem de resposta ao

filtro Log-Gabor é obtida por meio do cálculo da inversa da Transformada Rápida de Fourier (IFFT).

Após calcular todas as imagens de resposta, é realizada a combinação de todas elas para produzir um Mapa de Saída[2] M, conforme ilustrado na Figura 33. Esse Mapa de Saída é composto pelos índices onde a resposta do filtro foi a maior dentre todas as imagens de resposta. Mais especificamente, seja n o número de filtros utilizados, $j \in \{1, 2, \ldots, n\}$, o valor do pixel $M(x, y)$ é igual a j se o pixel da imagem de resposta $R_j(x, y)$ for o valor máximo de todas as imagens de resposta. Esse cálculo pode ser escrito conforme a Equação (13). A Figura 34 exemplifica as etapas realizadas na camada de mapeamento.

$$M(x,y) = \begin{cases} j, & \text{se } R_j(x,y) > R_i(x,y) \text{ para todo } i \neq j,\ i,j \in \{1,\ldots,n\}, \\ 0, & \text{caso contrário.} \end{cases} \tag{13}$$

Essa etapa de combinação de imagens de resposta é inspirada no método descritor EOH (*Edge Oriented Histogram*) (AGUILERA et al., 2012). Nesse método, de maneira geral, várias imagens são combinadas em uma única saída para produzir um mapa que contém apenas os índices das imagens onde as bordas foram mais relevantes. Esse processo de combinação de imagens também é utilizado de maneira similar em métodos recentes para descrição de características locais (FU et al., 2018a; FU et al., 2018b; LI; HU; AI, 2020; WU; ZHU, 2021). Vale ressaltar, no entanto, que apesar da camada de mapeamento incorporar técnicas conhecidas pela literatura, não há nenhuma camada em uma CNN típica com o propósito de resolver esse problema.

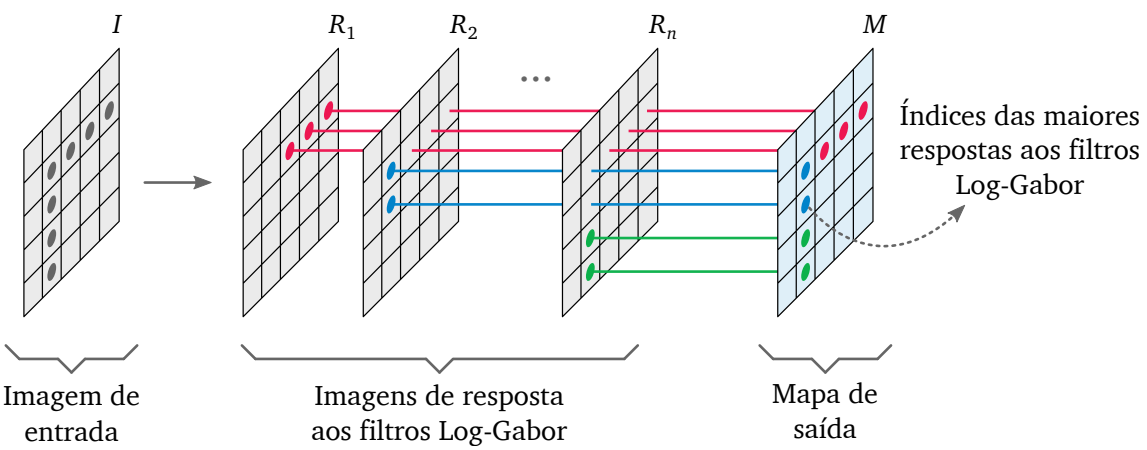


Figura 33 – Combinação das imagens na camada de mapeamento.

[2] Embora o Mapa de Saída possa ser representado em escalas de cinza, optou-se por usar uma escala de cores nas ilustrações para diferenciá-lo claramente das imagens de entrada no presente trabalho.

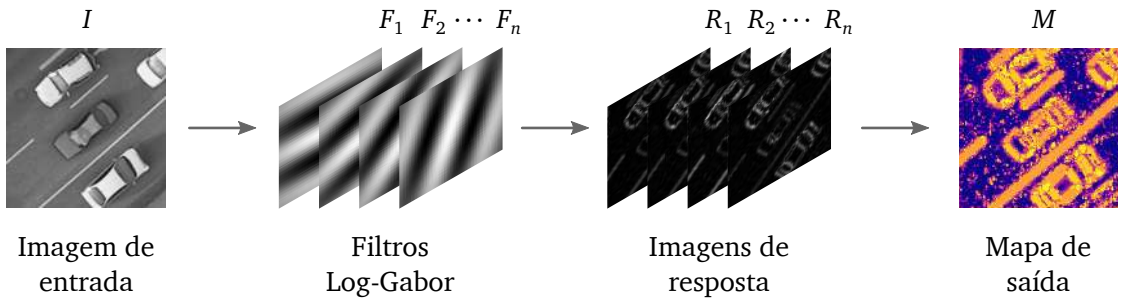


Figura 34 – Exemplo das etapas realizadas na camada de mapeamento.

4.2 Arquitetura da rede

Para a construção da arquitetura da rede deste trabalho, utilizou-se como ponto de partida uma CNN do estado da arte, também projetada para descrever características locais. À vista disso, foram incorporadas as camadas da arquitetura TFeat (BALNTAS et al., 2016b), visto que esse modelo possui propriedades desejáveis, por se tratar de uma arquitetura de poucas camadas e por se mostrar eficiente para produzir descritores de características locais em imagens obtidas do espectro visível. Para as camadas já existentes do modelo TFeat foram utilizados os mesmos parâmetros (conforme detalhados na Figura 36). Adicionalmente, a arquitetura da rede deste trabalho incorpora camadas de mapeamento, adicionadas antes de cada camada de convolução, visto que ela é responsável por identificar padrões, e o objetivo da camada de mapeamento seria de tornar similar a representação de padrões entre imagens obtidas de diferentes faixas do espectro eletromagnético.

Para verificar o comportamento da inserção da camada de mapeamento, foram comparadas três arquiteturas, quais sejam: (i) sem camadas de mapeamento, (ii) com apenas uma (na posição 1) e (iii) com duas camadas de mapeamento (nas posições 1 e 5). Assim sendo, realizou-se uma análise exploratória incorporando a camada de mapeamento sob diferentes maneiras, conforme ilustrada na Figura 35.

Vale ressaltar que, apesar de o método proposto neste trabalho ser baseado na arquitetura TFeat, por possuir propriedades desejáveis, a arquitetura proposta neste trabalho,

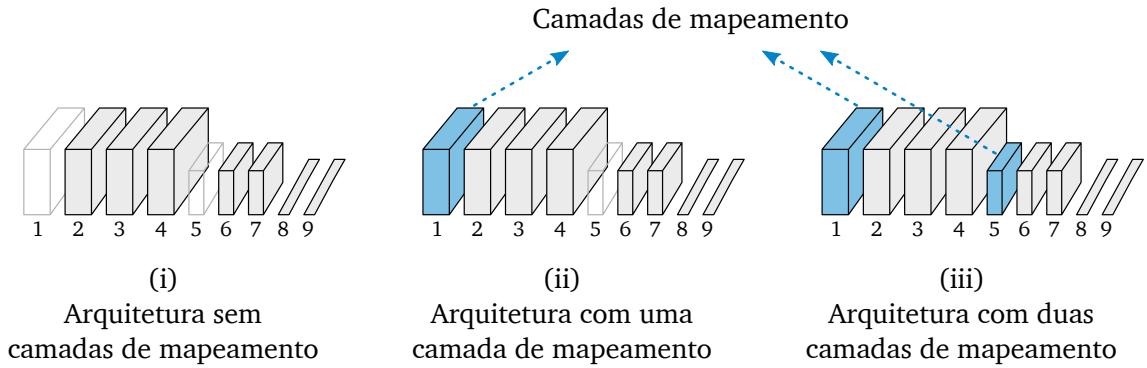


Figura 35 – Diferentes arquiteturas avaliadas.

intitulada MF-Net, foi desenvolvida para lidar com imagens VIS/IV. Portanto, o método aqui proposto não se trata de uma extensão de trabalhos anteriores, dado que este possui um funcionamento diferente e um escopo de aplicação distinto.

A Figura 36 exibe os parâmetros e os detalhes da arquitetura construída. O descritor final da arquitetura possui tamanho $k = 128$, conforme apresentado na camada 9 da Figura 36. Vale ressaltar que, o descritor produzido pode ser empregado em processos convencionais de correspondência entre imagens, por meio da utilização da distância euclidiana (norma L_2) como critério de similaridade.

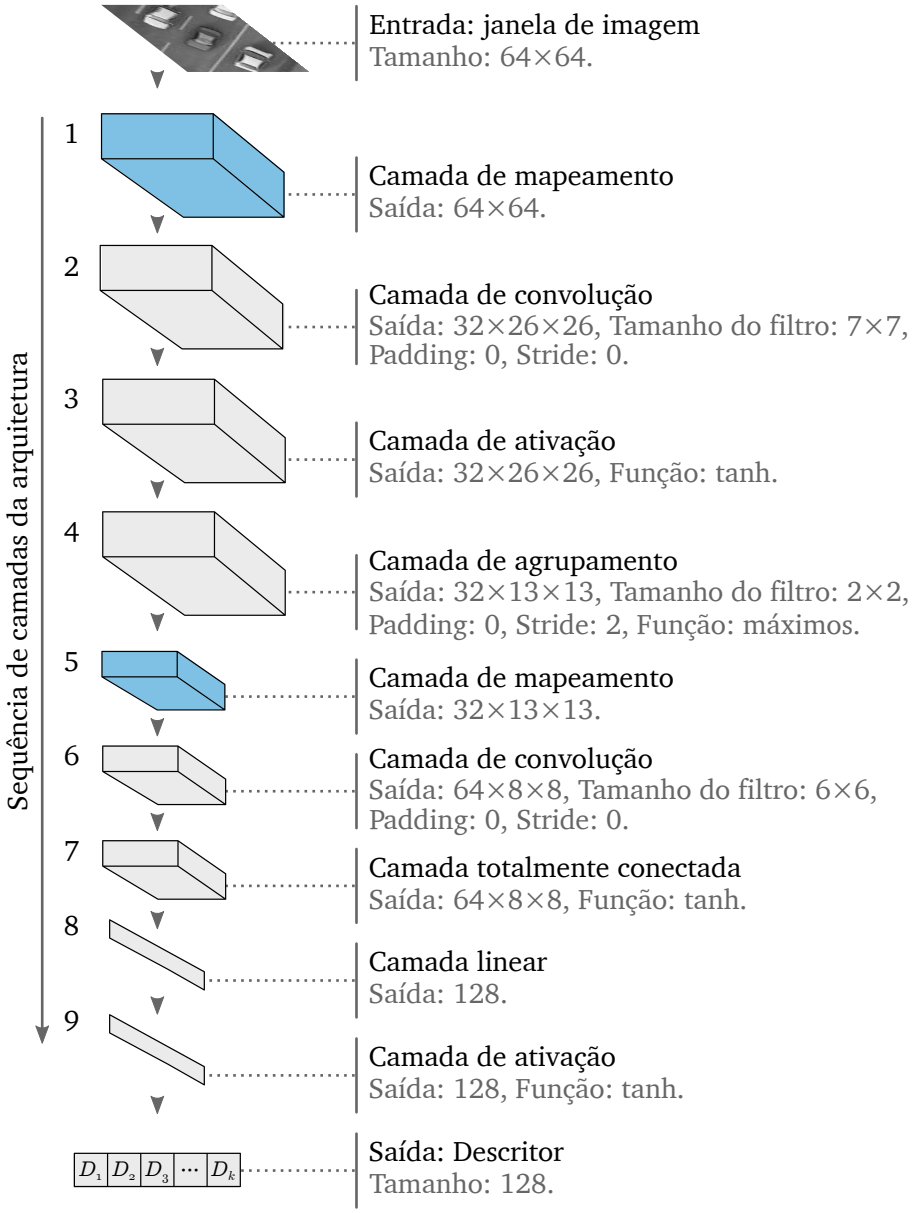


Figura 36 – Detalhes da arquitetura construída neste trabalho.

4.3 Treinamento da rede

Para o treinamento das redes, utilizou-se o mesmo protocolo definido no trabalho de Balntas et al. (2016b). Nesse sentido, otimizaram-se os parâmetros da rede por meio do algoritmo Gradiente Descendente Estocástico[3] (BOTTOU, 2010) e realizou-se o treinamento em lotes de 128 amostras com uma taxa de aprendizagem de 0,1. Esses valores seguem os mesmos definidos no trabalho de Balntas et al. (2016b), uma vez que a estrutura proposta é inspirada em seu trabalho e também se desejou manter uma comparação mais justa.

Para o processo de treinamento, utilizou-se a estrutura tripla, definida no trabalho de Balntas et al. (2016b), conforme ilustrado na Figura 37. Essa estrutura de treinamento é mais simples, comparado às outras estruturas do estado da arte, como no caso a estrutura de treinamento da arquitetura Q-Net. No treinamento, são utilizadas janelas de imagens, na forma $\{J_a, J_p, J_n\}$, que representa uma janela âncora J_a, uma janela positiva J_p e uma negativa J_n. A janela positiva consiste em uma região correspondente à janela âncora, já a negativa é constituída de uma região não-correspondente à âncora. Nos experimentos, as janelas J_a são janelas do espectro visível, J_p são do infravermelho e J_n são selecionadas aleatoriamente do espectro visível ou infravermelho, como indicado no trabalho de Aguilera et al. (2017), por ser a maneira mais apropriada para imagens VIS/IV.

A função de perda utilizada consiste na função definida no trabalho de Balntas et al. (2016b), conforme a Equação (14):

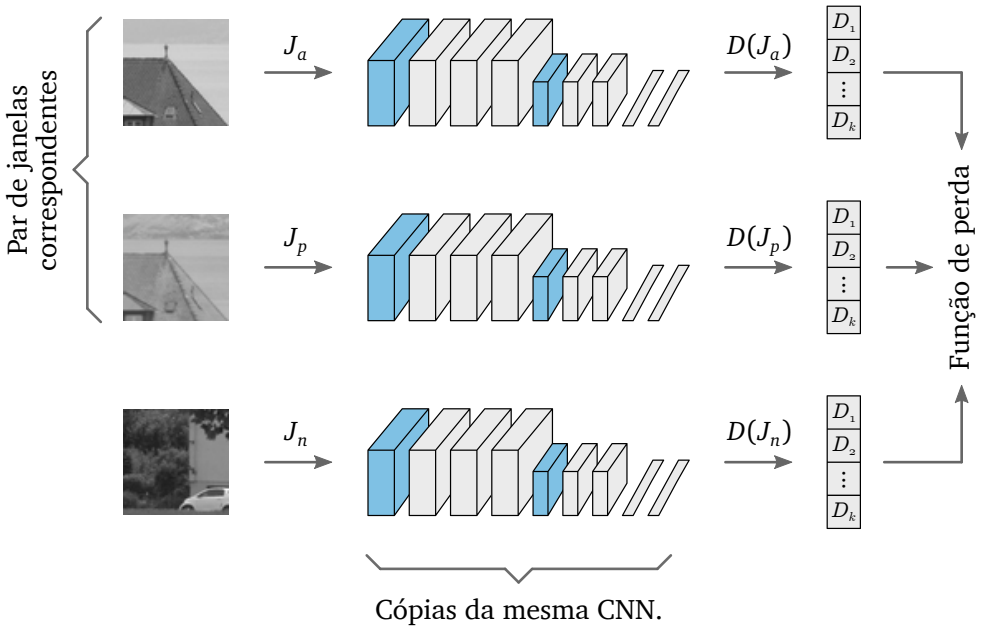


Figura 37 – Estrutura utilizada para o treinamento da rede.

[3] O método Gradiente Descendente Estocástico é inspirado no algoritmo Gradiente Descendente, no entanto, ele utiliza apenas amostras dos dados em vez de utilizar todos os dados. Essa amostragem é realizada de forma aleatória (BOTTOU, 2010).

$$L(\delta_p, \delta_n) = \max(0, \mu + \delta_p - \delta_n), \tag{14}$$

em que μ representa uma constante de margem ($\mu = 2$), δ_p corresponde à norma L_2 entre os descritores do par de janelas J_a e J_p, $\delta_p = \left\|D(J_a) - D(J_p)\right\|_2$ e δ_n corresponde à norma L_2 entre os descritores do par de janelas J_a e J_n, $\delta_n = \|D(J_a) - D(J_n)\|_2$.

Todas as janelas são embaralhadas de forma aleatória no início de cada época[4] (em um total de 100 épocas), e os valores de intensidade dos pixels de cada janela foram normalizados (média 0 e desvio padrão 1). Assim como em Aguilera et al. (2017), as janelas foram divididas em dois grupos, em que 80% foram utilizadas para o treinamento e 20% das janelas foram separadas para teste da rede. Todas as janelas utilizadas foram de tamanho 64×64.

Cada rede foi treinada 10 vezes para suprimir os efeitos de randomização na inicialização. Nenhum ajuste de parâmetros foi utilizado, todos os parâmetros utilizados foram os mesmos fornecidos pelos autores. Os valores das sementes dos geradores de números aleatórios foram fixos para garantir a reprodutibilidade dos resultados. Portanto, caso os experimentos fossem executados novamente, os mesmos resultados seriam obtidos.

Assim como no trabalho de Balntas et al. (2016b), este trabalho não utilizou nenhuma técnica para aumentar a quantidade de dados, como a adição cópias modificadas de dados já existentes (*data augmentation*). O objetivo dessa decisão foi tornar o processo de treinamento mais simples e demonstrar que as melhorias no resultado provém do método proposto e não do maior número de janelas utilizadas para o treinamento.

É importante ressaltar que foram tomadas precauções meticulosas para prevenir o sobreajuste dos dados (do inglês, *overfitting*), um problema comum em aprendizado de máquina em que o modelo se torna excessivamente adaptado aos dados de treinamento, comprometendo sua generalização em dados não vistos. Seguindo a metodologia do trabalho de Balntas et al. (2016b), a rede foi várias vezes treinada minimizando os efeitos da randomização, tanto para replicar fielmente o protocolo anterior quanto para garantir uma comparação mais justa. Adicionalmente, ao manter estrita separação entre os conjuntos de treinamento e teste, a metodologia reforça ainda mais a robustez e a generalização da rede, minimizando assim os riscos associados ao sobreajuste dos dados.

Em cada base de dados utilizada, o modelo foi treinado e avaliado individualmente. Isso significa que, para cada conjunto de dados, o treinamento e a subsequente avaliação foram realizados exclusivamente dentro dessa mesma base. Em nenhum momento houve a prática de treinar o modelo em uma base e testá-lo em outra diferente, garantindo assim a integridade e a especificidade dos resultados obtidos para cada conjunto de dados em questão.

[4] Uma época consiste em um ciclo no processo de treinamento de uma CNN, em que todo o conjunto de dados é apresentado uma única vez à rede para ajuste dos parâmetros (BUDUMA, 2017).

Neste ponto, é fundamental compreender a distinção entre treinar a rede neural e sua aplicação subsequente na descrição de características locais para produzir descritores. Durante o processo de treinamento, emprega-se uma estrutura tripla, como ilustrado anteriormente na Figura 37. Esta estrutura tripla é utilizada para garantir que a rede neural aprenda a diferenciar características distintas e forme uma representação robusta. No entanto, após o treinamento ser concluído, para a descrição de características locais em novas imagens (aplicação prática da rede), usa-se apenas uma única rede neural dessa estrutura tripla previamente treinada. Esta rede neural isolada é suficiente para capturar e descrever características locais com eficácia, uma vez que já internalizou os padrões necessários durante a fase de treinamento.

4.4 Metodologia de avaliação

A seguir, são explicados os passos e o processo de avaliação utilizado nos experimentos. A metodologia para avaliação utilizada aqui é inspirada nos trabalhos de Mouats et al. (2018) e Mikolajczyk e Schmid (2005), comumente empregada na literatura para avaliação de métodos descritores.

Os métodos do estado da arte, selecionados para a avaliação e comparação (*baselines*), foram aqueles que permitiam a utilização de seus descritores em processos convencionais de correspondência entre imagens. Nesse sentido, foram avaliados os métodos PN-Net, TFeat, Q-Net, bem como o método aqui proposto, intitulado MF-Net. Ademais, o método proposto não foi comparado aos métodos baseados em modelagem manual, dado que os experimentos visam mostrar a eficácia do método proposto em relação aos métodos baseados em CNN do estado da arte. Nesse ponto, cabe lembrar que, conforme apresentado no trabalho de Schonberger et al. (2017), ainda há uma discrepância sobre a eficácia entre tais abordagens para descrição de características locais.

A Figura 38 apresenta uma visão geral das etapas utilizadas para a avaliação dos métodos descritores. Nos experimentos foram utilizadas bases de dados, que serão detalhadas na próxima subseção, contendo pares de imagens de uma mesma cena, compostos por uma imagem do espectro visível e uma imagem do espectro infravermelho.

Conforme ilustrado na Figura 38, realizou-se a avaliação dos métodos descritores em um processo convencional de correspondência entre imagens, composto por três etapas principais, quais sejam: detecção, descrição e correspondência de características locais (MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; MIKOLAJCZYK; SCHMID, 2005).

A primeira etapa consistiu em extrair as características locais pela utilização de um método detector. Neste caso, foi adotado o método detector FAST, por ser indicado na literatura como um método de maior eficácia para imagens VIS/IV e de maior velocidade de detecção (MOUATS et al., 2018; SALEEM et al., 2017). Nessa etapa, para cada imagem do espectro visível, foram detectados os pontos-chave (centro das características locais)

e projetados estes pontos na imagem do espectro infravermelho, por meio das matrizes de homografia fornecidas pelas bases de dados. Em seguida, foram recortadas janelas de imagens em torno de cada ponto-chave. O tamanho definido das janelas foram de 64 × 64, por ser indicado, pela literatura, um tamanho adequado para descrever características locais em imagens VIS/IV (AGUILERA et al., 2017).

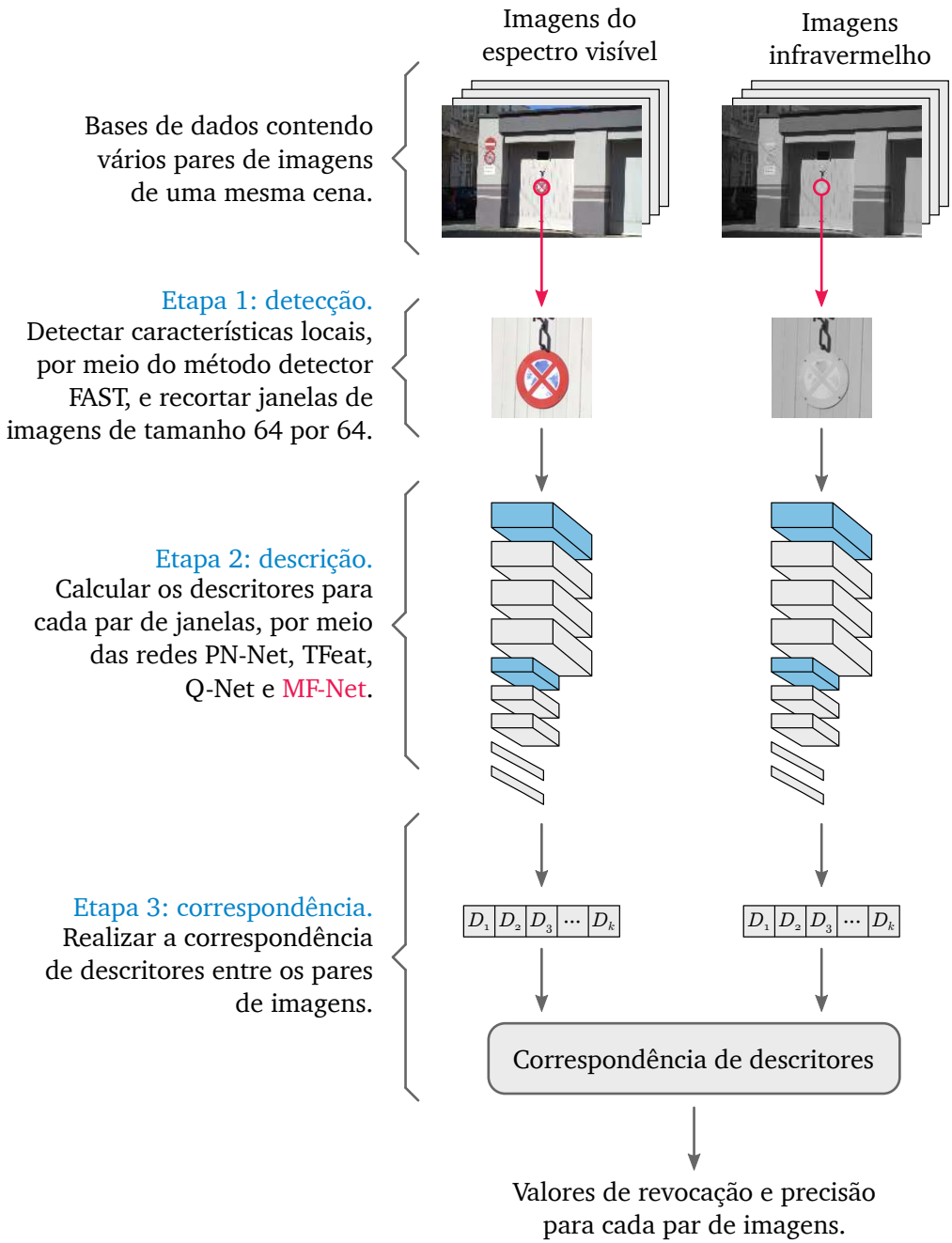


Figura 38 – Metodologia utilizada para avaliação dos métodos descritores baseados em aprendizagem profunda.

Na etapa de descrição, foram calculados os descritores para cada par de janelas de imagens. Nesse sentido, foram realizados experimentos com os métodos PN-Net, TFeat, Q-Net e o método proposto neste trabalho MF-Net. Nessa etapa, é necessária apenas uma CNN para descrever a janela de imagem durante a descrição de características locais, diferentemente da estrutura tripla utilizada durante o processo de treinamento da rede, conforme apresentada anteriormente na Figura 37.

Na última etapa, terceira etapa da Figura 38, é realizada a correspondência de cada descritor sobre o par de imagens por meio da estratégia NNDR, assim como amplamente utilizado na literatura (MOUATS et al., 2018; BALNTAS et al., 2016b; DELLINGER et al., 2015). Nos experimentos, a correspondência é realizada por diversas vezes, variando o valor NNDR (valor de τ, referente à Equação (3)) entre 0,8 e 1,0 em intervalos de 0,05, assim como realizado em alguns trabalhos da literatura[5] (YU; ZHAO, 2020; MA; MA; YU, 2019; FU et al., 2018b). Por fim, para cada valor do limiar, foram calculados valores de revocação e precisão e então produzidas as curvas de precisão por revocação. Também foram produzidas curvas de medida F_1 por valor NNDR, para uma melhor compreensão dos resultados.

As implementações dos métodos do estado da arte, selecionados nos experimentos para a avaliação e comparação, foram fornecidas pelos seus respectivos autores e encontram-se disponíveis publicamente[6]. Os valores dos parâmetros utilizados na execução desses métodos seguem aqueles sugeridos pelos próprios autores. Todos os parâmetros utilizados nos experimentos, bem como suas respectivas descrições, encontram-se no Apêndice E.

Para a implementação do método proposto neste trabalho utilizou-se a linguagem de programação Python com a biblioteca de Visão Computacional OpenCV[7] e a biblioteca de aprendizado de máquina PyTorch[8]. Ambas as bibliotecas são totalmente livres para o uso comercial e acadêmico.

Para os experimentos, utilizou-se uma estação de trabalho com CPU *Intel Core i7-8700* 3,20 GHz com 16 GB de memória RAM, rodando o sistema operacional Ubuntu 18.04 de 64 bits (GNU/Linux versão 4.15.0). Para o treinamento e execução dos métodos, foi utilizada uma GPU *GeForce GTX 1050 Ti* com 4 GB de memória RAM.

[5] Neste caso, poderiam também ser calculados limiares menores que 0,8, no entanto, para a comparação dos métodos não haveria tal necessidade, além de tornar os experimentos demorados. Também, valores próximos a 0,8 são mais adequados para aplicações práticas (LOWE, 2004; HEINLY; DUNN; FRAHM, 2012).

[6] O Apêndice F exibe uma listagem completa das implementações utilizadas nos experimentos.

[7] Página oficial: <https://opencv.org>.

[8] Página oficial: <https://pytorch.org>.

4.4.1 Bases de dados

Para demonstrar a eficácia e as limitações do método proposto, foram realizados experimentos em duas bases de dados distintas. As bases de dados estão disponíveis gratuitamente para garantir a reprodutibilidade dos resultados. As bases de dados são compostas por pares de imagens obtidas de diferentes faixas do espectro eletromagnético. Cada par é composto por uma imagem do espectro visível e uma imagem do espectro infravermelho.

A primeira base de dados, intitulada *Potsdam*, é amplamente utilizada na literatura em trabalhos para classificação e extração de características no contexto do sensoriamento remoto (LIU et al., 2021; MA; MA; YU, 2019; LIU; LI; PAN, 2019; WU et al., 2019; MARMANIS et al., 2018). Esta base, distribuída pela *International Society for Photogrammetry and Remote Sensing*[9], se encontra disponível publicamente[10] e é composta por 38 pares de imagens obtidas do espectro visível e do espectro NIR (infravermelho próximo) da vista aérea da cidade de Potsdam. Todas as imagens possuem o tamanho de 6000 × 6000 pixels e todos os pares de imagens se encontram alinhados espacialmente. A Figura 39 exibe algumas das imagens desta base de dados. Para esta base de dados, foram extraídos 624.139 pares de janelas de imagens.

A segunda base de dados, denominada *EPFL*, proposta por Brown e Susstrunk (2011), é comumente utilizada na literatura (SALEEM; BAIS, 2020; LIU; LI; PAN, 2019; FU et al., 2018a; SALEEM et al., 2017; AGUILERA et al., 2017) e se encontra disponível publicamente[11]. Esta base consiste em um conjunto de 477 pares de imagens obtidas do espectro visível e do espectro NIR, distribuídas em 9 categorias, quais sejam: *Country* (52 pares), *Field* (51

(a) Par de imagens *3_12*.

(b) Par de imagens *4_12*.

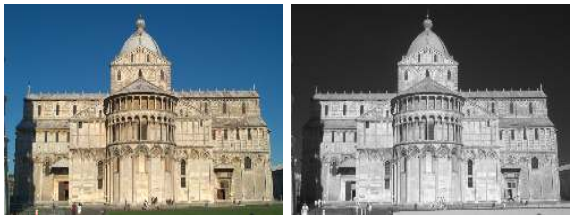
(c) Par de imagens *5_13*.

(d) Par de imagens *6_11*.

Figura 39 – Exemplos de pares de imagens da base de dados Potsdam.

[9] Página oficial: <https://www.isprs.org>.
[10] Disponível em: <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>.

(a) Par de imagens da categoria *water*.



(b) Par de imagens da categoria *oldbuilding*.

(c) Par de imagens da categoria *field*.

(d) Par de imagens da categoria *street*.

Figura 40 – Exemplos de pares de imagens da base de dados EPFL.

pares), *Forest* (53 pares), *Indoor* (56 pares), *Mountain* (55 pares), *Old Building* (51 pares), *Street* (50 pares), *Urban* (58 pares) e *Water* (51 pares). Tais categorias cobrem cenas de países, campos, florestas, ambientes internos, montanhas, edifícios, ruas, áreas urbanas e cenas contendo água. As imagens possuem diferentes tamanhos e todos os pares de imagens se encontram alinhados espacialmente. A Figura 40 exibe algumas das imagens desta base de dados. Para esta base de dados, foram extraídos 1.664.000 pares de janelas de imagens.

4.4.2 Procedimento estatístico

Nesta seção, é apresentado o método de análise estatística aplicado nos experimentos. Esse procedimento estatístico teve o objetivo de verificar a significância dos resultados obtidos durante a comparação dos métodos avaliados.

Nos experimentos deste trabalho utilizou-se o teste t de Student unilateral superior (MONTGOMERY, 2017). Como neste trabalho se desejou comparar diferentes métodos (algoritmos), o teste t foi utilizado para a média das diferenças pareadas entre dois métodos, comparando o método proposto com os demais (um contra todos, dois a dois).

É importante salientar que os testes para a média das diferenças pareadas são, geralmente, mais simples e requerem menos pré-requisitos. Esses testes também produzem resultados interpretáveis com mais facilidade e são consideravelmente mais simples de analisar. Nesse sentido, a análise de variância (ANOVA, do inglês *Analysis of Variance*) não é estritamente necessária ao realizar comparação de vários algoritmos, apesar de comumente utilizada na literatura (CAMPELO; WANNER, 2020).

[11] Disponível em: <https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html>.

No teste estatístico aplicado nos experimentos, a hipótese nula H_0 representa se a média das diferenças pareadas entre dois métodos μ_Δ é igual a 0, significando não haver diferença estatística no resultado entre os dois métodos. Já a hipótese alternativa H_1 representa se essa média é maior que 0, indicando que o resultado do método proposto é superior, conforme a Equação (15). Caso o teste estatístico rejeite a hipótese nula H_0, admite-se a alternativa H_1.

$$\begin{cases} H_0 : \mu_\Delta = 0, \\ H_1 : \mu_\Delta > 0. \end{cases} \tag{15}$$

em que $\mu_\Delta = \mu_{\text{método}_1} - \mu_{\text{método}_k}$, $k \in \{2,\ 3, \ldots,\ n\}$, em que n representa a quantidade de métodos avaliados.

Para todos os testes estatísticos, utilizou-se um nível de significância de $\alpha = 0{,}05$, e como se trata de múltiplas comparações, corrigiu-se esse valor para evitar a probabilidade de se obter uma conclusão falsa em uma série de testes de hipótese[12]. Para a correção desse nível de significância, utilizou-se o método de Bonferroni, por se tratar de uma abordagem simples e conservadora. Nesse método, divide-se o valor de α pelo número de comparações realizadas (MONTGOMERY, 2017).

Como pré-requisito desse procedimento estatístico, foi necessário verificar se os valores das diferenças entre as médias se aproximam de uma distribuição normal (CAMPELO; WANNER, 2020). Para essa verificação, utilizou-se o teste de normalidade Shapiro-Wilk.

4.5 Resultados experimentais

A seguir, são apresentados os resultados experimentais sobre a avaliação do método proposto para a descrição de características locais em imagens VIS/IV. Primeiro, na Subseção 4.5.1 são apresentados os resultados da análise exploratória das diferentes arquiteturas produzidas. Por fim, na Subseção 4.5.2 são apresentados os resultados experimentais sobre a comparação do método proposto com os demais métodos do estado da arte no que diz respeito à eficácia e eficiência. Todos os experimentos foram realizados para as duas bases de dados apresentadas previamente, quais sejam: Potsdam e EPFL.

4.5.1 Análise exploratória sobre as diferentes arquiteturas

Conforme explicado na Seção 4.2, para verificar o comportamento da inserção da camada de mapeamento, realizou-se uma análise exploratória incorporando a camada de mapeamento sob diferentes maneiras. Portanto, foram avaliadas três arquiteturas, quais

[12] Procurou-se reduzir a taxa do erro de família (FWER, do inglês *family-wise-error rate*), ou seja, obter erros do *Tipo I*, em que se rejeita uma hipótese nula quando ela é realmente verdadeira (MONTGOMERY, 2017).

sejam: (i) sem camadas de mapeamento, (ii) com apenas uma camada de mapeamento e (iii) com duas camadas de mapeamento.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados Potsdam. A Figura 41 exibe os valores de revocação e precisão, obtidos pelas diferentes arquiteturas nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Por meio deste experimento, observa-se que as arquiteturas (iii) e (ii), com camadas de mapeamento, obtiveram maiores eficácias tanto na precisão quanto na revocação, comparadas à arquitetura (i). A arquitetura (ii), que possui apenas uma camada de mapeamento, obteve um desempenho superior à arquitetura (iii), que possui duas camadas de mapeamento.

A Figura 42 apresenta os valores para a medida F_1 sob diferentes limiares NNDR (conforme a Equação (3) explicada na Seção 2.2). Nessa figura, são apresentados os valores médios e o desvio padrão (barras verticais no gráfico) da medida F_1, calculados para todos os pares de imagens da base de dados Potsdam[13]. Por estes resultados, é possível observar também que a arquitetura com apenas uma camada de mapeamento obteve um desempenho superior, comparada às demais arquiteturas. Adicionalmente, verificou-se para estes valores a significância dos resultados obtidos pelo procedimento estatístico usando o teste t, conforme descrito na Subseção 4.4.2. O resultado desse procedimento indicou que, existem evidências estatísticas de que a arquitetura (ii) é superior na média das medidas F_1, para cada valor NNDR.

Aqui, vale lembrar que um valor NNDR próximo a 1 retorna mais correspondências, mas com maior incidência de correspondências incorretas (conforme explicado na Seção 2.2). Elevando o valor NNDR, aumenta-se a revocação, mas diminui-se a precisão. Como a medida F_1 equilibra essas duas medidas, o gráfico apresenta curvas com formato de "U" invertido,

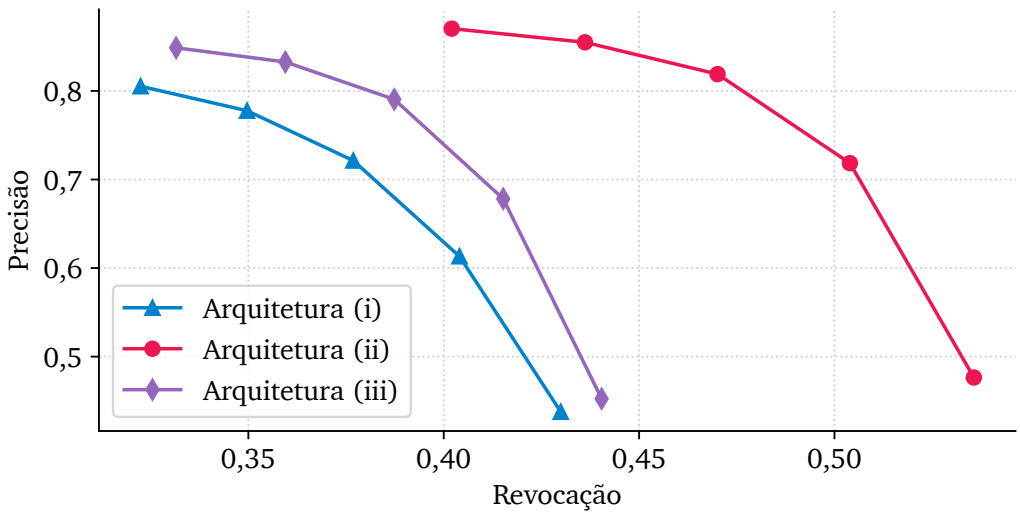


Figura 41 – Curvas de precisão por revocação para diferentes arquiteturas na base de dados Potsdam.

[13] Os valores médios e o desvio padrão da medida F_1 para a base de dados Potsdam estão reportados na Tabela 1 do Apêndice D.

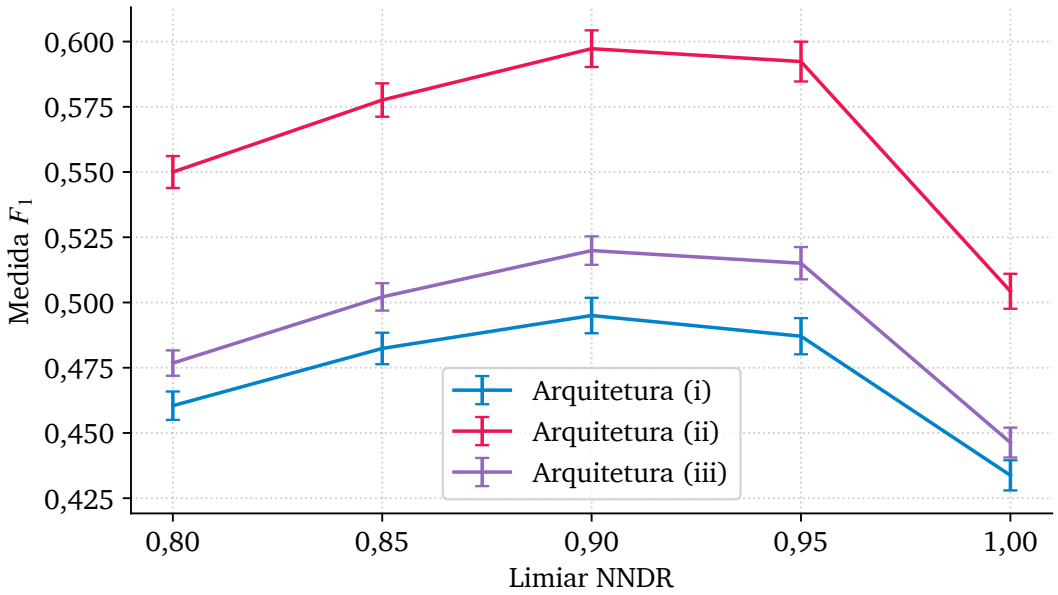


Figura 42 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam.

conforme observado na Figura 42.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados EPFL. A Figura 43 exibe os valores de revocação e precisão, obtidos pelas diferentes arquiteturas nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Por meio deste experimento, observa-se que as arquiteturas (iii) e (ii), com camadas de mapeamento, obtiveram maiores eficácias tanto na precisão quanto na revocação, comparadas à arquitetura (i). A arquitetura (ii), que possui apenas uma camada de mapeamento, obteve um desempenho superior à arquitetura (iii), que possui duas camadas de mapeamento.

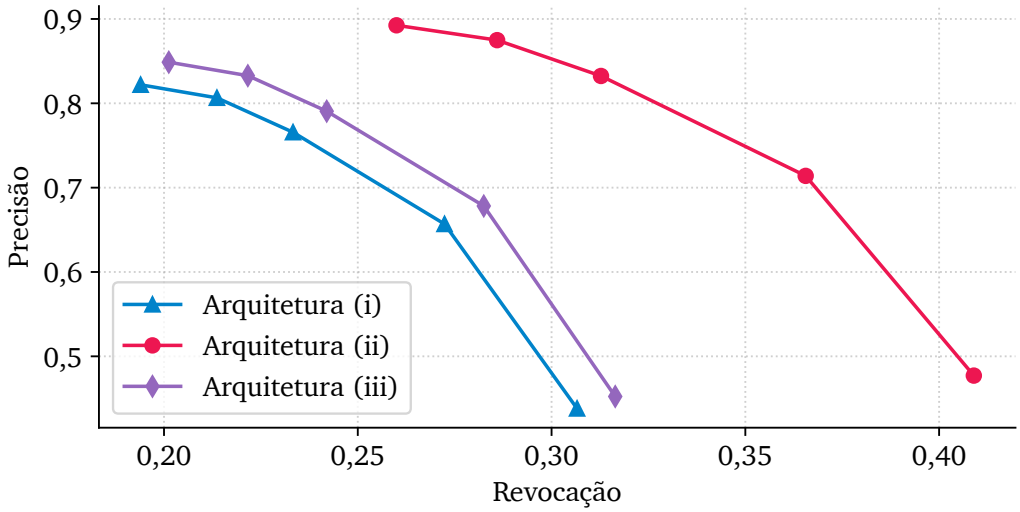


Figura 43 – Curvas de precisão por revocação para diferentes arquiteturas na base de dados EPFL.

A Figura 44 apresenta os valores para a medida F_1 sob diferentes limiares NNDR. Nessa figura, são apresentados os valores médios e o desvio padrão (barras verticais no gráfico) da medida F_1, calculados para todos os pares de imagens da base de dados EPFL[14]. Por estes resultados, é possível observar também que a arquitetura com apenas uma camada de mapeamento obteve um desempenho superior, comparada às demais arquiteturas. Também verificou-se para estes valores a significância dos resultados obtidos pelo procedimento estatístico usando o teste t. O resultado desse procedimento indicou que, existem evidências estatísticas de que a arquitetura (ii) é superior na média das medidas F_1, para cada valor NNDR.

Os experimentos apresentados nesta seção mostraram que a arquitetura com apenas uma camada de mapeamento possui maior eficácia para ambas as bases de dados. Nesse sentido, a inserção da camada de mapeamento no início da arquitetura foi responsável por obter um maior ganho de eficácia, comparada às outras arquiteturas. No entanto, a rede não se tornou mais eficaz ao adicionar mais uma camada de mapeamento. O comportamento desse resultado e suas implicações serão discutidas ao final deste capítulo, na discussão dos resultados (Seção 4.6).

Com base na eficácia da arquitetura com uma camada de mapeamento, esta arquitetura foi selecionada para os próximos experimentos, a qual foi intitulada MF-Net (*Multispectral Features Network*), para compará-la aos métodos do estado da arte que utilizam CNN para descrição de características locais.

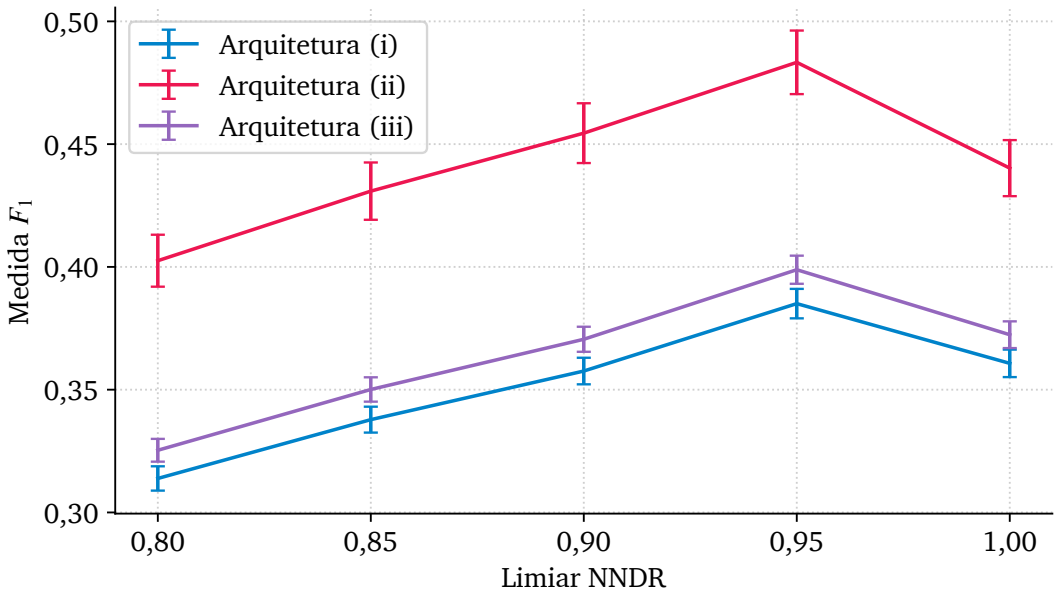


Figura 44 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL.

[14] Os valores médios e o desvio padrão da medida F_1 para a base de dados EPFL estão reportados na Tabela 2 do Apêndice D.

4.5.2 Avaliação de desempenho da arquitetura MF-Net

A arquitetura MF-Net e os demais métodos utilizados como referências (do inglês, *baselines*) foram avaliados em cada base de dados separadamente, com o intuito de se obter uma melhor compreensão dos resultados dos métodos ao utilizar diferentes escopos de aplicação.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados Potsdam. A Figura 45 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Neste resultado, observa-se que as arquiteturas TFeat e PN-Net obtiveram eficácia menor tanto na precisão quanto na revocação. A arquitetura MF-Net obteve maiores valores de revocação e precisão em relação aos demais métodos, incluindo a arquitetura Q-Net, que também é projetada para imagens VIS/IV.

A Figura 46 apresenta os valores para a medida F_1, sob diferentes limiares NNDR, dos experimentos realizados na base de dados Potsdam. Nessa figura, são apresentados os valores médios e o desvio padrão (barras verticais no gráfico) da medida F_1, calculados para todos os pares de imagens da base de dados Potsdam[15]. Estes resultados mostram que o método proposto MF-Net obteve um desempenho superior, comparado aos demais métodos. Adicionalmente, verificou-se para estes valores a significância dos resultados obtidos pelo procedimento estatístico usando o teste t, e o resultado obtido indicou haver evidências estatísticas de que os valores das medidas F_1 da arquitetura MF-Net são estatisticamente superiores comparados às demais médias.

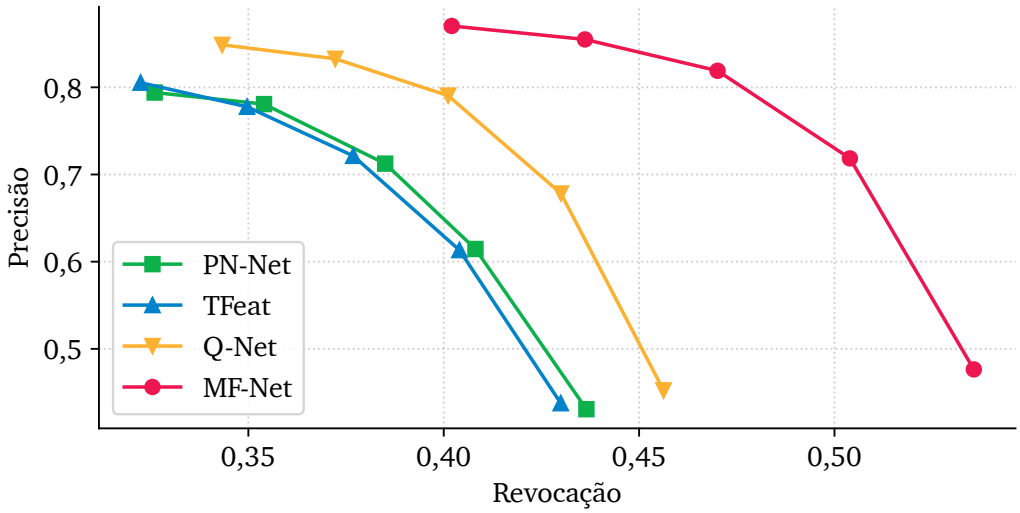


Figura 45 – Curvas de precisão por revocação para a base de dados Potsdam.

[15] Os valores médios e o desvio padrão da medida F_1 para a base de dados Potsdam estão reportados na Tabela 3 do Apêndice D.

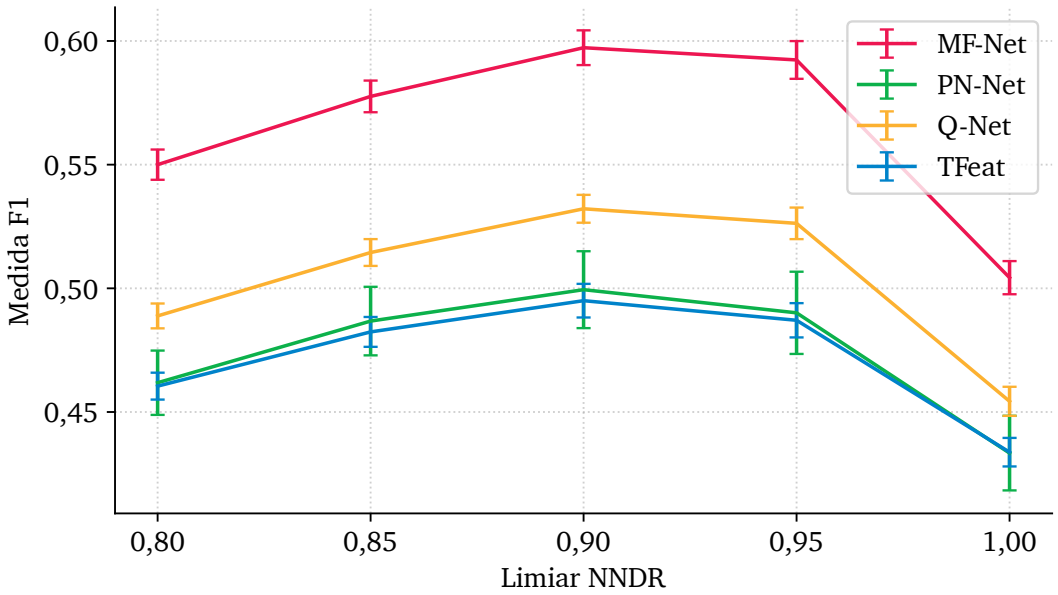


Figura 46 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados EPFL. A Figura 47 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Neste resultado, observa-se que as arquiteturas TFeat e PN-Net obtiveram eficácia menor tanto na precisão quanto na revocação. A arquitetura MF-Net obteve maiores valores de revocação e precisão em relação aos demais métodos, incluindo a arquitetura Q-Net, que também é projetada para imagens VIS/IV.

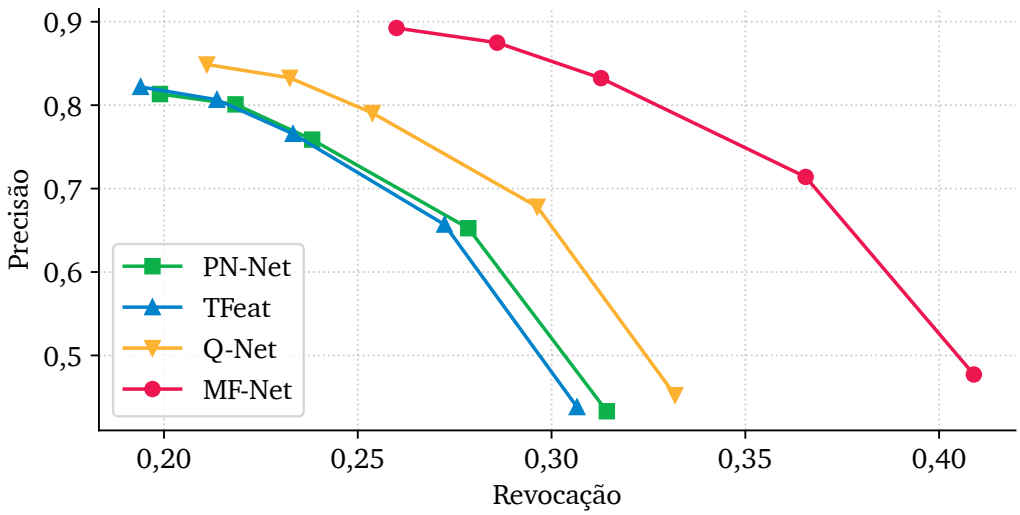


Figura 47 – Curvas de precisão por revocação para a base de dados EPFL.

A Figura 48 apresenta os valores para a medida F_1, sob diferentes limiares NNDR, dos experimentos realizados na base de dados EPFL. Nessa figura, são apresentados os valores médios e o desvio padrão (barras verticais no gráfico) da medida F_1, calculados para todos os pares de imagens da base de dados EPFL[16]. Estes resultados mostram que o método proposto MF-Net obteve um desempenho superior, comparado aos demais métodos. Também verificou-se para estes valores a significância dos resultados obtidos pelo procedimento estatístico usando o teste t, e o resultado obtido indicou haver evidências estatísticas de que os valores das medidas F_1 da arquitetura MF-Net são estatisticamente superiores comparados às demais médias.

A seguir, são apresentados os resultados sobre um estudo de tempo de processamento para os métodos avaliados, com o objetivo de verificar a eficiência de cada método no que diz respeito ao seu tempo de execução em milissegundos. Também foram avaliados o tempo durante o processo de treinamento de cada rede.

A Figura 49 exibe o gráfico de valores médios para o tempo de processamento (tempo de relógio) dos métodos avaliados, executados para cada par de imagens em todas as bases de dados. O tempo de processamento, nesse caso, representa o tempo de relógio que cada método levou para concluir a descrição das características locais em um par de imagens. Por este gráfico observa-se que os métodos PN-Net e TFeat obtiveram menores tempos, ambos similares. O método Q-Net obteve um tempo de, cerca de, 5% maior que os métodos PN-Net e TFeat. O método proposto, MF-Net, obteve um tempo de processamento ligeiramente superior aos demais métodos, de cerca de 8% maior que o método Q-Net e 12% maior que os métodos PN-Net e TFeat.

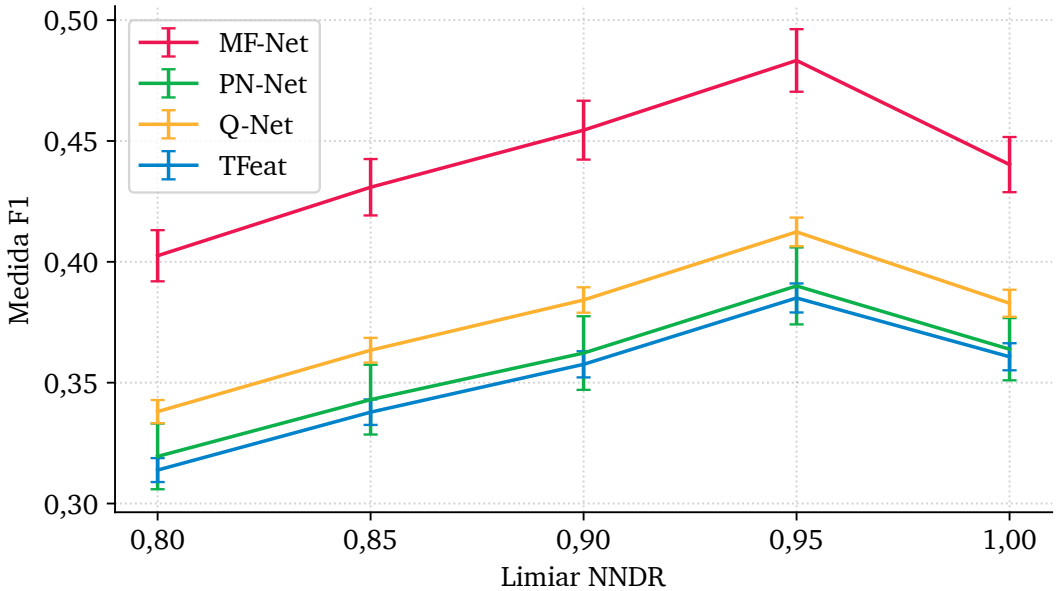


Figura 48 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL.

[16] Os valores médios e o desvio padrão da medida F_1 para a base de dados EPFL estão reportados na Tabela 4 do Apêndice D.

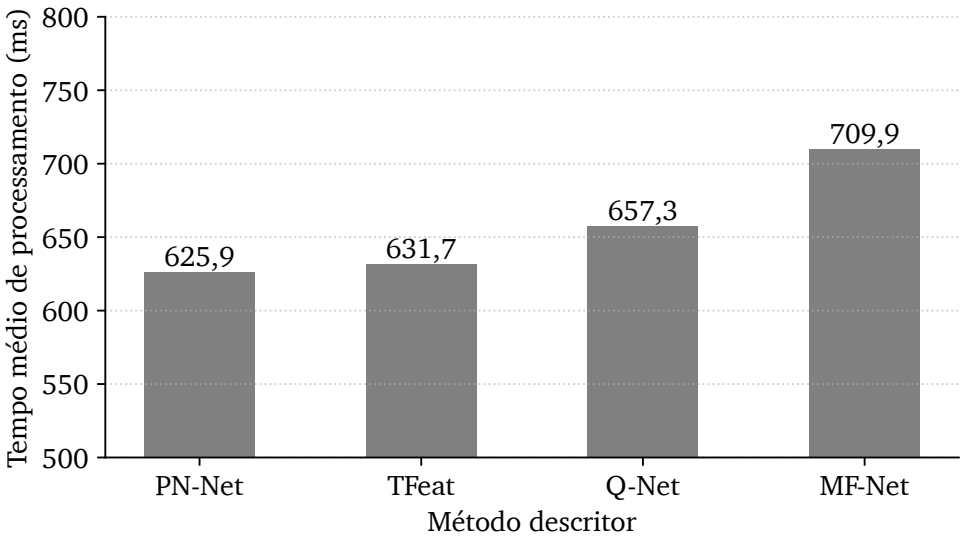


Figura 49 – Valores médios para o tempo de processamento em cada par de imagens.

A Figura 50 exibe o gráfico de valores médios para o tempo de treinamento das redes avaliadas, para cada conjunto de janelas de imagens em todas as bases de dados. O tempo de treinamento, nesse caso, representa o tempo de relógio que cada rede levou para concluir o treinamento do conjunto de janelas apresentadas à rede. Por este gráfico observa-se que os métodos PN-Net e TFeat obtiveram menores tempos, ambos similares. O método Q-Net obteve o maior tempo de, cerca de, 45% maior que os métodos PN-Net e TFeat. O método proposto, MF-Net, obteve um tempo de entre os métodos TFeat e Q-Net, sendo 13% maior que os métodos PN-Net e TFeat e 22% mais rápido que o método Q-Net.

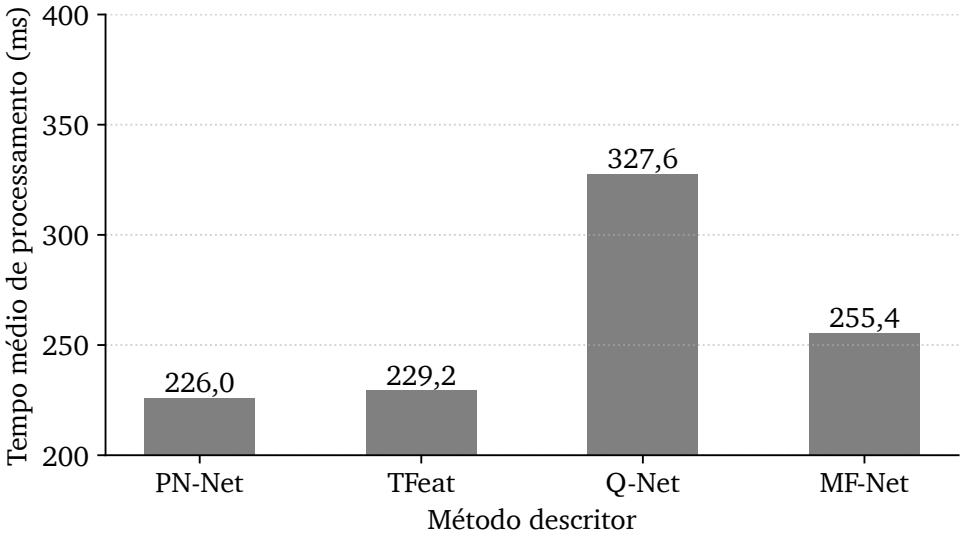


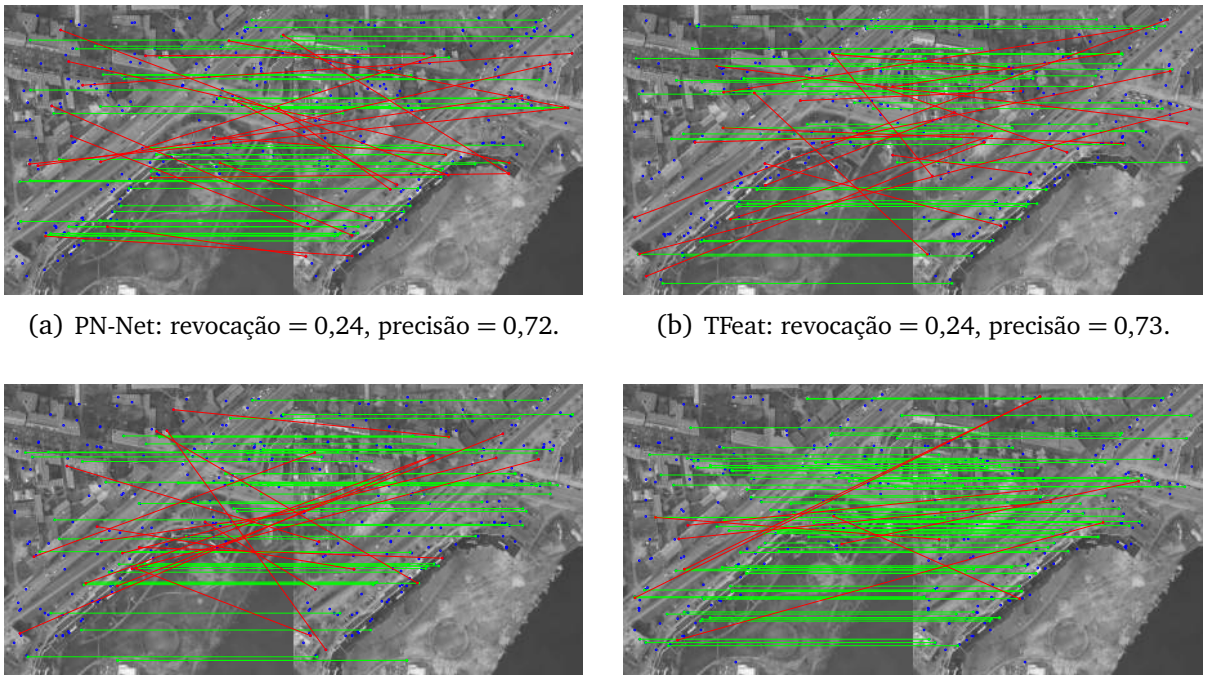
Figura 50 – Valores médios para o tempo de treinamento em cada conjunto de janelas.

Com o objetivo de apresentar os resultados obtidos de maneira qualitativa, a Figura 51 e a Figura 52 apresentam o resultado da correspondência para um par de imagens da base de dados Potsdam e EPFL, respectivamente. Por estas figuras, é possível observar que o método descritor MF-Net apresenta maior taxa de precisão, indicado pela maior quantidade de correspondências corretas (linhas verdes) comparadas às incorretas (linhas vermelhas), em relação aos demais métodos avaliados. Por esses resultados, é possível observar também que ambos os métodos apresentam baixa revocação, pois grande parte das características locais detectadas não foram correspondidas (correspondências reais não retornadas).

4.6 Discussão

Por meio dos resultados obtidos na análise exploratória, é possível observar a efetividade da camada de mapeamento ao incorporá-la a uma CNN do estado da arte para descrever características locais. Uma arquitetura com apenas uma camada de mapeamento obteve maior eficácia para ambas as bases de dados. Esse resultado sugere que a inserção da camada de mapeamento no início da arquitetura foi responsável por obter um maior ganho de eficácia, comparada às outras arquiteturas.

Um comportamento observado nos experimentos foi de que adicionar mais uma camada de mapeamento não tornou a rede mais eficaz, o que pode ser justificado com base na sua construção. O objetivo dessa camada é de possibilitar que imagens obtidas de diferentes faixas do espectro eletromagnético tenham uma representação similar na



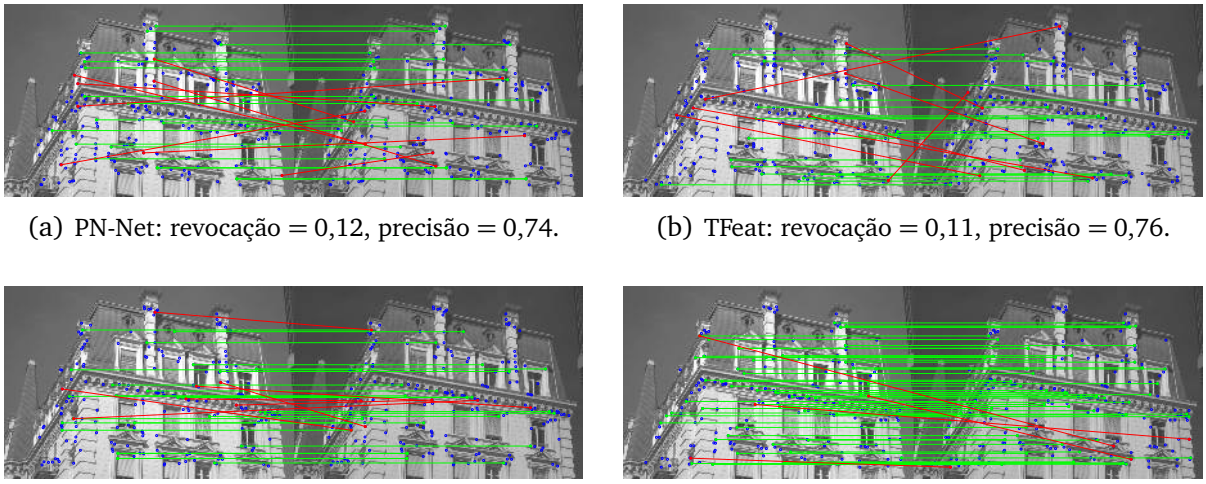
(a) PN-Net: revocação = 0,24, precisão = 0,72.

(b) TFeat: revocação = 0,24, precisão = 0,73.

(c) Q-Net: revocação = 0,26, precisão = 0,78.

(d) MF-Net: revocação = 0,44, precisão = 0,92.

Figura 51 – Resultados qualitativos para um par de images da base Potsdam.



(a) PN-Net: revocação = 0,12, precisão = 0,74.

(b) TFeat: revocação = 0,11, precisão = 0,76.

(c) Q-Net: revocação = 0,13, precisão = 0,79.

(d) MF-Net: revocação = 0,29, precisão = 0,94.

Figura 52 – Resultados qualitativos para um par de images da base EPFL.

rede, portanto apenas uma camada de mapeamento no início da rede seria suficiente para concretizar tal objetivo. Considerando a perda de informações na camada de mapeamento (como explicado na Seção 4.1), quando é inserida mais uma camada deste tipo, esta outra realizaria uma filtragem nos dados, de tal forma que o descritor final se torna menos distintivo e pode reduzir o desempenho da rede para descrição de características locais.

A eficácia obtida da arquitetura com camada de mapeamento, nos resultados obtidos na análise exploratória, corrobora o pressuposto de que a utilização de informações de bordas das imagens, no contexto das imagens VIS/IV, é uma decisão importante, como apontado nos trabalhos de Fu et al. (2018a) e Liu, Li e Pan (2019). Pois, apesar das variações consideráveis de intensidade entre as imagens VIS/IV, as aparências das formas dos objetos tendem a permanecer preservadas em ambas as imagens (JIANG et al., 2021; WU; ZHU, 2021; FANG et al., 2019).

Em relação aos resultados da avaliação de desempenho da arquitetura MF-Net, o método proposto possui maior eficácia quanto aos demais métodos avaliados, para ambas as bases de dados. O método proposto supera o método Q-Net, que também foi projetado para descrever imagens VIS/IV, em eficácia. Também nessa avaliação, conforme esperado, as arquiteturas PN-Net e TFeat apresentaram resultados parecidos, dado que tais arquiteturas são bastante parecidas, cujas diferenças são, de maneira geral, parâmetros no processo de treinamento, conforme explicado no Capítulo 3.

Ainda nessa avaliação, os resultados obtidos em ambas as bases de dados corroboram a observação feita no estudo de Schonberger et al. (2017), em que os autores mostraram que os métodos descritores, baseados em CNN, apresentam variações significativas em eficácia para diferentes bases de dados. Nesse sentido, os resultados de eficácia para a base de dados Potsdam são melhores, comparados aos resultados obtidos na base de dados EPFL para todos os métodos avaliados. Vale ressaltar que, a base de dados EPFL é mais desafiadora por

abranger categorias de cenas distintas, diferentemente da base de dados Potsdam, composta por um mesmo cenário em todas as imagens. Apesar das discrepâncias nos resultados entre as bases de dados, não é possível supor que os métodos avaliados tenham uma aplicação mais apropriada para um determinado cenário. Nesse caso, poderia haver outras circunstâncias que afetam os valores de revocação e precisão, incluindo diversidades de resolução e presença de ruído nas imagens (que não foram avaliadas neste trabalho).

Pelos resultados dos experimentos qualitativos é possível observar a precisão do método proposto comparado aos demais, por apresentar mais correspondências corretas. Nesses resultados também é possível observar baixas taxas de revocação para ambos os métodos, pois várias características detectadas não foram correspondidas, ou, em outras palavras, várias correspondências reais não foram retornadas. Em problemas práticos, no entanto, essa questão não é um fator crítico, visto que é mais importante que as correspondências sejam precisas (SCHONBERGER et al., 2017), indicando que as localizações das características foram determinadas de forma estável, comparada à quantidade de correspondências obtidas.

Em relação ao tempo de processamento, o método apresentou um aumento em relação aos demais métodos, cerca de 12% maior. Esse resultado é condizente com a construção do método, visto que a arquitetura do método proposto introduz uma camada adicional. No entanto, o método proposto é mais rápido que o método Q-Net para realizar o treinamento da rede (22% mais rápido). Cabe lembrar aqui que, o método Q-Net utiliza uma estrutura mais complexa para o treinamento, composta por quatro cópias de uma mesma CNN, durante o processo de treinamento. Enquanto os métodos PN-Net, TFeat e MF-Net utilizam uma estrutura mais simples, composta por uma estrutura tripla. Apesar do método MF-Net apresentar um tempo de processamento ligeiramente maior para descrever características em imagens VIS/IV ele possui uma eficácia maior e sua estrutura de treinamento é mais simples e mais rápida que a do método Q-Net, também projetado para imagens VIS/IV.

Em relação ao processo de avaliação, pode-se evidenciar a limitação de que a avaliação dos métodos se deu no espectro do infravermelho próximo (NIR) e no espectro visível. Nesse sentido, não se pode afirmar a eficácia do método proposto para toda a faixa do espectro infravermelho apenas por meio dos experimentos realizados, sendo necessário mais estudos. No entanto, como a construção do método se baseia na estratégia de se utilizar informações de bordas das imagens, como nos trabalhos de Li, Hu e Ai (2020), Ma, Ma e Yu (2019) e Fu et al. (2018b), especula-se que o método também tenha eficácia em outras faixas do espectro infravermelho. Tal hipótese se manifesta dado que trabalhos na literatura obtiveram sucesso na descrição de características locais em outras faixas do espectro infravermelho utilizando uma estratégia similar, porém por métodos que não utilizam a abordagem baseada em aprendizagem profunda.

Algumas limitações também podem ser listadas no que se refere ao método proposto. Primeiramente, as imagens de resposta aos filtros Log-Gabor para a obtenção de informações de bordas, em alguns casos, podem ter resultados diferentes entre imagens obtidas de

distintas faixas do espectro eletromagnético. Como exemplo, tanto no espectro visível quanto no infravermelho podem ocorrer casos em que as bordas das imagens são ambíguas devido às diferenças na reflexão da luz para diferentes faixas do espectro eletromagnético. Esse fenômeno, portanto, pode influenciar na extração de informações de bordas entre imagens de diferentes espectros. Outra limitação que pode ser inferida é que, assim como os métodos PN-Net, TFeat e Q-Net, o método descritor MF-Net, por sua construção, não trata diferentes escalas de características locais, conforme feito em alguns métodos projetados para o espectro visível, como os algoritmos SIFT e SURF. No entanto, cabe ainda avaliar o comportamento do método proposto nessas situações.

No que diz respeito à acurácia obtida pelo MF-Net, acredita-se que, ao incorporar a camada de mapeamento na arquitetura Q-Net (projetada para imagens VIS/IV), se obtenham resultados ainda melhores. Também, acredita-se que esta camada possa ser útil em abordagens que utilizam CNNs para processar outras categorias de dados que possuam uma relação não-linear ou não-monotônica, dado que esta camada é relativamente simples de ser incorporada a outras arquiteturas.

Por fim, alguns estudos futuros podem ser realizados com base nos resultados obtidos, como a avaliação do método proposto em imagens obtidas de outras faixas do espectro eletromagnético. Também, a avaliação em relação a desafios referentes às questões geométricas (variações em ponto de vista, escala e rotação) como também às questões fotométricas (variações de iluminação da cena, exposição da imagem, borrões ou ruídos). Adicionalmente, pode-se avaliar a camada de mapeamento proposta neste trabalho para outras categorias de imagens, como mapas de profundidade.

Capítulo 5

Scale-Orientation Robust Infrared Features (SORIF)

Conforme observado no capítulo de Trabalhos Relacionados, os métodos baseados em modelagem manual continuam sendo amplamente investigados. No contexto das imagens VIS/IV, uma das limitações dos métodos baseados em modelagem manual é a dependência predominante dos algoritmos SIFT e SURF. Esses métodos são conhecidos por serem computacionalmente custosos e complexos de implementar, o que estimula a busca por técnicas mais eficientes e simples (WANG; SHI; CAO, 2019; BARAJAS-GARCÍA; SOLORZA-CALDERÓN; GUTIÉRREZ-LÓPEZ, 2019).

É importante ressaltar que os métodos propostos para a descrição de características em imagens VIS/IV não abordam adequadamente os desafios geométricos, como variações de escala e rotação (KIM et al., 2020). Métodos mais recentes baseados em modelagem manual, como HODM, RIFT e HGEO, também não têm abordado essas questões de forma direta. Nesse contexto, um dos objetivos deste trabalho é também desenvolver um método baseado em modelagem manual que não se apoie nos algoritmos SIFT e SURF e possa lidar com os desafios geométricos em imagens VIS/IV.

Este capítulo aborda as principais etapas relacionadas à elaboração e avaliação do método proposto, intitulado *Scale-Orientation Robust Infrared Features* (SORIF). A seção subsequente, intitulada Método proposto (Seção 5.1), fornece uma explicação detalhada sobre a construção e o funcionamento desse método. Em seguida, na seção Metodologia de avaliação (Seção 5.2), são apresentados os passos e os critérios utilizados durante os experimentos realizados. Além disso, na seção Resultados experimentais (Seção 5.3), são apresentados os resultados obtidos a partir desses experimentos. Posteriormente, na seção Discussão (Seção 5.4), são analisados os resultados e destacados os principais aspectos do método proposto, bem como suas limitações.

5.1 Método proposto

Conforme mencionado no capítulo de Introdução deste trabalho, um dos principais desafios na descrição de características locais em imagens multiespectrais é a variação significativa das direções dos gradientes de intensidade associados a uma mesma região entre as diferentes faixas do espectro eletromagnético. No entanto, apesar dessas variações,

as características gerais dos objetos, como bordas e texturas, tendem a permanecerem visíveis (JIANG et al., 2021; WU; ZHU, 2021).

Os filtros de Gabor, também conhecidos como wavelets de Gabor, são amplamente utilizados para extrair características de textura e detectar bordas em imagens, principalmente no contexto das imagens VIS/IV (LIU et al., 2021; LI et al., 2021). No método descritor proposto, denominado SORIF, adotaram-se os filtros de Log-Gabor[1] para estimar a orientação da janela em torno do ponto-chave (detalhado na Subseção 5.1.1) e para extrair a textura e as bordas da imagem (detalhado na Subseção 5.1.2), seguindo uma abordagem semelhante ao método descritor MF-Net deste trabalho (apresentado no Capítulo 4).

5.1.1 Cálculo da orientação

No processo de descrição do método SORIF, recorta-se uma janela de tamanho $N \times N$ ao redor do ponto-chave detectado. Após recortada a janela, é estimado o ângulo principal ao qual esta janela poderia estar rotacionada por meio do cálculo de congruência de fase (KOVESI, 1999; KOVESI, 2000), também utilizado em trabalhos recentes, como no trabalho de Ma et al. (2022). A congruência de fase é útil para detectar características e bordas em uma imagem ou então estimar a orientação em partes da imagem (MA et al., 2022). Esta tarefa é realizada por meio dos passos a seguir:

- Decomposição da imagem: É realizada a convolução entre a imagem original e um conjunto de filtros Log-Gabor com diferentes parâmetros de escala e orientação (os detalhes desse conjunto serão explicados mais adiante na Subseção 5.1.2). Cada filtro produz uma imagem contendo números complexos para cada pixel onde o módulo desse número representa a intensidade da resposta ao filtro e o argumento desse número (fase) representa a orientação na imagem original.

- Cálculo da Congruência de Fase: Para cada pixel é calculada a diferença de fase entre a resposta do filtro para aquela orientação e a média das respostas dos filtros para todas as outras orientações. O resultado é uma medida de quão consistente (congruente) é a fase em todas as orientações.

- Construção da imagem de orientação: Uma imagem de orientação é então construída onde o valor de cada pixel é a orientação (ângulo) do filtro que teve a maior congruência de fase para aquele pixel. Isso resulta em uma imagem onde o valor numérico de cada pixel representa a orientação dominante naquele local na imagem original.

- Construção do histograma: Após a obtenção da imagem de orientação, é calculado um histograma sobre esta imagem. O pico do histograma representa o ângulo provável no qual esta janela poderia estar rotacionada.

[1] Para mais detalhes sobre os filtros Log-Gabor, veja o Apêndice A deste trabalho.

Após a detecção do ângulo da janela recortada, esta janela é rotacionada conforme o ângulo estimado e então se inicia o processo de cálculo do descritor.

5.1.2 Cálculo do descritor

A estrutura do método descritor SORIF foi inspirada no método descritor RIFT, conforme será detalhado mais adiante. Ao contrário de alguns métodos disponíveis para imagens VIS/IV, como SAR-SIFT, mencionados no capítulo de Trabalhos Relacionados, o método descritor SORIF distingue-se por não se fundamentar no método SIFT nem empregar o cálculo de gradientes para descrever pontos-chave em imagens de espectros diferentes. Essas características são especialmente relevantes no contexto de imagens VIS/IV, em que a utilização de métodos que não dependem de cálculos de gradientes pode ser mais vantajosa e desejável (HU et al., 2022). Isso torna o método SORIF uma opção promissora para aplicações específicas nesse campo de pesquisa.

Vale destacar que, como exemplificado na introdução deste trabalho (ver Figura 3), um dos grandes desafios na descrição de características locais em imagens multiespectrais é a notável variação das direções dos gradientes de intensidade numa mesma região entre as diferentes faixas do espectro eletromagnético. Por este motivo, o cálculo de gradientes não se mostra adequado para a construção de um método descritor baseado em modelagem manual neste contexto.

A janela recortada em torno do ponto-chave é dividida em 16^2 sub-regiões, sendo cada uma delas uma unidade básica para extrair informações sobre bordas. Então, para cada sub-região, determina-se qual borda é predominante. Esse processo inicial para descrição de características é definido pelo padrão MPEG-7[3], o qual é similarmente utilizado em outros descritores de características, como os métodos descritores EOH, LGHD, MFD e RIFT. O padrão MPEG-7 estipula uma série de descritores que podem ser aplicados para avaliar a semelhança entre imagens ou vídeos (MANJUNATH et al., 2001; WON; PARK; PARK, 2002).

Após determinar as bordas predominantes, os valores das colunas correspondentes a essas bordas no histograma são incrementados. No descritor SORIF, foram estabelecidos quatro filtros Log-Gabor direcionais com as orientações 0, 90, 180 e 270 graus. Esses ângulos foram escolhidos para abranger todos os 360 graus, mas com uma quantidade de filtros menor que o descritor RIFT, que utiliza seis filtros de orientação. A Figura 53 exemplifica a extração da sub-região, a aplicação de cada filtro para determinar a borda predominante e o cálculo do histograma.

[2] A quantidade de sub-regiões estabelecida segue aquela definida no padrão MPEG-7, também utilizado pelos descritores LGHD, MFD e HGEO. O descritor RIFT utiliza 36 sub-regiões, o que produz um descritor maior.

[3] O padrão MPEG-7, define metodologias para descritores, esquemas de descrição e linguagens. Essas metodologias possibilitam descrições estruturadas de dados audiovisuais em várias escalas (abrangendo regiões, imagens e vídeos) e em distintos domínios (como descrição de conteúdo, organização, navegação e interação do usuário) (CHANG; SIKORA; PURL, 2001).

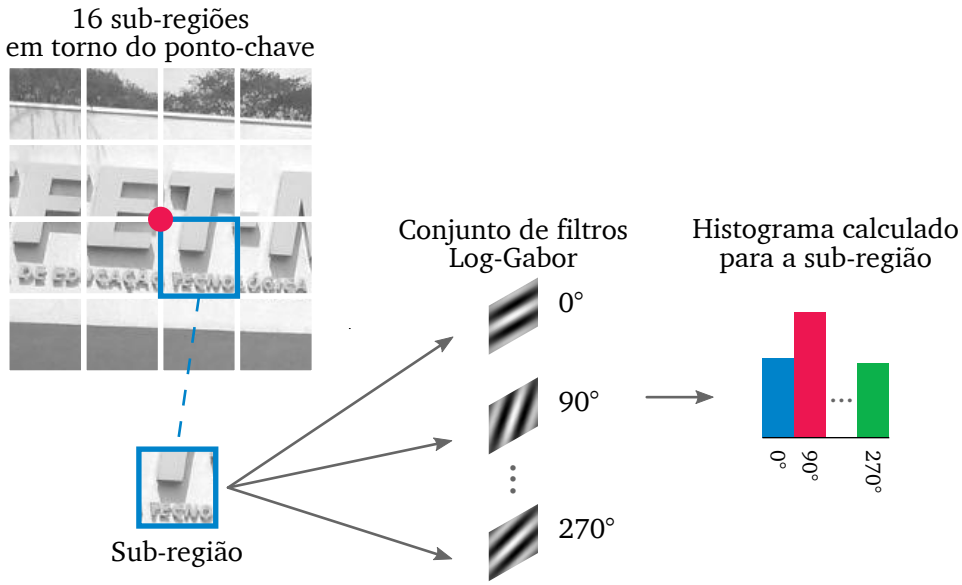


Figura 53 – Extração da sub-região e cálculo do histograma.

Para cada ângulo de orientação, foram estabelecidos 4 filtros de escala (valores de s da Equação (18) variando de 0 a 3), conforme estabelecido no método descritor RIFT. Tal definição foi adotada, pois os autores do método descritor RIFT evidenciaram melhores resultados na descrição de características utilizando essa configuração.

O tamanho da janela do método descritor SORIF foi estabelecido em 96×96 pixels, conforme empregado nos trabalhos citados de Li, Hu e Ai (2020) e Wu e Zhu (2021). A decisão de manter alguns parâmetros semelhantes teve o intuito de possibilitar uma comparação mais justa e controlada entre os métodos descritores. No entanto, é possível ajustar o tamanho da janela por parâmetros, permitindo que o algoritmo possa lidar com diferentes tamanhos de regiões na imagem.

No método SORIF, estabeleceram-se as constantes para os filtros Log-Gabor, conforme proposto no estudo de Walia e Verma (2016). Isso se deve ao fato de tais constantes demonstrarem resultados superiores para extração de informações de texturas quando empregados na descrição de imagens. Os valores definidos foram: $\lambda = 3{,}0$, $\kappa = 2{,}0$ e $\sigma = 0{,}65$ (ver Equação (18) para detalhes de cada constante).

Após aplicar o filtro Log-Gabor em cada sub-região e computar o histograma correspondente, os valores desses histogramas são unidos em um único vetor, como mostrado na Figura 54. Este vetor, que incorpora os histogramas de todas as sub-regiões, é então normalizado, dividindo cada um de seus valores pela norma L_2 do próprio vetor. O vetor normalizado se torna, por fim, o descritor final.

Embora o método descritor proposto neste trabalho compartilhe uma certa similaridade em sua construção com os métodos RIFT e HGEO, é pertinente destacar algumas das principais diferenças entre eles, quais sejam:

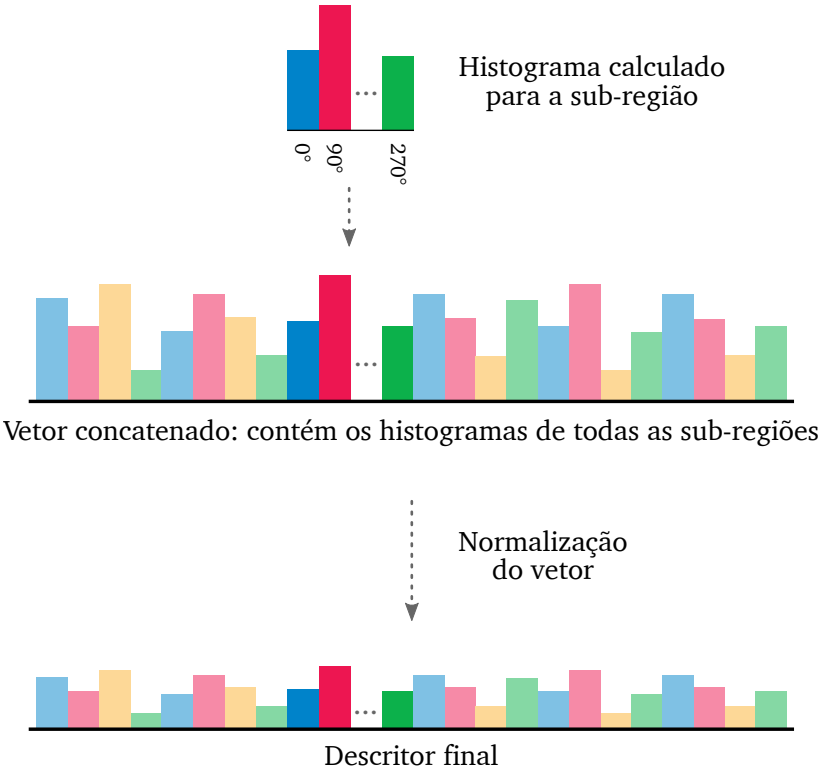


Figura 54 – Montagem do descritor SORIF.

- O algoritmo SORIF implementa uma técnica exclusiva para lidar com robustez à rotação, o que não é feito de forma direta pelos métodos anteriores;
- O método em questão emprega parâmetros para a função Log-Gabor ajustados de maneira aprimorada para a extração das texturas das imagens. Essa questão é importante nesse contexto de descrição de características em imagens VIS/IV, pois as aparências das formas dos objetos nessas imagens tendem a permanecer preservadas, como as texturas e bordas.
- Embora o método descritor SORIF e o método descritor HGEO compartilhem uma estrutura semelhante para realizar a descrição de regiões, seguindo o padrão MPEG-7, eles diferem no tipo de filtro utilizado para a extração de informações de bordas. O método descritor HGEO utiliza filtros Sobel, e consequentemente o método descritor SORIF utiliza filtros Log-Gabor;
- Ambos os métodos descritores SORIF e RIFT empregam filtros Log-Gabor para a extração de informações de bordas nas imagens. No entanto, eles diferem na quantidade de filtros utilizados. O RIFT usa 24 filtros Log-Gabor, enquanto o SORIF usa somente 16 (4 filtros de orientação multiplicados por 4 filtros de escala). Assim, o SORIF aplica um número menor de filtros comparativamente ao RIFT;

- O método em questão utiliza 16 sub-regiões em uma janela de imagem, divididas em uma configuração de 4×4, seguindo a mesma estratégia dos algoritmos MFD e HGEO. Isso resulta, indiretamente, em um descritor de dimensões menores do que aqueles gerados pelos algoritmos HODM e RIFT, que utilizam um número maior de sub-regiões.

Toda a estrutura de descrição do método descritor SORIF produz um descritor final de tamanho 64 (64 valores de ponto flutuantes). Este tamanho é menor, comparado a alguns descritores da literatura, como LGHD (384), MFD (180), HODM (1152) e RIFT (216). É importante ressaltar que um descritor de dimensão menor pode facilitar e tornar mais rápido o processo de correspondência, ao exigir uma comparação mais simples e rápida entre os descritores.

5.2 Metodologia de avaliação

A seguir, são explicados os passos e o processo de avaliação utilizado nos experimentos. A metodologia de avaliação aqui é semelhante à metodologia utilizada na Seção 4.4 deste trabalho e é também inspirada nos trabalhos de Mouats et al. (2018) e Mikolajczyk e Schmid (2005), comumente empregada na literatura para avaliação de métodos descritores.

Foram avaliados os métodos descritores do estado da arte para imagens VIS/IV que utilizam abordagem em modelagem manual, quais sejam: LGHD, MFD, HODM, RIFT e HGEO. Adicionalmente, avaliou-se o algoritmo descritor SIFT, por ser considerado um método de referência encontrado na maioria dos trabalhos apresentados durante a revisão bibliográfica, desenvolvida no Capítulo 3.

A Figura 55 apresenta uma visão geral das etapas utilizadas para a avaliação dos métodos descritores. Nos experimentos, foram utilizadas bases de dados (os detalhes das bases de dados serão explicitados na subseção seguinte) que contêm conjuntos de imagens, cada conjunto composto por uma imagem do espectro visível e uma do espectro infravermelho.

Conforme apresentado na Figura 55, a avaliação dos métodos descritores se deu em um processo convencional de correspondência entre imagens, composto pelas etapas de detecção, descrição e correspondência de características locais (MA et al., 2020; KUPPALA; BANDA; BARIGE, 2020; MIKOLAJCZYK; SCHMID, 2005).

A primeira etapa consistiu em extrair as características locais pela utilização de um método detector. Neste caso, foi adotado o método detector FAST, por ser indicado na literatura como um método de maior eficácia para imagens VIS/IV e de maior velocidade de detecção (MOUATS et al., 2018; SALEEM et al., 2017). Nessa etapa, para cada imagem do espectro visível, foram detectados os pontos-chave (centro das características locais) e projetados estes pontos na imagem do espectro infravermelho.

Na etapa de descrição, foram calculados os descritores para cada ponto-chave previamente detectado. Nesse sentido, foram realizados experimentos com os métodos descritores SIFT, LGHD, MFD, HODM, RIFT, HGEO e o método proposto SORIF.

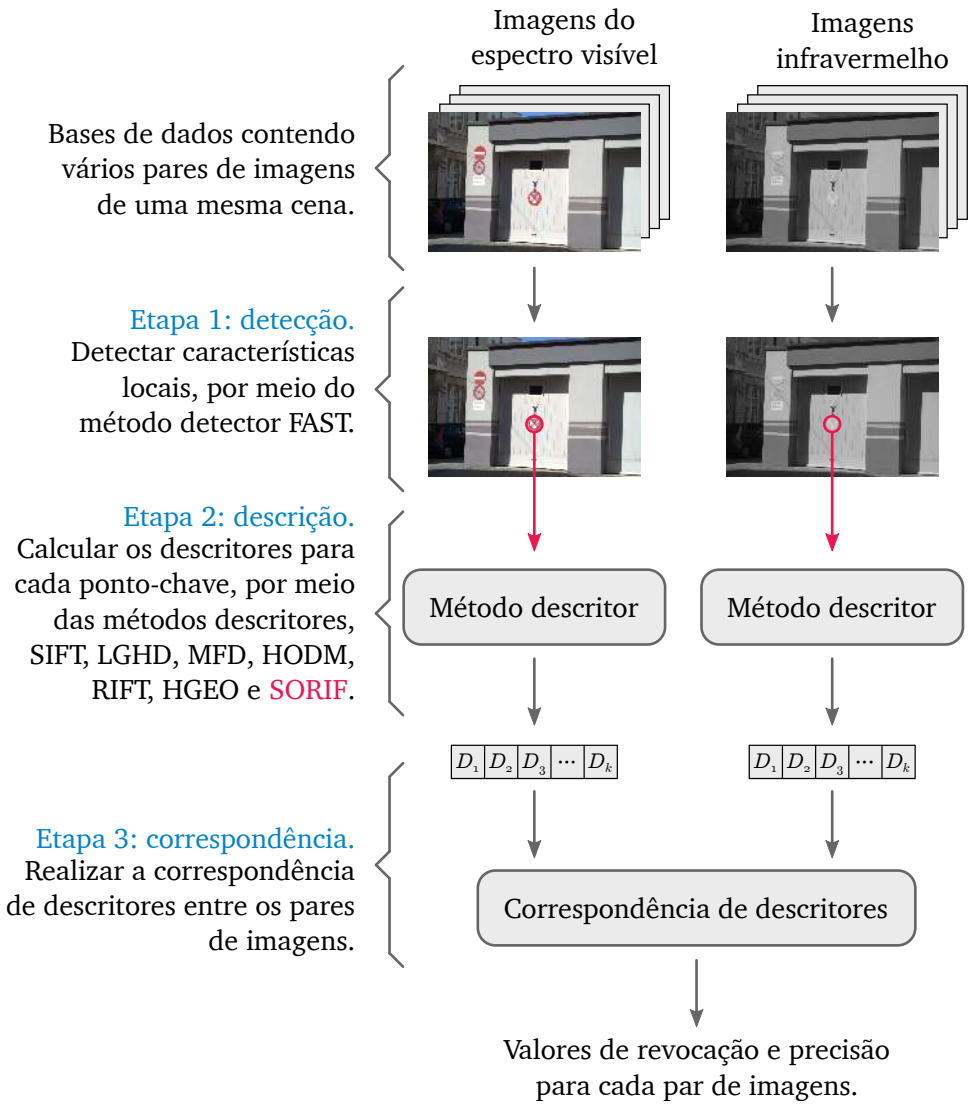


Figura 55 – Metodologia utilizada para avaliação dos métodos descritores baseados em modelagem manual.

Na última etapa, terceira etapa da Figura 38, foi realizada a correspondência de cada descritor sobre o par de imagens por meio da estratégia NNDR, assim como amplamente utilizado na literatura (MOUATS et al., 2018; BALNTAS et al., 2016b; DELLINGER et al., 2015). Nos experimentos, a correspondência é realizada por diversas vezes, variando o valor NNDR (valor de τ, referente à Equação (3)) entre 0,8 e 1,0 em intervalos de 0,05, assim como realizado em alguns trabalhos da literatura (YU; ZHAO, 2020; MA; MA; YU, 2019; FU et al., 2018b). Por fim, para cada valor do limiar, foram calculados valores de revocação e precisão e então produzidas as curvas de precisão por revocação.

As implementações dos métodos, selecionados nos experimentos para a avaliação e comparação, foram fornecidas pelos seus respectivos autores e encontram-se disponíveis publicamente[4]. Os valores dos parâmetros utilizados na execução desses métodos seguem aqueles sugeridos pelos próprios autores. Todos os parâmetros utilizados nos experimentos, bem como suas respectivas descrições, encontram-se no Apêndice E.

[4] O Apêndice F exibe uma listagem completa das implementações utilizadas nos experimentos.

Para a implementação do método proposto neste trabalho utilizou-se a linguagem de programação Python com a biblioteca de Visão Computacional OpenCV[5]. Ambas as bibliotecas são totalmente livres para o uso comercial e acadêmico.

Na análise dos métodos descritores, avaliou-se sua sensibilidade a variações de rotação e de escala das imagens. Ao examinar a sensibilidade a variações de rotação, para cada par de imagens das bases de dados, manteve-se fixa a imagem obtida do espectro visível, enquanto a imagem proveniente do espectro infravermelho sofreu rotações que variaram de 0 a 360 graus, com incrementos de 15 graus. No tocante à sensibilidade a mudanças de escala, novamente manteve-se fixa a imagem obtida do espectro visível de cada par de imagens da base de dados. Neste caso, foi redimensionada a escala da imagem obtida do espectro infravermelho, variando de 0,5 a 1,5 vezes o seu tamanho original, em incrementos de 0,1. Em ambas as situações, mensurou-se a precisão durante a correspondência entre seus descritores.

5.2.1 Bases de dados

Para evidenciar a eficácia, bem como as restrições do método proposto, experimentos foram executados em quatro bases de dados distintas compreendendo diversas faixas do espectro infravermelho e diferentes dimensões espaciais. As bases de dados estão disponíveis publicamente para garantir a reprodutibilidade dos resultados. Cada uma dessas bases de dados é constituída por conjuntos de imagens, as quais foram captadas de distintas faixas do espectro eletromagnético. Em cada conjunto, está incluída uma imagem do espectro visível e outra do espectro infravermelho.

A primeira base de dados, intitulada *Potsdam*, é amplamente utilizada na literatura em trabalhos para classificação e extração de características no contexto do sensoriamento remoto (LIU et al., 2021; SALEEM; BAIS, 2020; MA; MA; YU, 2019; LIU; LI; PAN, 2019). Esta base, distribuída pela *International Society for Photogrammetry and Remote Sensing*[6], se encontra disponível publicamente[7] e é composta por 38 pares de imagens obtidas do espectro visível e do espectro NIR (infravermelho próximo) da vista aérea da cidade de Potsdam. Todas as imagens possuem o tamanho de 6000×6000 pixels e todos os pares de imagens se encontram alinhados espacialmente.

A segunda base de dados, denominada *EPFL*[8], proposta por Brown e Susstrunk (2011), é comumente utilizada na literatura (LI et al., 2021; LIU et al., 2021; YU; ZHAO, 2020; MA; MA; YU, 2019). Esta base consiste em um conjunto de 477 pares de imagens obtidas do espectro visível e do espectro NIR, distribuídas em 9 categorias. Tais categorias cobrem cenas de países, campos, florestas, ambientes internos, montanhas, edifícios, ruas, áreas urbanas e

[5] Página oficial: <https://opencv.org>.
[6] Página oficial: <https://www.isprs.org>.
[7] Disponível em: <https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx>.
[8] Disponível em: <https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html>.

cenas contendo água. As imagens possuem diferentes tamanhos e todos os pares de imagens se encontram alinhados espacialmente.

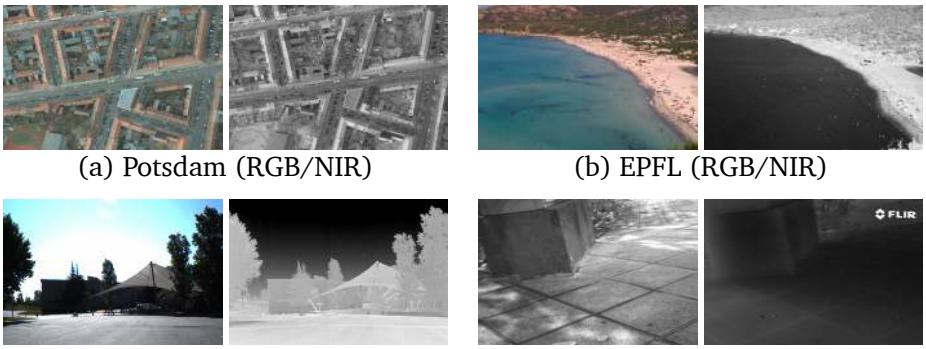
A terceira base de dados, intitulada *CVC*[9], proposta por Aguilera, Sappa e Toledo (2015), também é comumente utilizada na literatura (HU et al., 2022; LI et al., 2021; LIU et al., 2021; YU; ZHAO, 2020). Esta é composta por 44 pares de imagens dos espectros visível e LWIR (infravermelho de ondas longas), todas com dimensões de 639×431 pixels. Esta base de dados foi utilizada para complementar as duas primeiras bases de dados e verificar a eficácia do método proposto em outra faixa do espectro infravermelho, ou seja, infravermelho de ondas longas. As imagens correspondem a diversas cenas ao ar livre.

Por fim, a base *CVC2*[10], proposta por Aguilera et al. (2012), é comumente utilizada na literatura (HU et al., 2022; LIU et al., 2021; SALEEM; BAIS, 2020). Esta base inclui 100 pares de imagens nos espectros visível e LWIR (infravermelho de ondas longas), medindo 506×408 pixels. Estas imagens representam variadas cenas urbanas externas. A Figura 56 fornece exemplos de pares de imagens de cada uma das bases de dados previamente citadas.

5.3 Resultados experimentais

A seguir, são apresentados os resultados experimentais da avaliação do método descritor SORIF aplicado à descrição de características locais em imagens VIS/IV. Todos os experimentos foram executados nas quatro bases de dados previamente mencionadas, especificamente: Potsdam, EPFL, CVC e CVC2.

O método descritor SORIF, junto a outros métodos usados como referências (ou *baselines*, termo inglês), foram individualmente avaliados nas diferentes bases de dados. Isso foi realizado com o objetivo de adquirir um entendimento mais aprofundado sobre o desempenho desses métodos ao serem aplicados em contextos diferentes.



(a) Potsdam (RGB/NIR) (b) EPFL (RGB/NIR)

(c) CVC (RGB/LWIR) (d) CVC2 (VIS/LWIR)

Figura 56 – Exemplos de pares de imagens das bases de dados utilizadas nos experimentos.

[9] Disponível em: <http://adas.cvc.uab.es/projects/simeve/index539a.html>.
[10] Disponível em: <https://owncloud.cvc.uab.es/owncloud/index.php/s/1Wx715yUh6kDAO7>.

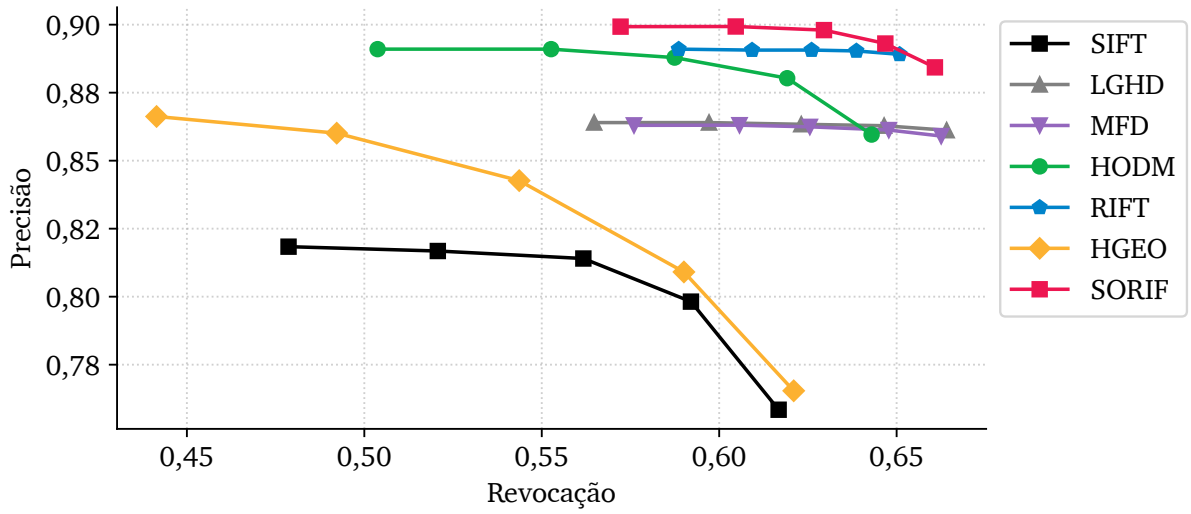


Figura 57 – Curvas de precisão por revocação para a base de dados Potsdam.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados Potsdam. A Figura 57 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Neste resultado, observa-se que os métodos descritores SIFT e HGEO obtiveram eficácia menor tanto na precisão quanto na revocação. O método descritor SORIF obteve maiores valores de revocação e precisão em relação aos demais métodos, incluindo os métodos descritores RIFT e HODM, que também foram projetados para imagens VIS/IV.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados EPFL. A Figura 58 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Esses valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens.

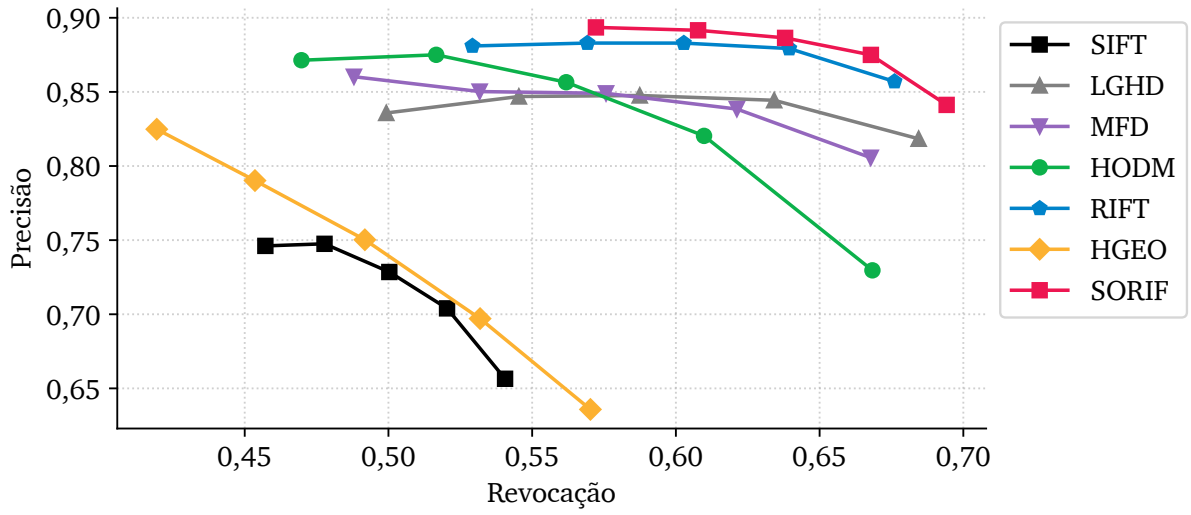


Figura 58 – Curvas de precisão por revocação para a base de dados EPFL.

Similarmente ao resultado anterior, para a base de dados Potsdam, o método descritor SORIF obteve maiores valores de revocação e precisão em relação aos demais métodos. Neste resultado observa-se também que os métodos descritores SIFT e HGEO obtiveram eficácia menor tanto na precisão quanto na revocação.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados CVC. A Figura 59 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Assim como nos outros resultados, os valores foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Neste resultado, observa-se uma distinção melhor entre os métodos descritores. O método descritor SORIF obteve maiores valores de revocação e precisão em relação aos demais métodos. Estes resultados também seguiram a mesma ordem de eficácia dos métodos descritores nas bases de dados Potsdam e EPFL.

A seguir, são apresentados os resultados referentes aos experimentos realizados na base de dados CVC2. A Figura 60 exibe os valores de revocação e precisão obtidos pelos métodos avaliados nessa base de dados. Os valores também foram obtidos por meio da média aritmética dos valores de precisão e revocação, calculados para todos os pares de imagens. Este resultado se assemelha ao resultado obtido na base de dados CVC, porém o resultado do método descritor SORIF se difere mais, com maior valor de precisão e revocação. Neste resultado, o método descritor SIFT obteve a menor eficácia no que diz respeito à precisão e à revocação.

A Figura 61 exibe as curvas de precisão por ângulo de rotação da imagem. Este resultado foi obtido após realizar a correspondência de características locais em todas as imagens de todas as bases de dados. Os valores foram obtidos por meio da média aritmética dos valores de precisão, calculados para todos os pares de imagens. Por esse resultado,

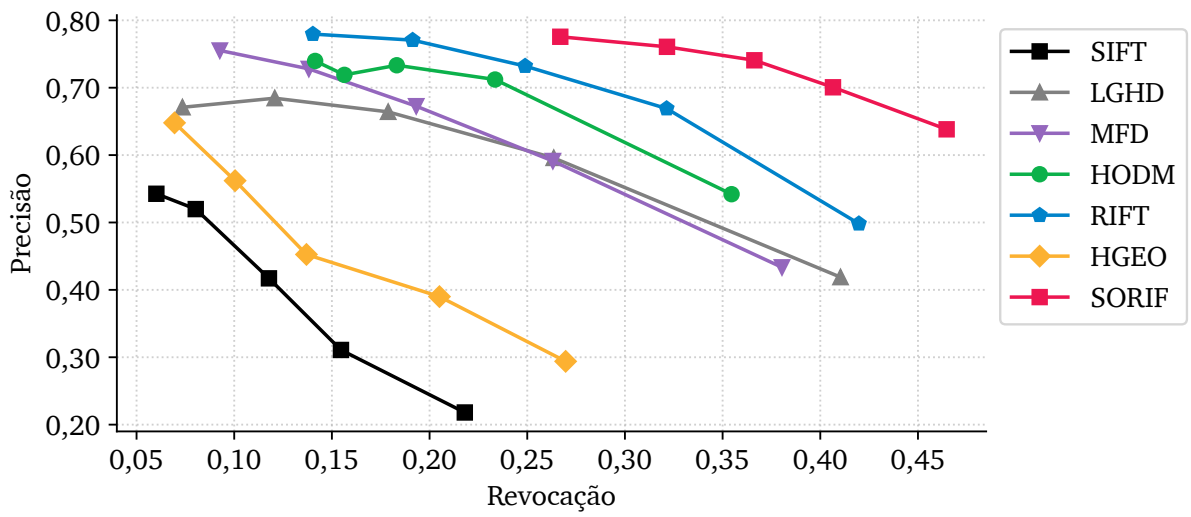


Figura 59 – Curvas de precisão por revocação para a base de dados CVC.

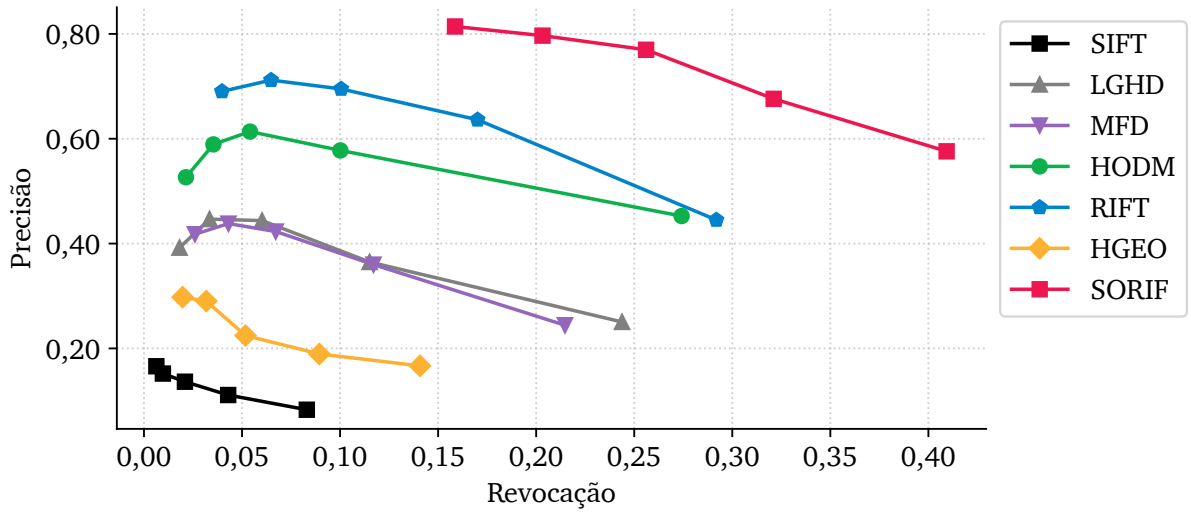


Figura 60 – Curvas de precisão por revocação para a base de dados CVC2.

é possível observar a sensibilidade às variações de rotação das imagens pelos métodos descritores. Em geral, os algoritmos descritores apresentaram comportamentos similares, com exceção do método descritor SORIF. Praticamente todos os métodos apresentam maiores valores de precisão quando não há rotação da imagem (referente a 0 graus), no entanto, a precisão é menor quando há alguma mudança na orientação da imagem.

Ainda na Figura 61 é possível observar que, além do método descritor SORIF possuir maiores valores de precisão, comparado aos demais métodos, ele apresenta alguns picos de precisão mais altos, situados nos ângulos de −90, 0, 90 e 180 graus. Esse comportamento se deve à construção do método em que foram empregados quatro filtros Log-Gabor direcionais com orientações prefixadas.

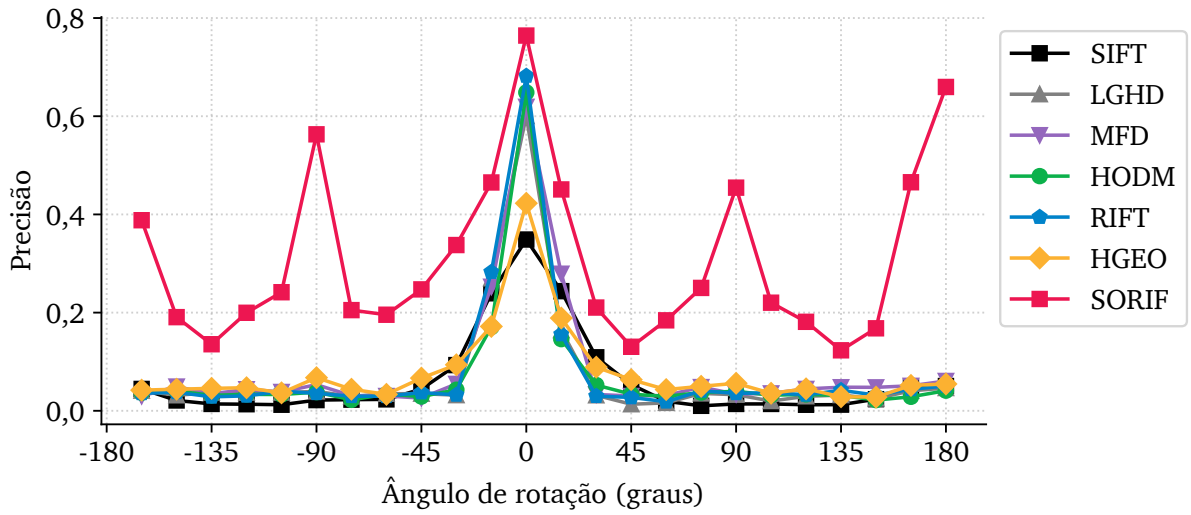


Figura 61 – Curvas de precisão por ângulo de rotação da imagem.

A Figura 62 exibe as curvas de precisão para diferentes valores de escala da imagem. Os valores foram obtidos por meio da média aritmética dos valores de precisão, calculados para todos os pares de imagens de todas as bases de dados. Em geral, os algoritmos descritores apresentaram comportamentos similares, com diferenças na média da precisão. À medida que o tamanho da imagem é reduzido ou aumentado, os valores de precisão diminuem para todos os métodos descritores. Neste resultado, é possível observar que o método descritor SIFT é pouco afetado para diferentes variações de escala. O método descritor SORIF apresentou resultados superiores aos demais métodos avaliados.

A seguir, são apresentados os resultados sobre o tempo de processamento, no qual foram avaliados os métodos descritores. O objetivo deste experimento consistiu em avaliar a eficiência de cada método, considerando o tempo de execução de cada um, expresso em milissegundos. As informações reunidas fornecem uma visão geral do desempenho computacional de cada abordagem.

A Figura 63 exibe o gráfico de valores médios para o tempo de processamento (tempo de relógio) dos métodos avaliados, executados para cada par de imagens em todas as bases de dados. Esses resultados foram obtidos pelo valor médio durante 30 execuções de cada método, e, o tempo de processamento, nesse caso, representa o tempo de relógio que cada algoritmo levou para concluir a descrição das características locais em um par de imagens. O método descritor LGHD apresentou o maior tempo médio de processamento, indicando que ele é o mais demorado entre os avaliados. Por outro lado, o método descritor HODM demonstrou ser o mais rápido. O método descritor SORIF apresentou um tempo de processamento próximo ao RIFT, mas ligeiramente maior devido à etapa de cálculo da orientação. Uma diferença de cerca de apenas 15% a mais no tempo de processamento em comparação com o algoritmo RIFT.

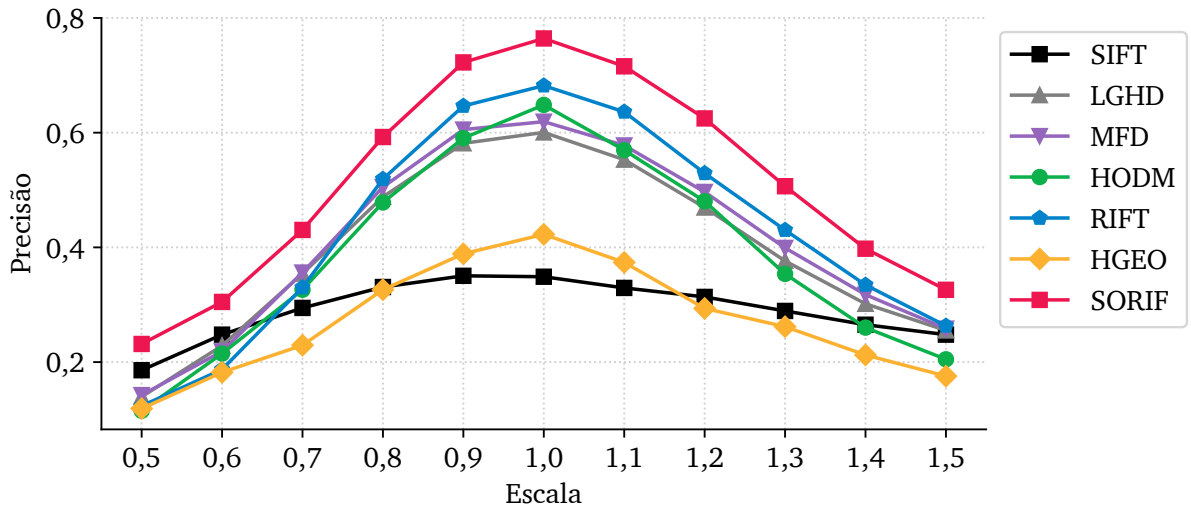


Figura 62 – Curvas de precisão por valores de escala.

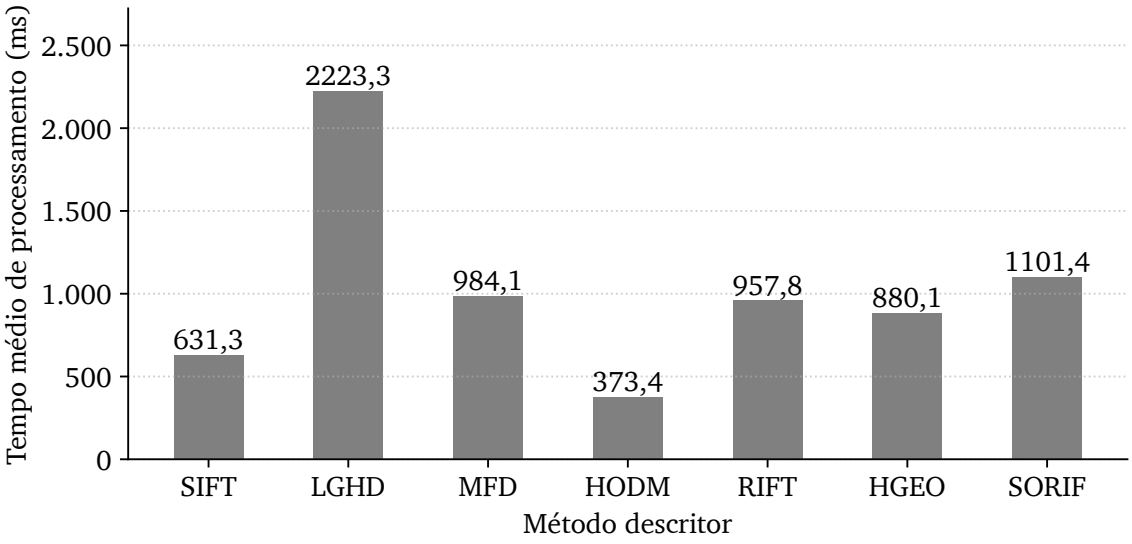


Figura 63 – Valores médios para o tempo de processamento em cada par de imagens.

5.4 Discussão

Conforme os resultados alcançados, o método proposto demonstra superioridade em efetividade em relação aos demais métodos avaliados, em ambas as bases de dados. Este supera os métodos RIFT e HGEO, em termos de eficácia, que foram igualmente desenvolvidos para a descrição de características locais em imagens VIS/IV.

Os métodos avaliados exibiram variações significativas entre diferentes bases de dados. Um aspecto que merece destaque é a distinção nos resultados entre as bases de dados que utilizam imagens de infravermelho próximo (NIR) e aquelas que recorrem a imagens de infravermelho de ondas longas (LWIR). Os métodos descritores mostraram-se superiores nas bases de dados de infravermelho próximo. Isto pode ser explicado pelo fato de que esta faixa do espectro de infravermelho contém mais detalhes sobre a cena em questão e se aproxima mais visualmente das imagens do espectro de luz visível. As imagens de infravermelho de ondas longas (LWIR) são mais desafiadoras, o que levou a resultados inferiores nessas bases de dados (CVC e CVC2). Esse comportamento é corroborado pela literatura, que também evidencia diferenças entre essas faixas do espectro infravermelho.

Apesar das diferenças nos resultados entre as bases de dados, não é possível julgar que os métodos avaliados sejam mais apropriados para um cenário específico. Nestes casos, por exemplo, poderiam existir outras variáveis que influenciam os valores de revocação e precisão, como a presença de ruído nas imagens, tipos de cenários, diferentes faixas do espectro infravermelho, fatores que não foram avaliados neste trabalho.

A implementação de técnicas para a utilização de bordas no cálculo do descritor SORIF, confirma a suposição de que o uso de informações de bordas das imagens, no contexto das imagens VIS/IV, é uma decisão adequada, como indicado nos estudos de Ma et al. (2022) e Liu, Li e Pan (2019). Isto ocorre, pois, apesar das variações significativas de intensidade

entre as imagens VIS/IV, a aparência das formas dos objetos tende a permanecer em ambas as imagens (JIANG et al., 2021; WU; ZHU, 2021).

Em relação ao experimento sobre a sensibilidade a variações de rotação da imagem, o resultado demonstrou a precisão dos algoritmos descritores em resposta às rotações de imagens. Os algoritmos apresentaram alta taxa de precisão quando não há rotação da imagem. No entanto, o algoritmo SORIF destacou-se ao demonstrar certa robustez à rotação, apresentando picos de precisão em certos ângulos. Essa peculiaridade é atribuída à implementação de filtros Log-Gabor direcionais na construção do SORIF.

No entanto, o desempenho do método descritor SORIF também revela uma limitação referente às orientações predefinidas dos filtros Log-Gabor. Essa rigidez restringe sua versatilidade e adaptabilidade ao lidar com rotações que divergem desses ângulos específicos. Esse resultado ilustra a necessidade de desenvolver algoritmos descritores mais robustos sobre a rotação e outros tipos de distorções.

Embora o SORIF tenha se mostrado uma exceção devido à sua robustez à rotação, sua limitação enfatiza a importância de explorar abordagens que oferecem maior flexibilidade e adaptabilidade. Isso poderia ser alcançado por meio da experimentação com diferentes orientações para filtros Log-Gabor ou pela integração de outras técnicas, necessitando de mais estudos nessa área.

No que tange aos experimentos sobre variações na escala da imagem, o método proposto demonstrou-se ligeiramente superior em comparação a outros métodos existentes. Cabe destacar que o método SORIF, devido à sua natureza intrínseca, não processa diferentes escalas de características locais, conforme realizado em alguns métodos desenvolvidos para o espectro visível, como o algoritmo SIFT. Neste ponto, é relevante enfatizar que também nenhum dos métodos avaliados do estado da arte, que foram projetados exclusivamente para imagens VIS/IV, incorpora uma etapa de processamento de escala. Essa lacuna pode ser objeto de estudo em trabalhos futuros.

Em relação ao tempo de processamento, o descritor SORIF apresentou um desempenho similar ao RIFT, embora ligeiramente superior, com uma diferença aproximada de 15% a mais. Esse resultado pode ser atribuído à estrutura do SORIF, que incorpora uma etapa adicional de cálculo de orientação das imagens.

O método descritor SORIF, apesar de apresentar um tempo de processamento mais alto que o RIFT para descrição de características, conta com um descritor de dimensão menor. Este aspecto é vantajoso para a realização de correspondências entre descritores, ao tornar mais rápido o processo de correspondência. O método proposto se mostrou aproximadamente duas vezes mais rápido do que o método descritor LGHD e quase equiparável ao MFD. É importante destacar que os métodos MFD e LGHD não contemplam variações geométricas, como alterações na orientação da imagem.

Com relação aos experimentos, cabe ressaltar que a avaliação dos métodos ocorreu apenas no âmbito do infravermelho próximo (NIR), infravermelho de ondas longas (LWIR)

e no espectro visível. Desta forma, a eficácia da abordagem sugerida para toda a gama do espectro infravermelho não pode ser assegurada unicamente pelos experimentos realizados, sendo necessário mais experimentos.

Contudo, dado que a concepção do método se apoia na estratégia de empregar dados de bordas das imagens, tal como nos estudos de Li, Hu e Ai (2020) e Ma, Ma e Yu (2019) há uma suposição de que a abordagem possa ser igualmente eficaz em outras partes do espectro infravermelho. Também, como as faixas NIR e LWIR cobrem uma grande área do espectro infravermelho, espera-se que o método também tenha resultados satisfatórios nas faixas intermediárias (faixas SWIR e MWIR).

Existem algumas limitações a serem observadas em relação ao método proposto. A primeira delas está relacionada com as imagens geradas em resposta aos filtros Log-Gabor, utilizados para extrair informações das bordas. Em determinados cenários, essas imagens podem apresentar variações de resultados, quando comparadas entre distintas faixas do espectro eletromagnético. Como exemplo, tanto no espectro visível quanto no infravermelho, podem ser identificadas situações em que as bordas das imagens se mostram ambíguas, devido às variações na reflexão da luz em diferentes faixas do espectro eletromagnético. Esta situação, portanto, pode impactar na eficácia do método ao extrair informações de bordas de imagens pertencentes a diferentes espectros.

Por último, é possível vislumbrar trabalhos futuros com base nos resultados alcançados, que poderiam incluir a análise do método proposto em imagens obtidas em outras bandas do espectro eletromagnético. Além disso, poderia ser realizada uma avaliação que considerasse os desafios relativos às questões geométricas, incluindo alterações no ponto de vista, assim como às questões fotométricas, como as variações na iluminação da cena e na presença de desfoques ou ruídos.

Além disso, conforme mencionado previamente, pode ser conduzida uma investigação para a incorporação de uma etapa de processamento de escala. Isso tem o intuito de incrementar a robustez do método proposto, especialmente no que se refere à sensibilidade a diferentes escalas, de maneira análoga ao que é realizado no algoritmo SIFT.

Capítulo 6

Discussão Geral

A escolha entre abordagens de modelagem manual ou aprendizagem profunda é um dilema comum em diversas áreas. No contexto deste estudo, torna-se essencial analisar e contrastar ambas as abordagens, especialmente quando aplicadas à descrição de características locais.

A eficácia dos métodos baseados em aprendizagem profunda é comparada à dos métodos projetados sob modelagem manual na Figura 64, que apresenta as curvas de precisão por revocação dos métodos descritores avaliados aqui neste trabalho. Este gráfico inclui apenas os métodos projetados para imagens VIS/IV e revela a superioridade da modelagem manual sobre os métodos descritores Q-Net e MF-Net, baseados em aprendizagem profunda.

Esta comparação evidencia que os métodos baseados em aprendizagem profunda ainda necessitam de mais estudos para se equiparar ou superar as técnicas de modelagem manual. Esta questão também é apontada em alguns trabalhos na literatura (CHEN; ROTTENSTEINER; HEIPKE, 2020; BELLAVIA; COLOMBO, 2020; KUPPALA; BANDA; BARIGE, 2020; JOSHI; PATEL, 2020). Isso ressalta que, mesmo com o crescente avanço em abordagens de aprendizagem profunda, as abordagens baseadas em modelagem manual ainda são valorizadas, especialmente em tarefas especializadas e desafiadoras, como a descrição de características locais em imagens VIS/IV.

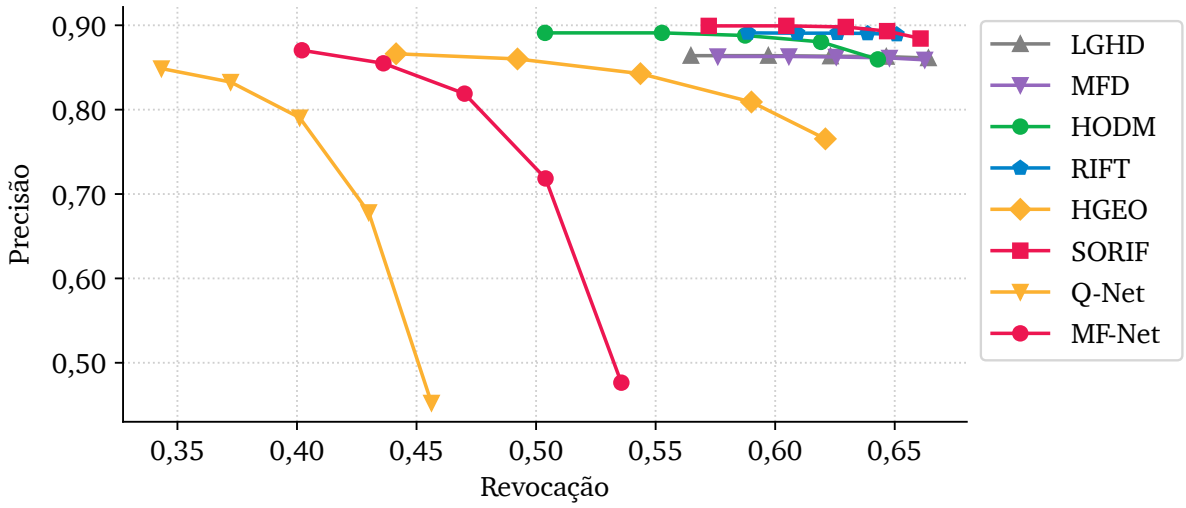


Figura 64 – Curvas de precisão por revocação para a base de dados Potsdam.

Conforme apresentado em Jiang et al. (2021), as redes neurais convolucionais (CNNs) ainda têm suas limitações na descrição de características locais. Quando se trata de representações simples, como características locais (pontos, cantos, bordas ou padrões específicos) os métodos baseados em modelagem manual têm uma vantagem natural. Eles são projetados com um entendimento específico do problema, sendo frequentemente mais aptos a representar formas simples, o que pode justificar sua superioridade em comparação às CNNs.

Os métodos fundamentados em modelagem manual, como o SORIF, apresentam outras vantagens importantes, sendo a interpretabilidade uma das principais delas. Estes métodos são projetados a partir do conhecimento e intuição humanos, o que facilita a compreensão dos mecanismos por trás de suas decisões. Isso contrasta com os métodos de aprendizagem profunda, que frequentemente possuem mecanismos internos menos transparentes e mais difíceis de se interpretar. Além disso, os métodos baseados em modelagem manual são, frequentemente, mais eficientes, uma vez que exigem menos recursos computacionais, conforme também apresentado pelos resultados de tempo de processamento dos métodos avaliados nesta tese. Outra vantagem significativa é que esses métodos não necessitam de dados para treinamento, tornando-os particularmente úteis em cenários em que os dados para treinamento são escassos ou então difíceis de se obter.

Por outro lado, as abordagens baseadas em aprendizagem profunda, como o MF-Net, são projetadas para aprender diretamente a partir dos dados. Isso lhes confere uma notável adaptabilidade, ao poderem se ajustar a variações nos dados que seriam desafiadoras de se modelar manualmente. Quando treinados, estes modelos podem oferecer uma robusta capacidade de generalização, reconhecendo padrões em dados nunca vistos anteriormente. No entanto, esses métodos exigem dados anotados para seu treinamento. Além disso, são mais exigentes em termos de custo computacional, tornando-os por vezes impraticáveis em dispositivos com poucos recursos computacionais.

Combinar ambas as abordagens em aplicações práticas para realizar a descrição de características locais pode ser uma estratégia promissora. Contudo, no contexto deste trabalho, as abordagens baseadas em aprendizagem profunda ainda não alcançaram o mesmo nível de desempenho das abordagens baseadas em modelagem manual. Assim, unir ambas as técnicas para trabalharem conjuntamente em um cenário prático, poderia não apenas diluir os benefícios da modelagem manual, mas também adicionar uma complexidade desnecessária e consumir mais recursos computacionais, o que poderia comprometer a eficiência e eficácia da solução final.

No contexto deste trabalho, o método MF-Net exemplificou como os elementos de modelagem manual, como a camada de mapeamento inspirada em filtros Log-Gabor, podem ser integrados em uma arquitetura de aprendizagem profunda, combinando o melhor de ambas as abordagens. Essa relevância é evidenciada no trabalho de Chen, Rottensteiner e Heipke (2020), que discute a integração do conhecimento de domínio especialista às abordagens baseadas em CNNs.

Vale enfatizar que as contribuições de ambas as abordagens são essenciais para o avanço da área de pesquisa. Os métodos descritores propostos neste trabalho, MF-Net e SORIF, demonstram essa complementaridade, em que a integração de técnicas de modelagem manual em estruturas de aprendizagem profunda pode oferecer novas oportunidades para avanços significativos, abrindo portas para futuras pesquisas e aplicações práticas em diferentes faixas do espectro eletromagnético.

No contexto das abordagens baseadas em aprendizagem profunda, pode-se dizer que este trabalho fez avanços significativos, principalmente no tratamento da relação não-monotônica dos pixels entre imagens obtidas de diferentes faixas do espectro eletromagnético empregando uma rede neural convolucional. Nesse contexto também é importante refletir sobre até que ponto é necessário introduzir engenharia manual nas abordagens de aprendizagem profunda para compensar suas limitações.

Por fim, futuros estudos devem avaliar como equilibrar os pontos fortes de cada abordagem, uma vez que nenhuma é inerentemente superior, pois se tratam de ferramentas distintas. Portanto, este trabalho serve como um ponto de partida para investigações adicionais que busquem entender a interação entre as diferentes abordagens neste contexto, buscando a integração mais eficaz e eficiente das duas abordagens tanto para futuras pesquisas quanto para suas utilizações em questões práticas.

Capítulo 7

Conclusão

Neste trabalho, dois métodos descritores de características locais foram desenvolvidos para abordar o desafio de correspondência entre imagens obtidas do espectro visível (VIS) e do espectro infravermelho (IV).

No contexto desta pesquisa, um dos principais obstáculos identificados durante a descrição de características locais em imagens VIS/IV reside na realidade de que os valores de intensidade, associados a pontos específicos, podem variar de maneira não previsível entre imagens adquiridas de diferentes faixas do espectro eletromagnético.

A elaboração do primeiro método descritor, denominado MF-Net, teve como motivação o recente sucesso de abordagens que empregam aprendizagem profunda (do inglês, *Deep Learning*). As soluções propostas na literatura para a descrição de características locais, fundamentadas em aprendizagem profunda, ainda apresentam limitações, uma vez que muitas delas não abordam a questão da não linearidade das imagens VIS/IV.

Dessa forma, propôs-se uma nova camada de aprendizagem profunda, intitulada como camada de mapeamento, para lidar com as relações não-monotônicas dos dados. Essa camada foi inspirada na estrutura de métodos descritores baseados em modelagem manual (do inglês, *handcrafted*), que empregam filtros Log-Gabor para executar a tarefa de extração de características e lidar com as especificidades das imagens VIS/IV.

A incorporação desta camada de mapeamento na arquitetura da CNN elevou a eficácia do modelo em descrever características locais no cenário das imagens VIS/IV. Além disso, em comparação com outros métodos baseados em aprendizagem profunda, o método descritor MF-Net evidenciou maior eficácia. Este resultado salienta a utilidade de informações de bordas das imagens no contexto das imagens VIS/IV. O método sugerido também apresentou maior eficiência em termos de tempo de processamento para treinamento.

Em relação ao resultado obtido sobre o método descritor MF-Net, acredita-se que, ao utilizar as camadas de mapeamento para novas arquiteturas (projetadas para imagens VIS/IV), possam ser obtidos resultados ainda melhores. Adicionalmente, supõe-se que a camada de mapeamento idealizada e construída neste trabalho possa ser útil em abordagens de aprendizagem profunda que lidam com outros tipos de dados, uma vez que é relativamente simples inserir e adaptar essa camada em outras arquiteturas já existentes.

Embora o método proposto MF-Net tenha superado os demais métodos baseados em aprendizagem profunda, em geral, sua eficácia ainda é inferior quando comparada aos métodos baseados em modelagem manual. Uma possível explicação para esse resultado é que as redes neurais convolucionais, quando empregadas na produção de descritores de características, falham em generalizar adequadamente as características locais, visto que estas são representações mais básicas em comparação com imagens de objetos complexos. Essa problemática ainda é um tema pendente que requer estudos mais aprofundados e não foi abordada neste trabalho.

O segundo método descritor proposto neste trabalho, intitulado SORIF, foi motivado pela lacuna na literatura para a descrição de características locais, em que métodos descritores baseados em modelagem manual apresentavam limitações quando confrontados com variações geométricas em imagens VIS/IV. Portanto, neste método, projetou-se uma etapa para o cálculo de orientação para estimar a orientação de parte da imagem a ser descrita.

Na avaliação do método SORIF, recorreu-se a quatro bases de dados compostas por imagens obtidas do espectro visível e do espectro infravermelho. O método demonstrou robustez às variações de rotação da imagem, graças à implementação da etapa de cálculo de orientação. No entanto, essa robustez à rotação revelou uma limitação do método, uma vez que os filtros Log-Gabor possuem orientações predefinidas e, portanto, podem não ser tão eficazes para lidar com rotações que divergem desses ângulos específicos. Isso sinalizou a necessidade de desenvolver algoritmos descritores mais robustos em relação à rotação e a outras formas de distorção.

No que concerne às variações na escala da imagem, o método descritor SORIF apresentou uma ligeira superioridade em relação a outros métodos. No entanto, o SORIF, devido à sua natureza, não possui uma etapa dedicada ao tratamento de diferentes escalas de características locais, ao contrário do que o algoritmo SIFT oferece. Esta é uma oportunidade para ser explorada em trabalhos futuros. Quanto ao tempo de processamento, o SORIF demonstrou um tempo similar ao RIFT, mas cerca de 15% mais lento. Entretanto, o SORIF possui um descritor de dimensão menor, o que torna o processo de correspondência mais rápido.

Para concluir, mesmo com os resultados promissores para ambos os métodos descritores propostos neste trabalho, há limitações a serem reconhecidas. Particularmente, as imagens produzidas em resposta aos filtros Log-Gabor podem apresentar variações nos resultados em diferentes faixas do espectro eletromagnético. Além disso, é imprescindível a realização de estudos complementares para corroborar a eficácia em todo o espectro infravermelho para ambos os métodos descritores. Também, pesquisas futuras poderiam focar nesses desafios e investigar a aplicação do método em outras faixas do espectro eletromagnético, assim como avaliar questões fotométricas e ruídos nas imagens.

Uma das hipóteses centrais deste estudo era que a utilização de funções matemáticas, tais como as funções Log-Gabor, e métodos apropriados para as peculiaridades das imagens VIS/IV, permitiria a obtenção de melhores resultados comparativamente aos algoritmos do estado da arte. Esta suposição foi validada com ambos os métodos descritores propostos, MF-Net e SORIF, e ambos superaram as técnicas anteriormente disponíveis em termos de eficácia.

Os avanços proporcionados pelos métodos descritores apresentados neste trabalho representaram progressos importantes na descrição de características locais em imagens VIS/IV. Apesar de cada método ter suas limitações, as oportunidades para aprimoramento e investigações adicionais são pertinentes. Com mais pesquisa e desenvolvimento, ambos os métodos possuem o potencial de aprimorar ainda mais a eficácia e a eficiência para descrição de características locais em imagens adquiridas de diferentes faixas do espectro eletromagnético, abrindo novas possibilidades para outras aplicações práticas.

Em relação à publicação científica, no decorrer deste trabalho, foram elaborados três artigos, dos quais dois foram publicados em periódicos. Há outro artigo que se encontra em processo de escrita e publicação. As informações dos artigos são apresentadas a seguir:

- NUNES, C. F. G.; PÁDUA, F. L. C. A convolutional neural network for learning local feature descriptors on multispectral images. **IEEE Latin America Transactions**, Institute of Electrical and Electronics Engineers (IEEE), v. 20, n. 2, p. 215-222, fev. 2022. DOI: 10.1109/tla.2022.9661460.

- SOUZA, P. V. de C.; NUNES, C. F. G.; GUIMARES, A. J.; REZENDE, T. S.; ARAUJO, V. S.; ARAJUO, V. J. S. Self-organized direction aware for regularized fuzzy neural networks. **Evolving Systems**, Springer Science and Business Media LLC, v. 12, n. 2, p. 303-317, mar. 2019. DOI: 10.1007/s12530-019-09278-5.

- NUNES, C. F. G.; PÁDUA, F. L. C. A scale-orientation robust local feature descriptor for visible-infrared image matching. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE). (A ser submetido).

Referências

AGUILERA, C.; BARRERA, F.; LUMBRERAS, F.; SAPPA, A. D.; TOLEDO, R. Multispectral image feature points. **Sensors**, MDPI AG, v. 12, n. 9, p. 12661–12672, set. 2012. DOI: 10.3390/s120912661. Citado 7 vezes nas páginas 55, 61, 63, 68, 98, 132 e 133.

AGUILERA, C.; SAPPA, A.; AGUILERA, C.; TOLEDO, R. Cross-spectral local descriptors via quadruplet network. **Sensors**, MDPI AG, v. 17, n. 4, p. 873, abr. 2017. DOI: 10.3390/s17040873. Citado 8 vezes nas páginas 58, 62, 63, 71, 72, 74, 76 e 139.

AGUILERA, C. A.; AGUILERA, F. J.; SAPPA, A. D.; AGUILERA, C.; TOLEDO, R. Learning cross-spectral similarity measures with deep convolutional neural networks. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.]: IEEE, 2016. DOI: 10.1109/cvprw.2016.40. Citado na página 58.

AGUILERA, C. A.; SAPPA, A. D.; TOLEDO, R. LGHD: A feature descriptor for matching across non-linear intensity variations. In: **2015 IEEE International Conference on Image Processing (ICIP)**. [S.l.]: IEEE, 2015. DOI: 10.1109/icip.2015.7350783. Citado 6 vezes nas páginas 32, 56, 62, 63, 98 e 139.

ALAHI, A.; ORTIZ, R.; VANDERGHEYNST, P. FREAK: Fast retina keypoint. In: **2012 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.]: IEEE, 2012. DOI: 10.1109/cvpr.2012.6247715. Citado 2 vezes nas páginas 50 e 63.

ANTONINI, M.; BARLAUD, M.; MATHIEU, P.; DAUBECHIES, I. Image coding using wavelet transform. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 1, n. 2, p. 205–220, abr. 1992. DOI: 10.1109/83.136597. Citado na página 125.

ANUTA, P. Spatial registration of multispectral and multitemporal digital imagery using fast fourier transform techniques. **IEEE Transactions on Geoscience Electronics**, Institute of Electrical and Electronics Engineers (IEEE), v. 8, n. 4, p. 353–368, out. 1970. DOI: 10.1109/tge.1970.271435. Citado na página 35.

APICELLA, A.; DONNARUMMA, F.; ISGRÒ, F.; PREVETE, R. A survey on modern trainable activation functions. **Neural Networks**, Elsevier BV, v. 138, p. 14–32, jun. 2021. DOI: 10.1016/j.neunet.2021.01.026. Citado 3 vezes nas páginas 40, 41 e 45.

BAJCSY, R.; KOVACIC, S. Multiresolution elastic matching. **Computer Vision, Graphics, and Image Processing**, Elsevier BV, v. 46, n. 1, p. 1–21, abr. 1989. DOI: 10.1016/s0734-189x(89)80014-3. Citado na página 44.

BALNTAS, V.; JOHNS, E.; TANG, L.; MIKOLAJCZYK, K. **PN-Net: Conjoined Triple Deep Network for Learning Local Image Descriptors**. 2016. arXiv: 1601.05030. Citado 4 vezes nas páginas 51, 52, 63 e 139.

BALNTAS, V.; RIBA, E.; PONSA, D.; MIKOLAJCZYK, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: **Procedings of the British Machine Vision Conference 2016**. [S.l.]: British Machine Vision Association, 2016. DOI: 10.5244/c.30.119. Citado 11 vezes nas páginas 52, 54, 62, 63, 69, 71, 72, 75, 96, 133 e 139.

BARAJAS-GARCÍA, C.; SOLORZA-CALDERÓN, S.; GUTIÉRREZ-LÓPEZ, E. Scale, translation and rotation invariant wavelet local feature descriptor. **Applied Mathematics and Computation**, Elsevier BV, v. 363, p. 124594, dez. 2019. DOI: 10.1016/j.amc.2019.124594. Citado 3 vezes nas páginas 22, 49 e 90.

BARUCH, E. B.; KELLER, Y. Joint detection and matching of feature points in multimodal images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2021. DOI: 10.1109/tpami.2021.3092289. Citado 3 vezes nas páginas 18, 21 e 64.

BAY, H.; TUYTELAARS, T.; GOOL, L. V. SURF: Speeded up robust features. In: **Computer Vision – ECCV 2006**. [S.l.]: Springer Berlin Heidelberg, 2006. p. 404–417. DOI: 10.1007/11744023_32. Citado 3 vezes nas páginas 22, 49 e 63.

BELLAVIA, F.; COLOMBO, C. Is there anything new to say about SIFT matching? **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 128, n. 7, p. 1847–1866, mar. 2020. DOI: 10.1007/s11263-020-01297-z. Citado 7 vezes nas páginas 16, 23, 30, 47, 54, 61 e 106.

BOTTOU, L. Large-scale machine learning with stochastic gradient descent. In: **Proceedings of COMPSTAT'2010**. [S.l.]: Springer, 2010. p. 177–186. DOI: 10.1007/978-3-7908-2604-3_16. Citado na página 71.

BRITO, D. N.; NUNES, C. F. G.; PÁDUA, F. L. C.; LACERDA, A. Evaluation of interest point matching methods for projective reconstruction of 3d scenes. **IEEE Latin America Transactions**, Institute of Electrical and Electronics Engineers (IEEE), v. 14, n. 3, p. 1393–1400, mar. 2016. DOI: 10.1109/tla.2016.7459626. Citado na página 50.

BROWN, M.; SUSSTRUNK, S. Multi-spectral SIFT for scene category recognition. In: **CVPR 2011**. [S.l.]: IEEE, 2011. DOI: 10.1109/cvpr.2011.5995637. Citado 2 vezes nas páginas 76 e 97.

BUDUMA, N. **Fundamentals of Deep Learning**. [S.l.]: O'Reilly UK Ltd., 2017. ISBN 1491925612. Citado 3 vezes nas páginas 45, 46 e 72.

CALONDER, M.; LEPETIT, V.; STRECHA, C.; FUA, P. BRIEF: Binary robust independent elementary features. In: **Computer Vision – ECCV 2010**. [S.l.]: Springer Berlin Heidelberg, 2010. p. 778–792. DOI: 10.1007/978-3-642-15561-1_56. Citado 3 vezes nas páginas 30, 50 e 63.

CAMPELO, F.; WANNER, E. F. Sample size calculations for the experimental comparison of multiple algorithms on multiple problem instances. **Journal of Heuristics**, Springer Science and Business Media LLC, v. 26, n. 6, p. 851–883, ago. 2020. DOI: 10.1007/s10732-020-09454-w. Citado 2 vezes nas páginas 77 e 78.

CANNY, J. A computational approach to edge detection. In: **Readings in Computer Vision**. [S.l.]: Elsevier, 1987. p. 184–203. DOI: 10.1016/b978-0-08-051581-6.50024-6. Citado 2 vezes nas páginas 55 e 132.

CHANG, S.-F.; SIKORA, T.; PURL, A. Overview of the MPEG-7 standard. **IEEE Transactions on Circuits and Systems for Video Technology**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, n. 6, p. 688–695, jun. 2001. DOI: 10.1109/76.927421. Citado na página 92.

CHEN, J.; TIAN, J.; LEE, N.; ZHENG, J.; SMITH, R. T.; LAINE, A. F. A partial intensity invariant feature descriptor for multimodal retinal image registration. **IEEE Transactions on Biomedical Engineering**, Institute of Electrical and Electronics Engineers (IEEE), v. 57, n. 7, p. 1707–1718, jul. 2010. DOI: 10.1109/tbme.2010.2042169. Citado 2 vezes nas páginas 55 e 63.

CHEN, L.; ROTTENSTEINER, F.; HEIPKE, C. Feature detection and description for image matching: from hand-crafted design to deep learning. **Geo-spatial Information Science**, Informa UK Limited, v. 24, n. 1, p. 58–74, nov. 2020. DOI: 10.1080/10095020.2020.1843376. Citado 9 vezes nas páginas 22, 23, 48, 49, 60, 61, 64, 106 e 107.

CHEN, M.; HABIB, A.; HE, H.; ZHU, Q.; ZHANG, W. Robust feature matching method for SAR and optical images by using gaussian-gamma-shaped bi-windows-based descriptor and geometric constraint. **Remote Sensing**, MDPI AG, v. 9, n. 9, p. 882, ago. 2017. DOI: 10.3390/rs9090882. Citado 3 vezes nas páginas 20, 28 e 56.

DEDE, M. A.; APTOULA, E.; GENC, Y. Deep network ensembles for aerial scene classification. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 16, n. 5, p. 732–735, maio 2019. DOI: 10.1109/lgrs.2018.2880136. Citado na página 50.

DELLINGER, F.; DELON, J.; GOUSSEAU, Y.; MICHEL, J.; TUPIN, F. SAR-SIFT: A SIFT-like algorithm for SAR images. **IEEE Transactions on Geoscience and Remote Sensing**, Institute of Electrical and Electronics Engineers (IEEE), v. 53, n. 1, p. 453–466, jan. 2015. DOI: 10.1109/tgrs.2014.2323552. Citado 6 vezes nas páginas 18, 19, 56, 63, 75 e 96.

DESTA, F.; BUXTON, M.; JANSEN, J. Fusion of mid-wave infrared and long-wave infrared reflectance spectra for quantitative analysis of minerals. **Sensors**, MDPI AG, v. 20, n. 5, p. 1472, mar. 2020. DOI: 10.3390/s20051472. Citado na página 35.

DETONE, D.; MALISIEWICZ, T.; RABINOVICH, A. SuperPoint: Self-supervised interest point detection and description. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.]: IEEE, 2018. DOI: 10.1109/cvprw.2018.00060. Citado na página 54.

DONG, Y.; JIAO, W.; LONG, T.; LIU, L.; HE, G.; GONG, C.; GUO, Y. Local deep descriptor for remote sensing image feature matching. **Remote Sensing**, MDPI AG, v. 11, n. 4, p. 430, fev. 2019. DOI: 10.3390/rs11040430. Citado 2 vezes nas páginas 54 e 63.

DUSMANU, M.; ROCCO, I.; PAJDLA, T.; POLLEFEYS, M.; SIVIC, J.; TORII, A.; SATTLER, T. D2-net: A trainable CNN for joint description and detection of local features. In: **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2019. DOI: 10.1109/cvpr.2019.00828. Citado na página 54.

FAN, C.; JIN, H.; WANG, F.; ZHANG, G.; LI, Y. Combining and matching keypoints and lines on multispectral images. **Infrared Physics & Technology**, Elsevier BV, v. 96, p. 316–324, jan. 2019. DOI: 10.1016/j.infrared.2018.12.004. Citado 4 vezes nas páginas 18, 29, 35 e 64.

FANG, B.; YU, K.; MA, J.; AN, P. EMCM: A novel binary edge-feature-based maximum clique framework for multispectral image matching. **Remote Sensing**, MDPI AG, v. 11, n. 24, p. 3026, dez. 2019. DOI: 10.3390/rs11243026. Citado 4 vezes nas páginas 35, 57, 66 e 87.

FIELD, D. J. Relations between the statistics of natural images and the response properties of cortical cells. **Journal of the Optical Society of America A**, The Optical Society, v. 4, n. 12, p. 2379, dez. 1987. DOI: 10.1364/josaa.4.002379. Citado 3 vezes nas páginas 66, 125 e 126.

FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. **Communications of the ACM**, Association for Computing Machinery (ACM), v. 24, n. 6, p. 381–395, jun. 1981. DOI: 10.1145/358669.358692. Citado 2 vezes nas páginas 29 e 56.

FU, Z.; QIN, Q.; LUO, B.; SUN, H.; WU, C. HOMPC: A local feature descriptor based on the combination of magnitude and phase congruency information for multi-sensor remote sensing images. **Remote Sensing**, MDPI AG, v. 10, n. 8, p. 1234, ago. 2018. DOI: 10.3390/rs10081234. Citado 9 vezes nas páginas 22, 32, 59, 62, 63, 66, 68, 76 e 87.

FU, Z.; QIN, Q.; LUO, B.; WU, C.; SUN, H. A local feature descriptor based on combination of structure and texture information for multispectral image matching. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–5, 2018. DOI: 10.1109/lgrs.2018.2867635. Citado 8 vezes nas páginas 35, 57, 59, 63, 68, 75, 88 e 96.

GABOR, D. Theory of communication. part 1: The analysis of information. **Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering**, Institution of Engineering and Technology (IET), v. 93, n. 26, p. 429–441, nov. 1946. DOI: 10.1049/ji-3-2.1946.0074. Citado na página 125.

GOODFELLOW, I.; BENGIO, J.; COURVILLE, A.; BACH, F. **Deep Learning**. [S.l.]: MIT Press Ltd, 2016. ISBN 0262035618. Citado 4 vezes nas páginas 38, 39, 44 e 64.

HAN, X.; LEUNG, T.; JIA, Y.; SUKTHANKAR, R.; BERG, A. C. MatchNet: Unifying feature and metric learning for patch-based matching. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2015. DOI: 10.1109/cvpr.2015.7298948. Citado 2 vezes nas páginas 50 e 63.

HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: **Procedings of the Alvey Vision Conference 1988**. [S.l.]: Alvey Vision Club, 1988. DOI: 10.5244/c.2.23. Citado na página 57.

HARTLEY, R. **Multiple view geometry in computer vision**. Cambridge, UK New York: Cambridge University Press, 2004. DOI: 10.1017/CBO9780511811685. ISBN 9780511811685. Citado na página 135.

HAYKIN, S. **Neural networks and learning machines**. New York: Prentice Hall/Pearson, 2009. ISBN 9780131471399. Citado 2 vezes nas páginas 39 e 40.

HE, K.; LU, Y.; SCLAROFF, S. Local descriptors optimized for average precision. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.]: IEEE, 2018. DOI: 10.1109/cvpr.2018.00069. Citado 4 vezes nas páginas 32, 34, 53 e 63.

HE, R.; CAO, J.; SONG, L.; SUN, Z.; TAN, T. Adversarial cross-spectral face completion for NIR-VIS face recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers (IEEE), v. 42, n. 5, p. 1025–1037, maio 2020. DOI: 10.1109/tpami.2019.2961900. Citado na página 20.

HEINLY, J.; DUNN, E.; FRAHM, J.-M. Comparative evaluation of binary features. In: **Computer Vision – ECCV 2012**. [S.l.]: Springer Berlin Heidelberg, 2012. p. 759–773. DOI: 10.1007/978-3-642-33709-3_54. Citado 5 vezes nas páginas 32, 50, 52, 61 e 75.

HU, H.; LI, B.; YANG, W.; WEN, C.-Y. A novel multispectral line segment matching method based on phase congruency and multiple local homographies. **Remote Sensing**, MDPI AG, v. 14, n. 16, p. 3857, ago. 2022. DOI: 10.3390/rs14163857. Citado 2 vezes nas páginas 92 e 98.

HU, S.; SHORT, N.; RIGGAN, B. S.; CHASSE, M.; SARFRAZ, M. S. Heterogeneous face recognition: Recent advances in infrared-to-visible matching. In: **2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)**. [S.l.]: IEEE, 2017. DOI: 10.1109/fg.2017.126. Citado na página 36.

JIANG, X.; MA, J.; XIAO, G.; SHAO, Z.; GUO, X. A review of multimodal image matching: Methods and applications. **Information Fusion**, Elsevier BV, v. 73, p. 22–71, set. 2021. DOI: 10.1016/j.inffus.2021.02.012. Citado 13 vezes nas páginas 17, 18, 21, 22, 29, 33, 34, 37, 66, 87, 91, 104 e 107.

JIN, Y.; MISHKIN, D.; MISHCHUK, A.; MATAS, J.; FUA, P.; YI, K. M.; TRULLS, E. Image matching across wide baselines: From paper to practice. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 129, n. 2, p. 517–547, out. 2020. DOI: 10.1007/s11263-020-01385-0. Citado 2 vezes nas páginas 16 e 17.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science (AAAS), v. 349, n. 6245, p. 255–260, jul. 2015. DOI: 10.1126/science.aaa8415. Citado na página 38.

JOSHI, K.; PATEL, M. I. Recent advances in local feature detector and descriptor: a literature survey. **International Journal of Multimedia Information Retrieval**, Springer Science and Business Media LLC, v. 9, n. 4, p. 231–247, out. 2020. DOI: 10.1007/s13735-020-00200-3. Citado 17 vezes nas páginas 17, 19, 20, 21, 22, 24, 28, 29, 30, 33, 35, 38, 50, 61, 62, 64 e 106.

KHAN, A.; SOHAIL, A.; ZAHOORA, U.; QURESHI, A. S. A survey of the recent architectures of deep convolutional neural networks. **Artificial Intelligence Review**, Springer Science and Business Media LLC, abr. 2020. DOI: 10.1007/s10462-020-09825-6. Citado 5 vezes nas páginas 40, 41, 42, 43 e 45.

KIM, S.; MIN, D.; LIN, S.; SOHN, K. Dense cross-modal correspondence estimation with the deep self-correlation descriptor. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2020. DOI: 10.1109/tpami.2020.2965528. Citado 2 vezes nas páginas 22 e 90.

KOVESI, P. Image features from phase congruency. **Videre: Journal of computer vision research**, Citeseer, v. 1, n. 3, p. 1–26, 1999. Citado na página 91.

KOVESI, P. Phase congruency: A low-level image invariant. **Psychological Research**, Springer Science and Business Media LLC, v. 64, n. 2, p. 136–148, dez. 2000. DOI: 10.1007/s004260000024. Citado 2 vezes nas páginas 91 e 125.

KRIG, S. **Computer Vision Metrics**. [S.l.]: Springer-Verlag GmbH, 2016. ISBN 9783319337623. Citado na página 27.

KUPPALA, K.; BANDA, S.; BARIGE, T. R. An overview of deep learning methods for image registration with focus on feature-based approaches. **International Journal of Image and Data Fusion**, Informa UK Limited, v. 11, n. 2, p. 113–135, jan. 2020. DOI: 10.1080/19479832.2019.1707720. Citado 15 vezes nas páginas 17, 23, 27, 29, 30, 48, 49, 50, 60, 61, 62, 64, 73, 95 e 106.

LE, H.; SMAILIS, C.; SHI, L.; KAKADIARIS, I. EDGE20: A cross spectral evaluation dataset for multiple surveillance problems. In: **2020 IEEE Winter Conference on Applications of Computer Vision (WACV)**. [S.l.]: IEEE, 2020. DOI: 10.1109/wacv45572.2020.9093573. Citado 2 vezes nas páginas 20 e 36.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Springer Nature, v. 521, n. 7553, p. 436–444, maio 2015. DOI: 10.1038/nature14539. Citado 3 vezes nas páginas 41, 42 e 43.

LENG, C.; ZHANG, H.; LI, B.; CAI, G.; PEI, Z.; HE, L. Local feature descriptor for image matching: A survey. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 7, p. 6424–6434, 2019. DOI: 10.1109/access.2018.2888856. Citado 4 vezes nas páginas 47, 48, 49 e 61.

LEUTENEGGER, S.; CHLI, M.; SIEGWART, R. Y. BRISK: Binary robust invariant scalable keypoints. In: **2011 International Conference on Computer Vision**. [S.l.]: IEEE, 2011. DOI: 10.1109/iccv.2011.6126542. Citado 2 vezes nas páginas 50 e 63.

LI, J.; HU, Q.; AI, M. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 29, p. 3296–3310, 2020. DOI: 10.1109/tip.2019.2959244. Citado 10 vezes nas páginas 16, 17, 47, 59, 63, 68, 88, 93, 105 e 139.

LI, R.; ZHAO, H.; ZHANG, X.; GE, X.; YUAN, Z.; ZOU, Q. Automatic matching of multispectral images based on nonlinear diffusion of image structures. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, Institute of Electrical and Electronics Engineers (IEEE), v. 14, p. 762–774, 2021. DOI: 10.1109/jstars.2020.3043379. Citado 5 vezes nas páginas 60, 66, 91, 97 e 98.

LI, Y.; JING, J.; JIN, H.; QIAO, W. Building keypoint mappings on multispectral images by a cascade of classifiers with a resurrection mechanism. **Sensors**, MDPI AG, v. 15, n. 5, p. 11769–11786, maio 2015. DOI: 10.3390/s150511769. Citado 2 vezes nas páginas 21 e 37.

LI, Y.; ZOU, J.; JING, J.; JIN, H.; YU, H. Establish keypoint matches on multispectral images utilizing descriptor and global information over entire image. **Infrared Physics & Technology**, Elsevier BV, v. 76, p. 1–10, maio 2016. DOI: 10.1016/j.infrared.2016.01.011. Citado na página 22.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. W. M. van der; GINNEKEN, B. van; SÁNCHEZ, C. I. A survey on deep learning in medical image analysis. **Medical Image Analysis**, Elsevier BV, v. 42, p. 60–88, dez. 2017. DOI: 10.1016/j.media.2017.07.005. Citado na página 42.

LIU, X.; AI, Y.; TIAN, B.; CAO, D. Robust and fast registration of infrared and visible images for electro-optical pod. **IEEE Transactions on Industrial Electronics**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2018. DOI: 10.1109/tie.2018.2833051. Citado na página 32.

LIU, X.; LI, J.-B.; PAN, J.-S. Feature point matching based on distinct wavelength phase congruency and log-gabor filters in infrared and visible images. **Sensors**, MDPI AG, v. 19, n. 19, p. 4244, set. 2019. DOI: 10.3390/s19194244. Citado 7 vezes nas páginas 27, 35, 66, 76, 87, 97 e 103.

LIU, X.; LI, J.-B.; PAN, J.-S.; WANG, S. An advanced gradient texture feature descriptor based on phase information for infrared and visible image matching. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 80, n. 11, p. 16491–16511, mar. 2021. DOI: 10.1007/s11042-020-10213-z. Citado 9 vezes nas páginas 16, 17, 47, 60, 66, 76, 91, 97 e 98.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, Springer Nature, v. 60, n. 2, p. 91–110, nov. 2004. DOI: 10.1023/b:visi.0000029664.99615.94. Citado 9 vezes nas páginas 22, 31, 49, 63, 75, 129, 130, 131 e 139.

LYON, R. F. A brief history of 'pixel'. In: SAMPAT, N.; DICARLO, J. M.; MARTIN, R. A. (Ed.). **Digital Photography II**. [S.l.]: SPIE, 2006. DOI: 10.1117/12.644941. Citado na página 16.

MA, J.; JIANG, X.; FAN, A.; JIANG, J.; YAN, J. Image matching from handcrafted to deep features: A survey. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 129, n. 1, p. 23–79, ago. 2020. DOI: 10.1007/s11263-020-01359-2. Citado 18 vezes nas páginas 16, 17, 18, 20, 22, 28, 29, 30, 32, 47, 48, 49, 54, 60, 61, 62, 73 e 95.

MA, T.; HUANG, W.; DU, L.; WANG, C.; GE, S.; LIU, Y.; YU, K. Multimodal remote sensing image registration based on local phase congruency and keypoint filtering. In: **2022 China Automation Congress (CAC)**. [S.l.]: IEEE, 2022. Citado 2 vezes nas páginas 91 e 103.

MA, T.; MA, J.; YU, K. A local feature descriptor based on oriented structure maps with guided filtering for multispectral remote sensing image matching. **Remote Sensing**, MDPI AG, v. 11, n. 8, p. 951, abr. 2019. DOI: 10.3390/rs11080951. Citado 11 vezes nas páginas 18, 22, 57, 62, 64, 75, 76, 88, 96, 97 e 105.

MAES, F.; COLLIGNON, A.; VANDERMEULEN, D.; MARCHAL, G.; SUETENS, P. Multimodality image registration by maximization of mutual information. **IEEE Transactions on Medical Imaging**, Institute of Electrical and Electronics Engineers (IEEE), v. 16, n. 2, p. 187–198, abr. 1997. DOI: 10.1109/42.563664. Citado na página 17.

MANJUNATH, B. S.; OHM, J.-R.; VASUDEVAN, V. V.; YAMADA, A. Color and texture descriptors. **IEEE Transactions on Circuits and Systems for Video Technology**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, n. 6, p. 703–715, jun. 2001. DOI: 10.1109/76.927424. Citado 2 vezes nas páginas 55 e 92.

MARMANIS, D.; SCHINDLER, K.; WEGNER, J. D.; GALLIANI, S.; DATCU, M.; STILLA, U. Classification with an edge: Improving semantic image segmentation with boundary detection. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier BV, v. 135, p. 158–172, jan. 2018. DOI: 10.1016/j.isprsjprs.2017.11.009. Citado na página 76.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, Springer Science and Business Media LLC, v. 5, n. 4, p. 115–133, dez. 1943. DOI: 10.1007/bf02478259. Citado na página 39.

MENG, L.; ZHOU, J.; LIU, S.; DING, L.; ZHANG, J.; WANG, S.; LEI, T. Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. **International Journal of Applied Earth Observation and Geoinformation**, Elsevier BV, v. 102, p. 102403, out. 2021. DOI: 10.1016/j.jag.2021.102403. Citado na página 20.

MIKOLAJCZYK, K.; SCHMID, C. A performance evaluation of local descriptors. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 27, n. 10, p. 1615–1630, out. 2005. DOI: 10.1109/tpami.2005.188. Citado 12 vezes nas páginas 27, 31, 32, 33, 34, 49, 50, 52, 57, 62, 73 e 95.

MIKOLAJCZYK, K.; TUYTELAARS, T.; SCHMID, C.; ZISSERMAN, A.; MATAS, J.; SCHAFFALITZKY, F.; KADIR, T.; GOOL, L. V. A comparison of affine region detectors. **International Journal of Computer Vision**, Springer Nature, v. 65, n. 1-2, p. 43–72, out. 2005. DOI: 10.1007/s11263-005-3848-x. Citado 3 vezes nas páginas 22, 49 e 57.

MONTGOMERY, D. **Design and analysis of experiments**. Hoboken, NJ: John Wiley & Sons, Inc, 2017. ISBN 9781119113478. Citado 2 vezes nas páginas 77 e 78.

MOUATS, T.; AOUF, N. Multimodal stereo correspondence based on phase congruency and edge histogram descriptor. In: **Proceedings of the 16th International Conference on Information Fusion**. [S.l.: s.n.], 2013. p. 1981–1987. Citado 2 vezes nas páginas 55 e 63.

MOUATS, T.; AOUF, N.; NAM, D.; VIDAS, S. Performance evaluation of feature detectors and descriptors beyond the visible. **Journal of Intelligent & Robotic Systems**, Springer Science and Business Media LLC, v. 92, n. 1, p. 33–63, fev. 2018. DOI: 10.1007/s10846-017-0762-8. Citado 10 vezes nas páginas 31, 32, 33, 57, 61, 62, 73, 75, 95 e 96.

NGO, L.; CHA, J.; HAN, J.-H. Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 29, p. 303–312, 2020. DOI: 10.1109/tip.2019.2931461. Citado na página 50.

NUNES, C. F. G.; PÁDUA, F. L. C. A local feature descriptor based on log-gabor filters for keypoint matching in multispectral images. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 14, n. 10, p. 1850–1854, out. 2017. DOI: 10.1109/lgrs.2017.2738632. Citado 7 vezes nas páginas 17, 19, 57, 62, 63, 67 e 139.

NUNES, C. F. G.; PÁDUA, F. L. C. A convolutional neural network for learning local feature descriptors on multispectral images. **IEEE Latin America Transactions**, Institute of Electrical and Electronics Engineers (IEEE), v. 20, n. 2, p. 215–222, fev. 2022. DOI: 10.1109/tla.2022.9661460. Citado na página 111.

PAL, S. K.; MITRA, S. Multilayer perceptron, fuzzy sets, and classification. **IEEE Transactions on Neural Networks**, Institute of Electrical and Electronics Engineers (IEEE), v. 3, n. 5, p. 683–697, 1992. DOI: 10.1109/72.159058. Citado 2 vezes nas páginas 40 e 41.

PAOLETTI, M. E.; HAUT, J. M.; PLAZA, J.; PLAZA, A. A new deep convolutional neural network for fast hyperspectral image classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier BV, v. 145, p. 120–147, nov. 2018. DOI: 10.1016/j.isprsjprs.2017.11.021. Citado na página 35.

PATTERSON, J.; GIBSON, A. **Deep learning: A practitioner's approach**. [S.l.]: O'Reilly Media, Inc., 2017. Citado 2 vezes nas páginas 43 e 46.

PAVEL, K.; DAVI, S. Algorithms for efficient computation of convolution. In: **Design and Architectures for Digital Signal Processing**. [S.l.]: InTech, 2013. DOI: 10.5772/51942. Citado na página 44.

PENEV, P. S.; ATICK, J. J. Local feature analysis: a general statistical theory for object representation. **Network: Computation in Neural Systems**, Informa UK Limited, v. 7, n. 3, p. 477–500, jan. 1996. DOI: 10.1088/0954-898x_7_3_002. Citado na página 28.

PETROU, M.; PETROU, C. **Image Processing: The Fundamentals**. [S.l.]: Wiley, 2010. DOI: 10.1002/9781119994398. ISBN 9780470745861. Citado na página 35.

PINCHON, N.; CASSIGNOL, O.; NICOLAS, A.; BERNARDIN, F.; LEDUC, P.; TAREL, J.-P.; BRÉMOND, R.; BERCIER, E.; BRUNET, J. All-weather vision for automotive safety: Which spectral band? In: **Advanced Microsystems for Automotive Applications 2018**. [S.l.]: Springer International Publishing, 2018. p. 3–15. DOI: 10.1007/978-3-319-99762-9_1. Citado na página 37.

PLUIM, J. P. W.; MAINTZ, J. B. A.; VIERGEVER, M. A. Mutual-information-based registration of medical images: a survey. **IEEE Transactions on Medical Imaging**, Institute of Electrical and Electronics Engineers (IEEE), v. 22, n. 8, p. 986–1004, ago. 2003. DOI: 10.1109/tmi.2003.815867. Citado na página 17.

QI, Q.; HUO, Q.; WANG, J.; SUN, H.; CAO, Y.; LIAO, J. Personalized sketch-based image retrieval by convolutional neural network and deep transfer learning. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 7, p. 16537–16549, 2019. DOI: 10.1109/access.2019.2894351. Citado na página 50.

QIN, Y.; CAO, Z.; ZHUO, W.; YU, Z. Robust key point descriptor for multi-spectral image matching. **Journal of Systems Engineering and Electronics**, Journal of Systems Engineering and Electronics, v. 25, n. 4, p. 681–687, ago. 2014. DOI: 10.1109/jsee.2014.00078. Citado 2 vezes nas páginas 21 e 55.

REN, L.; LU, J.; FENG, J.; ZHOU, J. Uniform and variational deep learning for RGB-d object recognition and person re-identification. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 28, n. 10, p. 4970–4983, out. 2019. DOI: 10.1109/tip.2019.2915655. Citado na página 50.

RICAURTE, P.; CHILÁN, C.; AGUILERA-CARRASCO, C.; VINTIMILLA, B.; SAPPA, A. Feature point descriptors: Infrared and visible spectra. **Sensors**, MDPI AG, v. 14, n. 2, p. 3690–3701, fev. 2014. DOI: 10.3390/s140203690. Citado 2 vezes nas páginas 35 e 36.

RICHARDS, J. A. **Remote Sensing Digital Image Analysis**. [S.l.]: Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-30062-2. ISBN 9783642441011. Citado na página 56.

ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. In: **Computer Vision – ECCV 2006**. [S.l.]: Springer Berlin Heidelberg, 2006. p. 430–443. DOI: 10.1007/11744023_34. Citado 3 vezes nas páginas 56, 57 e 132.

RUBLEE, E.; RABAUD, V.; KONOLIGE, K.; BRADSKI, G. ORB: An efficient alternative to SIFT or SURF. In: **2011 International Conference on Computer Vision**. [S.l.]: IEEE, 2011. DOI: 10.1109/iccv.2011.6126544. Citado 3 vezes nas páginas 22, 50 e 63.

SALEEM, S. Establishing feature point correspondences between multisensor images with a robust feature matching strategy. **Journal of Electronic Imaging**, SPIE-Intl Soc Optical Eng, v. 28, n. 05, p. 1, set. 2019. DOI: 10.1117/1.jei.28.5.053004. Citado na página 57.

SALEEM, S.; BAIS, A. Visible spectrum and infra-red image matching: A new method. **Applied Sciences**, MDPI AG, v. 10, n. 3, p. 1162, fev. 2020. DOI: 10.3390/app10031162. Citado 9 vezes nas páginas 20, 21, 24, 32, 57, 59, 76, 97 e 98.

SALEEM, S.; BAIS, A.; SABLATNIG, R.; AHMAD, A.; NASEER, N. Feature points for multisensor images. **Computers & Electrical Engineering**, Elsevier BV, v. 62, p. 511–523, ago. 2017. DOI: 10.1016/j.compeleceng.2017.04.032. Citado 8 vezes nas páginas 20, 28, 57, 61, 62, 73, 76 e 95.

SALEEM, S.; SABLATNIG, R. A robust SIFT descriptor for multispectral images. **IEEE Signal Processing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 21, n. 4, p. 400–403, abr. 2014. DOI: 10.1109/lsp.2014.2304073. Citado 4 vezes nas páginas 18, 19, 56 e 63.

SAXENA, S.; SINGH, R. K. A survey of recent and classical image registration methods. **International Journal of Signal Processing, Image Processing and Pattern Recognition**, Science and Engineering Research Support Society, v. 7, n. 4, p. 167–176, ago. 2014. DOI: 10.14257/ijsip.2014.7.4.16. Citado na página 48.

SCHONBERGER, J. L.; HARDMEIER, H.; SATTLER, T.; POLLEFEYS, M. Comparative evaluation of hand-crafted and learned local features. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2017. DOI: 10.1109/cvpr.2017.736. Citado 11 vezes nas páginas 17, 33, 34, 52, 59, 61, 62, 64, 73, 87 e 88.

SHAW, G. A.; BURKE, H. K. Spectral imaging for remote sensing. **Lincoln laboratory journal**, v. 14, n. 1, p. 3–28, 2003. Citado na página 36.

SHEN, X.; WANG, C.; LI, X.; YU, Z.; LI, J.; WEN, C.; CHENG, M.; HE, Z. RF-net: An end-to-end image matching network based on receptive field. In: **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2019. DOI: 10.1109/cvpr.2019.00832. Citado na página 54.

SHI, J.; TOMASI. Good features to track. In: **Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94**. [S.l.]: IEEE Comput. Soc. Press, 1994. DOI: 10.1109/cvpr.1994.323794. Citado na página 57.

SIKORA, T. The MPEG-7 visual standard for content description-an overview. **IEEE Transactions on Circuits and Systems for Video Technology**, Institute of Electrical and Electronics Engineers (IEEE), v. 11, n. 6, p. 696–702, jun. 2001. DOI: 10.1109/76.927422. Citado na página 132.

SIMO-SERRA, E.; TRULLS, E.; FERRAZ, L.; KOKKINOS, I.; FUA, P.; MORENO-NOGUER, F. Discriminative learning of deep convolutional feature point descriptors. In: **2015 IEEE International Conference on Computer Vision (ICCV)**. [S.l.]: IEEE, 2015. DOI: 10.1109/iccv.2015.22. Citado 2 vezes nas páginas 51 e 63.

SOBEL, I.; FELDMAN, G. A 3x3 isotropic gradient operator for image processing. A talk at the Stanford Artificial Project in. 1968. Citado na página 55.

SOUZA, C. L.; PÁDUA, F. L. C.; NUNES, C. F. G.; ASSIS, G. T.; SILVA, G. D. A unified approach to content-based indexing and retrieval of digital videos from television archives. **Artificial Intelligence Research**, Sciedu Press, v. 3, n. 3, set. 2014. DOI: 10.5430/air.v3n3p49. Citado na página 131.

SOUZA, P. V. de C.; NUNES, C. F. G.; GUIMARES, A. J.; REZENDE, T. S.; ARAUJO, V. S.; ARAJUO, V. J. S. Self-organized direction aware for regularized fuzzy neural networks. **Evolving Systems**, Springer Science and Business Media LLC, v. 12, n. 2, p. 303–317, mar. 2019. DOI: 10.1007/s12530-019-09278-5. Citado na página 111.

TEUTSCH, M.; SAPPA, A. D.; HAMMOUD, R. I. Computer vision in the infrared spectrum: Challenges and approaches. **Synthesis Lectures on Computer Vision**, Morgan & Claypool Publishers LLC, v. 10, n. 2, p. 1–138, out. 2021. DOI: 10.2200/s01127ed1v01y202109cov019. Citado 2 vezes nas páginas 20 e 36.

TIAN, Y.; FAN, B.; WU, F. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2017. DOI: 10.1109/cvpr.2017.649. Citado 2 vezes nas páginas 53 e 63.

TUYTELAARS, T.; MIKOLAJCZYK, K. Local invariant feature detectors: A survey. **Foundations and Trends® in Computer Graphics and Vision**, Now Publishers, v. 3, n. 3, p. 177–280, 2008. DOI: 10.1561/0600000017. Citado 3 vezes nas páginas 24, 27 e 28.

USS, M.; VOZEL, B.; LUKIN, V.; CHEHDI, K. Efficient discrimination and localization of multimodal remote sensing images using CNN-based prediction of localization uncertainty. **Remote Sensing**, MDPI AG, v. 12, n. 4, p. 703, fev. 2020. DOI: 10.3390/rs12040703. Citado na página 50.

VULTUR, O. M. Performance analysis of “drive me” - a human robot interaction system. In: **2018 International Conference on Communications (COMM)**. [S.l.]: IEEE, 2018. DOI: 10.1109/iccomm.2018.8484759. Citado na página 33.

WALIA, E.; VERMA, V. Boosting local texture descriptors with log-gabor filters response for improved image retrieval. **International Journal of Multimedia Information Retrieval**, Springer Nature, v. 5, n. 3, p. 173–184, abr. 2016. DOI: 10.1007/s13735-016-0099-2. Citado 5 vezes nas páginas 66, 67, 93, 125 e 127.

WANG, B.; ZHANG, J.; LU, L.; HUANG, G.; ZHAO, Z. A uniform SIFT-like algorithm for SAR image registration. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 12, n. 7, p. 1426–1430, jul. 2015. DOI: 10.1109/lgrs.2015.2406336. Citado na página 56.

WANG, Q.; GAO, X.; WANG, F.; JI, Z.; HU, X. Feature point matching method based on consistent edge structures for infrared and visible images. **Applied Sciences**, MDPI AG, v. 10, n. 7, p. 2302, mar. 2020. DOI: 10.3390/app10072302. Citado na página 60.

WANG, Q.; ZHU, G.; YUAN, Y. Multi-spectral dataset and its application in saliency detection. **Computer Vision and Image Understanding**, Elsevier BV, v. 117, n. 12, p. 1748–1754, dez. 2013. DOI: 10.1016/j.cviu.2013.07.002. Citado na página 34.

WANG, R.; SHI, Y.; CAO, W. GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra. **Pattern Recognition Letters**, Elsevier

BV, v. 127, p. 11–17, nov. 2019. DOI: 10.1016/j.patrec.2018.11.001. Citado 2 vezes nas páginas 22 e 90.

WON, C. S. W.; PARK, D. K. P.; PARK, S.-J. P. Efficient use of MPEG-7 edge histogram descriptor. **ETRI Journal**, Wiley, v. 24, n. 1, p. 23–30, fev. 2002. DOI: 10.4218/etrij.02.0102.0103. Citado na página 92.

WU, Q.; XU, G.; CHENG, Y.; WANG, Z.; DONG, W.; MA, L. Robust and efficient multi-source image matching method based on best-buddies similarity measure. **Infrared Physics & Technology**, Elsevier BV, v. 101, p. 88–95, set. 2019. DOI: 10.1016/j.infrared.2019.05.020. Citado na página 76.

WU, Q.; ZHU, S. Multispectral image matching method based on histogram of maximum gradient and edge orientation. **IEEE Geoscience and Remote Sensing Letters**, Institute of Electrical and Electronics Engineers (IEEE), p. 1–5, 2021. DOI: 10.1109/lgrs.2021.3077688. Citado 9 vezes nas páginas 18, 60, 63, 66, 68, 87, 91, 93 e 104.

XAUD, M. F. S.; LEITE, A. C.; FROM, P. J. Thermal image based navigation system for skid-steering mobile robots in sugarcane crops. In: **2019 International Conference on Robotics and Automation (ICRA)**. [S.l.]: IEEE, 2019. DOI: 10.1109/icra.2019.8794354. Citado na página 20.

XU, G.; WU, Q.; CHENG, Y.; YAN, F.; LI, Z.; YU, Q. A robust deformed image matching method for multi-source image matching. **Infrared Physics & Technology**, Elsevier BV, v. 115, p. 103691, jun. 2021. DOI: 10.1016/j.infrared.2021.103691. Citado na página 60.

YI, K. M.; TRULLS, E.; LEPETIT, V.; FUA, P. LIFT: Learned invariant feature transform. In: **Computer Vision – ECCV 2016**. [S.l.]: Springer International Publishing, 2016. p. 467–483. DOI: 10.1007/978-3-319-46466-4_28. Citado na página 54.

YU, G.; ZHAO, S. A new feature descriptor for multimodal image registration using phase congruency. **Sensors**, MDPI AG, v. 20, n. 18, p. 5105, set. 2020. DOI: 10.3390/s20185105. Citado 6 vezes nas páginas 32, 66, 75, 96, 97 e 98.

ZAGORUYKO, S.; KOMODAKIS, N. Learning to compare image patches via convolutional neural networks. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2015. DOI: 10.1109/cvpr.2015.7299064. Citado 3 vezes nas páginas 50, 51 e 63.

ZHANG, Z.; YAN, C.; KREBS, C. J.; STENSETH, N. C. Ecological non-monotonicity and its effects on complexity and stability of populations, communities and ecosystems. **Ecological Modelling**, Elsevier BV, v. 312, p. 374–384, set. 2015. DOI: 10.1016/j.ecolmodel.2015.06.004. Citado na página 18.

ZHUANG, Y.; GAO, K.; MIU, X.; HAN, L.; GONG, X. Infrared and visual image registration based on mutual information with a combined particle swarm optimization – powell search algorithm. **Optik**, Elsevier BV, v. 127, n. 1, p. 188–191, jan. 2016. DOI: 10.1016/j.ijleo.2015.09.199. Citado na página 17.

ZITOVÁ, B.; FLUSSER, J. Image registration methods: a survey. **Image and Vision Computing**, Elsevier BV, v. 21, n. 11, p. 977–1000, out. 2003. DOI: 10.1016/s0262-8856(03)00137-9. Citado 3 vezes nas páginas 16, 48 e 60.

Apêndices

APÊNDICE A

Filtros Log-Gabor

Os filtros de Gabor, propostos no trabalho de Gabor (1946), consistem em um grupo de *wavelets*[1], em que cada *wavelet* é responsável por uma determinada frequência e orientação. Tais filtros têm sido comumente utilizados na literatura para extração de textura e detecção de bordas em imagens. Também, esses filtros permitem utilizar diferentes parâmetros, como os parâmetros de escala e orientação, os quais os tornam bastante úteis para a análise de textura em imagens. No entanto, os filtros de Gabor produzem respostas altas para entradas constantes ou de baixas frequências. Esse comportamento gera um problema, pois geralmente as respostas altas são esperadas apenas em bordas para extração de características e não no caso de regiões com valores de intensidade constantes (WALIA; VERMA, 2016; KOVESI, 2000).

Uma alternativa aos filtros de Gabor são os filtros de Log-Gabor, propostos no trabalho de Field (1987), que eliminam o problema de alta resposta para entradas contínuas ou de baixas frequências dos filtros de Gabor. A Figura 65 ilustra as diferenças das funções Gabor e Log-Gabor no que diz respeito à frequência de resposta, e observa-se que, os filtros Log-Gabor possuem uma baixa resposta para baixas frequências de entrada.

O filtro Log-Gabor, no domínio da frequência e utilizando coordenadas polares, é definido conforme a Equação (16):

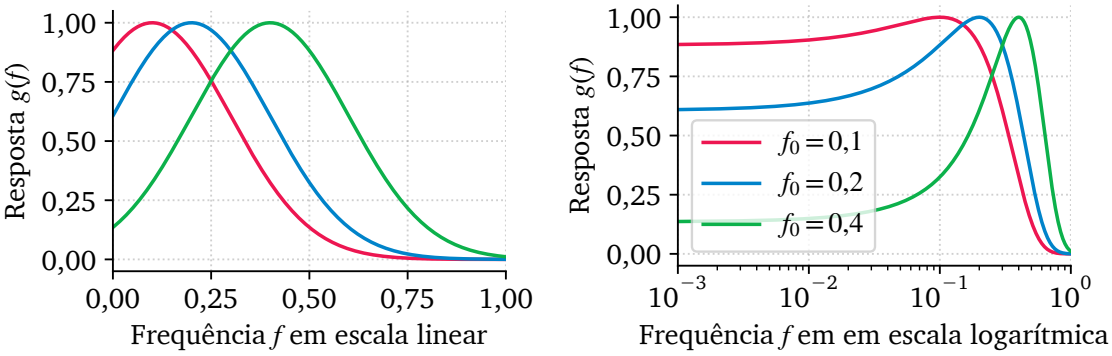
$$F_{s,\omega}(r,\theta) = \exp\left(-\frac{\log\left({}^{r}/_{f_0}\right)^2}{2\sigma_r^2}\right).\exp\left(-\frac{(\theta-\omega)^2}{2}\right), \tag{16}$$

em que r e θ representam o raio e o ângulo do filtro em coordenadas polares, ω consiste no ângulo de orientação do filtro, σ_r representa o desvio padrão gaussiano do filtro e f_0 representa a frequência central do filtro dada pela Equação (17):

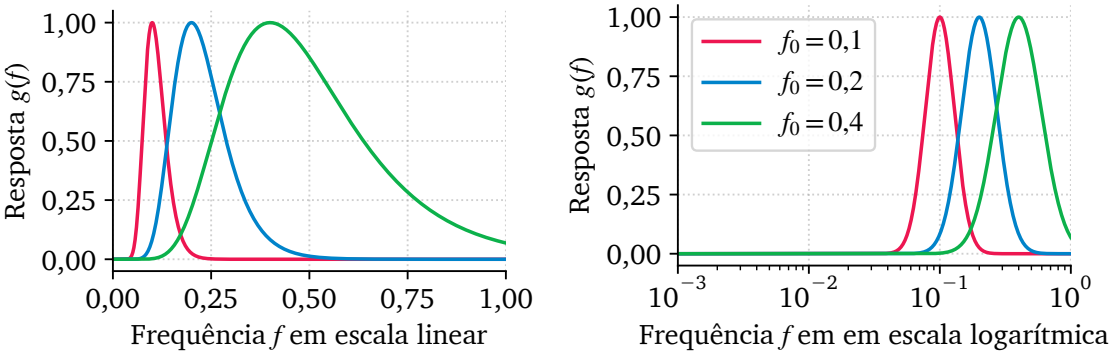
$$f_0 = \lambda\kappa^s, \tag{17}$$

em que λ consiste no comprimento de onda da menor escala do filtro, κ representa uma constante multiplicativa de escala entre filtros sucessivos e κ consiste na escala do filtro.

[1] Uma *Wavelet*, ou ondaleta, consiste em uma função capaz de representar qualquer outra função como superposição de ondas, de forma a decompô-las em diferentes níveis de frequência ou escala (ANTONINI et al., 1992).



(a) Respostas para o filtro Gabor. Função: $g(f) = \exp\left(-\frac{(f-f_0)^2}{2\sigma^2}\right)$.



(b) Respostas para o filtro Log-Gabor. Função: $g(f) = \exp\left(-\frac{\log(f-f_0)^2}{2\log(\sigma/f_0)^2}\right)$.

Figura 65 – Diferenças entre os filtros Gabor e Log-Gabor em relação à frequência de resposta.

Fonte: Adaptado de Field (1987).

Reescrevendo as Equações (16) e (17), tem-se:

$$F_{s,\omega}(r,\theta) = \exp\left(-\frac{\log\left({}^{r}/_{\lambda\kappa^s}\right)^2}{2\sigma_r^2}\right).\exp\left(-\frac{(\theta-\omega)^2}{2}\right), \tag{18}$$

A utilização dos filtros Log-Gabor é feita no domínio da frequência, dado que a construção desses filtros nesse domínio é mais simples. Então, para extrair as bordas de uma imagem I, essa imagem é transformada do domínio espacial para o domínio da frequência, por meio da Transformada Rápida de Fourier (FFT). Em seguida, é realizada a convolução da imagem I com o filtro $F_{s,\omega}$. Por fim, a imagem de resposta ao filtro Log-Gabor é obtida por meio do cálculo da inversa da Transformada Rápida de Fourier (IFFT), conforme a Equação (19):

$$R_{s,\omega} = |\,\mathrm{IFFT}(\mathrm{FFT}(I) * F_{s,\omega})|, \tag{19}$$

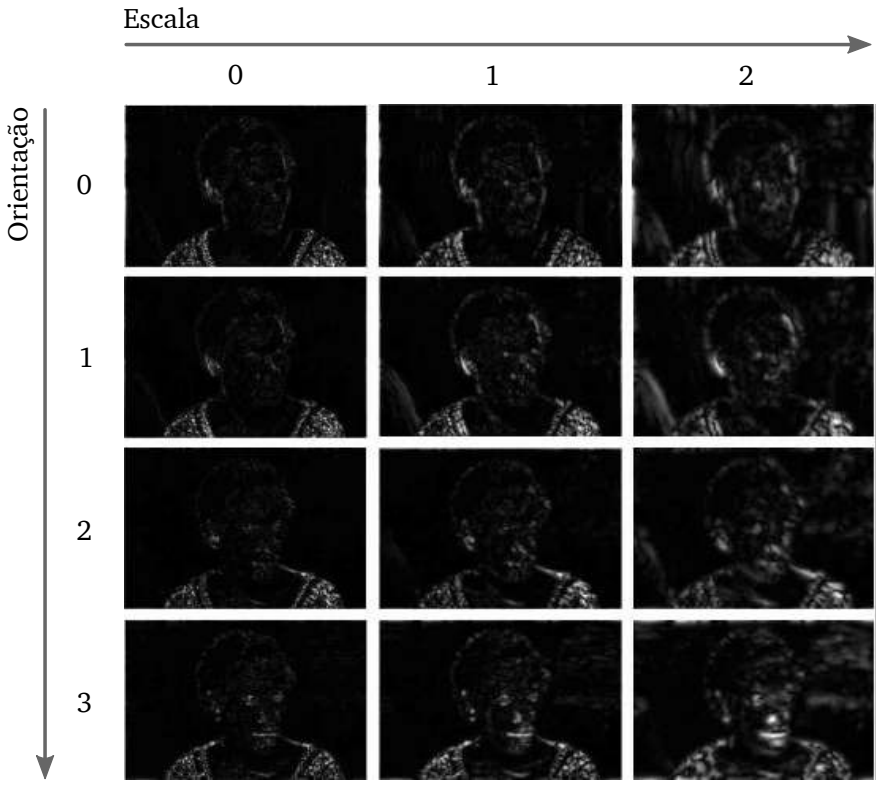
em que "||" representa o módulo (valor absoluto), FFT representa a Transformada Rápida de

Fourier e IFFT a sua função inversa. O filtro Log-Gabor é denotado por $F_{s,\omega}$ para a escala s e orientação ω (conforme a Equação (18)).

A Figura 66 exemplifica algumas imagens de resposta aos filtros Log-Gabor, representadas pelas bordas da imagem de entrada. Nesse exemplo, utilizou-se um conjunto de filtros formado pela associação de 4 valores de orientação (ω indo de 0 até 180 graus, igualmente divididos) por 3 valores de escala ($s \in \{0, 1, 2\}$, da Equação (18)), o que resulta em um conjunto de 12 filtros. Neste exemplo, foram definidos os parâmetros: $\lambda = 3{,}0$, $\kappa = 2{,}0$ e $\sigma_r = 0{,}65$.

As imagens de resposta aos filtros Log-Gabor podem, em alguns casos, ter resultados diferentes entre imagens obtidas de diferentes faixas do espectro eletromagnético. Tanto no espectro visível quanto no infravermelho podem ocorrer casos em que as bordas das imagens são similares, devido às diferenças na reflexão da luz para diferentes faixas do

(a) Imagem de entrada.



(b) Imagens de resposta aos filtros Log-Gabor.

Figura 66 – Exemplo de imagens de resposta aos filtros Log-Gabor.

Fonte: Adaptado de Walia e Verma (2016).

espectro eletromagnético. Esse fenômeno dificulta a extração de informações de bordas, sendo, portanto, uma limitação ao se utilizar informações de bordas das imagens VIS/IV. A Figura 67 ilustra essa diferença de bordas extraídas em uma imagem do espectro visível e uma imagem do espectro infravermelho para uma mesma cena.

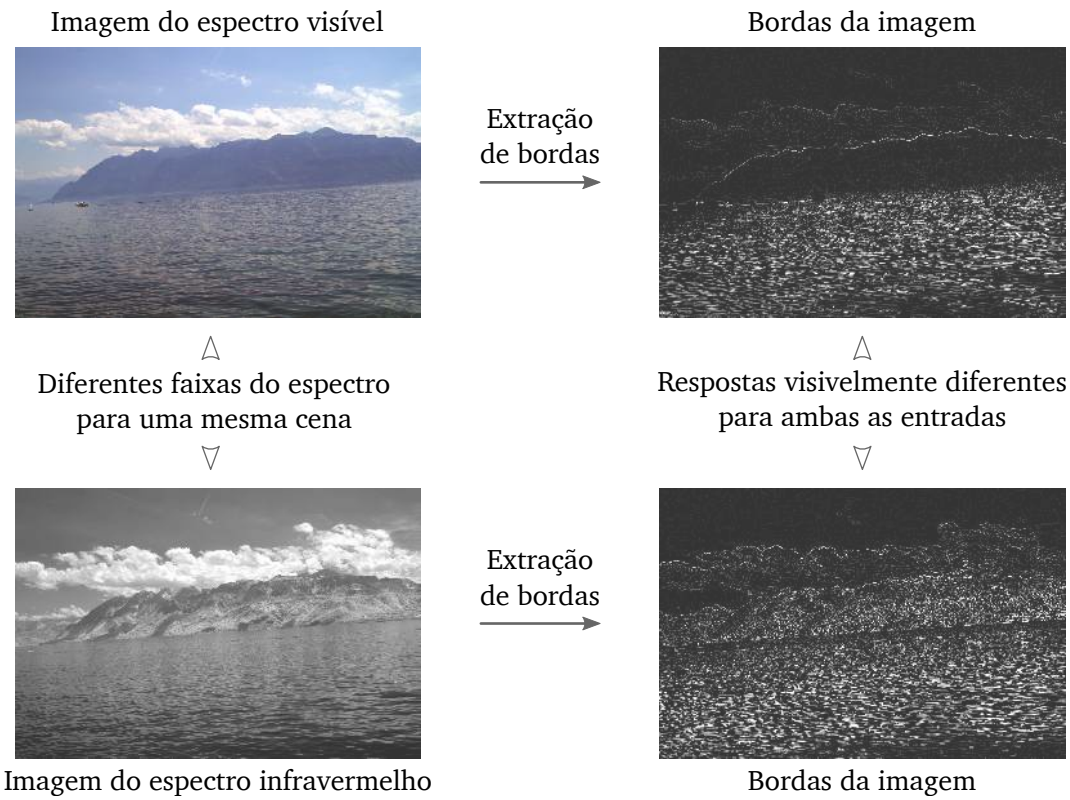


Figura 67 – Diferenças na extração de bordas entre uma imagem do espectro visível e uma imagem do espectro infravermelho para uma mesma cena.

APÊNDICE B

Métodos para descrição de características locais

B.1 Algoritmo SIFT

O algoritmo SIFT é composto por um método detector e um método descritor de características locais. Este algoritmo possui quatro etapas principais: (i) detecção de extremos, (ii) localização de pontos-chave, (iii) atribuição da orientação e (iv) descrição da característica local. As duas primeiras etapas correspondem ao seu método detector e as duas etapas seguintes correspondem ao seu método descritor.

A etapa de detecção de extremos tem o objetivo de encontrar os máximos e mínimos locais mediante uma filtragem em cascata dada pela função DoG (*Diferença-de-Gaussianas*). Primeiro, uma pirâmide de escala é construída conforme a Figura 68. Em várias escalas, é

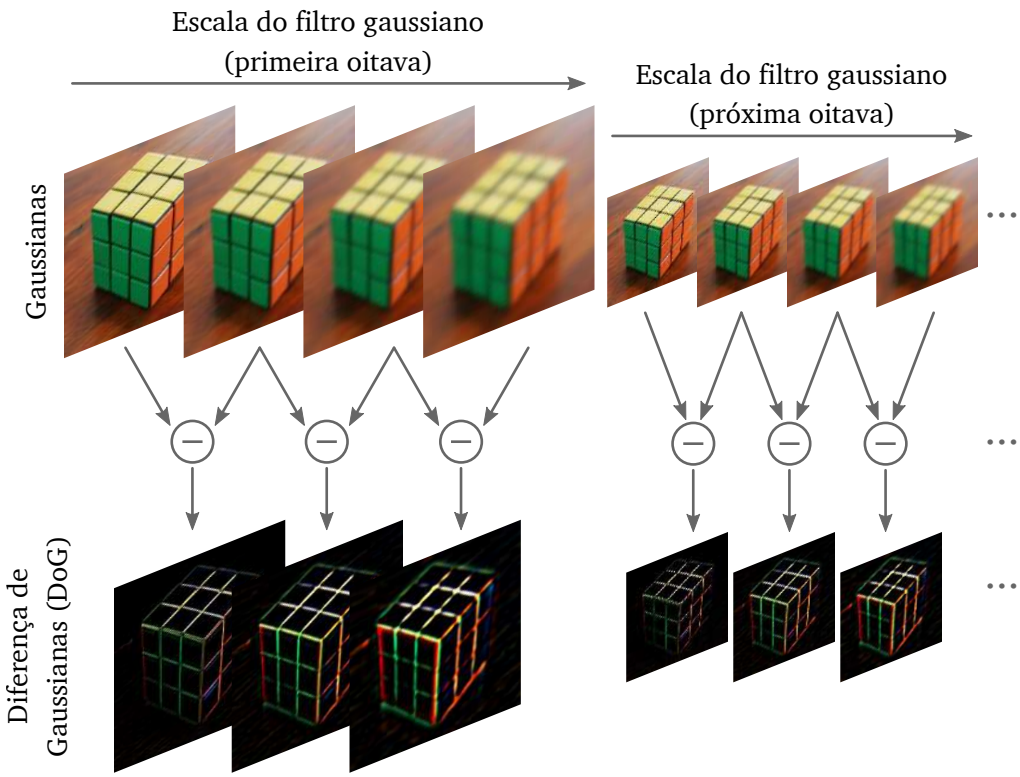


Figura 68 – Pirâmide de escala montada pelo algoritmo SIFT.

Fonte: Adaptado de Lowe (2004).

calculada a convolução entre a imagem original $I(x,y)$ com o operador DoG, que permite detectar variações nos valores de intensidade, como bordas. A função DoG é expressa pela Equação (20):

$$DoG_{\sigma}(x,y) = [G_{\sigma k}(x,y) - G_{\sigma}(x,y)] * I(x,y), \tag{20}$$

em que σ representa a escala atual do filtro gaussiano, k consiste em uma constante multiplicativa e $G_{\sigma}(x,y)$ é a função Gaussiana dada pela Equação (21):

$$G_{\sigma}(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right). \tag{21}$$

Após a geração da pirâmide, utilizando-se as imagens *Diferença-de-Gaussianas*, cada ponto é comparado com seus vizinhos, tanto na imagem atual quanto nas imagens adjacentes. Então, faz-se a seleção dos pontos candidatos a serem pontos-chave, como ilustrado na Figura 69. Um ponto é selecionado se seu valor de intensidade for maior ou menor do que todos os seus 26 vizinhos.

Na etapa de localização de pontos-chave, todos os pontos detectados na etapa anterior são avaliados. São rejeitados os pontos extremos com baixo contraste e os pontos mal localizados ao longo de uma borda. Todos os pontos restantes ao final desta etapa são definidos como pontos-chave.

A etapa de atribuição da orientação possibilita alcançar robustez à rotação da imagem. A escala em que o ponto-chave foi detectado é usada para selecionar a imagem Gaussiana de escala mais próxima, garantindo que os cálculos sejam realizados independente da escala. Para os pontos da região que contém a característica local, são calculadas a magnitude do gradiente $m(x,y)$, dada pela Equação (22), e a orientação $o(x,y)$, dada pela Equação (23). O tamanho desta região é determinado pela escala do ponto-chave.

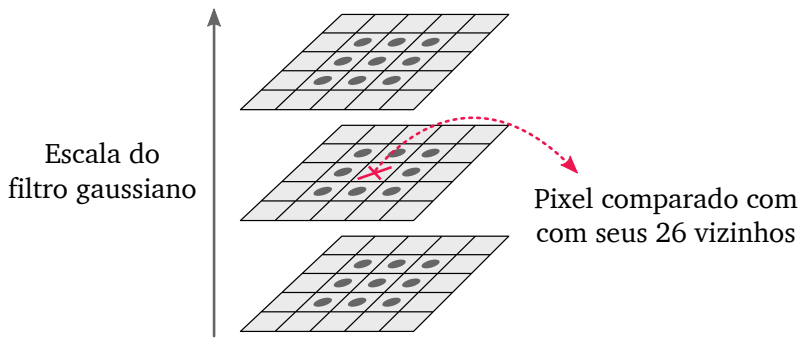


Figura 69 – Detecção de máximos e mínimos locais do algoritmo SIFT.

Fonte: Adaptado de Lowe (2004).

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}, \tag{22}$$

$$o(x,y) = \tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right). \tag{23}$$

É construído então um histograma de orientações para os pontos desta região. São definidos 36 níveis de $o(x,y)$ que abrangem 360° de orientações. Cada amostra na vizinhança do ponto-chave, adicionada ao histograma, é ponderada com base na magnitude $m(x,y)$ e em uma janela gaussiana circular em torno do ponto-chave. Picos na orientação do histograma correspondem a direções dominantes. Os picos que correspondem a níveis até 80% do maior pico, são usados para definir a orientação do ponto-chave.

Na etapa de descrição de características, são calculados os descritores que representam suas respectivas regiões. Para tanto, é selecionada a imagem $L(x,y)$ com a oitava referente ao ponto-chave. Em seguida, as orientações dos gradientes são rotacionadas em relação à orientação do ponto-chave. Após isto, é montada uma matriz de 16×16 pixels divididos em 4 regiões em torno do ponto-chave. Em cada região, é calculado um histograma de 8 direções, com base na magnitude dos pixels. É aplicada uma ponderação gaussiana para dar menos ênfase aos gradientes mais afastados do ponto-chave. Então, o descritor é construído a partir desse histograma. Este processo é ilustrado na Figura 70. Ao final, o descritor é normalizado para criar um descritor mais robusto às mudanças de iluminação.

O algoritmo SIFT é tradicionalmente reconhecido na literatura. Ele tem sido usado em diversas aplicações, incluindo segmentação de vídeos (SOUZA et al., 2014), em que é utilizado para identificar e rastrear pontos de interesse ao longo de sequências de frames, facilitando a análise e compreensão de cenas complexas.

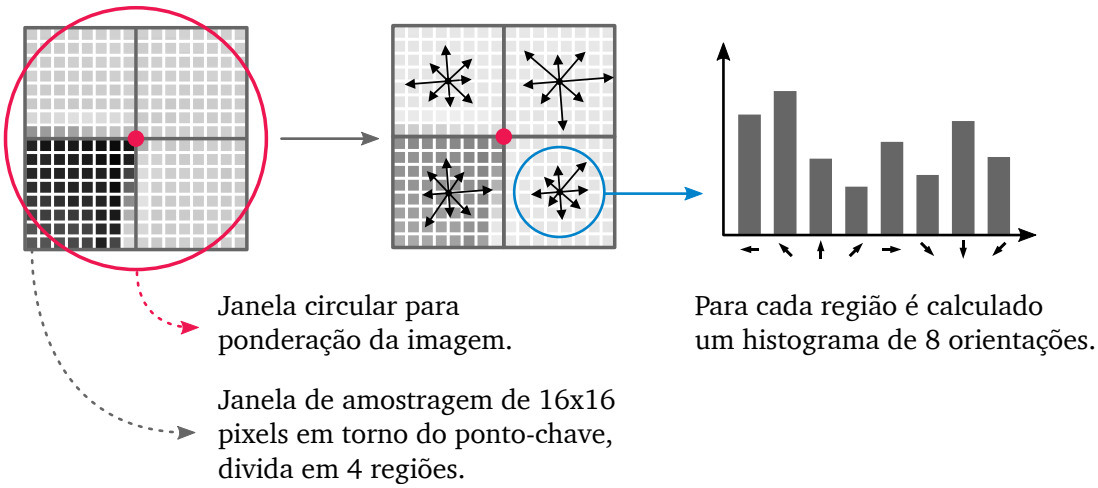


Figura 70 – Etapas para a construção do descritor SIFT.

Fonte: Adaptado de Lowe (2004).

B.2 Algoritmo FAST

O algoritmo FAST (*Features from Accelerated Segment Test*), proposto por Rosten e Drummond (2006), consiste em um método de detecção de cantos, que possui um desempenho superior aos métodos que se baseiam na *Diferença-de-Gaussianas*. Para definir se um ponto é realmente um canto, é utilizada uma circunferência de 16 pixels. Se 12 pixels contíguos sobre a circunferência possuem um brilho superior ao brilho do ponto em teste, somado a um valor limiar, ou se estes pixels são mais escuros, então o ponto em teste é definido como canto. A Figura 71 ilustra a detecção de canto do algoritmo FAST.

B.3 Algoritmo EOH

O algoritmo descritor EOH (*Edge Oriented Histogram*), proposto por Aguilera et al. (2012), é baseado no método descritor EHD (*Edge Histogram Descriptor*), definido no padrão MPEG-7 (SIKORA, 2001). Para realizar o cálculo do descritor, a imagem de entrada é convertida em uma imagem de bordas, por meio da utilização do detector Canny (CANNY, 1987). Visto que a imagem é representada mediante suas bordas, os descritores são calculados para cada característica local, como ilustrado na Figura 72.

Primeiro, uma janela de tamanho $n \times n$[1] é posicionada no centro da característica local detectada (seu ponto-chave). Essa janela é subdividida em 4×4 sub-regiões (total de 16 sub-regiões), e para cada sub-região é calculado o histograma de orientações. O histograma representa a distribuição espacial de quatro filtros de bordas direcionais (filtros Sobel) e um filtro não-direcional. Cada posição do histograma representa os ângulos de $0, 45, 90, 135$ graus e, adicionalmente, uma posição do histograma representa ausência de orientação (n.o.) para corresponder às áreas que não contêm bordas.

Canto detectado no pixel P, pois a circunferência passa sobre 12 pixels mais claros que o pixel P.

Figura 71 – Exemplo de funcionamento do detector de cantos FAST.

Fonte: Adaptado de Rosten e Drummond (2006).

[1] No trabalho de Aguilera et al. (2012) é definido o tamanho da janela como 80.

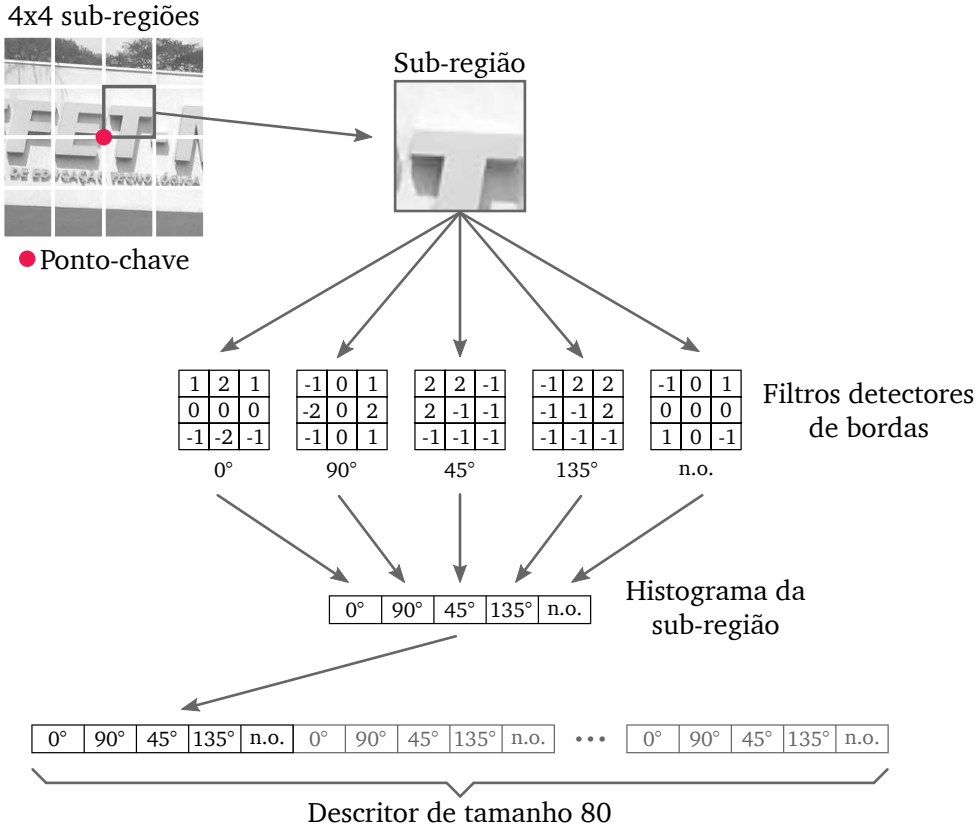


Figura 72 – Construção do descritor EOH.

Fonte: Adaptado de Aguilera et al. (2012).

Todo pixel de cada sub-região contribui para uma posição no histograma conforme os cinco filtros: o filtro com o maior valor de resposta é utilizado como critério para uma determinada posição do histograma. Após o processamento de todos os elementos das 16 sub-regiões obtém-se um descritor de tamanho 80 (referente aos 5 filtros para cada uma das 16 sub-regiões). Este descritor final então é normalizado, sendo este o descritor para a característica local detectada.

B.4 Algoritmo TFeat

O algoritmo TFeat, proposto por Balntas et al. (2016b), consiste em uma arquitetura de CNN desenvolvida para produzir descritores de imagens obtidas do espectro visível. Assim como o TFeat, outras soluções baseadas em CNN seguem uma mesma ideia de funcionamento: primeiro, a rede é treinada com grupos de janelas de imagens e, após o treinamento, a rede funciona da mesma forma que o algoritmo SIFT, produzindo um descritor no final de sua camada, conforme exemplificado na Figura 73. Dessa forma, essas soluções podem ser utilizadas como uma forma de substituição aos algoritmos tradicionais, como SIFT e SURF em processos de correspondência entre imagens. Os detalhes de cada camada dessa arquitetura são apresentados no Quadro 3.

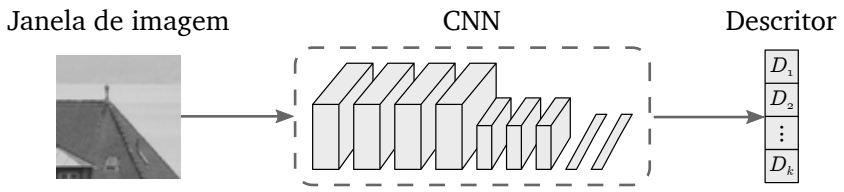


Figura 73 – Cálculo do descritor TFeat.

Durante a etapa de treinamento, são utilizadas três CNNs, com pesos compartilhados. Essas CNNs recebem um conjunto de janelas de imagens na forma $\{J_a, J_p, J_n\}$, em que J_a é intitulada de janela âncora, J_p de janela positiva e J_n negativa, conforme ilustrado na Figura 27. O par de janelas J_a e J_p correspondem a um mesmo ponto no mundo real, porém obtidos sob diferentes pontos de vista (par de janelas similares). Já a janela n corresponde a um diferente ponto no mundo real. A ideia do treinamento desse conjunto é tornar as janelas J_a e J_p mais próximas, no espaço de características, e deixar as janelas J_a e J_n mais distantes. A Equação (24) apresenta a função de perda utilizada para o treinamento da rede:

$$L(\delta_p, \delta_n) = \max(0, \mu + \delta_p - \delta_n), \tag{24}$$

em que μ representa uma constante de margem, δ_p corresponde à norma L_2 entre os descritores do par de janelas J_a e J_p, $\delta_p = \left\|D(J_a) - D(J_p)\right\|_2$ e δ_n corresponde à norma L_2 entre os descritores do par de janelas J_a e J_n, $\delta_n = \|D(J_a) - D(J_n)\|_2$.

Camada	**Tipo de camada**	**Característica da camada**
1	Convolução	32 filtros de tamanho 7×7.
2	Ativação	Função de ativação tangente hiperbólica (tanh).
3	Agrupamento	Valores máximos em blocos de tamanho 2×2.
4	Convolução	64 filtros de tamanho 6×6.
5	Ativação	Função de ativação tangente hiperbólica (tanh).
6	Totalmente conectada	Tamanho da camada de 128 valores.
7	Ativação	Função de ativação tangente hiperbólica (tanh).

Quadro 3 – Camadas da arquitetura TFeat.

APÊNDICE C

Homografia 2D

A Homografia 2D, ou transformação projetiva planar, consiste em uma transformação que mapeia os pontos de um plano para outro plano (HARTLEY, 2004). A Figura 74 ilustra este processo, em que o ponto P no plano π é mapeado para o seu ponto correspondente P' no plano π'.

Um ponto bidimensional em uma imagem $P = (x, y)$ pode ser representado em uma coordenada tridimensional na forma $P = (x_1, x_2, x_3)$, em que $x = x_1/x_3$ e $y = x_2/x_3$. Esta representação recebe o nome de *coordenada homogênea*. A homografia entre os planos pode ser escrita em coordenadas homogêneas como $P'_i = HP_i$ (Equação (25)), em que H é a matriz de homografia que descreve o mapeamento de um conjunto de pontos correspondentes $P_i \longleftrightarrow P'_i$ entre dois planos. A matriz de homografia H é representada por uma matriz 3×3 não-singular.

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \tag{25}$$

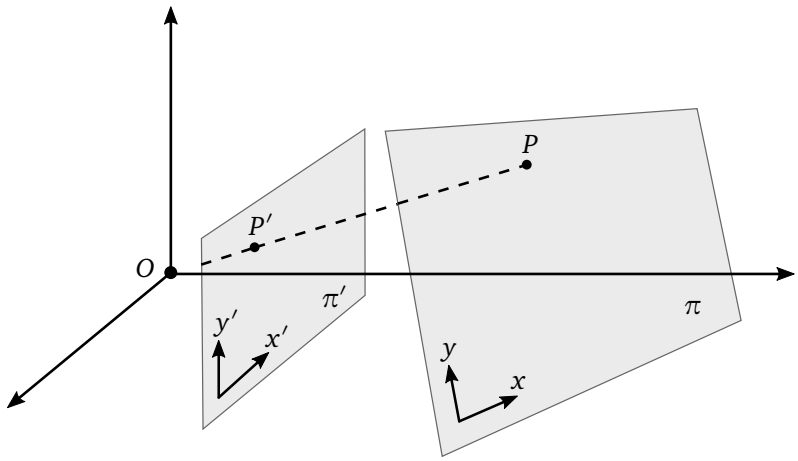


Figura 74 – Mapeamento de pontos entre dois planos.

Fonte: Hartley (2004).

APÊNDICE D

Tabelas de resultados do método MF-Net

NNDR	Arquiteturas		
	(i)	(ii)	(iii)
0,80	0,460±0,005	**0,550±0,006**	0,477±0,005
0,85	0,482±0,006	**0,578±0,006**	0,502±0,005
0,90	0,495±0,007	**0,597±0,007**	0,520±0,005
0,95	0,487±0,007	**0,592±0,008**	0,515±0,006
1,00	0,434±0,006	**0,504±0,007**	0,446±0,006

Tabela 1 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam.

NNDR	Arquiteturas		
	(i)	(ii)	(iii)
0,80	0.314±005	**0.403±011**	0.325±005
0,85	0.338±005	**0.431±012**	0.350±005
0,90	0.358±005	**0.454±012**	0.371±005
0,95	0.385±006	**0.483±013**	0.399±006
1,00	0.361±006	**0.440±011**	0.372±005

Tabela 2 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL.

NNDR	Arquiteturas			
	PN-Net	TFeat	Q-Net	MF-Net (Proposto)
0,80	0,462±0,013	0,460±0,005	0,489±0,005	**0,550±0,006**
0,85	0,487±0,014	0,482±0,006	0,514±0,005	**0,578±0,006**
0,90	0,499±0,016	0,495±0,007	0,532±0,006	**0,597±0,007**
0,95	0,490±0,017	0,487±0,007	0,526±0,006	**0,592±0,008**
1,00	0,433±0,015	0,434±0,006	0,454±0,006	**0,504±0,007**

Tabela 3 – Valores médios e desvio padrão para a medida F_1 na base de dados Potsdam.

NNDR	**Arquiteturas**			
	PN-Net	**TFeat**	**Q-Net**	**MF-Net (Proposto)**
0,80	0,319±0,014	0,314±0,005	0,338±0,005	**0,403±0,011**
0,85	0,343±0,014	0,338±0,005	0,363±0,005	**0,431±0,012**
0,90	0,362±0,015	0,358±0,005	0,384±0,005	**0,454±0,012**
0,95	0,390±0,016	0,385±0,006	0,412±0,006	**0,483±0,013**
1,00	0,364±0,013	0,361±0,006	0,383±0,006	**0,440±0,011**

Tabela 4 – Valores médios e desvio padrão para a medida F_1 na base de dados EPFL.

APÊNDICE E

Parâmetros utilizados nos experimentos

Parâmetro	Valor	Descrição
nFeatures	0	Quantidade de características para retornar. As características são classificadas por meio do contraste local. Nesta implementação, o valor nulo permite o algoritmo retornar todas as características.
nOctaveLayers	3	Quantidade de camadas em cada oitava. A quantidade de oitavas é calculada automaticamente a partir do tamanho da imagem de entrada.
contrastThreshold	0,04	Limiar de contraste utilizado para filtrar as características de baixo contraste. Quanto maior for este limiar, menor será a quantidade de características produzidas pelo detector.
edgeThreshold	3,0	Limiar utilizado para filtrar as características ao longo de uma borda. Quanto maior for este limiar, maior será a quantidade de características produzidas pelo detector.
sigma	1,6	O valor de σ da função Gaussiana aplicada à imagem de entrada na primeira oitava. Se a de entrada imagem estiver desfocada, pode-se reduzir esse valor.

Quadro 4 – Parâmetros do algoritmo SIFT.

Parâmetro	Valor	Descrição
threshold	64	Limiar da diferença de intensidade de pixels entre o pixel central do círculo definido.

Quadro 5 – Parâmetros do algoritmo FAST.

Parâmetro	Valor	Descrição
windowSize	80	Tamanho da janela em torno do ponto-chave para realizar a descrição.

Quadro 6 – Parâmetros dos algoritmos MFD e LGHD.

APÊNDICE F

Implementações utilizadas

Método	Trabalho	Implementação
SIFT	Lowe (2004)	<https://github.com/opencv/opencv>
LGHD	Aguilera, Sappa e Toledo (2015)	<https://github.com/ngunsu/LGHD>
PN-Net	Balntas et al. (2016a)	<https://github.com/vbalnt/pnnet>
TFeat	Balntas et al. (2016b)	<https://github.com/vbalnt/tfeat>
MFD	Nunes e Pádua (2017)	<https://github.com/cfgnunes/mfd>
Q-Net	Aguilera et al. (2017)	<https://github.com/ngunsu/qnet>
RIFT	Li, Hu e Ai (2020)	<https://github.com/cfgnunes/rift>

Quadro 7 – Implementações utilizadas nos experimentos deste trabalho.