UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL

DIULHIO CANDIDO DE OLIVEIRA

UMA ABORDAGEM PARA DETECÇÃO DE PESSOAS EM IMAGENS DE VEÍCULOS AÉREOS NÃO-TRIPULADOS

DISSERTAÇÃO

CURITIBA

2016

DIULHIO CANDIDO DE OLIVEIRA

UMA ABORDAGEM PARA DETECÇÃO DE PESSOAS EM IMAGENS DE VEÍCULOS AÉREOS NÃO-TRIPULADOS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de "Mestre em Ciências" – Área de Concentração: Engenharia de Computação.

Orientador: Marco Aurelio Wehrmeister

CURITIBA

2016

Dados Internacionais de Catalogação na Publicação

O48ab
2016

Oliveira, Diulhio Candido de
Uma abordagem para detecção de pessoas em imagens de veículos aéreos não-tripulados / Diulhio Candido de Oliveira .-- 2016.
103 f.: il.; 30 cm.

Texto em português, com resumo em inglês.
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Área de concentração: Engenharia de computação, Curitiba, 2016.
Bibliografia: f. 95-103.

1. Processamento de imagens - Técnicas digitais. 2. Aeronave não-tripulada. 3. Sistemas de reconhecimento de padrões. 4. Segmentação de imagens. 5. Termografia. 6. Redes neurais (Computação). 7. Convoluções (Matemática). 8. Visão por computador. 9. Métodos de simulação. 10. Operações de busca e salvamento. 11. Engenharia elétrica - Dissertações. I. Wehrmeister, Marco Aurelio, orient. II. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial. III. Título.

CDD: Ed. 22 -- 621.3

Biblioteca Central da UTFPR, Câmpus Curitiba

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Curitiba

Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial

Título da Dissertação Nº. ____

Uma Abordagem para Detecção de Pessoas em Imagens de Veículos Aéreos não Tripulados

por

Diulhio Candido de Oliveira

Orientador: Prof. Dr. Marco Aurelio Wehrmeister (*UTFPR*)

Esta dissertação foi apresentada como requisito parcial à obtenção do grau de MESTRE EM CIÊNCIAS – Área de Concentração: **Engenharia de Automação e Sistemas** do Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial – CPGEI – da Universidade Tecnológica Federal do Paraná – UTFPR, às **10h** do dia 14 de junho 2016. O trabalho foi aprovado pela Banca Examinadora, composta pelos professores doutores:

Prof. Dr. Marco Aurelio wehrmeister
(Presidente – UTFPR)

Prof. Dr. Hugo Vieira Neto
(UTFPR)

Prof. Dr. Alceu de Souza Britto Jr.
(PUC-PR)

Visto da coordenação:

Prof. Jean Carlos Cardozo da Silva, Dr.
(Coordenador do CPGEI)

A Folha de Aprovação assinada encontra-se na Coordenação do Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI).

Dedico à minha família e amigos!

AGRADECIMENTOS

Primeiramente, agradeço à minha querida mãe Jandira Candido de Oliveira, por todo apoio, carinho e compreensão. Agradeço por todo o seu esforço e dedicação, coisas que fizeram de mim o homem que sou hoje, obrigado pelo seu exemplo, com certeza você sempre será o pilar principal de qualquer sucesso de minha vida.

Agradeço ao meu orientador e com certeza amigo Marco Aurélio Wehrmeister, por toda confiança, sabedoria e amizade, além de estar sempre disponível para ajudar, até mesmo quando os prazos eram curtos! Espero que nossa parceria perdure nos trabalhos futuros.

A minha amada namorada Fabiane, agradeço por todo amor e compreensão, mesmo em momentos que eu não merecia, sempre esteve comigo. Espero que nosso amor e parceria dure eternamente.

Agradeço a toda minha família, com certeza vocês tem papel importante nisso. Especialmente aos meus padrinhos Nacir e Maria do Carmo, com certeza, meus segundos pais. Também ao meu padrasto Marcelo que sempre me auxiliou e juntamente com a minha mãe Jandira não mediu esforços para me proporcionar as melhores oportunidades.

Aos amigos Albino, Thiago Moreno, Maurício, Henrique e muitos outros pela amizade que dura e desde a graduação, além de todos os momentos de descontração.

Agradeço também a ajuda de todo o pessoal do laboratório BIOINFO, pelo conhecimento e equipamentos que me auxiliaram nos resultados obtidos neste trabalho.

A Deus, criador de tudo, por ter me dado as oportunidades, saúde e força para alcançar meus objetivos.

A CAPES e ao CNPq pelo suporte financeiro para a realização deste trabalho.

E também a todos que de alguma maneira me ajudaram e contribuíram com este sonho, muito obrigado!

RESUMO

OLIVEIRA, Diulhio Candido de. UMA ABORDAGEM PARA DETECÇÃO DE PESSOAS EM IMAGENS DE VEÍCULOS AÉREOS NÃO-TRIPULADOS. 105 f. Dissertação – Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2016.

Este trabalho tem como objetivo propor um método reconhecimento de pessoas em imagens aéreas obtidas a partir de Veículos Aéreos Não Tripulados de pequeno porte. Esta é uma aplicação de grande interesse, pois pode ser inserida em diversas situações tanto civis quanto militares como, por exemplo, missões de busca e salvamento. O uso de Veículos Aéreos Não Tripulados autônomos tende a aumentar com o barateamento desta tecnologia. Assim, esta tecnologia pode sobressair sobre outras utilizadas atualmente, como satélites e voos com grandes aeronaves. Para o reconhecimento de pessoas em imagens aéreas de forma autônoma, este trabalho propõe métodos na forma de Sistemas de Reconhecimento de Padrões (SRP) aplicados ao reconhecimento de imagens. Para este métodos, foram testadas quatro técnicas de aprendizado de máquina: Redes Neurais Convolucionais, HOG+SVM, Cascata Haar e Cascata LBP. Além disso, a fim de possibilitar o reconhecimento de pessoas em imagens aéreas em tempo real, foram testadas e avaliadas técnicas de detecção e segmentação de objetos: Mapa de Saliências e o Processamento de Imagens Térmicas de baixa resolução (PIT). Neste trabalho foram avaliadas as taxas de reconhecimento dos SRPs, além do seu tempo de processamento em um sistema embarcado de baixo custo e em uma Base de Controle Móvel (BCM). Os resultados de reconhecimento mostraram a efetividade das Redes Neurais Convolucionais, com uma acurácia de 0,9971, seguido do HOG+SVM com 0,9236, Cascata Haar com 0,7348 e por fim, Cascata LBP com 0,6615. Em situações onde foi simulado a oclusão parcial, as Redes Neurais Convolucionais atingiram Sensibilidade média 0,72, HOG+SVM de 0,50 e as Cascatas 0,20. Nos experimentos com os SRPs (algoritmos de segmentação e detecção juntamente com as técnicas de reconhecimento), o Mapa de Saliências pouco afetou as taxas de reconhecimento, quais ficaram muito próximas das obtidas no experimentos de reconhecimento. Já o Processamento de Imagens Térmicas de baixa resolução apresentou dificuldades em executar uma segmentação precisa, obtendo imagens com variação na translação, prejudicando a precisão do sistema. Por fim, este trabalho propõe uma nova abordagem para implementação de um SRP para o reconhecimento de pessoas em imagens áreas, utilizando Processamento de Imagens Térmicas juntamente com as Redes Neurais Convolucionais. Este SRP une altas taxas de reconhecimento com desempenho computacional de ao menos 1 fps na plataforma BCM.

Palavras-chave: Reconhecimento de pessoas, Imagens aéreas, Redes Neurais Convolucionais, Reconhecimento em Tempo-Real, Resgate e Salvamento

ABSTRACT

OLIVEIRA, Diulhio Candido de. AN APPROACH TO PEOPLE DETECTION IN UNMANNED AERIAL VEHICLES IMAGES. 105 f. Dissertação – Programa de Pós-graduação em Engenharia Elétrica e Informática Industrial, Universidade Tecnológica Federal do Paraná. Curitiba, 2016.

This work aims to propose a method for people recognition in Small Unmanned Aerial Vehicles aerial imagery. This is an application of high interest, it can be used in several situations, both civilian and military, as search and rescue missions. The use of Unmanned Aerial Vehicles autonomously tends to increase with the cheapening of this technology, supporting search and rescue missions. Thus, this technology can excel over others currently used, as satellites and flights with large aircraft. For autonomous people recognition, this work proposes new methods as Pattern Reconigition System (PRS) applied to image recognition, applying it in aerial images. Four Pattern Reconigition techniques were tested: Convolutional Neural Networks, HOG+SVM, Haar Cascade and LBP Cascade. Furthermore, in order to achieve recognition of people in aerial images in Real-Time target and detection techniques were tested and evaluated: Saliency Maps and Low-resolution Thermal Image Processing (TIP). In this work were considered recognition rates of the methods and their computational time in a low-cost embedded system and a Mobile Ground Control Station (MGCS). The recognition results shown the Convolutional Neural Network potential, where an accuracy of 0.9971 was achieved, followed by HOG + SVM with 0.9236, Haar Cascade with 0.7348 and LBP Cascade with 0.6615. In situations simulated partial occlusion, where was the CNNs achieved average Sensitivity of 0.72, HOG+SVM with 0.50 and both Cascades 0.20. In experiments with PRS (targeting and detection algorithms with the recognition techniques), the Saliency Map had little influence in recongition rates, it was close to the rates achieved in recognition experiments. While the Low-resolution Thermal Image Processing had difficulties in segmentation process, where translation variantions occured, it harmed the system precision. Lastly, this work proposes a new approach for PRS implementation for people recognition in aerial imagery, using TIP with CNN. This PRS combines high rates of recognition with an computational performace of, at least, 1 fps in MGCS plataform.

Keywords: People Recognition, Aerial imagery, Convolutional Neural Networks, Real-Time recognition , Search and Rescue

LISTA DE FIGURAS

FIGURA 1 – Etapas de um Sistema de Reconhecimento de Padrões 20
FIGURA 2 – Processo de criação do mapa de saliências 23
FIGURA 3 – Características Haar 27
FIGURA 4 – Exemplo de características Haar em uma imagem 27
FIGURA 5 – Pontos utilizados para cálculo de intensidade em imagens integrais 28
FIGURA 6 – Processo de extração do LBP 30
FIGURA 7 – Características HOG 31
FIGURA 8 – Classificadores em cascata 32
FIGURA 9 – Funcionamento do *boosting* com três classificadores 33
FIGURA 10 – Hiperplano produzido por um SVM para separar duas classes. 34
FIGURA 11 – Distribuição entre duas classes não lineares. 35
FIGURA 12 – Exemplo de uma CNN 37
FIGURA 13 – Exemplo de um filtro de convolução 38
FIGURA 14 – Exemplo de uma camada de *pooling* 39
FIGURA 15 – Exemplo de uma camada de *pooling* máximo 39
FIGURA 16 – Diagrama de Blocos empregado na metodologia. 52
FIGURA 17 – Raspberry Pi 2 *Model B* v1.1 54
FIGURA 18 – Câmeras utilizadas no trabalho 56
FIGURA 19 – Imagens da base GMVRT-v1 57
FIGURA 20 – Imagens da base GMVRT-v2 57
FIGURA 21 – Imagens de teste da base GMVRT-v2 58
FIGURA 22 – Imagens coletadas da base UCF-ARG *Data Set*. 58
FIGURA 23 – Imagens coletadas para os testes dos SRPs. 59
FIGURA 24 – Processo de *squash* das imagens da base GMVRT-v1. 61
FIGURA 25 – Exemplos do Processo de Aumento de Dados. 62
FIGURA 26 – Arquitetura da CNN1 utilizada no trabalho. 65
FIGURA 27 – Arquitetura da CNN2 utilizada no trabalho. 67
FIGURA 28 – Exemplo de Gráfico ROC. 70
FIGURA 29 – Amostras do experimento de oclusão. 71
FIGURA 30 – Sistemas de Reconhecimento de Padrões utilizados neste trabalho. 72
FIGURA 31 – Treinamento da CNN1. 76
FIGURA 32 – Treinamento da CNN2. 76
FIGURA 33 – Gráfico ROC dos resultados dos testes de reconhecimento. 79

LISTA DE TABELAS

TABELA 1 – Técnicas de segmentação e detecção do método RUQ. 45
TABELA 2 – Número de imagens após o PAD 62
TABELA 3 – Distribuição das bases de imagem através do *Holdout*. 63
TABELA 4 – Detalhes das camadas de convolução da arquitetura da CNN1. 66
TABELA 5 – Detalhes das camadas de convolução da arquitetura da CNN2. 67
TABELA 6 – Estrutura da Matriz de Confusão. 69
TABELA 7 – Matriz de confusão da Cascata Haar. 77
TABELA 8 – Matriz de confusão da Cascata LBP. 77
TABELA 9 – Matriz de confusão de HOG + SVM. 77
TABELA 10 – Matriz de confusão da CNN1. 77
TABELA 11 – Matriz de confusão da CNN2. 78
TABELA 12 – Medidas extraídas a partir das Matrizes de Confusão. 78
TABELA 13 – Desempenho dos classificadores em imagens com oclusão parcial superior. 80
TABELA 14 – Desempenho dos classificadores em imagens com oclusão parcial inferior. 80
TABELA 15 – Desempenho dos classificadores em imagens sem oclusão parcial e média dos casos de oclusão. 81
TABELA 16 – Número de objetos encontrados por cada técnica de segmentação e detecção. 83
TABELA 17 – Matriz de confusão da CNN1 com Mapa de Saliências e Proc. Imagem Térmica. 83
TABELA 18 – Matriz de confusão da CNN2 com Mapa de Saliências e Proc. Imagem Térmica. 83
TABELA 19 – Matriz de confusão do HOG + SVM com Mapa de Saliências e Proc. Imagem Térmica. 84
TABELA 20 – Matriz de confusão da Cascata Haar com Mapa de Saliências e Proc. Imagem Térmica. 84
TABELA 21 – Matriz de confusão da Cascata LBP com Mapa de Saliências e Proc. Imagem Térmica. 84
TABELA 22 – Medidas retiradas a partir das Matrizes de Confusão dos testes dos SRPs. 84
TABELA 23 – Desempenho Computacional dos métodos de detecção e classificação executados no Raspberry Pi. 87
TABELA 24 – Desempenho Computacional Total dos métodos executados no Raspberry Pi. 87
TABELA 25 – Desempenho Computacional dos métodos de detecção e classificação métodos executados na BCM. 88
TABELA 26 – Desempenho Computacional Total dos métodos executados na BCM. .. 88

LISTA DE SIGLAS

AdaBoost	*Adaptative Boosting*
BLAS	*Basic Linear Algebra Subprograms*
CNN	*Convolutional Neural Networks*
CPGEI	Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial
DPM	*Discriminately Trained Models*
ECTM	Estação de Controle Terrestre Móvel
FPS	*Frames per second*
FWIR	*Far Wavelenght Infrared*
GPS	*Global Position System*
HOG	*Histogram of Oriented Gradient*
ICF	*Integral Channel Features*
k-NN	*k-Nearest-Neighbors*
LASER	Laboratório Avançado de Sistemas Embarcados e Robótica
LBP	*Local Binary Pattern*
LCN	*Local Contrast Normalization*
LRN	*Local Response Normalization*
LWIR	*Long Wavelenght Infrared*
MLP	*Multi-Layer Percepton*
MWIR	*Mid Wavelenght Infrared*
NIR	*Near Infrared*
OpenCV	*Open Source Computer Vision Library*
PAD	Processo de Aumento de Dados
PBD	*Poselet Based Detection*
PD	*Pitch-Trained Detector*
PPGCA	Programa de Pós-Graduação em Computação Aplicada
PRD	*Pitch and Roll-trained Detector*
PS	*Pictorial Structures*
QP	Questões de Pesquisa
RBM	Reconhecimento Baseado em Movimento
RE	Reconhecimento Estático
ReLU	*Rectified Linear Units*
RNC	Redes Neurais Convolucionais
ROC	*Receiver Operating Characteristics*
RD	*Roll-trained Detector*
SIFT	*Scale Invariant Feature Transform*
SRP	Sistema de Reconhecimento de Padrões
SURF	*Speeded Up Robust Features*
SVM	*Support Vector Machine*
SWIR	*Short Wavelenght Infrared*
VANT	Veículo Aéreo Não Tripulado
VLWIR	*Very Long Wavelenght Infrared*

SUMÁRIO

1 INTRODUÇÃO 13
1.1 MOTIVAÇÃO 14
1.2 OBJETIVOS 16
1.2.1 Objetivo geral 16
1.2.2 Objetivos específicos 16
1.3 CONTRIBUIÇÕES DESTE TRABALHO 17
1.4 ESTRUTURA DO TRABALHO 18
2 FUNDAMENTAÇÃO TEÓRICA 19
2.1 SISTEMAS DE RECONHECIMENTO DE PADRÕES 19
2.2 AQUISIÇÃO DE IMAGENS TÉRMICAS 20
2.3 SEGMENTAÇÃO E DETECÇÃO 21
2.3.1 Janela deslizante 21
2.3.2 Mapa de saliências 22
2.4 EXTRAÇÃO DE CARACTERÍSTICAS 26
2.4.1 Características Haar 26
2.4.2 Características LBP 29
2.4.3 Características HOG 29
2.5 CLASSIFICADORES 31
2.5.1 Classificadores em Cascata 31
2.5.2 SVM 34
2.6 REDES NEURAIS CONVOLUCIONAIS 35
2.6.1 Camada de convolução 37
2.6.2 *Pooling* 38
2.6.3 Unidades lineares retificadas 40
2.6.4 Normalização 40
2.6.5 Camada totalmente conectada 41
2.6.6 *Dropout* 41
2.6.7 *Loss* 42
2.7 DISCUSSÃO 42
3 ESTADO DA ARTE 43
3.1 MÉTODOS DE RECONHECIMENTO LEVANTADOS 43
3.2 AS TÉCNICAS EMPREGADAS NO RECONHECIMENTO 44
3.3 DESAFIOS E PROBLEMAS NO RECONHECIMENTO 46
3.4 SOLUÇÕES PARA PROBLEMAS DE RECONHECIMENTO 47
3.5 TRABALHOS RELEVANTES 49
3.6 CONSIDERAÇÕES 51
4 MATERIAIS E MÉTODOS 52
4.1 MATERIAIS 54
4.1.1 RaspBerry Pi 54
4.1.2 Base de Controle Móvel 55
4.1.3 Câmeras 55

4.1.4 Bases de imagens 56
4.1.5 OpenCV 59
4.1.6 Caffe 59
4.1.7 OpenBLAS 60
4.2 MÉTODOS 60
4.2.1 Definição dos dados 60
4.2.1.1 Processo de Aumento de Dados 61
4.2.1.2 Divisão da Base de Imagens 62
4.2.2 Treinamento dos classificadores 63
4.2.2.1 Classificadores em Cascata 63
4.2.2.2 HOG + SVM 64
4.2.2.3 Redes Neurais Convolucionais 64
4.2.3 Avaliação 68
4.2.3.1 Testes de Reconhecimento dos Classificadores 68
4.2.3.2 Testes de Oclusão 70
4.2.3.3 Análise do Sistema de Reconhecimento de Padrões 71
4.3 CONSIDERAÇÕES 74
5 RESULTADOS E DISCUSSÃO 75
5.1 TREINAMENTO DOS CLASSIFICADORES 75
5.2 TESTES DE RECONHECIMENTO DOS CLASSIFICADORES 77
5.3 TESTES DE OCLUSÃO 79
5.4 ANÁLISE DO SISTEMA DE RECONHECIMENTO DE PADRÕES 82
5.4.1 Análise da segmentação e detecção 82
5.4.2 Análise da segmentação e detecção com classificadores 83
5.4.3 Análise de desempenho computacional 85
5.5 AMEAÇAS À VALIDADE DOS RESULTADOS 89
5.6 DISCUSSÃO 90
6 CONCLUSÕES E TRABALHOS FUTUROS 92
REFERÊNCIAS 95

1 INTRODUÇÃO

Recentemente pode-se perceber a popularização do uso de Veículos Aéreos Não Tripulados (VANTs) de pequeno porte devido ao avanço tecnológico e o barateamento dos seus componentes. Eles têm sido aplicados em diversas pesquisas no campo da visão computacional e robótica. O VANT tem diversos benefícios em comparação com veículos terrestres (GASZCZAK et al., 2011), como: (i) tem mais graus de liberdade no movimento, dessa forma possui mais recursos para desviar de obstáculos, (ii) por ser um veículo aéreo, pode cobrir uma área maior em um período de tempo menor (SAIF et al., 2013). Os VANTs também apresentam vantagens, dependendo da aplicação, em relação ao monitoramento utilizando satélites, que não oferece a riqueza de detalhes que algumas aplicações exigem (MATESE et al., 2015). Entretanto, os VANTs tem a capacidade de voar em menores altitudes e proporcionar um monitoramento com maiores detalhes.

Várias pesquisas científicas estão sendo conduzidas utilizando VANTs nas mais diversas áreas de interesse civis e militares. Muitas pesquisas utilizam imagens como foco principal, tais como: (i) detecção e rastreamento (ANDRILUKA et al., 2010; GASZCZAK et al., 2011; MORANDUZZO; MELGANI, 2014b; REILLY et al., 2013), (ii) uso na agricultura (TOKEKAR et al., 2013; GRENZDÖRFFER et al., 2008; GUO et al., 2012) e modelagem 3D (SUN; SALVAGGIO, 2013) ou (iii) mapeamento de regiões (REMONDINO et al., 2011; ZONGJIAN, 2008).

Dentre as principais aplicações do VANT utilizando a captura de imagens está a detecção e rastreamento de objetos. Nesta aplicação incluem-se missões de resgate em locais de difícil acesso, grandes áreas de buscas, ou áreas atingidas por desastres naturais (MOLINA et al., 2012). Tragédias recentes como o acidente nuclear em Fukushima (MATSON, 2016) e o rompimento da barragem de rejeitos de minério em Mariana são exemplos práticos (TERRA NOTICIAS, 2015; SIMÕES, 2016). Nestes dois casos o contato humano com o ambiente era de alto risco. Entretanto era necessário tomar medidas rápidas de resposta para não gerar mais vítimas ou para buscar possíveis sobreviventes. No caso de Mariana havia uma grande área de busca a ser coberta. Seriam necessários diversos VANTs sobrevoando, em paralelo, partes desta

grande área, de forma a cobrir a área devastada em menos tempo. Todos estes VANTs gerariam uma quantidade gigantesca de imagens que deveriam ser analisadas pela equipe da defesa civil. Em problemas como estes, a detecção autônoma de objetos de interesse pode minimizar a necessidade de análise manual das imagens, diminuindo o tempo necessário para identificar vítimas, o que pode ser determinante para o sucesso de uma missão de resgate.

A análise de imagens obtidas através de VANTs traz diversos desafios, como: (i) a instabilidade da imagem, (ii) o tamanho pequeno dos objetos de interesse e (iii) as mudanças de pose que os objetos de interesse sofrem devido aos graus de liberdade do VANT (GASZCZAK et al., 2011). A instabilidade da imagem se deve ao fato de os VANTs serem plataformas de movimentação rápida e de pouca estabilidade, o que pode gerar distorções na imagem. Por outro lado as mudanças visuais do objeto, dependendo da altitude de voo e da orientação da câmera podem alterar drasticamente, a aparência dos objetos no ambiente, tornando a detecção automática de objetos uma tarefa desafiadora.

As questões apresentadas anteriormente afetam a detecção automática pelo fato de aumentarem a variabilidade dos objetos. Muitas vezes um sistema de classificação não é capaz de generalizar o problema (ZHANG; MA, 2012). Assim, se o sistema não for treinado com exemplos de uma dada variação na imagem, não será capaz de classificar corretamente os objetos afetados pela variação. Por isso, ao se projetar um sistema de classificação é necessário avaliar todas essas variáveis, considerando as diversas variações. Mesmo assim, muitas vezes o sistema não é capaz de generalizar o problema. Um exemplo disso é um sistema de reconhecimento de pedestres, que provavelmente, seria ineficiente para o problema de reconhecimento de pessoas em imagens aéreas. Apesar do objeto de interesse ser o mesmo nos dois problemas, o reconhecimento de pedestres usando câmeras fixas tem menos graus de liberdade, dessa forma a variação dentro da classe é menor.

1.1 MOTIVAÇÃO

Apesar das diversas pesquisas para a automatização e apoio ao rastreamento de pessoas em imagens aéreas realizadas pela comunidade científica (ANDRILUKA et al., 2010; BLONDEL et al., 2014b; FLYNN; CAMERON, 2013; GASZCZAK et al., 2011; REILLY et al., 2013), ainda há a dificuldade de generalizar o problema, visto que existem muitas variações nele. Isto se deve, principalmente, pelo fato de que algumas técnicas ainda serem muito restritas contemplando apenas algumas situações, não sendo capazes de generalizar as variações de ângulo sofridas no uso de VANTs ou em casos onde o objeto sofre oclusão parcial. Para a aplicação de missões de busca e salvamento, são necessárias técnicas robustas que tenham

bom desempenho de classificação em diversos ambientes e situações. Dessa forma a busca pode ser realizada tanto em ambientes simples (neve, campos abertos, desertos, etc), quanto em ambientes mais complexos (florestas, ambientes urbanos, etc).

Ao avaliar um sistema de reconhecimento é importante considerar duas questões: (i) desempenho de classificação e (ii) desempenho computacional (DUDA et al., 2012). O desempenho de classificação leva em conta a capacidade do sistema de detectar corretamente os objetos submetidos ao processo de classificação. Muitos sistemas exigem precisão na classificação, pois devem tomar decisões críticas que podem acarretar problemas de planejamento ou alocação incorreta da equipe de salvamento em casos de erros. Já o desempenho computacional deve suprir as exigências de tempo envolvidas no problema. Se um sistema é crítico ao ponto de exigir decisões em tempo-real (i.e. navegação autônoma de veículos), o classificador também deve ser capaz de tomar decisões em tempo-real, exigindo assim um alto desempenho computacional.

As pesquisas voltadas à detecção de pessoas tem recebido considerável atenção nos últimos anos, principalmente voltadas à detecção de pedestres em plataformas terrestres, onde os resultados obtidos tem sido bem sucedidos (DALAL; TRIGGS, 2005; LUVIZON et al., 2014). Essas abordagens costumam utilizar objetos maiores, com muitas informações de textura e formas. Em contraste, a detecção de pessoas em imagens aéreas é um problema devido ao pequeno tamanho que o corpo humano apresenta em imagens de plataformas aéreas, onde apenas uma pequena quantidade de características é apresentada em comparação à detecção de pessoas em plataformas terrestres. Isto se dá pelo fato dos VANTs voarem em alturas maiores. Além disso, a visão aérea limita a visão da câmera, onde nem sempre todas as partes do corpo humano são visíveis.

Além da questão de desempenho na detecção e reconhecimento de pessoas em imagens aéreas, nesta aplicação é necessário considerar o desempenho computacional das técnicas empregadas. Em problemas como missões de resgate e salvamento, respostas em tempo real são exigidas para que as medidas para planejamento e salvamento sejam executadas o mais rápido possível. Neste contexto, duas formas de implementação do sistema de detecção podem ser consideradas adequadas: (i) embarcar o sistema de detecção e reconhecimento de pessoas no próprio VANT, ou (ii) indicar o uso de uma Base de Controle Móvel (BCM) com hardware que forneça maior capacidade computacional para atingir o reconhecimento em tempo real.

O problema envolvido nesta pesquisa é focado no reconhecimento de pessoas em imagens aéreas para missões de salvamento e resgate. Segundo o levantamento do Estado da Arte (ver Capítulo 3), este é um problema que ainda carece de técnicas que sejam capazes de ge-

neralizar o problema, principalmente na questão de desempenho de classificação. Além disso, poucos autores levam em conta o desempenho computacional do sistema, um fator determinante para missões de salvamento, onde se exige informações em tempo real. Cumprir estes requisitos é importante para que o sistema envolvido nas missões de resgate e salvamento seja eficiente para o planejamento da missão, provendo informações consistentes e em tempo hábil para a tomada de decisões.

É importante ressaltar que este trabalho faz parte do conjunto de esforços empreendidos pela equipe de pesquisa do Laboratório Avançado de Sistemas Embarcados e Robótica (LASER), que envolve integrantes do programas de Pós-Graduação em Engenharia Elétrica e Informática Industrial (CPGEI) e Pós-Graduação em Computação Aplicada (PPGCA), ambos da Universidade Tecnológica do Paraná (UTFPR). Sendo este um dos primeiros trabalhos científicos voltados ao uso de Veículos Aéreos Não Tripulados voltados para aplicações de busca e salvamento, onde o ponto de partida é a questão de reconhecimento de pessoas. Esse trabalho visa contribuir no auxílio a tomada de decisão em missões de busca e salvamento, através do uso de VANTs equipados com câmeras, onde a imagem é processada automaticamente por um sistema de detecção de pessoas.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

Propor um método para detecção de pessoas em imagens aéreas, utilizando técnicas de visão computacional e reconhecimento de padrões.

1.2.2 OBJETIVOS ESPECÍFICOS

- Testar e avaliar as técnicas de detecção de candidatos a pessoas em imagens: Mapa de Saliências e Processamento de Imagens Térmicas, para reduzir o espaço de busca dos classificadores.
- Treinar, testar e avaliar técnicas de aprendizado de máquina para o reconhecimento dos candidatos a pessoas: HOG+SVM, Classificadores em Cascata Haar e LBP e Redes Neurais Convolucionais, visando buscar uma técnica robusta para o reconhecimento de pessoas em diversos ambientes e situações.
- Avaliar as técnicas de aprendizado de máquina para o reconhecimento dos candidatos a pessoas em situações de oclusão parcial das pessoas.

- Propor um método que englobe a união de técnicas de detecção e de classificação (Sistema de Reconhecimento de Padrões).
- Avaliar o desempenho computacional dos métodos (união de detecção e reconhecimento) em um sistema embarcado de baixo custo Raspberry Pi 2, considerando a resposta em tempo real de no mínimo 1 Hz (1 FPS).
- Construir uma base de imagens aéreas reunindo imagens de pessoas.

1.3 CONTRIBUIÇÕES DESTE TRABALHO

Este trabalho contribui propondo métodos de reconhecimento de pessoas em imagens aéreas em tempo real utilizando técnicas de aprendizado de máquina: Máquina de Vetor de Suporte, Classificadores em Cascata e Redes Neurais Convolucionais. Atualmente, existem diversos trabalhos voltados à este tema, entretanto, grande parte considera apenas as taxas de reconhecimento. Dessa forma, o desempenho computacional das técnicas é deixado em segundo plano. Muitas vezes, algumas das técnicas utilizadas tem alto tempo de processamento, impossibilitando o uso em aplicações que exigem informações em tempo real. Neste trabalho, além de avaliarmos a eficiência do reconhecimento, o desempenho computacional das técnicas também foi avaliado. Os métodos também foram avaliados em hardware de baixo custo, com baixa capacidade de processamento. A avaliação do funcionamento do hardware abre o direcionamento para mais pesquisas neste sentido, onde hardware com maior capacidade de processamento pode ser utilizado.

Além disso, também avaliamos o uso das Redes Neurais Convolucionais para a tarefa de reconhecimento de pessoas em imagens aéreas em tempo real, comparando-as com técnicas mais utilizadas no Estado da Arte. Até onde sabemos, este é o primeiro trabalho a usar as RNCs para esta tarefa. Isto enfatiza a importância deste trabalho, visto o destaque que as Redes Neurais Convolucionais tem no Estado da Arte atual.

Também se destaca o trabalho realizado com as câmeras térmicas de baixa resolução. A maioria dos trabalhos encontrados na literatura que trabalha com câmeras térmicas, utiliza câmeras com resolução maior. O estudo do uso de câmeras com menor resolução é importante dado o alto custo desta tecnologia, o qual costuma ser alto conforme a resolução das imagens térmicas.

Durante a execução deste trabalho foi necessário a preparação de uma base de imagens consistente e de alta quantidade de dados para o treinamento dos classificadores. As Redes

Neurais Convolucionais exigem grande quantidade de dados para um treinamento eficiente. Foram necessários processos de extração de exemplos, pois algumas bases não disponibilizavam as imagens prontas e rotuladas para o treinamento de classificadores. Além disso, um processo de aumento de dados foi executado a fim de aumentar a base de imagens e aumentar a variabilidade dos exemplos em: pequenos graus de rotação, escala e translação. Apesar disso, durante a realização deste trabalho não foi possível adquirir imagens de VANTs para a avaliação, então foram coletadas imagens do quarto piso de um edifício, simulando o voo de VANTs a cerca de 15 metros de altura. Entretanto, as imagens utilizadas no treinamento dos classificadores foram obtidas por VANTs.

Durante o desenvolvimento deste trabalho também foi possível contribuir com trabalhos científicos. Um destes trabalhos foi apresentado em congresso internacional no mês de maio de 2016 (OLIVEIRA; WEHRMEISTER, 2016). Além de um trabalho em fase de revisão a ser submetido à periódico internacional.

1.4 ESTRUTURA DO TRABALHO

Este trabalho é dividido em 5 capítulos.

A fundamentação teórica (capítulo 2) apresenta os conhecimentos relativos aos conceitos das técnicas utilizadas neste trabalho, tais como: câmeras e imagens térmicas, Mapas de Saliências, Características Haar, LBP e HOG, Classificadores em Cascata, SVM e Redes Neurais Convolucionais.

O capítulo 3 apresenta o Estado da Arte, evidenciando os trabalhos relacionados ao tema de reconhecimento de pessoas em imagens aéreas, destacando as metodologias empregadas tratadas em outros trabalhos.

Os materiais e métodos são apresentados no capítulo 4, onde são descritos o contexto e quais as técnicas desenvolvidas neste trabalho para a solução do problema de reconhecimento de pessoas.

O capítulo 5 discute os resultados obtidos, onde os resultados são apresentados, analisados e são feitas as devidas considerações.

Por fim, no capítulo 6, são apresentadas as conclusões gerais do trabalho, detalhando o cumprimento dos objetivos propostos. Além disso, propõe-se abordagens para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SISTEMAS DE RECONHECIMENTO DE PADRÕES

O Reconhecimento de Padrões é uma área da ciência da computação cujo o objetivo é identificar padrões e classificá-los em diferentes categorias. Este reconhecimento pode ser realizado sobre as mais diversas fontes de informações, como: imagens, sinais (forma de ondas), dados entre outros. As aplicações podem estar contidas nos mais diversos ramos do conhecimento, tais como Medicina, Psicologia, Ciência da Computação, etc.

Diversas questões estão envolvidas no Reconhecimento de Padrões, o que pode tornar bastante complexa a tarefa de reconhecer e classificar padrões. Segundo Duda et al. (2012), um Sistema de Reconhecimento de Padrões (SRP) pode ser dividido em cinco etapas: **aquisição**, **segmentação**, **extração de características**, **classificação** e **pós processamento**. A Figura 1 mostra um diagrama com todas as etapas.

A etapa de **aquisição** refere-se à aquisição dos dados submetidos ao sistema, onde um sensor ou dispositivo capta conjuntos de dados como sinais de rádio, imagens convencionais, imagens térmicas, sinais de áudio entre outros. A **segmentação** é a etapa que faz a separação dos dados, procurando candidatos à objeto de interesse do sistema. Nesta etapa, o candidato é separado do restante dos dados de entrada, por exemplo, em imagens o objeto pode ser extraído utilizando um blob, produzindo assim uma sub-imagem. A partir do candidato é possível **extrair as características** que podem revelar um padrão do objeto. O **classificador** então utiliza as características extraídas em conjunto com outras que foram aprendidas *a priori* através do treinamento para determinar à classe qual pertence o objeto. Por fim, o **pós-processamento** pode agir com base da decisão do classificador, ponderando esta classificação para tomar uma decisão final.

A maioria dos Sistemas de Reconhecimento de Padrões segue a linha descrita em Duda et al. (2012), entretanto ela não é uma regra geral. Alguns sistemas podem pular algumas etapas, outros fazem a união de duas ou mais delas, como no caso de alguns sistemas que usam técnicas de *Deep Learning*, que unem a extração de caracteristicas ao classificador (AREL et

al., 2010). O fluxo do sistema também pode variar, em alguns casos ele não possui sentido único, havendo retro-alimentação entre as etapas (DUDA et al., 2012). Geralmente há retro-alimentação quando há mais de um classificador no sistema, assim o classificador que está em uma etapa mais avançada pode utilizar informações obtidas pelo classificador anterior.

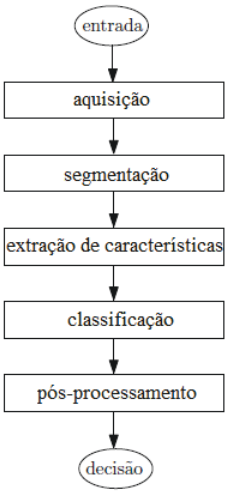

Figura 1: Etapas de um Sistema de Reconhecimento de Padrões.
Fonte: Adaptado de (DUDA et al., 2012)

Para que o sistema tenha sucesso na classificação é necessário extrair características pertinentes ao problema e com elas treinar um classificador. A etapa de treinamento de um classificador requer amostras de todas as classes do problema, de modo que o classificador tenha um conhecimento prévio de cada uma das classes, possibilitando a separação correta dos objetos. As amostras do problema devem abranger o maior número possível de classes e variações delas. Após o treinamento, o classificador terá uma configuração que guarda as informações obtidas. A configuração é utilizada no SRP tornando-o capaz de separar objetos de interesse em classes previamente definidas. Quando há a rotulação dos exemplos de treinamento conforme sua classe, o treinamento é conhecido como supervisionado. Já quando os exemplos de treinamento são submetidos ao classificador sem rotulação, este é chamado treinamento não-supervisionado. No caso do treinamento não-supervisionado, o classificador separa as classes conforme a similaridade entre os exemplos, segundo algum critério pré-definido.

2.2 AQUISIÇÃO DE IMAGENS TÉRMICAS

O espectro eletromagnético é a distribuição de radiação de um conjunto de ondas eletromagnéticas (RICHARDS, 2001) formada por: raios gamma, raios-X, ultravioleta, luz visível,

infravermelho, micro-ondas e ondas de rádio. A diferença entre estes tipos de onda está no comprimento de onda e frequência, sendo que em todos os casos a propagação se dá na velocidade da luz.

O espectro infravermelho é dividido em seis tipos de onda (RICHARDS, 2001): próxima (*Near Infrared*, NIR - 700nm a 1μm), curtas (*Short Wavelenght Infrared*, SWIR - 1μm a 2.5μm), médias (*Mid Wavelenght Infrared*, MWIR - 3μm a 5μm), longas (*Long Wavelenght Infrared*, LWIR - 8μm a 14μm), muito longas (*Very Long Wavelenght Infrared*, VLWIR - 14μm a 25μm) e extremamente longas (*Far Wavelenght Infrared*, FWIR - 25μm a 1mm).

A principal fonte de radiação infravermelha é a radiação térmica. Dessa forma, qualquer objeto que emita calor, irradiará radiação infravermelha conforme a sua temperatura. Um objeto muito quente, pode chegar a emitir luz visível, quando atingir faixa do espectro da luz visível. As câmeras térmicas capturam radiação térmica do ambiente, produzindo imagens com informação térmica.

Dentre os seis tipos de onda infravermelha, quatro são possíveis de ser capturadas em câmeras térmicas: NIR, SWIR, MWIR e LWIR (RICHARDS, 2001). As ondas NIR e SWIR dependem de iluminação infravermelha para a visualização de objetos nas imagens. As ondas MWIR fornecem informações detalhadas, sem a necessidade de luz adicional. Entretanto, há limitações de distância conforme a tecnologia utilizada. Já as ondas LWIR tem grande alcance e fornecem grande quantidade de informações. Em uma aplicação utilizando imagens aéreas, há a necessidade da utilização de ondas LWIR, devido à grande distância entre os objetos e a câmera térmica.

Existem dois tipos de sensores que capturam as ondas de radiação térmica: criogenicamente refrigerados e dispositivos termais (ROGALSKI, 2007). O sensor criogenicamente refrigerado é um sensor selado a vácuo que deve ser constantemente refrigerado. É um tipo de sensor com grande resolução e alcance, porém o seu custo é elevado. Já os dispositivos termais trabalham com mudanças de tensão, resistência ou corrente quando atingidos pela radiação térmica. São sensores simples e baratos, entretanto possuem baixa resolução.

2.3 SEGMENTAÇÃO E DETECÇÃO

2.3.1 JANELA DESLIZANTE

O reconhecimento de objetos em imagens é um problema que envolve diversas etapas, dentre elas existe a fase de segmentação de objetos, que também pode ser chamada de detecção. Entretanto, em uma imagem, um objeto pode estar presente em várias regiões. É necessário

então que um sistema de reconhecimento extraia esse objeto para que este seja submetido às próximas etapas (ver secção 2.1). Uma das técnicas utilizadas para a detecção é a janela deslizante, a mais utilizada e a mais simples das técnicas de detecção de objetos (FORSYTH; PONCE, 2002).

A Janela deslizante é utilizada na fase de classificação, onde haja um classificador treinado com amostras positivas e negativas pertinentes ao problema. O procedimento consiste em uma janela de tamanho *n* x *m* que constitui uma região da imagem, que percorre a imagem do início ao fim, ou seja, da esquerda para a direita e de cima para baixo. A cada passo da janela, a imagem de tamanho *n* x *m* é submetida ao classificador, onde é rotulada.

A técnica tem a vantagem de ser simples e de fácil implementação, entretanto apresenta problemas, principalmente quanto à questão computacional. A técnica submete diversos objetos ao classificador, em alguns casos milhares ou até milhões, o problema é que o classificador costuma ser a fase de um Sistema de Reconhecimento de Padrões com o maior custo computacional. Milhões de instâncias sendo classificadas podem tornar inviável o uso em uma aplicação, podendo levar horas para processar uma única imagem. Outro problema é que nem todos os objetos vão ter sempre o mesmo tamanho, assim o uso de uma janela de tamanho fixo pode levar informações incompletas ao classificador, induzindo-o ao erro.

2.3.2 MAPA DE SALIÊNCIAS

O mapa de saliência (ITTI et al., 1998) é uma técnica de detecção de objetos baseada no conceito de atenção visual. Esse conceito é inspirado na visão humana, que pode localizar objetos de interesse rapidamente em uma cena através de estímulos causados pelas diferenças de intensidade, contraste, cor e orientação dos elementos da imagem.

O modelo proposto por Itti et al. (1998) é baseado em características visuais, como intensidade, cor e orientação. O modelo é apresentado, inicialmente, para imagens com resolução *640* x *480* pixels, pirâmides Gaussianas de 9 níveis de escala, representados por $\sigma = \{0,1,..,8\}$. A escala vai de 1 para 1 (no nível zero) até 1 para 256 (no nível oito). O número de níveis varia conforme a resolução da imagem, sendo que quanto maior a imagem, maior será o número de níveis. A Figura 2 ilustra todo o processo de obtenção do mapa de características.

Segundo Itti et al. (1998), primeiramente é criado o canal de intensidade para cada escala, dado pela equação 1, sendo r, g e b, respectivamente, os canais vermelho, verde e azul.

$$I = \frac{(r+g+b)}{3} \tag{1}$$

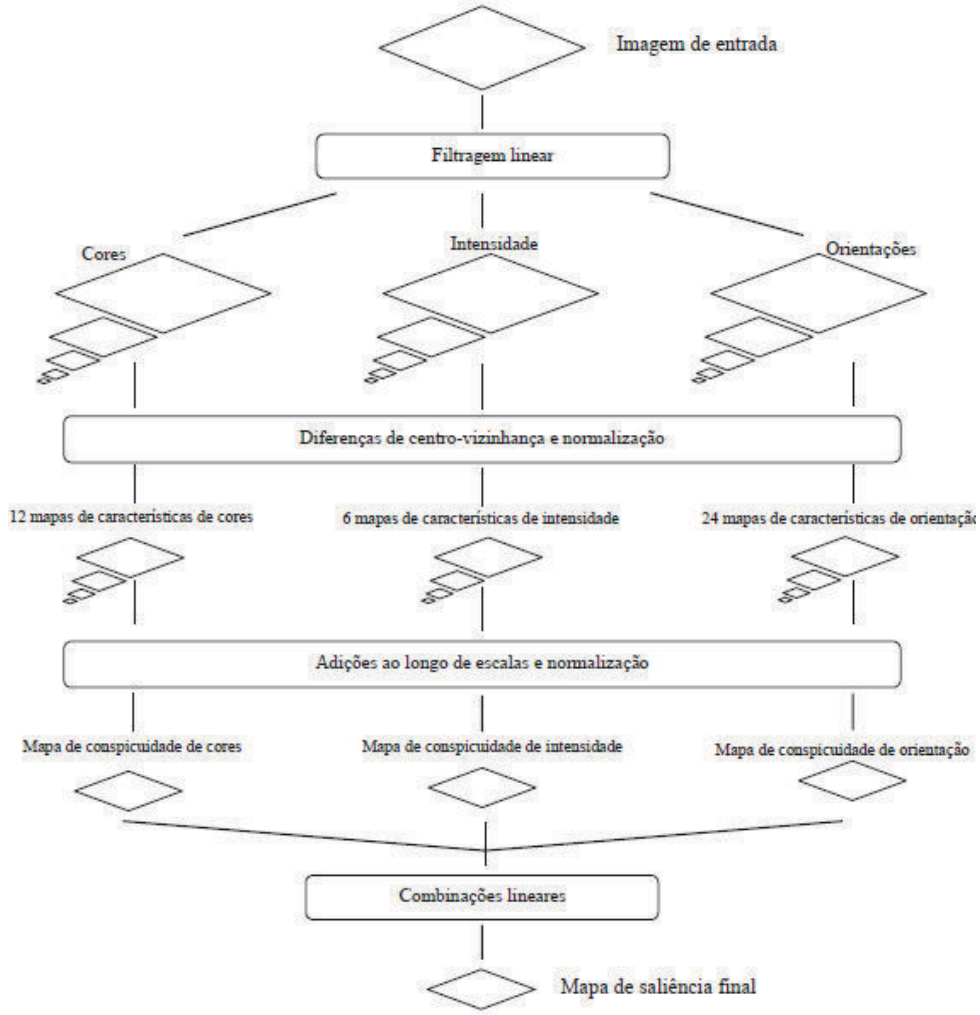

Figura 2: Processo completo de criação do mapa de saliências.

Fonte: Adaptado de (ITTI et al., 1998)

Através das cores r, g e b são criados quatro canais de cores, sendo: vermelho (R), verde (G), azul(B) e amarelo (Y). Cada canal é mostrado pelas equações 2, 3, 4 e 5.

$$R = r - \frac{(g+b)}{2} \tag{2}$$

$$G = g - \frac{(r+b)}{2} \tag{3}$$

$$B = b - frac(r+g)2 \tag{4}$$

$$Y = \frac{(r+g)}{2} - \frac{|r-g|}{2} - b \tag{5}$$

Através desses mapas são criados outros dois canais de cores. Estes combinam dois pares de cores vermelho/verde e azul/amarelo. Os canais são definidos pelas equações 6 e 7

$$RG = R - G \tag{6}$$

$$BY = B - Y \tag{7}$$

Para cada um dos canais de cores R, G, B e Y e o canal de intensidade I são criadas as pirâmides gaussianas R(σ), G(σ), B(σ), Y(σ) e I(σ), onde σ corresponde aos níveis das pirâmides.

Os mapas de orientação são obtidos através de um processo de convolução (ITTI et al., 1998), utilizando os níveis da pirâmide de intensidade I(σ) com filtros de Gabor. O cálculo é feito através da equação 8 (WALTHER; KOCH, 2006), sendo M_I o nível atual da pirâmide de intensidade e G o filtro de Gabor com fases 0^o e 90^o.

$$M_\theta = ||M_I(\sigma) \otimes G_0(\theta)|| + ||M_I(\sigma) \otimes G_{\pi/2}(\theta)||, \text{com} \theta \in 0^o, 45^o, 90^o e 135^o \tag{8}$$

Os filtros de Gabor são definidos pela equação 9 (WALTHER; KOCH, 2006). Onde γ é a proporção, δ o desvio padrão, λ o comprimento de onda, Ψ a fase da função de Gabor nas coordenadas (x',y'), transformadas através da orientação das equações 10 e 11.

$$G_\Psi(x,y,\theta) = e^{-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}} cos(2\pi\frac{x'}{\lambda} + \Psi), com \Psi = \left\{0, \frac{\pi}{2}\right\}; \tag{9}$$

$$x' = xcos(\theta) + ysin(\theta) \tag{10}$$

$$y' = -xsin(\theta) + ycos(\theta) \tag{11}$$

Após a obtenção de todas as pirâmides descritas anteriormente, para se obter os mapas de características são feitas as diferenças de centro-vizinhança (ITTI et al., 1998). As diferen-

ças de centro-vizinhança baseiam-se na ideia do processo biológico de aferir contraste. (ITTI, 2000). Assim é possível obter diferenças que sobressaem à vizinhança. A operação de centro-vizinhança é implementada sobre a diferença entre as maiores e menores escalas das pirâmides, chamadas escalas finas (maiores) e grossas (menores). A interpolação bilinear é utilizada para efetuar a diferença entre as escalas ($\ominus$), expandindo ou reduzindo a escala das imagens. Onde uma das escalas é interpolada para a dimensão da outra. O centro é o pixel na escala c $\in$ {2,3,4} e a vizinhança é o pixel correspondente a uma escala s = c + δ, onde δ = {3,4}.

Os mapas de intensidade estão relacionados aos neurônios sensíveis ao brilho no centro e ao escuro na vizinhança e pelos sensíveis ao escuro no centro e ao brilho na vizinhança. Estes dois tipos de contraste são calculados simultaneamente pela equação 12. O operador N normaliza os diferentes mapas para que uma característica não predomine. Assim os mapas com maior pico de atividade não dominam os mapas de menor atividade (ITTI et al., 1998).

$$I(c,s) = N(|I(c) \ominus I(s)|) \tag{12}$$

Os mapas de cores são construídos sobre o conceito da oponência de cores (ITTI et al., 1998). No caso das cores, os neurônios centrais e da vizinhança referem-se às cores oponentes, por exemplo, no centro teremos vermelho e na vizinhança verde, sendo o inverso também verdadeiro. O mesmo acontece com a oponência azul e amarelo. Assim serão formados dois mapas de cores representados pelas equações 13 e 14.

$$RG(c,s) = N(|RG(c) \ominus RG(s)|) \tag{13}$$

$$BY(c,s) = N(|BY(c) \ominus BY(s)|) \tag{14}$$

Os mapas de orientação são obtidos através do contraste da orientação local entre o centro com a escala das vizinhanças. O contraste é dado pela equação 15

$$O(c,s,\theta) = N(|O(c) \ominus O(s)|) \tag{15}$$

Após o cálculo de todos os mapas, obtêm-se 6 mapas de intensidade, 12 mapas de cores e 24 mapas de orientação, totalizando 42 mapas de características (*feature maps*).

Os mapas resultantes são combinados em três mapas de conspicuidade para intensidade (equação 16), cores (equação 17) e orientação (equação 18). Essa relação de conspicuidade é

a adição entre escalas ($\oplus$), onde faz-se a redução ou expansão dos mapas para a escala 4 e efetua-se a soma pixel a pixel.

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \tag{16}$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [N(RG(c,s)) + N(BY(c,s))] \tag{17}$$

$$\overline{O} = \sum_{\theta \in \{0^o, 45^o, 90^o, 135^o\}} N[\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta))] \tag{18}$$

Por fim, o mapa de saliências é formado pela combinação entre os três mapas de conspicuidade resultantes, mostrado na equação 19. A combinação é feita, porque apenas características iguais competem entre si, já características diferentes são combinadas para formar o mapa de saliências.

$$S = \frac{\overline{I} + \overline{C} + \overline{O}}{3} \tag{19}$$

2.4 EXTRAÇÃO DE CARACTERÍSTICAS

Nesta seção serão apresentados os fundamentos dos extratores de características utilizadas no Sistema de Reconhecimento de Padrões empregados neste trabalho.

2.4.1 CARACTERÍSTICAS HAAR

As características Haar-*like* são atributos extraídos de imagens, baseados nas *wavelets* de Haar (HAAR, 1910). Este tipo de característica para imagens foi proposto por Papageorgiou et al. (1998) e então utilizado para o reconhecimento de faces (VIOLA; JONES, 2001).

A Figura 3 mostra as características de Haar, onde cada retângulo é um filtro contendo regiões pretas e brancas. As regiões brancas são subtraídas das pretas, categorizando as sub-regiões de uma imagem utilizando apenas um valor numérico. Uma imagem pode conter milhares de caracteríticas Haar, este número pode ultrapassar até mesmo o número de pixels da imagem (VIOLA; JONES, 2001). Cada uma dessas características pode ser utilizada por um classificador. Entretanto, sozinha, uma característica Haar não é capaz de ter um desempenho satisfatório, chegando próximo de uma classificação aleatória (FORSYTH; PONCE, 2002).

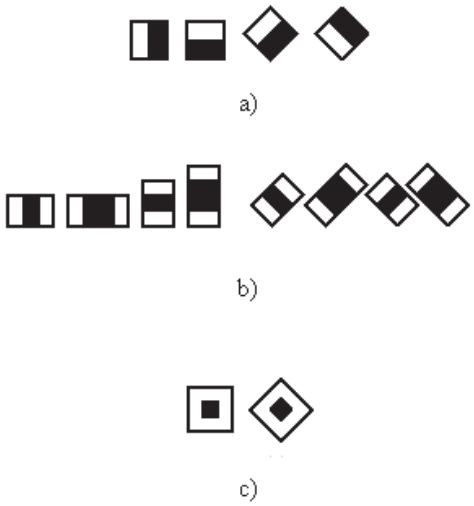

Figura 3: Características Haar, sendo características de (a) borda, (b) linha e (c) ponto.
Fonte: (LIENHART et al., 2003)

Um exemplo de características de Haar presentes em uma imagem pode ser visualizado na Figura 4, sendo que cada uma delas representa uma forma presente na imagem, variando entre características representadas por bordas, linhas e pontos.

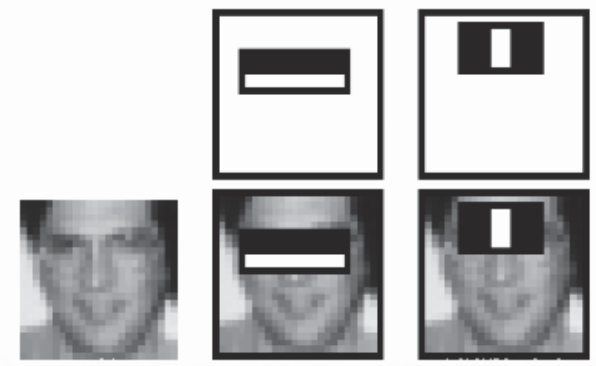

Figura 4: Exemplo de características Haar aplicadas ao reconhecimento de faces.
Fonte: (VIOLA; JONES, 2001)

O uso deste tipo de características apresenta diversas vantagens (FORSYTH; PONCE, 2002), como: alta eficiência computacional, invariância de intensidade de cor e escala, baixa variação intra-classes, alta variação inter-classes e orientação a diferenças de intensidades locais.

A eficiência computacional dessa abordagem se deve ao uso de uma representação intermediária da imagem utilizada por Viola e Jones (2001), chamada imagem integral. Através dessa representação é possível calcular a área de uma região com baixo número de operações.

Por exemplo, para calcular a intensidade de uma região de pixels 200x200 no modo convencional seriam necessárias 40000 operações de soma, já com o uso de imagens integrais este cálculo cai para 3 operações.

A imagem integral, também conhecida como tabela de soma de áreas, foi proposta por Crow (1984). O cálculo dessa imagem é simples, considerando um pixel na posição (x,y) de uma imagem qualquer, sua integral será dada pelo somatório de todos os pixels acima e à esquerda de (x,y) e o próprio pixel. Essa operação pode ser expressa pela equação 20, onde Ii(x,y) corresponde à imagem integral nas coordenadas (x,y), e I é o pixel da imagem original nas coordenadas (i,j).

$$Ii(x,y) = \sum_{i \leq x}^{i=0} \sum_{j \leq y}^{j=0} I(i,j) \tag{20}$$

Este cálculo também pode ser simplificado pela equação 21 (VIOLA; JONES, 2001), dessa forma não há a necessidade de refazer somas a cada cálculo de intensidade de um pixel, visto que os pixels acima e à esquerda já foram calculados previamente. Neste caso, é necessário que as bordas superiores e à esquerda sejam preenchidas com valores 0, para que o cálculo dos pixels das bordas seja possível.

$$Ii(x,y) = I(i,j) + Ii(x-1,y) + Ii(x,y-1) - Ii(x-1,y-1) \tag{21}$$

Assim, com a imagem integral é possível calcular a intensidade de área rapidamente, utilizando apenas os pontos mostrados na Figura 5 na equação 22.

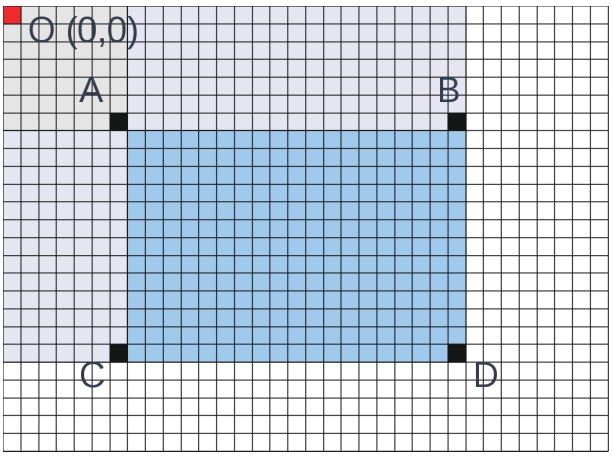

Figura 5: Pontos utilizados para o cálculo de intensidade de um pixel em imagens integrais. Cada pixel (x,y) representa a imagem integral do pixel (i,j) da imagem original.

Fonte: Própria autoria

$$Ii(x,y) = Ii(A) + Ii(D) - Ii(B) - Ii(C) \tag{22}$$

2.4.2 CARACTERÍSTICAS LBP

As características LBP (*Local Binary Pattern*) são baseadas em textura, que é definida como uma função de variação espacial da intensidade dos pixels. Foi inicialmente proposto por Wang e He (1990) e adaptado por Ojala et al. (1994). São características com alta robustez a mudanças de iluminação e têm baixa complexidade computacional. Para isso, atribui um valor numérico de intensidade à um pixel através de um cálculo envolvendo sua vizinhança.

O LBP utiliza uma vizinhança 3x3, onde faz a limiarização todos os pixels V_i da vizinhança pelo valor do pixel central, chamado V_0. Através de uma regra simples é gerada uma matriz com valores binários. Esta operação pode ser vista na equação 23.

$$s(x) = \begin{cases} 0, & V_i < V_0 \\ 1, & V_i \geq V_0 \end{cases} \tag{23}$$

Através da matriz de valores binários é aplicada a equação 24, sendo LBP(x,y) o valor qual se aplica o LBP, i o valor na matriz de valores binários na posição n.

$$LBP(x,y) = \sum_{n=1}^{8} i_n 2^n \tag{24}$$

A Figura 6 demostra o processo descrito pelas equações 23 e 24, sendo em Figura 6(c) demonstrada a ordem n seguida para o funcionamento da equação 24.

Após realizar o cálculo do operador LBP em cada pixel de toda a imagem, o descritor LBP é obtido através do histograma de várias regiões da imagem. A união desses histogramas, gera o descritor LBP da imagem.

2.4.3 CARACTERÍSTICAS HOG

O extrator de características HOG (*Histogram of Oriented Gradient*) foi proposto por Dalal e Triggs (2005). As características extraídas por este método retiram informações sobre a orientação de gradientes locais de intensidade. É um método de extração de características muito utilizado em reconhecimento de pessoas, principalmente com o classificador SVM (*Support Vector Machine*) (REILLY et al., 2013; TUERMER et al., 2013; FLYNN; CAMERON,

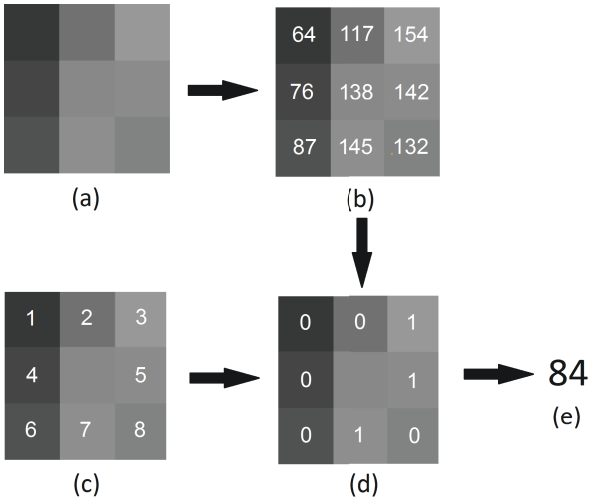

Figura 6: Exemplo de cálculo do LBP, em (a) uma vizinhança 3x3 de uma imagem arbitrária, (b) os valores numéricos de cada pixel, (c) a ordem em que será aplicada a equação 24, (d) a matriz de binários após a aplicação da equação 23 e (e) o valor do operador LBP para o pixel central da vizinhança.

Fonte: Própria autoria

2013).

O HOG foi construído com base no em outro algoritmo de extração de características, o SIFT (*Scale Invariant Feature Transform*) (LOWE, 1999). Enquanto o SIFT trabalha apenas com alguns pontos chave da imagem, o HOG considera regiões inteiras (DALAL; TRIGGS, 2005). Para isto, o algoritmo HOG passa por algumas etapas: normalização de níveis de cinza, cálculo de gradientes, criação de histogramas, divisão de blocos e normalização.

Na etapa de **normalização de níveis de cinza** é feita a normalização dos níveis de cinza de cada pixel da imagem, conforme os valores máximos e mínimos presentes na imagem. A segunda etapa, **cálculo de gradientes** é realizada através dos filtros horizontal e vertical [-1 0 1] e $[-1\ 0\ 1]^T$ ou filtro de Sobel. Na **criação de histogramas** os gradientes calculados são então agrupados em regiões chamadas células (um parâmetro do algoritmo), que por sua vez são armazenados em histogramas conforme o número de orientações definido como parâmetro (chamado *bins*). O número de orientações varia entre 0 e 180 graus, espaçados igualmente. A **divisão de blocos** é feita através da formação de conjunto de células, onde o tamanho destes blocos também é um parâmetro do algoritmo. Por fim, estes blocos sofrem uma **normalização**, que segundo Dalal e Triggs (2005) pode ser feita de várias formas: norma-L2, norma-L2-Hys (LOWE, 2004), norma-L1 e norma-L1-sqrt (raiz quadrada da norma-L1). A Figura 7 mostra um exemplo de características HOG extraidas da imagem de uma pessoa.

O número n de características extraídas de uma imagem é definido pela equação 25,

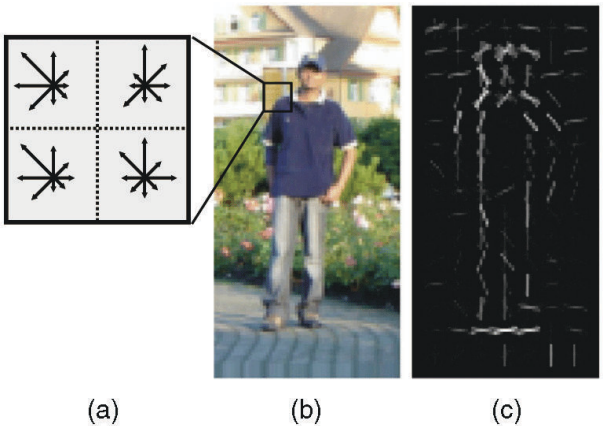

Figura 7: Características HOG. (a) bloco descritor, (b) o bloco em uma imagem de exemplo, e (c) as características HOG extraídas.

Fonte: Reproduzido de (GERÔNIMO et al., 2010)

onde B corresponde ao bloco, C à célula, J à janela e S ao passo dos blocos, sendo w a largura e h a altura de cada elemento.

$$n_{hog} = bins(\frac{B_w}{C_w})(\frac{B_h}{C_h})(\frac{J_w - B_w}{S_w - 1})(\frac{J_h - B_h}{S_h - 1}) \tag{25}$$

2.5 CLASSIFICADORES

Nesta seção serão apresentados os Classificadores em Cascata e a Máquina de Vetor de Suporte (SVM), ambos classificadores utilizados neste trabalho.

2.5.1 CLASSIFICADORES EM CASCATA

A técnica de classificadores em cascata consiste em unir diversos classificadores para atingir um objetivo (VIOLA; JONES, 2001). Cada classificador tem uma função distinta no processo de classificação, sendo responsável por uma porção do espaço de características. A Figura 8 mostra uma representação de uma cascata de classificadores, onde dada uma entrada, cada classificador tomará uma decisão. A decisão pode ser rejeitar a amostra ou submete-la ao próximo classificador.

Cada classificador da cascata pode ser considerado um classificador fraco (ZHANG; MA, 2012). Classificadores fracos não tem uma alta taxas de classificação, chegando próximos de uma classificação aleatória. Entretanto a união de vários classificadores fracos produz um classificador forte, que é o objetivo da cascata de classificadores.

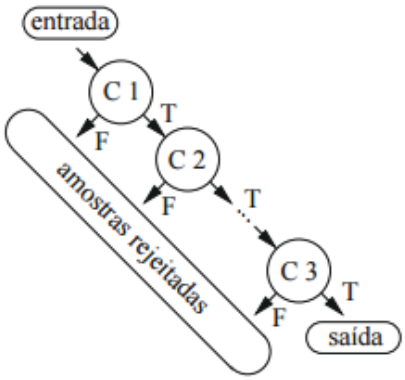

Figura 8: Classificador em cascata. Onde T representa que um conjunto características passaram pela classificação e F representa que este conjunto foi rejeitado.

Fonte: Própria autoria.

O uso de classificadores complexos no reconhecimento de objetos, resulta em um processamento lento. Assim, com uma cascata de classificadores fracos de baixa complexidade é possível unir velocidade de processamento com baixas taxas de erro. Essa abordagem utiliza a premissa de que nas fases iniciais da cascata estarão os classificadores mais rápidos, estes eliminarão com facilidade objetos que certamente sejam negativos (ZHANG; MA, 2012). Já em fases mais avançadas da cascata, os classificadores tem uma complexidade um pouco maior e com isso confirmarão se uma região é ou não um objeto de interesse.

A cascata de classificadores trabalha com a ideia de que em uma imagem existem mais regiões negativas do que positivas. Dessa forma a cascata pode rapidamente excluir vários objetos negativos e gastar mais tempo de processamento naqueles que tem grande chance de serem um objeto de interesse.

Para um classificador, quanto mais características forem utilizadas, melhor será o resultado, aumentando assim a taxa de detecção e diminuindo o número de alarmes falsos. Entretanto, quanto mais características, maior será o tempo de treinamento, devido à grande complexidade envolvida em encontrar um modelo ideal. Encontrar uma cascata que seja uma combinação ótima entre número de estágios, características e limiares dentro de cada estágio, é uma tarefa complicada e árdua que acontece durante o treinamento (VIOLA; JONES, 2001).

O treinamento de cada classificador da cascata se dá normalmente com algoritmos de *Boosting*, onde o *AdaBoost* (*Adaptative Boosting*) (FREUND; SCHAPIRE, 1996) é o mais utilizado (VIOLA; JONES, 2001; LIENHART et al., 2003). Durante o treinamento, o algoritmo

tenta atingir metas de falsos positivos e verdadeiros positivos previamente estabelecidas, utilizando para isso o menor número de características possível. O algoritmo *AdaBoost* se baseia em classificadores fracos, atribuindo pesos a amostras, onde a cada iteração aumenta os pesos das amostras classificadas incorretamente. No final do treinamento, o classificador pode ser visto como uma média ponderada de vários classificadores fracos (CRUZ, 2014). A Figura 9 mostra o funcionamento do *boosting* com apenas três classificadores fracos, mostrando o ajuste de pesos em cada etapa.

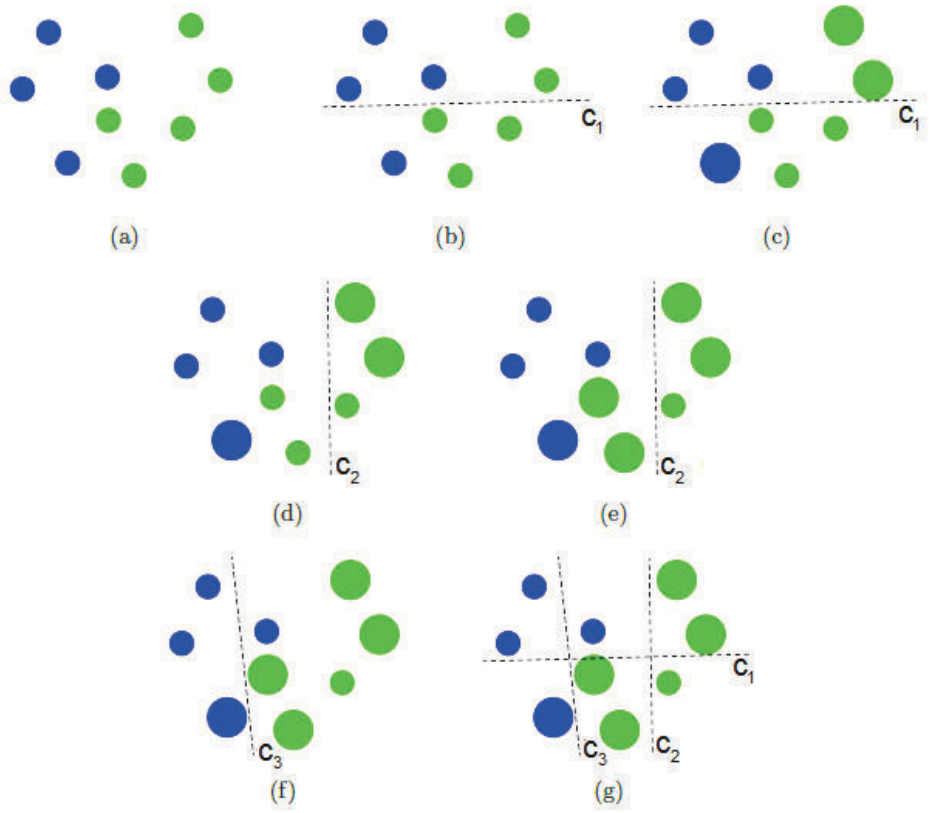

Figura 9: Funcionamento do *boosting* com três classificadores: (a) espaço de características, onde os pesos iniciais são iguais, (b) (d) (f) respectivamente, classificadores fracos c_1, c_2 e c_3, (c) (e) respectivamente, ajuste dos pesos baseado nas classificações incorretas de c_1 e c_2 , (g) classificador final.

Fonte: Reproduzido de (SZELISKI, 2010).

A tarefa do *AdaBoost* é simplesmente minimizar os erros, dessa forma, ele pode atingir altas taxas de detecção e ao mesmo tempo alta taxas de falsos positivos (VIOLA; JONES, 2001). Para ser adaptado à classificação em cascata, os limiares dos classificadores fracos são alterados (CRUZ, 2014). Assim, limiares altos produzirão classificadores com baixa taxa de detecção e de falsos positivos; limiares baixos produzirão alta taxa de detecção e de falsos positivos (VIOLA; JONES, 2001).

2.5.2 SVM

O SVM ou Máquina de Vetor de Suporte é um classificador linear, independemente da distribuição dos dados ser linear ou não (DUDA et al., 2012). É um classificador de aprendizado supervisionado utilizado em diversas aplicações de reconhecimento de padrões, como: imagens, fala e previsões de séries temporais.

Considerando dados de um treinamento de duas classes, em um modelo linear o objetivo do SVM é definir um hiperplano que classifique corretamente as classes (THEODORIDIS; KOUTROUMBAS, 2008). Para esta tarefa podem existir diversos hiperplanos que satisfaçam a classificação. Entretanto, a escolha do hiperplano deve ser feita com o intuito de aumentar a generalização do classificador, dando a capacidade de classificar dados que não estavam presentes no treinamento. Assim o SVM modela um hiperplano que cria a maior margem possível entre duas classes. Este hiperplano tem a mesma distância em relação aos pontos mais próximos entre as duas classes. A Figura 10 mostra a separação de um hiperplano, onde a margem entre duas classes é a mesma.

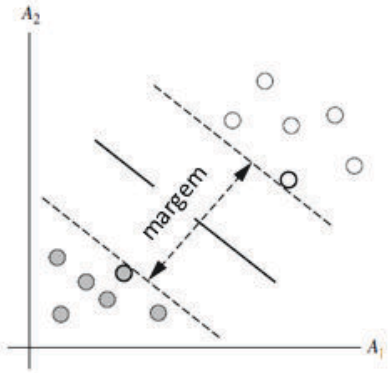

Figura 10: Hiperplano produzido por um SVM para separar duas classes, onde a margem entre é igual para os pontos mais próximos entre as duas classes.

Fonte: Adaptado de (HAN et al., 2011)

Existem casos em que a distribuição das classes envolvidas não é linear. Nestes casos a separação linear não é uma boa solução, sendo necessário utilizar recursos extras chamados funções de *Kernel* (SZELISKI, 2010). O uso da função de *Kernel* mapeia os dados em um espaço para outro definido pela função. A Figura 11 representa essa operação.

Dessa forma, através das funções de *Kernel* o SVM pode realizar a diversos problemas de classificação. Diversos *Kernels* podem ser utilizados, dentre os mais utilizados estão: polinomial, sigmoidal e gaussiano CBF (SZELISKI, 2010).

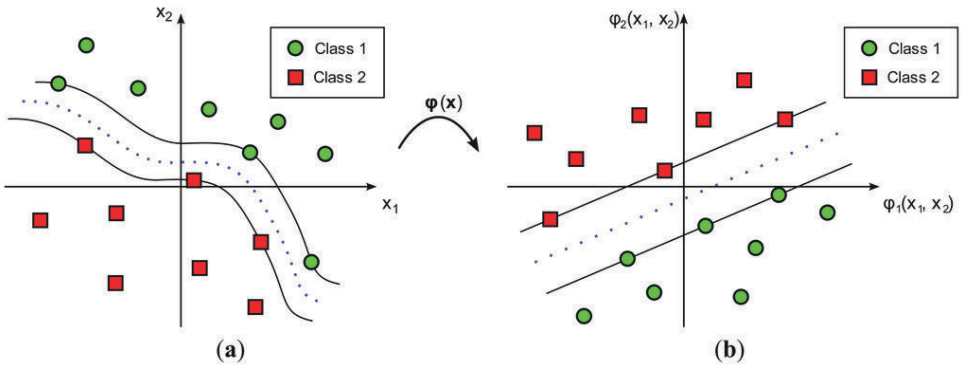

Figura 11: Distribuição entre duas classes não lineares. Em (a) classes em uma distribuição não-linear, e (b) classes após a aplicação de uma função de *Kernel*.

Fonte: Reproduzido de (RUIZ-GONZALEZ et al., 2014)

2.6 REDES NEURAIS CONVOLUCIONAIS

As Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Networks*), como muitas técnicas de aprendizado de máquina, são baseadas em arquiteturas biólogicas encontradas na natureza (ALPAYDIN, 2014). As CNNs são baseadas em um trabalho realizado nos anos 60 por Hubel e Wiesel que, através de experimentos com gatos e primatas, descobriram que a capacidade visual é organizada em um conjunto de células sensíveis a regiões chamadas campos receptivos. Estas células podem ser classificadas em três tipos: (i) simples, (ii) complexas e (iii) super complexas. Tal classificação depende da complexidade do padrão que estimula as células. As células simples são ativadas por padrões simples e as complexas e super complexas por padrões de alta complexidade. A partir deste estudo é possível inferir que a combinação de padrões gera os objetos e cenas, produzindo assim a representação visual (LECUN et al., 1998). Dessa forma, as CNNs surgiram e tentam representar esta ideia através de dados em duas dimensões, assim como as imagens. Sendo assim, as CNNs se tornam técnicas promissoras para o reconhecimento de imagens (AREL et al., 2010).

A CNN é uma técnica de aprendizado de máquina introduzida por (LECUN et al., 1998), mostrando excelente desempenho em tarefas de classificação de digitos manuscritos e reconhecimento de faces. Atualmente é um assunto de grande interesse na ciência da computação, principalmente para o reconhecimento de imagens (ZEILER; FERGUS, 2014), dada a quantidade de trabalhos recentes envolvendo desafios de classificação visual. A eficiência das CNN vem sendo demonstrado principalmente em competições de reconhecimento de imagens. Nestas competições há milhões de imagens que devem ser classificadas em milhares de classes. Já Krizhevsky et al. (2012), demonstrou excelentes resultados com uma diferença de quase

9.7% para o segundo colocado na competição de classificação de imagens ILSVRC 2012. Já o trabalho de Ciresan et al. (2012) demonstrou o desempenho da CNN comparado ao estado da arte nas bases de imagens NORB e CIFAR-10. Girshick et al. (2014) mostrou o desempenho da CNN na base PASCAL VOC. Por fim, mais recentemente Simonyan e Zisserman (2014) e Szegedy et al. (2015) mostraram mais uma vez a eficiência da CNN no desafio ILSVRC 2014, com a GoogleNet e a VGGNet Além destas aplicações, pode-se destacar o desempenho da CNN no reconhecimento de pedestres (ANGELOVA et al., 2015) e reconhecimento de faces (LEVI; HASSNER, 2015).

O sucesso das CNNs se deve ao fato delas garantirem alguns graus de liberdade para mudança, escala e distorção (LECUN et al., 1998). Isto torna o aprendizado robusto e preparado para diversas situações de variação. Outro fato importante é a capacidade da CNN em extrair suas próprias características (uma questão bastante levantada no aprendizado de máquina), como pode ser observado nos diversos trabalhos que utilizam variados tipos de extratores de características para diferentes situações (ANDRILUKA et al., 2010; GASZCZAK et al., 2011; MORANDUZZO et al., 2015; RUDOL; DOHERTY, 2008). Dessa forma, diferentemente das técnicas tradicionais de aprendizado de máquina (onde há pelo menos um extrator de características e um classificador), a extração de características e reconhecimento são feitos pela CNN. No caso do reconhecimento de imagens a CNN recebe a imagem como um dado bruto e extrai as características conforme o treinamento.

As CNNs são treinadas em múltiplos estágios que são chamados de camadas, com a finalidade de usar as características extraídas por elas mesmas. Dessa forma elas tem a capacidade de combinar as características obtidas nas camadas iniciais com características obtidas na camadas mais avançadas. Nas camadas inicias a CNN é capaz de extrair características visuais elementares, tais como: bordas orientadas, pontos finais, cantos ou características similares. Essas características são então combinadas nas camadas subsequentes, gerando características mais complexas (LECUN et al., 1998).

Na primeira camada, a entrada corresponde à imagem submetida à rede; já nas camadas mais profundas, a entrada é a saída da camada anterior. Estes tipos de dados recebidos pelas camadas, a partir da segunda, são chamados mapas de caracteríticas (*feature maps*) (LECUN et al., 1998). Para formar estes mapas de características, a imagem passa por diversas camadas e subcamadas que formam imagens de onde cada uma das características é obtida. Estas camadas e subcamadas têm diversas configurações e funções, as quais servem para a obtenção e combinação das características e também para o aprendizado. Os tipos de camadas são detalhadas nas próximas subsecções deste trabalho. Dentre as principais, se destacam as camadas de: convo-

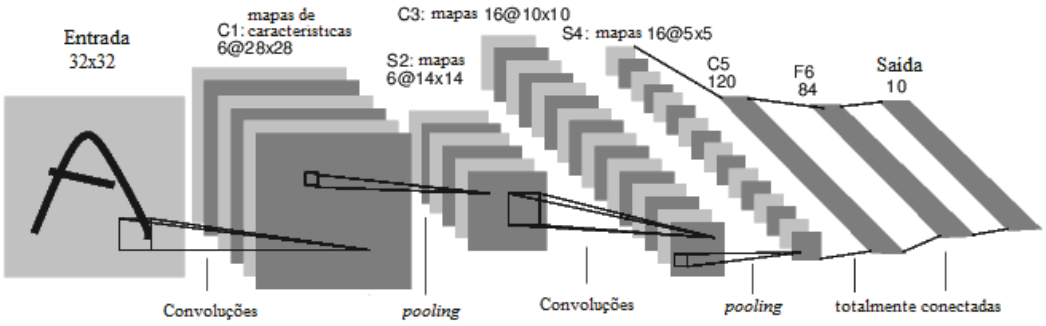

Figura 12: Exemplo de Rede Neural Convolucional, usada para o tarefas de classificação de digitos manuscritos.

Fonte: Adaptado de (LECUN et al., 1998)

lução, *pooling*, normalização, Unidades Lineares Retificadas (do inglês *Rectified Linear Units* ou ReLU) e camada totalmente conectada. A Figura 12 mostra um exemplo de CNN que será utilizada para esclarecer o uso de cada camada nas próximas seções.

É importante frisar que os neurônios da CNN funcionam da mesma forma que nas redes neurais regulares. Nas camadas intermediárias, estes neurônios têm funções não-lineares de ativação. As funções de ativação determinam se uma informação será propagada ou não, com base na informação da camada anterior (ou camada de entrada) multiplicada pelos pesos definidos durante o treinamento. Estes pesos representam o conhecimento da rede, geralmente são definidos em um treinamento utilizando algoritmo de *backpropagation* (ALPAYDIN, 2014), treinamento que também é utilizado pela CNN.

2.6.1 CAMADA DE CONVOLUÇÃO

A camada de convolução é a camada de construção da rede (LECUN et al., 1998). Nesta camada os parâmetros são filtros (ou kernels), com baixa receptividade, mas que se estendem por toda a imagem. Cada um desses filtros, aplicado a toda a imagem, consiste em um mapa de características. Os filtros são muito similares aos utilizados no processamento de imagem, como o mostrado na Figura 13.

A equação 26 (LECUN et al., 2010) expressa a operação realizada pela camada de convolução, onde h é a saída da camada de convolução k sobre a entrada x com filtros W^k.

$$h_{ij}^k = f((W^k \otimes x)_{ij}) + b) \tag{26}$$

Onde f é uma função não linear (máximo, sigmoidal, gaussiana, etc), b é um *bias*

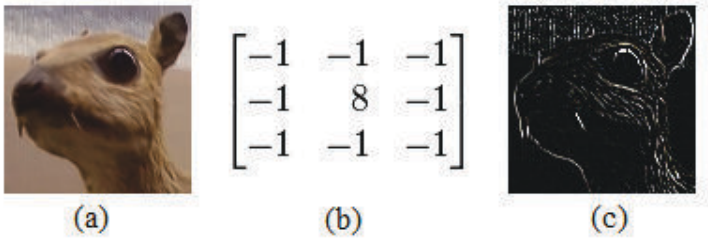

Figura 13: Exemplo de filtro de convolução aplicado a uma imagem: (a) imagem de entrada, (b) filtro de convolução e (c) mapa de características.

Fonte: Adaptado de (FISHER; KORYLLOS, 1998).

escalar e $\otimes$ representa a operação de convolução.

A saída da camada vai depender de quatro parâmetros dados na definição da arquitetura da rede. Estes parâmetros controlam a saída da camada de convolução, sendo: (i) tamanho do filtro, (ii) profundidade (*depth*), (iii) passo (*stride*) e (iv) preenchimento de zeros (*zero-padding*). O **tamanho do filtro** é dado pela largura e altura que os filtros da camada terão, estes são definidos na construção da arquitetura da rede, sendo largura e altura de tamanhos iguais. Um filtro de tamanho K x K em uma entrada de tamanho D x D, gera uma saída de tamanho (D-K+1) x (D-K+1). O tamanho da saída também é afetado pelo **passo**, parâmetro que controla a movimentação do filtro na entrada, assim um passo P gera uma saída de tamanho (((D-K)/P)+1) x (((D-K)/P)+1). A **profundidade** se refere à quantidade de mapas de características que serão extraídos da entrada. Já o **preenchimento de zeros** é usado para preencher as bordas, dessa forma não há diminuição do tamanho da entrada e assim a saída terá o mesmo tamanho da entrada.

Na Figura 12, a camada C1 é uma camada de convolução com profundidade 6, que processa uma imagem de tamanho 32x32 com um filtro de 5x5, com passo 1 e sem preenchimento de zeros. A camada C1 gera uma saída de 28x28x6. Já a camada de convolução C3 com profundidade 16, recebe como entrada a saída de tamanho 14x14x6 da camada S2, com um filtro de 5x5, passo 1 e sem preenchimento de zeros, gerando uma saída de 10x10x6. É na camada C3 que as 6 características da camada C1 são combinadas, formando novas características diferentes (LECUN et al., 1998).

2.6.2 *POOLING*

O *pooling*, assim como a camada de convolução, é um conceito importante da CNN. Na maioria das arquiteturas está presente após a camada de convolução. Este tipo de camada resume as saídas de grupos de neurônios vizinhos no mesmo filtro (KRIZHEVSKY et al., 2012),

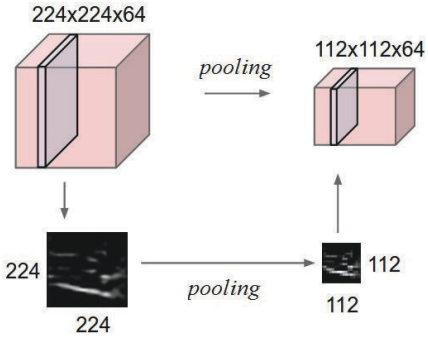

Figura 14: Exemplo de uma camada de *pooling* sobre uma imagem 224x224, com uma saída de 112x112.

Fonte: Adaptado de (KARPATHY, 2016)

ou seja, esta camada serve para reduzir progressivamente o tamanho espacial da representação intermediária. Assim, o uso do *pooling* reduz a quantidade de parâmetros e computação, portanto, também combate o *overfitting* (KRIZHEVSKY et al., 2012). A Figura 14 mostra a saída de uma camada de *pooling*.

Os parâmetros para a camada de *pooling* são: (i) tipo de operação, (ii) tamanho do filtro e (iii) passo. Geralmente, este tipo de camada trabalha com uma **operação** de máximo (Figura 15), média ou norma-L2. O **tamanho do filtro** pode variar, conforme a necessidade de reduzir os mapas de característica. O **passo** é o parâmetro que define o deslocamento da operação de *pooling*. A maioria das arquiteturas utiliza o *pooling* sem sobreposição, neste caso, o tamanho do filtro é igual ao tamanho do passo (na Figura 15 temos tamanho do filtro 2 e passo 2). Caso estes parâmetros tenham valores diferentes, haverá sobreposição. Segundo Krizhevsky et al. (2012), o *pooling* com sobreposição reduz o *overfitting*.

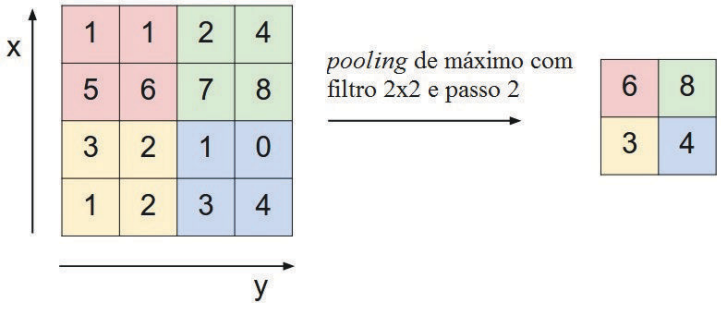

Figura 15: Exemplo de uma camada de *pooling* com função de máximo sem sobreposição.

Fonte: Adaptado de (KARPATHY, 2016)

Na Figura 12, as camadas S2 e S4 são camadas de *pooling* (chamadas de *subsampling*

por LeCun et al. (1998)). Nos dois casos temos filtros de 2 x 2 e passo 2, sem sobreposição. Dessa forma, o tamanho do mapa de características é reduzido pela metade.

2.6.3 UNIDADES LINEARES RETIFICADAS

Unidades Lineares Retificadas (do inglês *Rectified Linear Units* ou ReLU) são camadas que aplicam funções de ativação de não-saturação. Essas funções aumentam a não linearidade da função de decisão da rede, sem afetar os campos receptivos. Assim, após uma camada *ReLU*, a saída terá o mesmo tamanho da entrada. As funções de ativação mais utilizadas são: máximo, tangente hiperbólica e sigmoidal.

Este tipo de camada não costuma ser representada nos esquemas da arquitetura das redes. Ela tem poucas vantagens na generalização da rede. Entretanto a *ReLU* tem significativa contribuição no treinamento, com a redução do tempo de treinamento (KRIZHEVSKY et al., 2012).

2.6.4 NORMALIZAÇÃO

Existem diferentes tipos de camadas de normalização: Normalização com Resposta Local (*Local Response Normalization* - LRN) e Normalização com Contraste Local (*Local Contrast Normalization* - LCN). Em ambos os casos, a tarefa da camada de normalização é forçar a competição entre neurônios adjacentes em mapas de características e neurônios no mesmo espaço local de características. Este efeito melhora a invâriancia e otimização das características extraídas e aprendidas (KRIZHEVSKY et al., 2012).

As equações 27 e 28 se referem, respectivamente, a LRN e LCN (KRIZHEVSKY et al., 2012). Sendo $a^i_{x,y}$ um neurônio sobre o qual é aplicado o filtro i na posição (x,y), a resposta normalizada é dada por $b^i_{x,y}$. Em ambas as equações a soma acontece em n mapas de características adjacentes na mesma posição espacial, e N é o número total de *kernels* na camada. As constantes α e β são parâmetros a serem determinados, sendo o número de mapas de características adjacentes n também um parâmetro. Na equação 28, $m^{x,y}_i$ corresponde à média dos $a^{x,y}_i$ na vizinhança.

$$b^i_{x,y} = \frac{a^{x,y}_i}{(1+\alpha\sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)}(a^{x,y}_i)^2)^\beta} \tag{27}$$

$$b^i_{x,y} = \frac{a^{x,y}_i}{(1+\alpha\sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)}(a^{x,y}_i - m^{x,y}_i)^2)^\beta} \tag{28}$$

Através da das equações 27 e 28 é possível notar que a diferença das duas normalizações: o denominador da LRN computa a soma dos quadrados, enquanto da LCN a variância.

2.6.5 CAMADA TOTALMENTE CONECTADA

Em todas as camadas anteriores, e.g. convolução, *pooling*, *ReLU* (este tipo de camada também pode ser usado em conjunto com a camada totalmente conectada) e normalização, o objetivo da CNN era extrair e combinar características. Nas camadas totalmente conectadas (*fully-connected*), também conhecidas com *Inner Product*, inicia-se o aprendizado da rede. Nesta camada os mapas de características são convertidos de 2 dimensões para 1 e, assim como nas redes Multi Camada (do inglês *Multi-Layer Percepton* - MLP), todos os neurônios estão conectados a todos os neurônios ativos da camada anterior. Assim, é possivel afimar que esta parte da rede é similar às Redes Neurais regulares (LECUN et al., 1998). Consequentemente, suas ativações podem ser computadas como a multiplicação de matrizes somadas a um *bias*.

Na Figura 12 a camada totalmente conectada corresponde a F6, que está ligada à camada de convolução C5. Neste caso, a camada C5 funciona como uma camada totalmente conectada, visto que ela tem tamanho 1 x 1 e profundidade 120. Sendo assim, nesta camada existem 120 neurônios formando uma camada totalmente conectada.

2.6.6 *DROPOUT*

Como as camadas totalmente conectadas contêm a maioria dos parâmetros da CNN, elas estão propensa ao *overfitting*. Devido a isso, houve a necessidade de introduzir uma camada de *Dropout* (HINTON, 2014). A cada estágio do treinamento, alguns neurônios são retirados da rede com uma probabilidade de 1 - p. Assim, a rede é treinada sem os neurônios retirados. Após o fim do estágio de treinamento, estes neurônios são reinseridos na rede com o mesmo peso que tinham antes da sua retirada. Geralmente a probabilidade de um neurônio ser retirado é de 0,5. Entretanto, para neurônios de entrada a probabilidade de ser retirado deve ser menor, pois caso eles sejam retirados, a perda de informação pode afetar o treinamento da rede (HINTON, 2014). Consequentemente, evitando treinar todos os neurônios das camadas totalmente conectadas, a rede evita o *overfitting*, e também diminui o tempo de treinamento, visto que o número de neurônios é menor.

2.6.7 *LOSS*

A camada de *Loss* ou função de Perda é utilizada para guiar o aprendizado, sendo a última camada da rede. Assim é como o treinamento da rede penaliza o desvio entre a saída e como a entrada foi classificada, sendo a expressão do erro da classificação, onde o objetivo é a maior minimização possível. Existem diverss funções de *Loss*, algumas das mais utilizadas são: (i) *Softmax*, (ii) Sigmoidal com cruzamento de entropia e (iii) euclidiana. A função *Softmax* é utilizada para separar uma classe de várias. A função Sigmoidal com cruzamento de entropia serve para prever vários valores probabilísticos de 0 a 1. Já a função euclidiana prevê valores infinitos e reais.

2.7 DISCUSSÃO

Neste capítulo foram apresentados diversos conceitos e técnicas voltados ao desenvolvimento de Sistemas de Reconhecimento de Padrões. Todos estes conceitos são importantes para o bom entendimento deste trabalho, o qual consiste na definição de um SRP focado no reconhecimento de pessoas em imagens aéreas.

Para a definição deste SRP, os conceitos apresentados na Seção 2.1 são utilizados. Além do método comum de obtenção de imagens RGB, são utilizadas **Imagens térmicas** na etapa de **Aquisição**. Já a etapa de **Segmentação e detecção**, utiliza de técnicas de detecção como **Mapa de saliências** e processamento das Imagens térmicas para detectar possíveis objetos. Os objetos detectados são então reconhecidos através de **Classificadores em Cascata** e **SVM**, através de características **Haar**, **LBP** e **HOG**. Além disso, **Redes Neurais Convolucionais**, que unem a extração de características e o aprendizado, também são utilizadas para a classificação dos objetos.

Os conceitos de **Redes Neurais Convolucionais** apresentados foram empregados na construção das arquiteturas utilizadas neste trabalho. As camadas de **convolução**, juntamente com as camadas de ***pooling*** e **normalização** foram utilizadas na extração das características. Já as camadas **totalmente conectadas**, tiveram a tarefa de aprender com as informações extraídas. As camadas de **ReLU** auxiliaram em todos as etapas do treinamento das CNNs, reduzindo drasticamente o tempo de treinamento. Já o processo de ***Dropout***, auxiliou na generalização durante o treinamento, impedindo que a rede chegasse ao *overfitting*.

3 ESTADO DA ARTE

Para o levantamento do Estado da Arte foram levados em consideração trabalhos com o objetivo de reconhecer objetos de pequeno porte em imagens aéreas. Existem diversos trabalhos focados no reconhecimento de regiões, como: plantações, florestas, cidades entre outros, isto pode ser considerado como reconhecimento de objetos, mas constitui um problema menos complexo, pois estas regiões são formadas por uma grande quantidade de pixels, o que facilita sua identificação. Assim, neste estudo, foram considerados estudos que trabalham com objetos pequenos, como: pessoas, carros, árvores e animais.

O levantamento desses estudos foi conduzido com o objetivo de responder cinco questões de pesquisa (QP):

- **QP1**: Quais os métodos para o reconhecimento de objetos em imagens aéreas de VANTs?
- **QP2**: Quais as técnicas mais usadas em cada método?
- **QP3**: Quais os desafios e problemas em cada método?
- **QP4**: Como os desafios e problemas dos métodos são resolvidos?

Nas próximas seções essas questões de pesquisa serão respondidas através dos estudos encontrados no estado da arte. Elas tem como objetivo reunir informações a respeito do problema de reconhecimento de pessoas em imagens aéreas, e assim suprir um dos objetivos do trabalho que é levantar as dificuldades e melhorias futuras para a continuação do projeto com VANTs no grupo de pesquisa.

3.1 MÉTODOS DE RECONHECIMENTO LEVANTADOS

Dois métodos principais de reconhecimento foram identificados: Reconhecimento Baseado em Movimento (RBM) e Reconhecimento Estático (RE). A diferença entre os dois métodos se dá na maneira em que os candidatos a objeto de interesse são encontrados. Entretanto,

os dois métodos utilizam técnicas baseadas no reconhecimento de características, necessitando de um descritor de características e um classificador que atribuirá uma classe ao objeto.

No RBM apenas objetos que se movem entre vários quadros são reconhecidos. Neste método o reconhecimento é limitado, pois os objetos estáticos serão ignorados, tornando esse método inviável em diversas aplicações. A quantidade de trabalhos utilizando este método é menor, sendo eles Teutsch et al. (2011), Sekmen et al. (2009), Chen et al. (2011), Xiao et al. (2008) e Liang et al. (2013).

O RE usa apenas as informações de um único quadro, independentemente dos outros quadros. Esse método requer técnicas de segmentação ou detecção de objetos, como janelas deslizantes, ou ainda técnicas que diminuam o espaço de busca da imagem, submetendo apenas regiões que contenham alta probabilidade de conter um objeto. Uma técnica bastante popular utiliza câmeras térmicas (TEUTSCH et al., 2014; RUDOL; DOHERTY, 2008; PORTMANN et al., 2014; GASZCZAK et al., 2011; FLYNN; CAMERON, 2013). Este tipo de tecnologia oferece informações de temperatura, que podem ser utilizadas em casos onde a temperatura do objeto de interesse tende a ser constante, como, por exemplo, no reconhecimento de pessoas.

Há duas maneiras de realizar o RE (ANDRILUKA et al., 2010; BLONDEL et al., 2014b) usando: modelos monolíticos ou modelos baseados em partes. O modelo monolítico trata o objeto como um elemento, ou seja, o objeto é um todo e a detecção e reconhecimento será sobre o objeto inteiro. O modelo baseado em partes decompõe o objeto em várias partes, onde essas partes podem ser independentes ou interconectadas. No modelo com partes interconectadas é definida uma parte principal, à qual as outras devem estar conectadas; no modelo independente não é necessária nenhuma conexão entre as partes, basta elas existirem em qualquer configuração. Este modelo em partes é vantajoso em casos em que o objeto pode sofrer oclusão, mas requer modelos robustos e uma boa quantidade de informação sobre as partes (ANDRILUKA et al., 2010).

3.2 AS TÉCNICAS EMPREGADAS NO RECONHECIMENTO

Diversos algoritmos e métodos foram encontrados na literatura. Eles são definidos em três tipos, conforme a definição de Sistemas de Reconhecimento de Padrões (ver Seção 2.1): segmentação ou detecção, extrator de características e classificadores.

Para a segmentação e detecção no método RBM, os algoritmos mais utilizados são: subtração de fundo (CHEN et al., 2011; LIANG et al., 2013; TEUTSCH et al., 2011; XIAO et al., 2008; LIANG et al., 2013), previsão temporal (LIANG et al., 2013), algoritmo de Lucas-

Kanade (SEKMEN et al., 2009) e Fluxo ótico (XIAO et al., 2008).

Já no método de RE, as técnicas de segmentação ou detecção utilizadas são diversas. Por isso, elas são apresentadas na Tabela 1. Em adição à tabela, o trabalho de Tuermer et al. (2013) utiliza informações geográficas e de GPS, que indicam a posição geográfica de ruas, limitando o espaço de busca com conhecimento prévio de locais onde existe alta probabilidade de haver objetos de interesse. Há ainda um estudo que foca na detecção de ruas, para assim procurar veículos (ZHAO; NEVATIA, 2003).

Tabela 1: Técnicas de segmentação e detecção do método RUQ.

Técnicas	Trabalhos relatados
Janelas Deslizantes	Andriluka et al. (2010), Blondel et al. (2014a), Breckon et al. (2009), Cao et al. (2011), Chen et al. (2014), Chen e Meng (2013), Grabner et al. (2008), Lavigne et al. (2010), Ma'sum et al. (2013), Malek et al. (2014), Moranduzzo e Melgani (2014a, 2013), Moranduzzo et al. (2013), Morse et al. (2012), Tongphu et al. (2009), Yang et al. (2011), Zheng et al. (2012)
Imagens Térmicas	Davis e Keck (2005), Flynn e Cameron (2013), Gaszczak et al. (2011), Portmann et al. (2014), Rudol e Doherty (2008), Teutsch et al. (2014)
Sombras	Hung et al. (2012), Reilly et al. (2013, 2010), Wang (2011)
Mapa de Saliências	Blondel et al. (2014b)
Detecção de cantos	Gleason et al. (2011), Pingting et al. (2012)
Limiarização	Gururatsakul et al. (2010), Lin et al. (2009), Yang et al. (2012)
Template Matching	Khan et al. (2010)

Quanto às técnicas de extração de características, a variedade é grande, por isso serão destacados os mais utilizados e os que apresentaram melhores resultados. Sendo elas: Histograma de Gradientes Orientados (HOG) (DALAL; TRIGGS, 2005), Características Haar (VIOLA; JONES, 2001), Padrão Binário Local (LBP) (OJALA et al., 2002), Características Invariantes à Transformação de Escala (SIFT) (LOWE, 1999), características robustas rápidas (SURF) (BAY et al., 2006), Momentos Invariantes (HU, 1962) e Análise morfológica (SERRA, 1983). Destaca-se as características de Histograma de Gradientes e Haar como sendo as mais utilizadas.

Assim como os extratores de características, há diversos classificadores aplicados ao reconhecimento de objetos, como: Máquina de Vetores Suporte (SVM - *Support Vector Machine*) (CORTES; VAPNIK, 1995), Classificadores em Cascata (VIOLA; JONES, 2001) e k-Vizinhos-Próximos (k-NN - *k-Nearest-Neighbors*). Destaca-se a popularidade do SVM e dos Classificadores em Cascata, os quais estão presente em cerca de 75% dos trabalhos atuais sobre

reconhecimento de objetos em imagens aéreas.

3.3 DESAFIOS E PROBLEMAS NO RECONHECIMENTO

Os problemas e desafios identificados são apresentados das mais diversas formas. Alguns desses problemas são referentes à obtenção das imagens, outros estão atrelados às características do objeto de interesse e alguns ao ambiente de captura. Dessa forma, cada problema pode necessitar de uma solução diferente, que pode ou não causar consequências que também podem ser tratadas. Assim, dificilmente haverá uma abordagem que poderá tratar todos os problemas, mas sim amenizar os que mais prejudicam a detecção. Os problemas mais interessantes atrelados a um Sistema de Reconhecimento de pessoas são:

- Objetos pequenos (REILLY et al., 2010; TEUTSCH et al., 2014): câmeras com baixa resolução costumam fornecer imagens com objetos pequenos. Estes objetos são representados em uma área com uma pequena quantidade de pixels, o que dificulta o reconhecimento de padrões, pois poucas informações podem ser extraídas do objeto.

- Objetos estáticos (REILLY et al., 2010): em métodos RBM, objetos estáticos não são detectados, pois o método só é capaz de identificar objetos que se movem entre quadros. Entretanto, em plataformas móveis, como os VANTs, é possível que mesmo objetos em movimento não sejam detectados.

- Sombras (REILLY et al., 2013, 2010; ZHAO; NEVATIA, 2003; WANG, 2011): as sombras costumam afetar alguns sistemas de reconhecimento, visto que, em alguns casos, a sombra e o objeto estão unidos e algumas técnicas de segmentação não são capazes de separar os dois elementos. Por sua vez, no resultado de métodos que utilizam técnicas baseadas em extração de características, com o treinamento apropriado, as sombras não costumam ter muito impacto.

- Oclusão parcial de objetos (SEKMEN et al., 2009; ANDRILUKA et al., 2010): em ambientes complexos é comum ter um objeto que sofre oclusão por outro maior. Métodos de reconhecimento monolítico são afetados negativamente por tal problema.

- Baixo contraste entre o alvo e o fundo (GASZCZAK et al., 2011; ZHAO; NEVATIA, 2003): em alguns casos as cores e textura do objeto podem ser muito similares ao fundo, isto torna a tarefa do reconhecimento mais difícil.

- Multidões (GASZCZAK et al., 2011; TEUTSCH et al., 2014): um grupo de objetos de interesse forma uma multidão, o que pode tornar o reconhecimento mais difícil, visto que os classificadores costumam ser treinados para identificar apenas um objeto.

- Variabilidade intra-classe (GASZCZAK et al., 2011; TEUTSCH et al., 2014): a classe de um mesmo objeto pode apresentar variabilidade. Se o classificador não é preparado para identificar esta variabilidade, a taxa de reconhecimento pode cair.

- Variedade de poses de objetos (BRECKON et al., 2009; ANDRILUKA et al., 2010; TEUTSCH et al., 2014; BLONDEL et al., 2014b; REILLY et al., 2013): devido aos graus de liberdade que uma câmera ou sensor está sujeito, pode haver variedade na pose dos objetos, o que pode causar muitas mudanças nas características deste objeto.

- Mudanças nas condições do ambiente (BLONDEL et al., 2014b; TEUTSCH et al., 2014): mudanças inesperadas no ambiente podem afetar o reconhecimento, tais como chuva, mudanças de luz entre outros.

- Imagens borradas (CHEN; MENG, 2013; REILLY et al., 2013): algumas câmeras colocadas em plataformas móveis podem apresentar imagens borradas. Como a plataforma (assim como a câmera) está em constante movimento, a câmera pode registrar a imagem se movimentando, o que gera imagens instáveis.

- Recursos computacionais escassos (CHEN; MENG, 2013; TEUTSCH et al., 2014): em aplicações envolvendo sistemas embarcados, como VANTs, os recursos computacionais geralmente são limitados afetando algoritmos com alto custo computacional pois eles requerem mais recursos para o funcionamento adequado.

3.4 SOLUÇÕES PARA PROBLEMAS DE RECONHECIMENTO

Diversos desafios e problemas foram levantados, mostrando a complexidade envolvida no reconhecimento de pessoas em imagens aéreas. Muitos desses problemas envolvem soluções complicadas ou que podem acabar acarretando outros problemas. Dessa forma, encontrar uma solução para todos os casos não é uma tarefa trivial. Entretanto, a literatura mostra que esses "problemas"também podem ser usados como solução em alguns casos.

Como mostrado na seção anterior (ver 3.3), **Sombras** podem ser um problema durante o processo de detecção. Um exemplo de problema transformado em solução é o algoritmo de detecção de sombras utilizado por Wang (2011) para encontrar possíveis candidatos na imagem.

Enquanto Reilly et al. (2010) e Reilly et al. (2013) utilizam um algoritmo de detecção de sombras, somado à informações geográficas fornecidas por sensores no VANT. O sistema estima a altura das pessoas e juntamente com dados geográficos de longitude e latitude estima a posição do sol, assim é possível determinar a posição aproximada da pessoa em relação à sombra. Este tipo de aplicação previne problemas com sombras. Entretanto, algoritmos de detecção de sombras costumam ter alto custo computacional. Além disso, dado que este tipo de técnica depende das sombras, em ambientes sem iluminação solar a técnica é ineficaz.

Para o problema de oclusão parcial, Andriluka et al. (2010) utilizaram modelos baseados em partes. Caso alguma parte sofra oclusão, é possível identificar pessoas através das outras partes não oclusas. Chen et al. (2014) também usaram o modelo baseado em partes, mas para o reconhecimento de veículos.

No problema com objetos pequenos, a solução é de menor complexidade. Basta utilizar imagens com resolução maior (MORANDUZZO; MELGANI, 2014a, 2012; LAVIGNE et al., 2010), dessa forma os objetos terão uma área de pixels maior. Entretanto, isso afeta o custo computacional da aplicação, visto que o sistema terá imagens com mais informação para processar.

O problema de objetos estáticos só aparece em trabalhos com métodos RBM, a única solução encontrada para esse problema é utilizar métodos RE ou uma combinação das duas técnicas (YANG et al., 2012).

A utilização de imagens térmicas (DAVIS; KECK, 2005; DOHERTY; RUDOL, 2007; FLYNN; CAMERON, 2013; PORTMANN et al., 2014; RUDOL; DOHERTY, 2008; TEUTSCH et al., 2014) oferece a solução para vários problemas, como: baixo contraste, multidões, mudanças nas condições do ambiente e recursos computacionais escassos. Essa é uma abordagem que requer hardware específico e conhecimento prévio dos objetos. Em aplicações onde a temperatura dos objetos é constante ou tem pouca variação, o processo de segmentação é rápido e simples. Esta tecnologia também oferece a possibilidade de reconhecimento noturno, visto que não necessita de luz, como no caso de câmeras convencionais. Quanto à questão computacional, essa abordagem consegue detectar prováveis objetos sem muito esforço, visto que técnicas básicas de processamento de imagens podem ser empregadas na detecção. A grande desvantagem é a necessidade do hardware específico, que tem custo elevado.

A utilização de uma base de imagens robusta e consistente é a solução para as variabilidades de objetos e poses. Uma base de imagens que oferece diversas situações possibilita um treinamento mais eficiente do classificador, trazendo diversas situações próximas do mundo real. Em estudos focados no reconhecimento de pessoas em imagens aéreas, existem poucas

bases de imagens disponíveis. Assim, vários autores produzem suas próprias imagens e não disponibilizam este conteúdo. Por isto é difícil comparar os resultados obtidos. Blondel et al. (2014b) oferecem uma base de imagens para o estudo de reconhecimento de pessoas em imagens aéreas. Em seu trabalho a nova base é comparada com a base INRIA (DALAL; TRIGGS, 2005).

A opção mais utilizada para o problema de custo computacional, de alguns classificadores, é a utilização de técnicas para diminuir o espaço de busca. Dessa forma, uma menor quantidade de regiões da imagem serão analisadas, o que pode reduzir significativamente o custo computacional. No caso da detecção de veículos, é possível delimitar o espaço de busca às ruas e estradas, visto que veículos costumam transitar apenas em ruas e estradas (MORANDUZZO; MELGANI, 2014a; ZHAO; NEVATIA, 2003; GLEASON et al., 2011; MORANDUZZO; MELGANI, 2014b; KOZEMPEL et al., 2011; TUERMER et al., 2013). Já em problemas de detecção de pessoas, o problema de custo computacional é mais complicado, visto que uma pessoa pode estar em qualquer região da imagem.

3.5 TRABALHOS RELEVANTES

Os trabalhos do Estado da Arte mais recentes, com grande influência sobre outros analisados e com resultados que mostraram técnicas robustas, foram reunidos nesta secção para uma análise mais profunda. Os trabalhos aqui analisados exemplificam o uso das técnicas identificadas como mais utilizadas para a solução dos desafios e problemas, além de mostrar algumas técnicas que utilizam variações.

O reconhecimento de vitimas em imagens aéreas é abordado por Andriluka et al. (2010), entretanto o problema é focado em ambientes fechados e vítimas deitadas no chão, produzindo assim um trabalho próximo à detecção de pedestres. O método utilizado é o de RE com modelos baseados em partes. Três abordagens são comparadas: *Pictorial Structures* (PS) (ANDRILUKA et al., 2009), *Discriminately Trained Models* (DPM) (FELZENSZWALB et al., 2010) e *Poselet Based Detection* (PBD) (BOURDEV; MALIK, 2009). O PS divide o corpo em várias partes e é treinado com o AdaBoost (FREUND; SCHAPIRE, 1997). O DPM também decompõe o corpo em partes, definindo uma parte raiz, na qual todas as outras partes estão conectadas, treinado então por um classificador SVM (CORTES; VAPNIK, 1995). No PBD o corpo é dividido em diversas partes conectadas por um sistema de votação, que decide se as partes são similares a um modelo pré-definido. Diversas situações de oclusão são simuladas, mostrando resultados de reconhecimento de 51,5% com o DPM, 42,5% com o PS e 31,5% com o PBD. Já em testes combinando as técnicas, a combinação de PS e DPM obteve o melhor

resultado, com 66%.

No trabalho de Blondel et al. (2014a) o foco é o reconhecimento de pessoas em imagens aéreas, levando em consideração os graus de liberdade em um VANT: *pitch*, *roll* e *yaw*. A técnica Características de Canal Integral (ICF - *Integral Channel Features*) (DOLLÁR et al., 2009) é utilizada em conjunto com um classificador *Cluster Boosting Tree* (CBT) (WU; NEVATIA, 2007), pois necessita de uma detecção rápida em situações em que há variabilidade de ângulos durante a captura da imagem. Para se adaptar à variabilidade imposta pelo problema, são propostas adaptações no ICF e no CBT, chegando a uma solução chamada *Pitch and Roll-trained detector*(PRD). Os testes conduzidos buscaram analisar três aspectos: desempenho geral, tempo de processamento e capacidade de reconhecimento em diferentes ângulos. Para o teste de desempenho foram usados os algoritmos ICF, *Pitch-trained detector* (PD), *Roll-trained detector* (RS) e o PRD. Neste teste os melhores resultados foram obtidos pelo PD e o PRD. No teste de em relação aos ângulos, PD e PRD foram os melhores nas variações, onde o ICF se saiu melhor entre -20 e 20 graus. Já o ICF e RS, ficaram estáveis de -90º a 90º. Para o tempo de processamento, foram testados ICF, PD e PRD. Neste teste, o PRD se mostrou 1,75 vezes mais lento que os outros, devido ao grande número de classificadores fracos.

Através de algoritmos de detecção de sombras e metadados geográficos, Reilly et al. (2013) estimou a posição de pessoas e veículos. Para isso são feitas algumas suposições: (i) pessoas estarão em pé e terão uma sombra, e (ii) veículos terão uma sombra retangular. Para detectar essas sombras, a técnica de janela deslizante é usada para fazer a busca de objetos que estejam na orientação que a sombra se encontra em relação ao objeto. Sabendo onde a sombra está e a sua direção é possível estimar a posição do objeto que projeta esta sombra. Em casos em que os metadados não estão disponíveis, a abordagem muda um pouco. Neste caso são encontradas todas as sombras com a mesma orientação e classificados todos os objetos em uma varredura de 360º da sombra. É feito o reconhecimento dos objetos através de um SVM com características de Haar. Os resultados mostram a vantagem da abordagem de detecção com sombra, onde o número de falsos positivos cai drasticamente, cerca de 50%.

O estudo apresentado por Gaszczak et al. (2011) faz uso de imagens térmicas para auxiliar o reconhecimento de pessoas e veículos. No caso dos veículos, primeiramente uma imagem convencional é submetida a Classificadores em Cascata treinados com características de Haar, no qual os objetos classificados como veículos são então submetidos a uma segunda classificação através de uma imagem térmica. Para pessoas, a detecção é feita em imagens térmicas, em um classificador similar ao utilizado nos veículos, combinada com uma técnica *matching* de forma Gaussiana multivariada (BELONGIE et al., 2002). Os resultados mostraram

um técnica eficiente, tanto para ambientes urbanos como rurais, com baixas taxas de falsos positivos. A detecção de pessoas ficou em torno de 70%, enquanto para veículos fica próxima de 80%.

3.6 CONSIDERAÇÕES

Esta dissertação focou em técnicas de limitação do espaço de busca, principalmente por utilizar plataformas com recursos computacionais escassos. Para isso, são utilizados os métodos: Mapa de Saliências e Processamento de Imagens Térmicas. O Processamento de Imagens Térmicas utilizado neste trabalho se difere do empregado por Gaszczak et al. (2011), principalmente, na tecnologia utilizada. A câmera utilizada neste trabalho produz imagens de baixa resolução (80 x 60), por isso apenas a etapa de detecção dos objetos é feita com informações térmicas, diferente de Gaszczak et al. (2011) que utiliza as informações térmicas na etapa de reconhecimento.

Para melhor eficiência no processo de reconhecimento, uma das técnicas em evidência na literatura (Redes Neurais Convolucionais) foi utilizada a fim de buscar soluções para a variação nos graus de liberdade, enfatizada no trabalho de Blondel et al. (2014a). Os resultados obtidos com as Redes Neurais Convolucionais foram comparados com técnicas HOG+SVM e Cascatas Haar e LBP, técnicas comumente utilizadas na literatura. Diferente da técnica utilizada por Andriluka et al. (2010) que utiliza modelos baseados em partes, nesta dissertação o reconhecimento é feito com modelos monolíticos.

4 MATERIAIS E MÉTODOS

Este trabalho aborda o reconhecimento de pessoas em imagens aéreas. Conforme exposto na seção 2.1, o uso de um Sistema de Reconhecimento de Padrões (SRP), requer a utilização de um classificador eficaz. Para tal é necessário um treinamento igualmente eficaz cobrindo as possibilidades envolvidas no problema. Para isso, seguiu-se uma metodologia de pesquisa baseada na definição de dados, abrangendo diversas bases de imagens a fim de criar uma base de imagens consistente. Por sua vez essa base de dados é utilizada para o treinamento, validação e teste dos classificadores empregados neste trabalho. Por fim, as técnicas propostas passam por testes de desempenho a fim de validar seu desempenho nas situações propostas. O roteiro para a condução deste método é descrito na Figura 16.

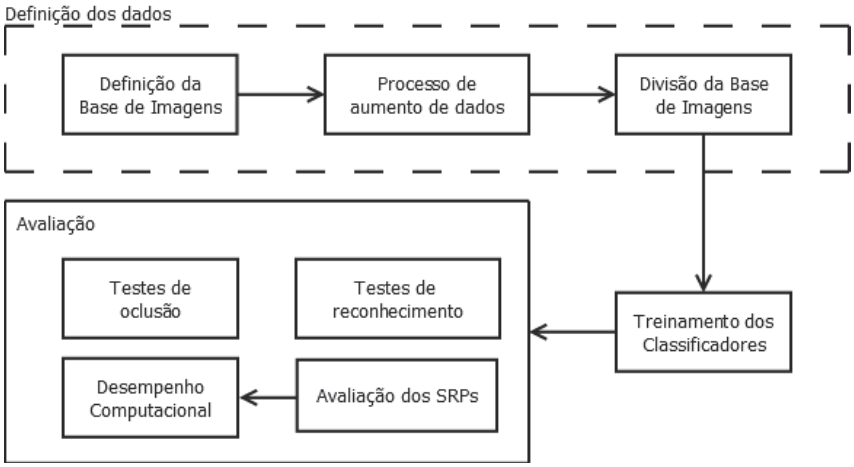

Figura 16: Fluxograma empregado na metodologia.
Fonte: Própria autoria

O primeiro passo consiste em definir uma base de dados consistente para o reconhecimento de pessoas em imagens aéreas. A base de dados deve conter todos os tipos de objetos que contemplam o problema no reconhecimento, diversificando poses, ângulos e forma dos objetos. No caso da aplicação envolvida neste trabalho, a base de dados é constituída por uma base de imagens. Além da definição, também é necessário separar algumas imagens para o treinamento, validação e teste, visto que alguns classificadores utilizam essas três etapas. As imagens de trei-

namento servem para treinar o classificador. As imagens de validação também são utilizadas no treinamento, entretanto apenas como forma de avaliar o treinamento conforme ele evolui. Já as imagens de teste são utilizadas para avaliar a eficácia do classificador ao final da etapa de treinamento.

O treinamento dos classificadores foi realizado conforme os parâmetros que melhor se ajustam ao problema de reconhecimento de pessoas em imagens aéreas. Neste trabalho, foram avaliados Classificadores em Cascata com características Haar e LBP (ou Cascata Haar e Cascata LBP), SVM com características HOG (HOG+SVM) e Redes Neurais Convolucionais, sendo todos os classificadores treinados e testados com o mesmo conjunto de imagens. Os Classificadores em Cascata com características de Haar e LBP foram selecionados devido à sua popularidade na literatura, onde foram identificados diversos estudos utilizando estas técnicas para o reconhecimento de objetos em imagens aéreas (MA'SUM et al., 2013; CAO et al., 2011; DOHERTY; RUDOL, 2007). Já as características HOG com classificador SVM é uma combinação continuamente utilizada na literatura para o reconhecimento de pessoas (SU et al., 2015; CHEN et al., 2011; REILLY et al., 2010). Por fim, as Redes Neurais Convolucionais foram selecionadas devido aos resultados obtidos em trabalhos da literatura com outra aplicação, além da motivação de não haver trabalho abordando o reconhecimento de pessoas em imagens aéreas utilizando esta técnica (PERLIN; LOPES, 2015; SIMONYAN; ZISSERMAN, 2014; KRIZHEVSKY et al., 2012).

Por fim, foram realizados testes de reconhecimento (utilizando as imagens de teste), de oclusão e desempenho em situações reais. Os testes de reconhecimento têm o intuito de comparar os classificadores treinados, determinando assim o mais eficaz para a tarefa de reconhecimento de pessoas em imagens aéreas. Os testes de oclusão têm como objetivo determinar o desempenho dos classificadores sob condições de oclusão parcial, uma situação comum em imagens aéreas. Por fim, os testes em condições reais têm como objetivo avaliar o desempenho dos classificadores juntamente com algoritmos de segmentação e detecção. Nesta etapa final, foi avaliada a união das técnicas de detecção e classificação visando identificar o fornecimento de dados de reconhecimento em Tempo Real. Este teste de desempenho computacional foi realizado em duas plataformas: plataforma embarcada Raspberry Pi 2 e uma Base de Controle Móvel (BCM).

4.1 MATERIAIS

4.1.1 RASPBERRY PI

O Raspberry Pi é um computador com dimensões próximas a de um cartão de crédito. Foi desenvolvido no Reino Unido pela Fundação Raspberry Pi. O hardware permite acoplar diversos módulos comercializados separadamente, como: câmera e sistema para monitor. Diversos sistemas operacionais são disponibilizados para uso do Raspberry Pi, entretanto o mais utilizado é Raspbian, um sistema operacional livre baseado em Linux Debian e otimizado para o Raspberry Pi. A plataforma não inclui disco rígido como nos computadores convencionais. Contudo, fornece entrada para cartão de memória não volátil micro SD, onde o sistema operacional é instalado.

Existem diversos modelos de Raspberry Pi, sendo que o modelo utilizado nesta dissertação é o Raspberry Pi 2 *Model B* v1.1 (Figura 17), que contém memória RAM de 1GB e CPU quad-core ARM Cortex-A7 900MHz, ou seja, quatro processadores físicos. Além disso, oferece quatro portas USB, 40 pinos GPIO, interface de vídeo, porta *Ethernet* e interface para câmera e monitor. Este modelo foi escolhido por conter quatro processadores, possibilitando o uso de múltiplas *threads* para processamento de imagens.

A plataforma foi utilizada para avaliar o desempenho do método proposto em um sistema embarcado. O Raspberry Pi 2 foi escolhido por ser uma plataforma de baixo custo, além de ter um hardware potente para um hardware embarcado de pequeno porte. Também era a plataforma embarcada disponível no Laboratório Avançado de Sistemas Embarcados e Robótica (LASER) com maior capacidade de processamento.

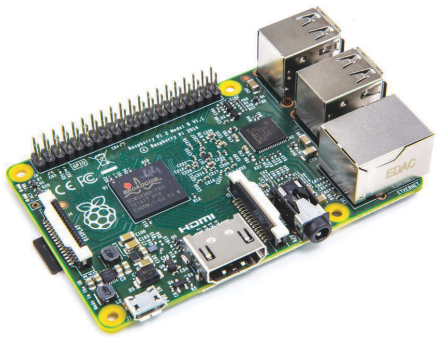

Figura 17: Raspberry Pi 2 *Model B* v1.1.
Fonte: Reproduzido de (UPTON; HALFACREE, 2014)

4.1.2 BASE DE CONTROLE MÓVEL

A Base de Controle Móvel (BCM) utilizada neste trabalho consiste em um laptop comum, com configurações comuns a máquinas populares do mercado. A escolha deste tipo de tecnologia foi feita devido a mobilidade que um laptop permite, pois no uso do VANT juntamente com uma BCM, será necessário um método de comunicação entre as duas plataformas. Assim faz-se necessário que as plataformas não estejam a uma distância muito grande. Além disso, o uso de um laptop comum não exigirá tecnologias de alto custo para o bom funcionamento do sistema.

A máquina utilizada contém memória RAM de 4GB e CPU Intel Core i5-3210M 2.5GHz, com dois processadores físicos e dois lógicos (INTEL, 2012). O Sistema Operacional Linux Ubuntu 14.04 de 64 bits foi escolhido para os experimentos por ser um SO popular e de fácil utilização, além de ser software livre.

4.1.3 CÂMERAS

Para os experimentos conduzidos neste trabalho, foram utilizados dois tipos de câmera para a coleta de imagens de teste: uma *Raspberry Pi Camera* v1.3 (Raspicam) e uma *FLIR Lepton Long Wave Infrared* (FLIR, 2014). A primeira foi usada para a captura de imagens convencionais, enquanto a segunda para a captura de imagens térmicas.

A Raspicam é uma câmera específica para utilização com o Raspberry Pi, e funciona através de interface serial. A câmera pode capturar imagens de 24 bits com resolução máxima de 2592 x 1944 a 15 fps. As imagens podem ser feitas em RGB ou YUV.

A câmera *FLIR Lepton Long Wave Infrared* fornece imagens térmicas de baixa resolução (80 x 60). Esta tecnologia captura ondas *LWIR*, ou seja, ondas de comprimento de onda longo (de 8 a 14 μm), sendo uma ótima opção para captura de imagens aéreas. A câmera contém uma interface serial compatível, tornando possível a utilização em conjunto com o Raspberry Pi, operando em temperaturas entre -10^{o} e 65^{o} Celsius, com possibilidade de operação até -20^{o} C ou 75^{o} C, onde há certa degradação das informações (FLIR, 2014).

Para a captura das imagens foi necessária improvisação de um suporte que comportasse as duas câmeras de forma que fizessem imagens da mesma cena, para que, posteriormente, fosse possível relacionar os pixels da imagem capturada pela Raspicam com os capturados pela câmera térmica. A Figura 18 mostra as câmeras utilizadas no suporte improvisado.

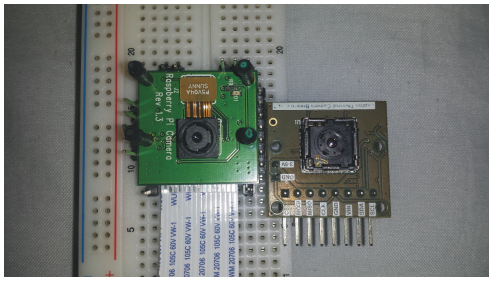

Figura 18: Câmeras utilizadas no trabalho montadas em um suporte improvisado, à esquerda a Raspicam e à direita a câmera térmica.

Fonte: Própria autoria

4.1.4 BASES DE IMAGENS

Neste trabalho foram utilizadas diversas bases de imagens aéreas de pessoas, com o objetivo de fornecer uma grande quantidade de exemplos ao treinamento dos classificadores. Dentre as bases presentes no estado da arte foram selecionadas três bases: GMVRT-v1 (BLONDEL et al., 2014b), GMVRT-v2 (BLONDEL et al., 2014a), UCF-ARG *Data Set* (NAGENDRAN et al., 2010). As bases para treinamento, validação e teste possuem apenas imagens visuais positivas e negativas. Enquanto as imagens empregadas nos testes do Sistema de Reconhecimento de Padrões são imagens visuais e térmicas.

A base de imagens GMVRT-v1 (BLONDEL et al., 2014b) possui uma grande variedade de imagens positivas (de pessoas) com variados ângulos de câmera, envolvendo os três graus de liberdades (*yaw*, *roll* e *pitch*). As imagens positivas disponibilizadas têm a câmera em posições variando entre 0^o e 90^o nos três graus de liberdade. As pessoas presentes nas imagens apresentam diversas poses, além de vestirem roupas em cores variadas. Já o conjunto de imagens negativas, possui variedade de imagens de ambientes urbanos e rurais. Originalmente, todas as imagens são de 3 canais RGB (24 bits), fornecidas no formato JPEG com dimensões de 64 x 128 pixels (largura x altura). A base é fornecida com 4223 imagens positivas e 8461 negativas. A Figura 19 mostra exemplos positivos e negativos extraídos da base GMVRT-v1.

A base GMVRT-v2 (BLONDEL et al., 2014a) possui 3846 instâncias rotuladas como pessoas em variadas poses, além de conter alta variedade de ângulos em três graus de liberdade. Diferente da GMVRT-v1, a versão 2 tem uma grande variedade de pessoas diferentes em ambientes variados. A base também contém imagens rotuladas como negativas, sendo 13821 exemplos de imagens em ambientes variados, desde urbanos a rurais. Todas as imagens têm as mesmas dimensões 128 x 128 pixels, 24 bits e o formato PNG. A Figura 20 mostra exemplos positivos e negativos da base GMVRT-v2.

Figura 19: Imagens da base GMVRT-v1: (a) imagens positivas e (b) imagens negativas.
Fonte: Reproduzido de (BLONDEL et al., 2014b)

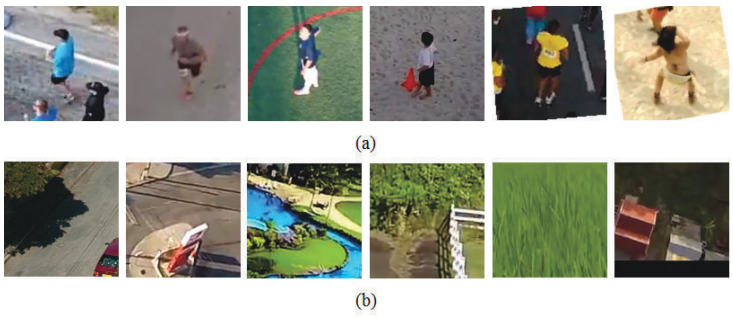

Figura 20: Imagens da base GMVRT-v2: (a) imagens positivas e (b) imagens negativas.
Fonte: Própria autoria

Além dos exemplos preparados para o treinamento, a base GMVRT-v2 traz imagens para teste (BLONDEL et al., 2014a), totalizando 210 imagens. Essas imagens contêm três ou quatro pessoas em ambientes rurais em diversos ângulos de câmera entre 45 e 90 graus nos três graus de liberdade. Com o intuito de aumentar o número de exemplos de pessoas, foram extraídos manualmente dois exemplos positivos de cada imagem, totalizando 420 imagens de pessoas. As imagens originais estão no formato JPEG, com dimensões 1280 x 720 e 24 bits, já as imagens foram extraídas com dimensões 128 x 128 pixels e no formato PNG. A Figura 21 traz uma amostra das imagens de teste disponibilizadas na base e como um exemplo foi extraído.

A base de imagens UCF-ARG *Data Set* (NAGENDRAN et al., 2010) originalmente é disponibilizada em forma de vídeos, em um total de 1440. Os vídeos foram produzidos em três posições diferentes de câmera: aérea, terrestre e sobre o topo de um edifício. Cada

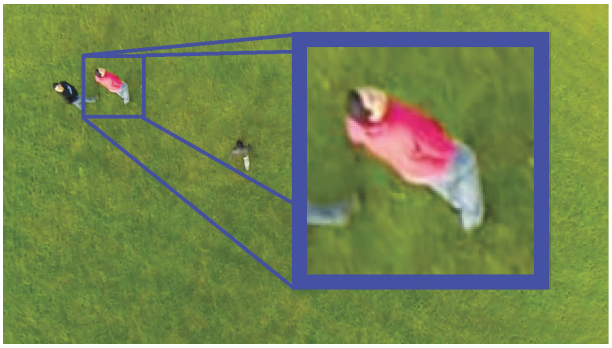

Figura 21: Imagens de teste da base GMVRT-v2, mostrando a extração da imagem de uma pessoa.
Fonte: Reproduzido de (BLONDEL et al., 2014a)

posição de câmera reúne 480 vídeos com resolução 960 x 540 a 30 fps. Para cada câmera, são disponibilizados 480 vídeos onde pessoas executam 10 ações diferentes, como: socar o ar, carregar objetos, andar, cavar entre outros. A partir da base UCF-ARG foi coletado um quadro aleatório de cada vídeo totalizando 480 imagens, onde 410 foram selecionados. Devido a problemas com ruído, algumas imagens foram excluídas do processo. A partir dos quadros 410 imagens de pessoas foram extraídas, sendo que o processo utilizado foi similar ao realizado na base de testes da base GMVRT-v2, resultando 410 em imagens de dimensões 128 x 128 pixels e 24 bits no formato PNG. Esta base é interessante ao presente estudo por conter pequenas variações de escala, mudanças e poses. A Figura 22 contém amostras das imagens coletadas na base UCF-ARG *Data Set*.

Figura 22: Imagens coletadas da base UCF-ARG *Data Set*.
Fonte: Própria autoria

Para os experimentos utilizando câmera térmica, foram coletadas imagens do quarto piso de um edifício, simulando assim VANTs a cerca de 15 metros de altura. O ambiente onde a coleta foi realizada é um local mais próximo do rural, em um espaço amplo com gramado. Foram coletadas 17 imagens convencionais e 17 térmicas, cada par (imagem RGB e imagem térmica) representando o mesmo instante de tempo. As imagens convencionais tem dimensões de 1280 x 720, 24 bits e o formato PNG, já as imagens térmicas dimensões 80 x 60, 24 bits e formato PNG. A Figura 23 mostra as imagens coletadas para os experimentos dos Sistemas de Reconhecimento de Padrões. As imagens RGB foram coletadas utilizando a Raspicam,

enquanto as imagens térmicas foram obtidas com a câmera térmica FLIR.

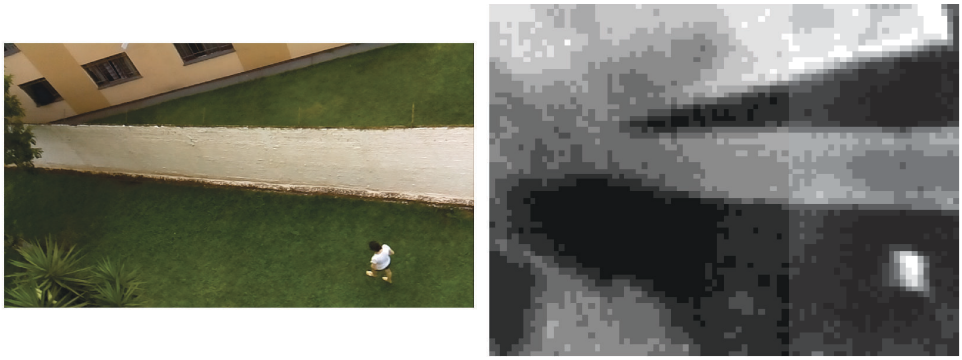

Figura 23: Imagens coletadas para os testes dos SRPs: (a) Imagem RGB coletada com a Raspicam e (b) Imagem térmica coletada com a FLIR. Neste exemplo as imagens foram redimensionadas para melhor visualização, além da intensidade da Imagem Térmica ter sido aumentada.

Fonte: Própria autoria

4.1.5 OPENCV

A OpenCV (*Open Source Computer Vision Library*) (BRADSKI et al., 2000) é uma biblioteca voltada para a programação de sistemas de visão computacional. Originalmente desenvolvida pela Intel, tem suporte para as linguagens de programação C, C++, Python e Java. Atualmente, reúne mais de 500 funções de processamento de imagens, reconhecimento de padrões e aprendizagem de máquina.

Na OpenCV estão contidas as funcionalidades para treinamento, teste e utilização dos Classificadores em Cascata, tanto com características de Haar quanto LBP. Também através da OpenCV é possível extrair características HOG e usá-las no treinamento do SVM. Além disso, através do opencv_contrib, são disponibilizadas diversos algoritmos recentes desenvolvidos pela comunidade do OpenCV. Dentre as funcionalidades está uma versão do Mapa de Saliências disponibilizado por Montabone e Soto (2010). Neste trabalho o OpenCV é utilizado para o treinamento e teste dos Classificadores em Cascata, HOG+SVM e uso do Mapa de Saliências, além do processamento de imagens necessário nas imagens térmicas.

4.1.6 CAFFE

O *Framework* Caffe (JIA et al., 2014) é uma ferramenta de código aberto para o estudo de *Deep Learning*, no caso deste trabalho, para Redes Neurais Convolucionais (CNNs). O

código é escrito em C++ e possui suporte para a biblioteca NVIDIA CUDA, ferramenta usada para o processamento com placas de vídeo NVIDIA. Além de C++, o Caffe tem suporte para Python/Numpy e MATLAB. Diversos tipos de camadas são disponibilizadas para a criação de arquiteturas de redes neurais, além de ferramentas para treinamento e teste de CNNs com imagens. Através do Caffe também é possível visualizar gráficos do aprendizado do treinamento e mostrar em forma de imagens os mapas de características utilizadas pela rede para o aprendizado.

Além de todas essas vantagens, o Caffe é bastante utilizado em trabalhos acadêmicos e aplicações industriais em diversas áreas, possuindo um extenso material para consulta, e uma comunidade de usuários ativa. Nesta dissertação, o *Framework* foi utilizado para o treinamento, teste e avaliação de uma CNN para o reconhecimento de pessoas em imagens aéreas.

4.1.7 OPENBLAS

O OpenBLAS é uma implementação de código aberto do BLAS (*Basic Linear Algebra Subprograms*) (LAWSON et al., 1979), uma especificação que estabelece uma série de rotinas de baixo nível para realizar operações de álgebra linear, como adição de matrizes, multiplicações e produtos escalares, combinações lineares e multiplicação de matrizes. O OpenBLAS é otimizado para diversas arquiteturas, incluindo ARM, tendo suporte para o uso de multi-*threads*. Sendo assim o OpenBLAS é uma ferramenta adequada para otimização de sistemas, principalmente em situações onde há diversos processadores disponíveis. Nesta dissertação, o OpenBLAS foi utilizado para otimizar o desempenho dos experimentos na placa Raspberry Pi e na BCM.

4.2 MÉTODOS

Nesta seção serão apresentados os métodos utilizados neste trabalho, sendo: A definição da base de imagens, o treinamento executado no classificadores e os métodos de avaliação dos experimentos.

4.2.1 DEFINIÇÃO DOS DADOS

Durante os experimentos, quatro bases de imagens foram utilizadas para treinamento e avaliação dos classificadores. A base de imagens total é composta pela união destas quatro bases: GMVRT-v1 (BLONDEL et al., 2014b), GMVRT-v2 (BLONDEL et al., 2014a), GMVRT-v2-*test* (BLONDEL et al., 2014a) e UCF-ARG (NAGENDRAN et al., 2010). O objetivo de

usar quatro diferentes bases de imagens é aumentar a generalização dos classificadores visando aproximar os experimentos das aplicações no mundo real.

Como todos os classificadores foram projetados para imagens de 256 x 256 pixels, o tamanho de todas as imagens teve de ser padronizado. Para esta padronização, na base GMVRT-v1 as imagens sofreram um processo de *squash*, ou seja, suas dimensões originais foram alteradas, produzindo imagens de objetos que parecem "espremidos". A Figura 24 mostra um exemplo deste processo. Todas as outras bases apenas sofreram o processo de redimensionamento, passando de 128 x 128 pixels para 256 x 256.

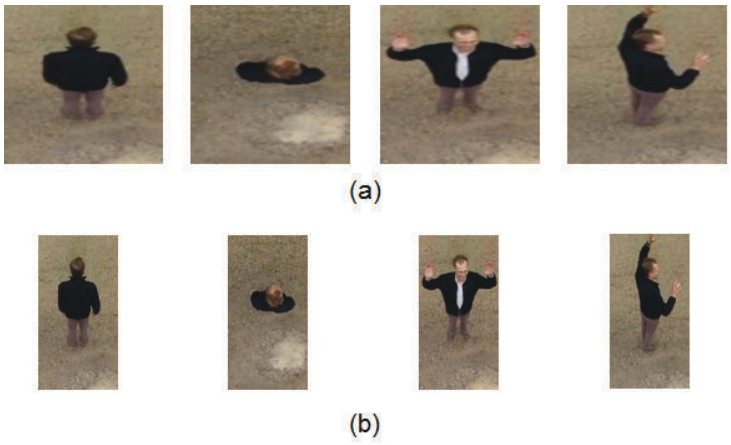

Figura 24: Processo de *squash* das imagens da base GMVRT-v1: (a) Imagens após o processo de *squash*, e (b) Imagens originais.

Fonte: Própria autoria

4.2.1.1 PROCESSO DE AUMENTO DE DADOS

Como a Rede Neural Convolucional exige um grande número de imagens para treinamento (KRIZHEVSKY et al., 2012), foi necessário conduzir um processo de aumento de dados (PAD), similar ao utilizado por Perlin e Lopes (2015). Além de aumentar a quantidade de imagens, através do PAD, também foi necessário aumentar o número de exemplos positivos, equilibrando assim as classes positivas e negativas. Este PAD tem como premissa expandir os exemplos positivos da base de imagens aplicando pequenas variações nas imagens originais. As transformações são aplicadas de forma automática e aleatória, sendo as transformações: (i) translação (20 pixels para eixos x e/ou y), (ii) escala (entre 0.98 e 1.1) e/ou (iii) rotação (entre -20 e 20 graus). Durante o processo cada transformação teve 50% de chance de acontecer, sendo que uma imagem deve passar por pelo menos uma transformação. A Figura 25 mostra exemplos dessas transformações.

Figura 25: Exemplos do Processo de Aumento de Dados: (a) rotação, (b) translação, (c) escala, (d) rotação e translação, e (e) rotação, translação e escala.

Fonte: Própria autoria

Após o processo de PAD, o número total de imagens passou de 31184 para 44765, um aumento de 13381 imagens positivas. O processo foi conduzido uma vez na base GMVRT-v1 e duas vezes nas bases GMVRT-v2, GMVRT-v2-*test* e UCF-ARG. Dessa forma, foi possível chegar à um equilíbrio entre as classes positivas e negativas. A Tabela 2 mostra a distribuição das classes conforme o PAD que resultou na base de imagens utilizada neste trabalho.

Tabela 2: Comparação da quantidade de imagens pré e Pós Processo de Aumento de Dados, resultado na base de imagens utilizada neste trabalho.

	Imagens originais		Pós Processo de Aumento de Dados		
Base	Nº positivos	Nº negativos	Nº de PADs	Nº positivos	Nº negativos
GMVRT-v1	4223	8461	1	8446	8461
GMVRT-v2	3846	13821	2	11538	13821
GMVRT-v2-test	420	0	2	1260	0
UCF-ARG	413	0	2	1239	0
Total	8902	22282	-	22483	22282

Fonte: Própria Autoria.

4.2.1.2 DIVISÃO DA BASE DE IMAGENS

Em aplicações de Reconhecimento de Padrões é comum a divisão da base de dados em treinamento, validação e teste (PERLIN; LOPES, 2015; GURURATSAKUL et al., 2010; SU et al., 2015; TEUTSCH et al., 2014). Esta divisão é chamada *Holdout*, e tem como objetivo avaliar a generalização de um modelo (KOHAVI et al., 1995), no caso, de um classificador. O conjunto de treinamento tem como objetivo oferecer exemplos para o aprendizado do algoritmo, enquanto o conjunto de validação verifica a generalização do classificador durante o treinamento. Já o conjunto de testes é utilizado após o fim do treinamento do classificador, a

fim de avaliar se a generalização desejada foi atingida. Na implementação dos algoritmos utilizadas nesta dissertação, o SVM e os Classificadores em Cascata utilizam apenas as etapas de treinamento e teste. Já as Redes Neurais Convolucionais utilizam os três conjuntos, treinamento e validação na etapa de treinamento e teste para a etapa a avaliação do treinamento.

Para a execução desta dissertação, a aplicação requer a divisão da base de imagens nas três etapas da *Holdout* (treinamento, validação e teste). Entretanto, no treinamento dos Classificadores em Cascata e o SVM o conjunto de validação é descartado. Assim, as imagens separadas para validação não são utilizadas em nenhuma etapa, seja treinamento ou teste. Na literatura diversas formas de divisão são aplicadas, variando conforme a necessidade do problema. Para este trabalho a base é dividida com as seguintes proporções: 80% das imagens para o conjunto de treinamento, 10% para o conjunto de validação e 10% para o conjunto de teste. Para manter a proporção de cada base de imagens utilizada, a divisão dos conjuntos foi feita conforme a Tabela 3.

Tabela 3: Distribuição das bases de imagem através do *Holdout*

	Treinamento			Validação / Teste		
Bases	%	Positivos	Negativos	%	Positivos	Negativos
GMVRT-v1	80%	6755	6769	10%	845	846
GMVRT-v2		9228	11057		1154	1382
GMVRT-v2-test		1008	0		126	0
UCF-ARG		991	0		124	0
Total	-	17982	17826	-	2248	2228

4.2.2 TREINAMENTO DOS CLASSIFICADORES

4.2.2.1 CLASSIFICADORES EM CASCATA

O treinamento dos Classificadores em Cascata foi conduzido utilizando a biblioteca OpenCV. Os classificadores foram treinados separadamente, ambos com o mesmo conjunto de treinamento. Para os Classificadores em Cascata, não é necessário definir conjuntos de validação do treinamento, portanto o conjunto de validação é descartado neste momento.

Os dois Classificadores em Cascata (Haar e LBP) foram treinados com a versão clássica do AdaBoost, também conhecida como *Gentle* Adaboost (FREUND; SCHAPIRE, 1997). O treinamento dos dois classificadores foi configurado com uma cascata de 15 níveis, com uma taxa mínima de falsos alarmes de 0,5 e taxa de verdadeiros positivos de 0,95. Os parâmetros foram ajustados conforme os treinamento realizados posteriormente, sendo os parâmetros inicias

os definidos como padrão pela OpenCV.

4.2.2.2 HOG + SVM

Assim como os Classificadores em Cascata, o treinamento do SVM também foi conduzido com a biblioteca OpenCV. O SVM também não necessita do conjunto de validação, então apenas o conjunto de treinamento foi utilizado durante o treinamento do algoritmo.

O SVM foi treinado com sua forma mais simples: *kernel* Linear, com critério de parada de 1^{-6} ou 100 mil iterações. Para isto foram utilizadas as características HOG. Neste trabalho, as características HOG foram definidas com número de *bins* 9, tamanho de bloco 32 x 32, tamanho de célula 16 x 16 e passo 16 x 16, totalizando 8100 características por imagem. Estes parâmetros foram definidos e ajustados conforme os treinamentos realizados, sendo o ponto de partida os parâmetros iniciais da OpenCV.

4.2.2.3 REDES NEURAIS CONVOLUCIONAIS

Duas arquiteturas de Redes Neurais Convolucionais foram utilizadas nesta dissertação. As duas arquiteturas são diferentes, principalmente quanto ao estágio de extração de características.

A primeira arquitetura de Rede Neural Convolucional utilizada foi a proposta por Perlin e Lopes (2015), neste trabalho ela será chamada CNN1. Esta arquitetura foi utilizada para identificação biométrica na forma de roupas e gênero. O método proposto chegou a uma boa capacidade de generalização, classificando três diferentes atributos com taxas de acurácia acima de 70% (PERLIN; LOPES, 2015). Esta arquitetura foi escolhida por já estar inserida no problema de reconhecimento de pessoas, dessa forma tem grande potencial para a aplicação foco deste trabalho. Além disso, a arquitetura da CNN1 proposta tem um número de camadas pequeno, o que pode resultar em um processamento rápido, atingindo melhores resultados em tempo computacional. O desempenho computacional desta arquitetura de rede não foi avaliado por Perlin e Lopes (2015).

A arquitetura da CNN1 é mostrada na Figura 26, dividida em seis camadas, onde três são camadas convolucionais e três totalmente conectadas. A diferença da arquitetura aplicada neste trabalho é quanto as saídas. No trabalho desenvolvido por Perlin e Lopes (2015), duas CNN são utilizadas, tendo duas ou três classes, neste trabalho são apenas duas classes, dessa forma a última camada terá dois neurônios de saída. Inicialmente essa arquitetura foi projetada para imagens de dimensões 128 x 128, entretanto neste trabalho a arquitetura interna original

foi mantida para imagens de entrada com dimensões 227 x 227. Assim, a arquitetura utilizada se difere da proposta por Perlin e Lopes (2015) nas camadas de entrada e saída.

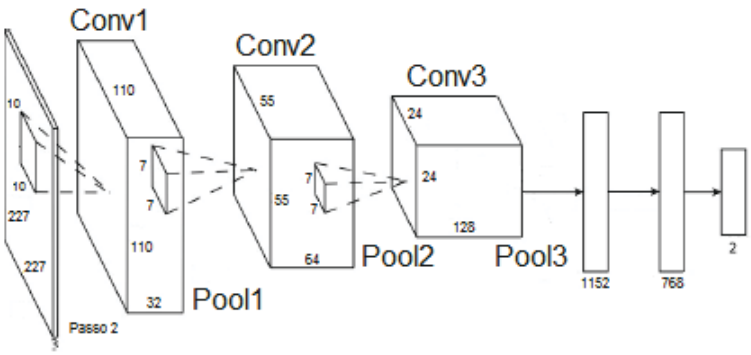

Figura 26: Arquitetura da CNN1 utilizada no trabalho.
Fonte: Adaptado de (PERLIN; LOPES, 2015).

A Tabela 4 mostra detalhes da arquitetura da CNN1 utilizada neste trabalho. A entrada da rede é uma matriz 3D, ou seja, uma imagem de dimensões 227 x 227 com canais vermelho, verde e azul (RGB). As imagens de entrada, como definido na seção 4.1.4, são de 256 x 256, entretanto quando essas imagens entram na rede elas sofrem um pré-processamento, onde as bordas são cortadas. Assim, as bordas são cortadas, como as pessoas estão centralizadas na imagem, elas sofrem poucas alterações com o processo.

A primeira camada convolucional (Conv1) filtra a entrada de 227 x 227 x 3 em 32 filtros de 10 x 10 com passo 2 x 2, resultando em uma saída de 110 x 110 x 32. Após a aplicação de uma função de ReLU $mx(x,0)$, o próximo passo é a aplicação do *pooling* (Pool1) com uma função de máximo. O tamanho de Pool1 é de 2 x 2 e passo 2 x 2, reduzindo a saída da primeira camada para 55 x 55 x 32. A camada convolucional Conv2 é composta por 64 mapas de características e filtros de 7 x 7 e passo 1 x 1. Novamente, após a aplicação do ReLU máximo, há a aplicação de uma função de *pooling* máximo (Pool2) com filtro 2 x 2 e passo 2 x 2, a saída será de 24 x 24 x 64. A terceira camada convolucional (Conv3) e última etapa de extração de características, gera 128 mapas de características através de um filtro 7 x 7. Da mesma forma que as camadas anteriores, o ReLU máximo é aplicado, seguido de um *pooling* máximo (Pool3) com filtros 2 x 2 e passo 2 x 2. A saída para a camada totalmente conectada é de 6 x 6 x 128, totalizando 4608 parâmetros.

A quarta camada é onde inicia a etapa de classificação da CNN1 através de uma camada totalmente conectada. Por fim, a quinta camada é outra camada totalmente conectada, contendo 768 neurônios. Nas duas camadas também há funções de ReLU máximo a fim de diminuir a

saturação da rede. Durante a fase de treinamento, a quarta e quinta camadas ainda sofrem o processo de *Dropout*, a fim de garantir a generalização da rede sem que ela atinja a situação de *overfitting*. A cada iteração de treinamento, 50% dos neurônios de cada camada são desativados e não participam do treinamento naquela iteração.

Tabela 4: Detalhes das camadas de convolução da arquitetura da CNN1.

	Conv1	Pool1	Conv2	Pool2	Conv3	Pool3
Mapas de Entrada	3	32	32	64	64	128
Entradas	227x227	110x110	55x55	49x49	24x24	18x18
Mapas de Saída	32	32	64	64	128	128
Saída	110x110	55x55	49x49	24x24	18x18	9x9
Filtros	10x10	2x2	7x7	2x2	7x7	2x2
Passo	2x2	2x2	1x1	2x2	1x1	2x2

A segunda arquitetura Rede Neural Convolucional utilizada foi a proposta por Krizhevsky et al. (2012), chamada aqui de CNN2. Esta arquitetura foi a vencedora da competição ILSVRC 2012 de reconhecimento de objetos, onde venceu com uma diferença de 9.7% para o segundo colocado. Recentemente outras arquiteturas foram propostas por Simonyan e Zisserman (2014) e Szegedy et al. (2015) no ILSVRC 2014, inclusive com melhores resultados. A arquitetura de Krizhevsky et al. (2012), tem um desempenho inferior, mas foi escolhida para uso neste trabalho, pelo fato de requerer um tempo de treinamento menor, visto que é uma rede menor em comparação com as outras mencionadas.

A arquitetura da CNN2 é mostrada na Figura 27, dividida em oito camadas, onde cinco são camadas convolucionais e três totalmente conectadas. A diferença da CNN utilizada no trabalho com a original é a saída, na original a saída é formada por 1000 classes, enquanto neste trabalho são apenas duas saídas.

A Tabela 5 mostra detalhes da arquitetura da CNN2 utilizada neste trabalho. A primeira camada convolucional (Conv1) filtra a entrada (uma imagem RGB) de 227 x 227 x 3 em 96 filtros de 11 x 11 com passo 4, resultando em uma camada de 55 x 55 x 96. A primeira camada passa por camadas de ReLU, LRN e *pooling* máximo (Pool1) de tamanho 5x5 e passo 2. A camada convolucional Conv2 filtra o resultado do *pooling* 27 x 27 x 96, com 256 filtros de 5 x 5, passo 1 e *zero-padding* 2, sendo que o resultado é uma camada de 27 x 27 x 256. O resultado da segunda camada também passa por camadas de ReLU, LRN e *pooling* (Pool2) de 3 x 3 e passo 2, reduzindo o tamanho dos mapas de características para 13 x 13 x 256. Já a camada convolucional Conv3, filtra a entrada com 384 filtros de 3 x 3, passo 1 e *zero-padding* 1, passando por uma camada de ReLU. A quarta camada convolucional (Conv4) tem as mes-

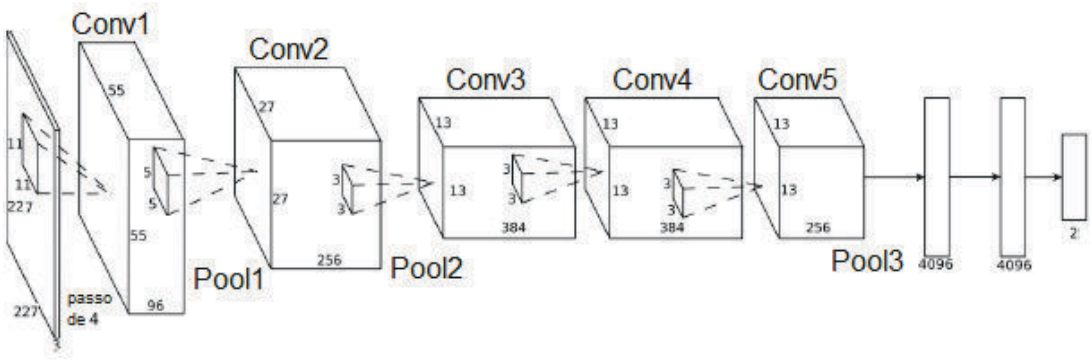

Figura 27: Arquitetura da CNN2 utilizada no trabalho.
Fonte: Adaptado de (KRIZHEVSKY et al., 2012)

mas configurações da terceira. A quinta camada (Conv5), utiliza 256 filtros de 3 x 3, passo 1 e *zero-padding* 1, que passa por ReLU e *pooling* (Pool3) 3 x 3 e passo 2. As três últimas camadas são totalmente conectadas, onde a sexta e sétima tem saída de 4096 neurônios, ambas passam por ReLU e sofrem processo de *dropout* para evitar *overfitting*.

Tabela 5: Detalhes das camadas de convolução da arquitetura da CNN2.

	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool3
Mapas de Entrada	3	96	96	256	256	384	384	256
Entradas	227x227	55x55	27x27	27x27	13x13	13x13	13x13	13x13
Mapas de Saída	96	96	256	256	384	384	256	256
Saída	55x55	27x27	27x27	13x13	13x13	13x13	13x13	6x6
Filtros	11x11	3x3	5x5	3x3	3x3	3x3	3x3	3x3
Passo	4x4	2x2	1x1	2x2	1x1	1x1	1x1	2x2
Preench. Zeros	0	0	2	0	1	1	1	0

Para o treinamento das CNNs, o conjunto de treinamento e validação foi utilizado. O ajuste dos pesos da rede é feito através de um algoritmo de *back-propagation* com *SoftMax* como função de métrica de *Loss*. Além disso, o algoritmo SGD é utilizado como estratégia de otimização. A taxa de aprendizagem inicia em 0,005, que é multiplicada por 0,1 a cada 30 épocas. O momento é configurado em 0,9 para auxiliar a convergência do método. Por fim, o número máximo de épocas de treinamento foi configurado como 1000. Estes parâmetros são indicados nos trabalhos de Perlin e Lopes (2015) e Krizhevsky et al. (2012).

O treinamento da CNN é um processo que exige grande tempo de processamento e recursos computacionais. Isto se deve à grande quantidade de camadas de neurônios envolvidos no processo. Por este motivo, a etapa de treinamento foi conduzida em um computador

equipado com GPU, um hardware que tem o poder de reduzir consideravelmente o tempo de treinamento. Para melhorar a eficiência do treinamento, um mini lote de 128 imagens foi definido, conforme sugerido por Krizhevsky et al. (2012). A cada iteração de treinamento, o número de imagens configurado no mini lote é carregado na memória da GPU. Todas as imagens do mini-lote são então submetidas ao treinamento da CNN. Após passar o mini lote pela rede, os pesos da rede são alterados conforme as respostas obtidas na CNN e a rotulação dos exemplos. Segundo Krizhevsky et al. (2012), o mini lote influencia a função de Perda do treinamento. Quanto menor o mini lote maior será a instabilidade da Perda no treinamento. Já quanto maior o mini lote menor será a instabilidade da função de Perda, entretanto mais memória será exigida da GPU.

4.2.3 AVALIAÇÃO

4.2.3.1 TESTES DE RECONHECIMENTO DOS CLASSIFICADORES

Para avaliar a robustez e generalização dos classificadores, diversos experimentos foram executados. Estes experimentos são executados com o conjunto de testes, submetendo cada uma das imagens à classificação. O principal objetivo é comparar os classificadores propostos, avaliando qual atingiu o melhor desempenho para o reconhecimento de pessoas em imagens aéreas. Todos os classificadores aqui treinados são binários, ou seja, há apenas duas classificações possíveis: positivo (é uma pessoa) ou negativo (não é uma pessoa). Em classificadores binários, quatro medidas básicas podem ser extraídas: (i) Verdadeiros Positivos (VP), (ii) Verdadeiros Negativos (VN), (iii) Falsos Positivos (FP) e (iv) Falsos Negativos (FN).

Os **Verdadeiros Positivos** correspondem aos objetos rotulados como positivos e classificados como positivos. Os **Verdadeiros Negativos** são objetos rotulados como negativos e classificados como negativos. Já os **Falsos Positivos** são objetos negativos classificados como positivos, enquanto os **Falsos Negativos** são exemplos positivos classificados como negativos. Dessa forma simples, os VPs e VNs são acertos e FPs e FNs são erros.

Para facilitar a visualização dos dados, as medidas são expressas através de uma tabela chamada Matriz de Confusão. Nesta matriz, os acertos estão na diagonal principal, enquanto os erros na diagonal contrária. A Matriz de Confusão de um classificador ideal deve ter erros igual a zero, ou seja, os elementos da diagonal contrária devem ter valor 0. A Tabela 6 mostra a estrutura de uma Matriz de Confusão, onde as colunas representam a rotulação feita *a priori*, enquanto as linhas representam a classificação das imagens segundo o classificador treinado.

Através das quatro medidas da Matriz de Confusão é possível calcular diversas me-

Tabela 6: Estrutura da Matriz de Confusão.

	Positivos	Negativos
Positivos	VP	FP
Negativos	FN	VN

didas de desempenho. A Acurácia é uma medida que leva em conta os acertos sob todos os exemplos do conjunto de teste. A medida de acurácia é uma boa métrica quando os exemplos do conjunto de treinamento são balanceados, ou seja, todas as classes do conjunto de treinamento tem a mesma porcentagem de dados. O conjunto de treinamento desta dissertação é balanceado, visto que 50,22% do conjunto é formado de exemplos positivos e 49,78% são negativos. A Equação 29 mostra o cálculo da Acurácia.

$$Acc = \frac{VP+VN}{VP+VN+FP+FN} \tag{29}$$

Outra medida utilizada neste trabalho é a Sensibilidade, *Recall* ou Taxa de VP. Esta medida avalia a taxa de classificação dos positivos como positivos. Com esta medida é possível avaliar o desempenho na classificação dos positivos isoladamente. A mesma ideia é seguida pela Especificidade. Entretanto, esta avalia o desempenho da classificação dos negativos. As Equações 30 e 31 mostram, respectivamente, o cálculo da Sensibilidade e Especificidade.

$$Sensibilidade = \frac{VP}{VP+FN} \tag{30}$$

$$Especificidade = \frac{VN}{VN+FP} \tag{31}$$

Além dessas análises através de cálculos, um método gráfico para avaliação foi utilizado: o gráfico ROC (*Receiver Operating Characteristics*) (FAWCETT, 2006). Este é um método originalmente utilizado na detecção de sinais, para avaliar a qualidade de transmissão de um sinal em um canal com ruído (EGAN, 1975). Esta análise é indicada em casos onde as classes são desbalanceadas ou quando se deve levar em consideração custo (1 - Especificidade) e benefício (Sensibilidade) (PRATI et al., 2008).

No Gráfico ROC, o classificador ideal é visto como o mais próximo possível do ponto (0,1). De maneira geral, quanto mais próximo do ponto (0,1) está um classificador, melhor ele é. Classificadores mais ao lado esquerdo são chamados "conservadores", pois fazem classificações positivas somente com forte evidência, portanto tendem a produzir menos Falsos Positivos.

Já classificadores à direita são chamados "liberais", visto que fazem classificações positivas com poucas evidências e com isso produzem mais Falsos Positivos. No Gráfico ROC existe uma linha diagonal que cruza todo o gráfico, esta linha representa a classificação aleatória. A Figura 28 mostra um exemplo de Gráfico.

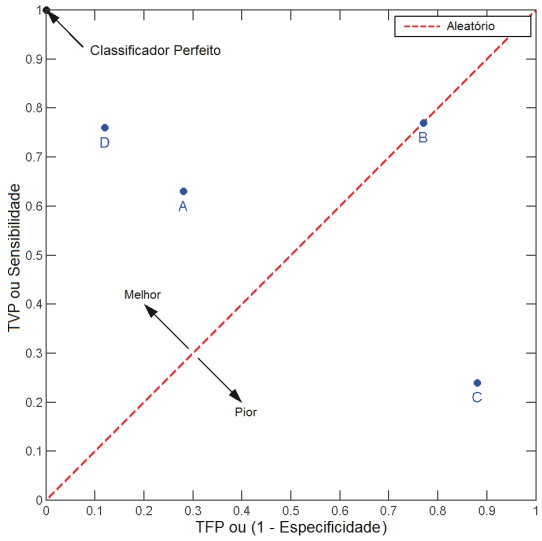

Figura 28: Exemplo de Gráfico ROC. O Classificador D pode ser considerado o melhor, seguido pelo A que é mais liberal. Já o Classificador B pode ser considerado aleatório, enquanto C pode ser considerado ruim.

Fonte: Própria autoria.

4.2.3.2 TESTES DE OCLUSÃO

Além dos testes utilizando o conjunto de testes, foram realizados experimentos de oclusão para avaliar o desempenho dos classificadores em situações em que a pessoa esta parcialmente encoberta por algum elemento da cena. Para isso, se utilizou uma técnica simples, onde 25% da imagem foi encoberta por um quadrado preto. Dessa forma, a imagem é dividida em quatro quadrantes, gerando quatro situações de oclusão diferentes: superior direito, superior esquerdo, inferior direito e inferior esquerdo. Para esta etapa, 8067 imagens positivas da base GMVRT-v1 e GMVRT-v2 utilizadas no treinamento foram selecionadas. As bases foram selecionadas para esta tarefa por terem todos os exemplos de pessoas centralizados, isto assegura que a oclusão dos objetos será próxima de 25%. A Figura 29 exemplifica as imagens que simulam estas situações.

As medidas utilizadas nos Testes de Oclusão são: Número de acertos e Sensibilidade. A Acurácia e a Especificidade não são utilizadas, visto que apenas exemplos positivos são submetidos a este teste. O Número de acertos mostra o valor absoluto de acertos atingidos no

experimento. Enquanto que a Sensibilidade mostra o número de acertos em relação ao total de exemplos. Assim, a Sensibilidade mostra a efetividade do classificador em relação aos exemplos positivos.

Figura 29: Amostras do experimento de oclusão.
Fonte: Própria autoria

4.2.3.3 ANÁLISE DO SISTEMA DE RECONHECIMENTO DE PADRÕES

Nesta etapa do trabalho, os classificadores trabalham em conjunto com uma de duas técnicas de detecção de objetos: mapas de saliência e imagem térmica. Sendo cinco classificadores e duas técnicas de detecção, serão dez combinações diferentes de experimentos. Imagens convencionais de dimensões 1280 x 720 e imagens térmicas de 80 x 60 são as entradas do sistema. A Figura 30 ilustra a relação das técnicas e classificadores, baseado na ideia do processo visto na Seção 2.1.

A etapa de aquisição envolve os dois dispositivos de captura: Raspicam (câmera convencional) e a *FLIR Lepton Long Wave Infrared* (câmera térmica). A aquisição das imagens foi feita simultaneamente. Assim, as imagens capturadas têm informações referentes ao mesmo instante de tempo. Contudo, as informações das duas câmeras não são completamente sincronizadas pelo fato das câmeras funcionarem em taxas de captura diferentes. A Raspicam funciona a 15 fps, enquanto a FLIR a 40 fps. Entretanto, a diferença de sincronia tem pouca (ou quase nenhuma) influência no reconhecimento, visto que a diferença entre os quadros é próxima de 41 milissegundos.

A segmentação e detecção é feita por dois algoritmos: Mapa de Saliências e Processamento de Imagens. O Mapa de Saliências processa apenas a imagem RGB, fornecendo uma imagem em tons de cinza onde destacam-se as saliências visuais da imagem. Esta imagem sofre um processo de Binarização, facilitando assim a extração dos objetos a serem submetidos ao classificador. Os pixels com valor 1 e que estiverem interconectados formarão um objeto de interesse. Estes objetos são pré-avaliados por uma regra simples: caso tenham largura ou altura maior que 256 pixels, são descartados, senão são submetidos à classificação. Por fim, os objetos encontrados são extraídos com tamanho fixo de 256 x 256 (conforme o treinamento dos classificadores) e submetidos ao classificador.

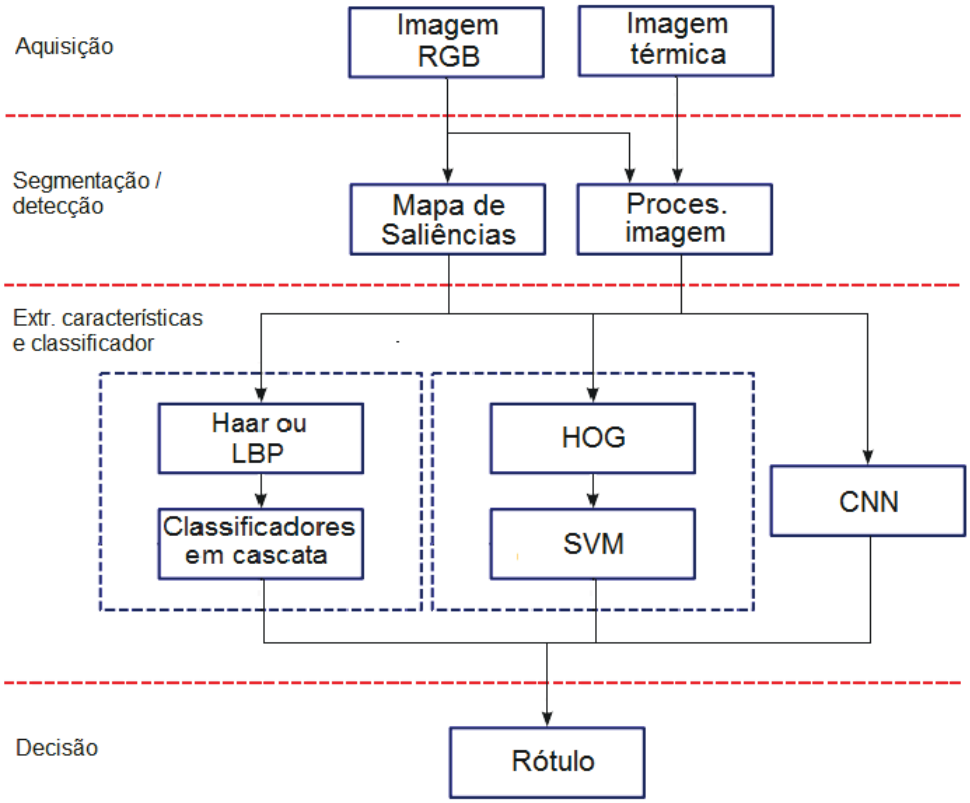

Figura 30: Sistemas de Reconhecimento de Padrões utilizados neste trabalho.

Fonte: Própria autoria

O Processamento de Imagens identifica possíveis objetos através das informações térmicas. A imagem térmica é fornecida em tons de cinza e sofre um processo de limiarização com um intervalo de limiares de 95 e 105. Através de experimentos constatou-se, empiricamente, que o corpo humano estava contido em regiões da imagem onde os valores dos pixels estavam dentro do intervalo de 95 e 105 estipulado. Dessa forma, todos os pixels que estiverem no intervalo de limiares receberão valor 1, do caso contrário, receberão 0. Com esse processo de limiarização a detecção de objetos é feita da mesma forma que no processo com Mapa de Saliências, através da Binarização e detecção dos objetos. A diferença nesta abordagem é quanto a imagem submetida ao classificador, o objeto a ser classificado é extraído da imagem RGB através relação entre imagens visual e térmica. Assim, a extração da imagem submetida ao classificador depende da informação contida nas imagens térmicas. A relação estipula a que ponto na imagem RGB um pixel da imagem térmica pertence. Esta relação é feita para o eixo x e y de um pixel (x,y) da imagem térmica definidas, respectivamente, pelas Equações 32 e 33. Onde x' e y' correspondem as posições da imagem RGB, W é a largura das imagens, H a altura das imagens, T é a imagem térmica e RGB a imagem RGB.

$$x' = (\frac{W_{RGB}}{W_T})x \tag{32}$$

$$y' = (\frac{H_{RGB}}{H_T})y \tag{33}$$

Em alguns objetos detectados nas imagens térmicas haviam áreas maiores que um pixel, neste caso todos os pixels conectados foram tratados como apenas um objeto. Assim, para o cálculo das equações 32 e 33 o pixel mais ao centro foi utilizado. Os objetos submetidos a classificação foram extraídos com dimensões 256 x 256, conforme o tamanho das imagens utilizadas no treinamento dos classificadores.

Os objetos extraídos pelas técnicas de segmentação e detecção são submetidos então aos classificadores. Através dos parâmetros ajustados durante o treinamento, os classificadores executam o processo de classificação de cada um dos objetos, alocando cada um sua respectiva classe. Por fim, os objetos reconhecidos como pessoas são marcados com quadrados vermelhos de dimensões 220 x 220 na imagem original.

Além de avaliar o desempenho de classificação de todo o sistema, nesta fase também é avaliado o desempenho computacional de cada uma das combinações. Esta avaliação é feita em uma plataforma embarcada Raspberry Pi 2 *Model* B v1.1 e um *notebook* com configurações com processador Intel Core i5-3210M 2.5GHz e RAM de 4GB. Os sistemas operacionais utilizados são: Raspbian para o Raspberry Pi e Linux Ubuntu 14.04 para o notebook.

Para mensurar o tempo de execução das técnicas de segmentação e detecção, extratores de características e classificadores, foi utilizada a função *GetTickCount()* disponibilizada no OpenCV. Ela retorna o tempo de execução de um trecho de programa em milissegundos. A função *GetTickCount()* foi inserida no código fonte antes e após a execução de cada etapa. Para uma única imagem, nas etapas de extração de características e classificação o tempo fornecido é a média entre todas as execuções, visto que várias execuções de cada etapa são feitas na imagem. Juntamente com as médias, também foram coletados os valores de desvio padrão. Já na etapa de segmentação e detecção, o tempo computado é o absoluto, visto que a execução é feita apenas uma vez por imagem.

A análise computacional dos experimentos foi feita através do maior tempo. É importante levar em conta o maior tempo, pois através dele é mais provável que o Sistema de Reconhecimento de Padrões sempre atinja os requisitos de tempo especificados para o bom funcionamento do sistema. Além disso, os menores tempos e tempos médios também são avaliados, visto que podem fornecer informações valiosas para a melhorias futuras nas técnicas.

4.3 CONSIDERAÇÕES

Na metodologia empregada neste trabalho, algumas das contribuições especificas para o reconhecimento de pessoas em imagens aéreas (ver Seção 1.3) ficam evidentes:

- A utilização de Redes Neurais Convolucionais para o reconhecimento de pessoas em imagens aéreas (ver Seção 4.2.2).
- O método utilizando câmeras térmicas de baixa resolução juntamente com classificadores (ver Seção 4.2.3).
- A avaliação do desempenho de diferentes sistemas de reconhecimento de padrões implementados em um sistema embarcado (ver Seção 4.2.3).

5 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados os resultados obtidos nos experimentos especificados na Metodologia (Capítulo 4). Primeiramente são mostrados os resultados do treinamento dos classificadores (Subseção 4.2.2), avaliando questões como convergência do treinamento para uma solução e *overfitting*. Em seguida, são mostrados os resultados obtidos no reconhecimento dos classificadores, conforme os experimentos metodológicos da Seção 4.2.3.1. Os resultados dos experimentos de oclusão também são demonstrados (conforme Seção 4.2.3.2 da Metodologia). Os resultados dos experimentos em Situações Reais são demonstrados, analisando cada uma das etapas do Sistema de Reconhecimento de Padrões, conforme especificado na Metodologia (ver Seção 4.2.3.3). Por fim, como os experimentos demonstrados neste Capítulo são empíricos, as ameaças que podem invalidar os resultados são levantadas e discutidas.

5.1 TREINAMENTO DOS CLASSIFICADORES

O treinamento dos Classificadores em Cascata com características Haar e LBP foi conduzido utilizando a OpenCV. Apesar das configurações iniciais terem definido 15 cascatas, os classificadores finais foram treinados até 8 níveis. O número reduzido de cascatas se deve aos altos valores definidos para as taxas de falsos positivos e verdadeiros positivos. As taxas de FP e VP foram mantidas com esses valores para tentar reduzir as taxas de erro de classificação. O ajuste dos parâmetros ao final da etapa de treinamento é salvo pela OpenCV em um arquivo XML. O arquivo XML contém todas informações de cada nível da cascata para o funcionamento do classificador. Assim, para utilizar o classificador treinado basta carregar os dados contidos no arquivo XML.

O processo de treinamento do SVM foi realizado em 95570 iterações, como o critério de parada era de 100 mil iterações, o critério de parada atingido foi a acurácia definida como 10^{-6}. Assim como o Classificador em Cascata, a OpenCV gera um arquivo XML com todos os parâmetros do SVM, possibilitando que os parâmetros sejam carregados posteriormente.

O treinamento das Redes Neurais Convolucionais foi realizado no *Framework* Caffe.

O critério de parada para treinamento das CNNs foi definido como 1000 épocas. Entretanto, ao final do treinamento verificou-se que por volta de 116 épocas as CNNs já convergiram para o mesmo resultado atingido em 1000 épocas. As Figuras 31 e 32 mostram a evolução dos treinamentos nas CNNs.

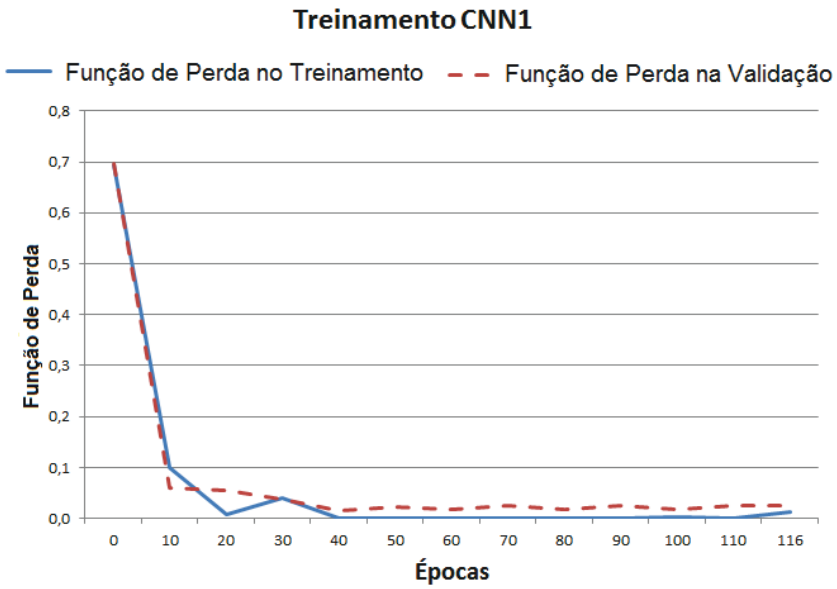

Figura 31: Gráfico de evolução do treinamento da CNN1.

Fonte: Própria autoria

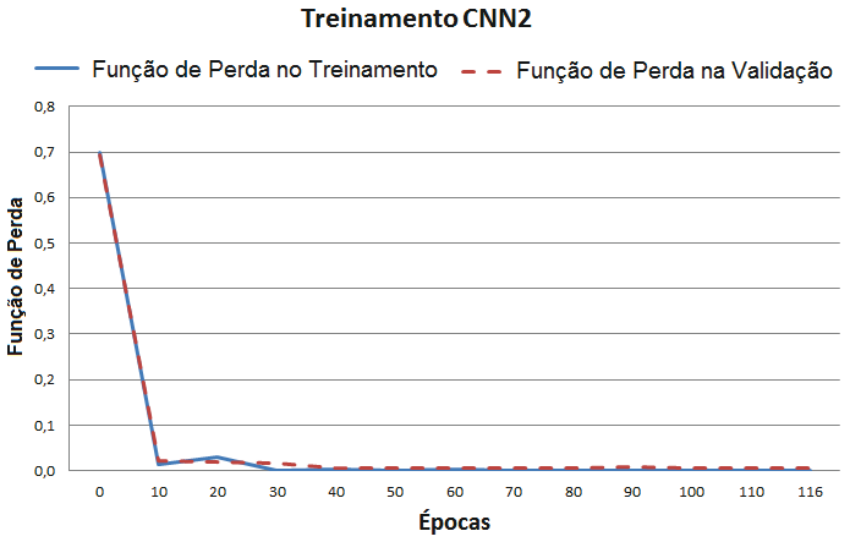

Figura 32: Gráfico de evolução do treinamento da CNN2.

Fonte: Própria autoria

Através dos gráficos obtidos dos treinamentos é possível observar que as CNNs convergiram para um bom resultado, com Função de perda muito próxima de 0. Já em aproxima-

damente 40 épocas as duas redes já chegaram em baixos valores de função de Perda, a partir de quando o treinamento evoluiu lentamente até 116 épocas. É importante destacar que apesar das redes terem sido originalmente sido treinadas até 1000 épocas, não houve *overfitting* devido ao uso de *Dropout* nas camadas totalmente conectadas.

5.2 TESTES DE RECONHECIMENTO DOS CLASSIFICADORES

A avaliação de Reconhecimento dos classificadores sob o conjunto de testes foi realizado com processamento em lote. O arquivo de lote contém a lista de arquivos do conjunto de testes juntamente com seu devido rótulo. Assim é possível comparar o resultado obtido nos processos de classificação com o rótulo atribuído à imagem. Dessa forma, os dados da matriz de confusão podem ser preenchidos permitindo o cálculo das medidas de Acurácia, Sensibilidade, Especificidade e Gráfico ROC. As matrizes de confusão dos classificadores são mostradas nas Tabelas 7, 8, 9, 10 e 11.

Tabela 7: Matriz de confusão da Cascata Haar.

	Positivo	Negativo
Positivo	1705	644
Negativo	543	1584

Tabela 8: Matriz de confusão da Cascata LBP.

	Positivo	Negativo
Positivo	1740	1007
Negativo	508	1221

Tabela 9: Matriz de confusão de HOG + SVM.

	Positivo	Negativo
Positivo	2068	162
Negativo	180	2066

Tabela 10: Matriz de confusão da CNN1.

	Positivo	Negativo
Positivo	2245	29
Negativo	3	2199

Como o conjunto de treinamento tem classes balanceadas, a análise da Acurácia mostra claramente o desempenho dos classificadores. Através das medidas de Acurácia mostradas

Tabela 11: Matriz de confusão da CNN2.

	Positivo	Negativo
Positivo	2248	13
Negativo	0	2215

Tabela 12: Medidas extraídas a partir das Matrizes de Confusão.

	CNN1	CNN2	HOG + SVM	Cascata Haar	Cascata LBP
Acurácia	0,9929	0,9971	0,9236	0,7348	0,6615
Sensibilidade	0,9987	1,0000	0,9199	0,7585	0,7740
Especificidade	0,9870	0,9942	0,9273	0,7110	0,5480
1-Especificidade	0,0130	0,0058	0,0727	0,2890	0,4520

na Tabela 12, os classificadores tiveram a seguinte ordem de desempenho (do melhor para o pior): CNN2, CNN1, HOG + SVM, Cascata Haar e Cascata LBP. Apesar da CNN2 ter obtido o melhor resultado, a CNN1 mostrou um resultado muito próximo, com uma diferença de apenas 0,0042. Essa diferença, sob uma primeira perspectiva, mostra que a arquitetura da CNN1 é o suficiente para cumprir a tarefa de reconhecimento de pessoas em imagens aéreas, mesmo que esta arquitetura extraia um número menor de características. Isto é uma questão importante, principalmente considerando o desempenho computacional do sistema. Arquiteturas de CNNs com muitas camadas tendem a ter um alto custo computacional, enquanto arquiteturas com poucas camadas tem um custo computacional menor, devido a um número reduzido de neurônios. Além disso, o número de mapas de características reduzido também influencia no desempenho computacional. A análise da Especificidade também mostra que as CNNs erram muito pouco, se mostrando como classificadores muito conservadores.

O HOG + SVM, apesar de ter atingido uma acurácia menor (0,9236), também obteve bons resultados, indicando os motivos de ser bastante usado em aplicações de reconhecimento de pessoas. Apesar da grande quantidade de imagens com diversas poses, pequenos exemplos de rotação e translação que aumentam a complexidade da classificação, o SVM foi capaz de gerar um modelo eficaz. A Especificidade atingida no SVM, mostra um classificador conservador, que atinge altas taxas de Verdadeiros Positivos sem abrir mão de baixa taxa de Falsos Positivos. A grande vantagem da combinação HOG + SVM é a simplicidade do método, o treinamento é simples e rápido.

Os Classificadores em Cascata mostraram certa dificuldade para a tarefa de reconhecimento de pessoas em imagens aéreas. A Acurácia de ambos ficou próxima de 0,70, o que indica uma alta taxa de erros. Além disso, a Especificidade teve valores baixos, principalmente o Cas-

cata LBP com apenas 0,5480. Estes resultados mostram que os Classificadores em Cascata são liberais, ou seja, para ter altas taxas de Verdadeiros Positivos o classificador abre mão das baixas taxas de Falsos Positivos. Dessa forma, ao mesmo tempo que o classificador acerta, ele também erra muito. A baixa eficácia destes classificadores deve-se a questão de que as características de Haar e LBP são descritores de textura, poucos atrativos para o reconhecimento de pessoas.

A Figura 33 mostra visualmente a diferença entre os classificadores. As CNNs chegaram a índices muito próximos do Classificador Perfeito. Já o HOG + SVM se mostrou um método conservador com bons resultados. Os Classificadores em Cascata se mostraram mais liberais, atingindo taxas consideráveis de erros. Com estes resultados, se confirma o melhor desempenho de classificação para as CNNs.

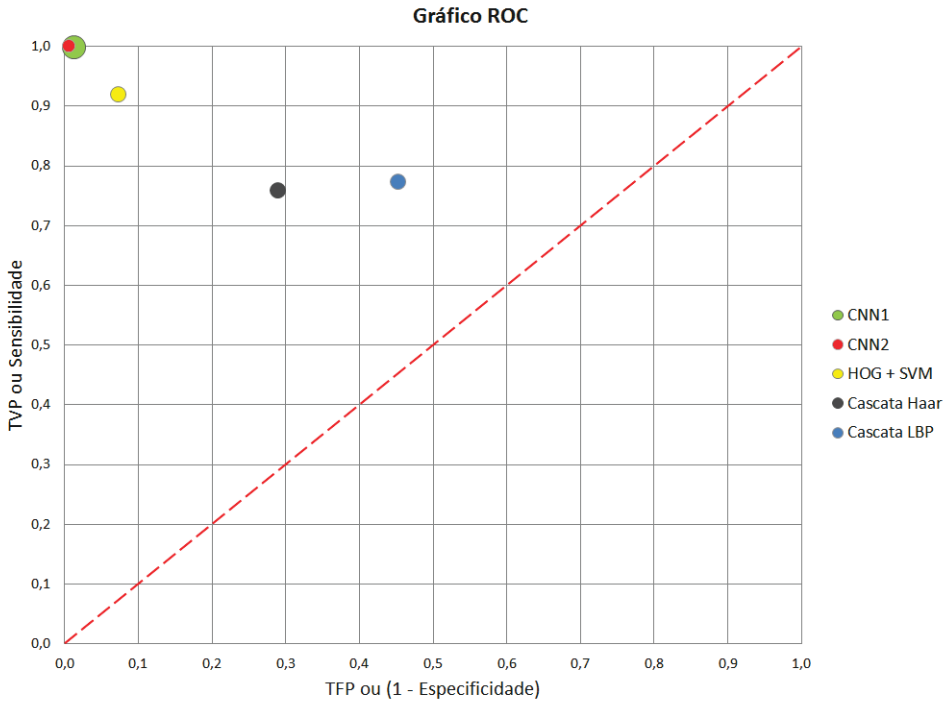

Figura 33: Gráfico ROC dos resultados dos testes de reconhecimento.

Fonte: Própria autoria

5.3 TESTES DE OCLUSÃO

Os testes de oclusão não deixam de ser testes de reconhecimento dos classificadores. Dessa forma, se utilizam do mesmo processo empregado nos testes de Reconhecimento dos classificadores, ou seja, um arquivo de lote contendo uma lista de arquivos a serem submetidos à classificação. A diferença destas imagens é que elas sofreram um processo de oclusão, onde

quadrados pretos foram inseridos em um dos quatro quadrantes da imagem: (i) superior direito, (ii) superior esquerdo, (iii) inferior direito e (iv) inferior esquerdo. Com isso, quatro situações diferentes de oclusão são simuladas afim de estudar o comportamento dos classificadores nestes casos. Este teste ajuda a avaliar a generalização dos classificadores, mostrando que, mesmo com a falta de alguns elementos do objeto de interesse o reconhecimento ainda é possível. Além disso, as situações de oclusão parcial podem ocorrer com frequência em situações reais, principalmente quando o ambiente tem muitos elementos altos, como árvores e edifícios.

Como neste experimento apenas as imagens positivas foram submetidas ao teste, nem todas as medidas utilizadas nos testes de reconhecimento de classificadores são calculadas. Na avaliação da oclusão, temos apenas Verdadeiros Positivos e Falsos Negativos. Dessa forma, o medida de avaliação empregada para este experimento é a Sensibilidade, que na prática (para este caso) é a mesma coisa que a Acurácia. Para uma comparação justa, o desempenho dos classificadores para as imagens sem oclusão também foi computado. A Tabela 15 mostra o desempenho de classificação para as imagens sem oclusão. Enquanto as Tabelas 13 e 14 mostram, respectivamente, o desempenho nas situações de oclusão superiores e inferiores para os lados direito e esquerdo.

Tabela 13: Desempenho dos classificadores em imagens com oclusão parcial superior.

	Direito		Esquerdo	
	Acertos	Sensibilidade	Acertos	Sensibilidade
CNN1	3815	0,47	4298	0,53
CNN2	5668	0,70	6453	0,80
HOG + SVM	6072	0,75	4688	0,58
Cascata Haar	2040	0,25	2240	0,28
Cascata LBP	2173	0,27	2278	0,28

Tabela 14: Desempenho dos classificadores em imagens com oclusão parcial inferior.

	Direito		Esquerdo	
	Acertos	Sensibilidade	Acertos	Sensibilidade
CNN1	2367	0,29	1852	0,23
CNN2	5062	0,63	5890	0,73
HOG + SVM	2045	0,25	3404	0,42
Cascata Haar	2053	0,14	1921	0,13
Cascata LBP	2146	0,14	2178	0,15

A média dos casos de oclusão mostra os melhores (do melhor para o pior) métodos conforme a ordem: CNN2, HOG + SVM, CNN1, Cascata LBP e Cascata Haar.

Tabela 15: Desempenho dos classificadores em imagens sem oclusão parcial e média dos casos de oclusão.

	Original		Médias dos casos de oclusão	
	Acertos	Sensibilidade	Acertos	Sensibilidade
CNN1	8055	0,99	3083	0,38
CNN2	8067	1	5768	0,72
HOG + SVM	7511	0,93	4052	0,50
Cascata Haar	6104	0,76	2064	0,20
Cascata LBP	6731	0,83	2194	0,21

O classificador HOG + SVM manteve as taxas de reconhecimento com melhores resultados, isto mostra a eficiência do método em situações de oclusão. No teste onde o HOG + SVM teve o pior desempenho, para as imagens utilizadas neste trabalho, pode-se notar a tendência do classificador a utilizar mais as características da porção superior da imagem no processo de classificação.

Neste teste a disparidade entre as CNNs aparece. A diferença de Sensibilidade entre as duas redes é praticamente o dobro na média dos casos de oclusão. Isso evidencia como o número de características extraídas pode influenciar a classificação. Além disso, com mais camadas, a CNN é capaz de fazer mais combinações entre características, o que pode aumentar a independência entre as características. Dessa forma, com mais características e mais combinações caso parte das características não esteja presente, existem outras que podem tornar a classificação possível. Os resultados também mostram que para a CNN2 as características presentes na porção inferior do objeto tem leve impacto na classificação, visto que na oclusão superior os experimentos tiveram as maiores taxas. Na CNN1 a dependência da porção inferior do objeto se mostra maior, devido as taxas de classificação maiores para os experimentos de oclusão onde esta região não esta encoberta.

O classificador HOG + SVM manteve as taxas de reconhecimento melhores que a CNN1, com Sensibilidade média de 0,5. A técnica mostrou alta dependência das informações contidas na porção inferior direita, visto que as taxas de classificação foram baixas quando a região foi encoberta pelo quadrado preto. Quando esta região sofreu oclusão, o classificador atingiu a Sensibilidade de apenas 0,25. No geral, a região que menos contribui para o reconhecimento é a porção superior direita. Na situação de oclusão da região superior direito, o classificador atingiu 0,75 de Sensibilidade, taxa melhor que a atingida pela CNN2 na situação de oclusão.

Já para os Classificadores em Cascata a dificuldade evidenciada nos testes de reconhe-

cimento de classificadores se mantém. Tanto com características de Haar quanto LBP tiveram os piores resultados em todos os casos de oclusão. Os classificadores se mostraram dependentes de todas as regiões do objeto, visto que as taxas de classificação ficaram muito próximas em todos os casos.

5.4 ANÁLISE DO SISTEMA DE RECONHECIMENTO DE PADRÕES

5.4.1 ANÁLISE DA SEGMENTAÇÃO E DETECÇÃO

As análises do SRP foram realizadas com o intuito de simular a eficácia de um Sistema de Reconhecimento de Padrões completo. A etapa de aquisição de imagens não foi realizado através do uso de um drone. Contudo, o fato das imagens previamente capturadas serem utilizadas em um processo de lote, simula uma situação próxima da real. Desse modo, este experimento pode ser visto como uma simulação, visto que nem todos os elementos de uma situação real estão efetivamente presentes. Essa simulação foi necessária pois, durante a execução dos experimentos, o projeto de pesquisa o qual este trabalho faz parte, não havia recebido os recursos financeiros necessários para a aquisição do VANT, o que impossibilitou a aquisição das imagens para teste.

No experimento, as etapas de segmentação ou detecção e classificação são avaliadas. Quanto ao seu desempenho computacional, o objetivo é entender a influência de técnicas de segmentação e detecção sob a etapa de classificação, visto que técnicas como Janelas Deslizantes, dificilmente cumprem requisitos de aplicações de tempo real. Assim, técnicas que diminuem o espaço de busca podem ter grande impacto no desempenho computacional do sistema. Este benefício também se estende ao desempenho da classificação, pois quanto menos exemplos de negativos forem submetidos ao classificador (devido aos critérios usados na delimitação do espaço de busca), menor será a probabilidade de uma classificação errada. Entretanto, a premissa inversa também pode influenciar: se a técnica de segmentação e detecção não for eficaz ao procurar possíveis objetos, afetará consideravelmente o classificador.

Os experimentos foram conduzidos com 17 imagens, todas no mesmo ambiente e com a mesma pessoa como exemplo. A grande diferença entre as imagens é a posição da pessoa na imagem, além das poses serem variadas. O primeiro experimento focou nas técnicas de segmentação e detecção, a fim de quantificar a quantidade de objetos encontrados em cada técnica. Além disso, levantando estes números é possível montar as matrizes de confusão nos experimentos envolvendo os classificadores. A Tabela 16 mostra a quantidade de exemplos de cada tipo encontrados por cada técnica, com a média e desvio padrão do número de detecções

negativas em cada imagem.

Tabela 16: Número de objetos encontrados por cada técnica de segmentação e detecção.

	Total Positivos	Total Negativos	Média de negativos
Mapa de Saliências	17	137	8,06 $\pm$ 1,78
Proc. Imagem Térmica	17	43	2,53 $\pm$ 1,33

Os resultados da Tabela 16 mostram uma tendência do Mapa de Saliências em encontrar mais objetos para classificação, algo que acontece com menos frequência no Processamento de Imagens Térmicas (PIT). Dessa forma, é esperado que a segmentação e detecção influencie no desempenho dos classificadores, visto que a probabilidade do classificador errar, em valores absolutos, será maior. O mesmo acontece com o desempenho computacional, será necessário classificar mais objetos com o Mapa de Saliências, o que pode aumentar o total de tempo necessário para o processamento de cada imagem.

5.4.2 ANÁLISE DA SEGMENTAÇÃO E DETECÇÃO COM CLASSIFICADORES

Os experimentos conduzidos com a união das técnicas de segmentação ou detecção e classificação são mostradas nas Matrizes de Confusão das Tabelas 17, 18, 19, 20 e 21. As medidas de Sensibilidade e Especificidade feitas a partir das Matrizes de Confusão são mostradas na Tabela 22.

Tabela 17: Matriz de confusão da CNN1 com Mapa de Saliências e Proc. Imagem Térmica.

	Mapa de Saliências		Proc. Imagem Térmica	
	positivo	negativo	positivo	negativo
positivo	17	0	16	0
negativo	0	137	1	43

Tabela 18: Matriz de confusão da CNN2 com Mapa de Saliências e Proc. Imagem Térmica.

	Mapa de Saliências		Proc. Imagem Térmica	
	positivo	negativo	positivo	negativo
positivo	17	0	15	0
negativo	0	137	2	43

Os resultados mostrados nas Tabelas confirmam as medidas atingidas no Testes de Reconhecimento dos Classificadores (Seção 5.2). Observa-se que todos os classificadores atingiram uma Sensibilidade menor com a técnica de segmentação e detecção com Imagens Térmicas, isto se deve as variações maiores de translação quando os objetos são segmentados. Pelo

Tabela 19: Matriz de confusão do HOG + SVM com Mapa de Saliências e Proc. Imagem Térmica.

	Mapa de Saliências		Proc. Imagem Térmica	
	positivo	negativo	positivo	negativo
positivo	14	15	11	2
negativo	3	122	6	41

Tabela 20: Matriz de confusão da Cascata Haar com Mapa de Saliências e Proc. Imagem Térmica.

	Mapa de Saliências		Proc. Imagem Térmica	
	positivo	negativo	positivo	negativo
positivo	12	35	8	12
negativo	5	102	9	31

Tabela 21: Matriz de confusão da Cascata LBP com Mapa de Saliências e Proc. Imagem Térmica.

	Mapa de Saliências		Proc. Imagem Térmica	
	positivo	negativo	positivo	negativo
positivo	17	70	15	10
negativo	0	67	2	33

Tabela 22: Medidas retiradas a partir das Matrizes de Confusão dos testes dos SRPs.

		CNN1	CNN2	HOG + SVM	Cas. Haar	Cas. LBP
Mapa de	Sensibilidade	1,00	1,00	0,82	0,71	1,00
Saliências	Especificidade	1,00	1,00	0,89	0,74	0,49
Proc. Imagem	Sensibilidade	0,94	0,88	0,65	0,47	0,88
Térmica	Especificidade	1,00	1,00	0,95	0,72	0,72

fato das Imagens Térmicas terem baixa resolução, ao fazer a relação com a imagem RGB, a segmentação não é totalmente precisa, assim os objetos submetidos a classificação sofrem pequenas translações. Já quanto à Especificidade, a tendência da técnica de segmentação com Mapa de Saliências aumentar o número de Falsos Positivos se confirmou (ver Tabela 16). O Mapa de Saliências fornecem mais objetos a serem classificados, assim a tendência é que os classificadores errem mais. Entretanto, em casos onde há um classificador mais robusto, como o caso das CNNs, são poucos (ou nenhum) erros. Isto se deve à invariância a translação das CNNs, justamente o efeito negativo que os objetos sofrem na segmentação com Mapa de Saliências.

Quanto a classificação, as CNNs confirmaram a robustez para a tarefa de reconhecimento de pessoas em imagens aéreas, mesmo com as variações de translação provocadas na

segmentação com PIT, as técnicas foram pouco afetadas, mantendo baixas taxas de erro. Já a Cascata LBP surpreendeu, atingindo altas taxas de reconhecimento, diferente dos Testes de Reconhecimento dos Classificadores (ver Seção 5.2). Entretanto isto mostra a falta de robustez do classificador de Cascata LBP, pois diferentes situações trazem resultados variados. Já o HOG + SVM teve resultados inferiores aos obtidos anteriormente, mostrando as dificuldades da técnica em variações de translação provocadas na segmentação com PIT. O problema do HOG + SVM com translação dos objetos é evidenciado na segmentação e detecção com Processamento de Imagens Térmicas, onde a Sensibilidade ficou em apenas 0,65. Por fim, a Cascata Haar continuou com dificuldades no reconhecimento, assim como as outras técnicas, teve grandes variações nas taxas de reconhecimentos executadas até aqui. Isto evidência o problema das imagens aéreas, onde os classificadores não foram capazes de identificar texturas relevantes de pessoas em imagens aéreas.

5.4.3 ANÁLISE DE DESEMPENHO COMPUTACIONAL

A análise de desempenho computacional foi executada a fim de verificar a eficiência computacional dos métodos aplicados neste trabalho. A aplicação de resgate e salvamento de pessoas utilizando VANTs requer que informações sejam fornecidas em tempo real. Isto facilita a tomada de decisão da equipe envolvida no resgate, agilizando a missão, aumentando assim a possibilidade de encontrar pessoas que necessitam de auxílio. As informações a serem utilizadas são coletadas pelo VANT, com o objetivo de fornecer informações em tempo real essas informações podem ser processadas de duas maneiras: (i) por um sistema embarcado no próprio VANT, que transmite apenas as informações geográficas à equipe de resgate ou (ii) por uma Base de Controle Móvel (BCM), um dispositivo (*notebook*, por exemplo) que fica a disposição da equipe de resgate recebendo as informações do VANT em tempo real e processando-as.

O uso do sistema embarcado no VANT tem a grande vantagem de fornecer apenas os dados geográficos de possíveis objetos a serem encontrados, isto é uma grande vantagem, visto que pode haver dificuldade da transmissão de grande quantidade de dados (imagens por exemplo) entre VANT e BCM. Entretanto, se o Sistema de Reconhecimento de Padrões não for eficiente, o sistema embarcado pode transmitir informações incorretas, o que pode acabar atrapalhando a missão de salvamento, visto que equipes seriam deslocadas com informações incorretas. Além disso, poucas plataformas computacionais têm poder de processamento suficiente para processar em tempo real imagens com algoritmos complexos de reconhecimento de padrões.

O BCM tem a vantagem de ser uma plataforma com mais recursos, visto que pode ser um *notebook*, que fornece mais capacidade de processamento que o sistema embarcado. Além disso, transmitindo as imagens do VANT em tempo real o processamento pode ser feito no BCM, mostrando para a equipe de resgate os resultados encontrados pelo SRP. Dessa forma, os resultados do SRP podem ser avaliados por um ser humano, que por sua vez, confirma as decisões do SRP. Isto aumenta a confiabilidade da tomada de decisão, visto que não é apenas a máquina que toma as decisões. Por outro lado, o BCM tem desvantagens: os dados nem sempre são transmitidos de forma confiável e as imagens podem chegar à BCM com ruído. Isto pode gerar interpretações erradas tanto do SRP quanto do humano que estará operando o sistema.

Os testes de desempenho computacional dos métodos foram conduzidos para avaliar o seu desempenho nas duas situações. As seguintes premissas foram estabelecidas: (i) a comunicação entre o VANT e o BCM é em tempo real, e (ii) a transmissão dos dados é confiável (sem ruídos). Outra questão envolvida é a repetição dos testes, que foram realizados 20 vezes seguidas. Como os Sistemas Operacionais executam rotinas continuamente, pode haver disputa por recursos, isto pode afetar o desempenho do SRP. Assim, o objetivo da repetição dos experimentos é evitar que rotinas dos Sistemas Operacionais interfiram no testes de desempenho computacional, fornecendo resultados mais confiáveis do desempenho do SRP.

Os testes foram realizados foram em duas plataformas: Raspberry Pi 2 (4.1.1) e um *notebook*. O Raspberry Pi 2 representa o sistema embarcado no VANT. Ele foi escolhido por ser uma placa muito popular no mercado, apresentando baixo custo. Já o notebook representa a BCM, com a configuração: processador Intel Core i5-3210M 2.5GHz e 4GB de memória RAM. O objetivo destes testes é avaliar sistemas de baixo custo aplicados ao problema aqui descrito. As Tabelas 23, 24,25 e 26 mostram os resultados obtidos em cada uma das plataformas.

Os testes do sistema embarcado mostram uma grande dificuldade de atingir tempos de processamento em tempo real, não sendo capaz de fornecer informações a 1 Hz (1 fps), que é um dos objetivos deste trabalho (ver Tabelas 23 e 24). Mesmo com técnicas visando a redução do espaço de busca para os algoritmos de classificação, o hardware do Raspberry Pi 2 não possui poder de processamento suficiente para as etapas de classificação, o que costuma tomar mais recursos do sistema. É importante salientar que nesta etapa o processamento foi realizado utilizando apenas a CPU do hardware, sem a utilização da GPU. Mesmo utilizando a detecção através do Processamento de Imagens Térmicas, a combinação das técnicas não atinge resultados em tempo real. A exceção dos experimentos é a combinação das características HOG com SVM com Processamento de Imagens Térmicas, elas obtiveram surpreendentes resultados, atingindo a taxa de 14 fps no pior caso. Entretanto, os resultados de desempenho de classificação

Tabela 23: Desempenho Computacional dos métodos de detecção e classificação executados no Raspberry Pi. MS se refere ao Mapa de Saliências, PIT ao Processamento de Imagens Térmicas, HOG ao classificador HOG+SVM, Haar à Cascata Haar e LBP à Cascata LBP.

	Detecção		Classificação	
	Média (s)	Máx (s)	Média (s)	Máx (s)
MS + CNN1	3,67±0,57	4,21	0,73±0,1	0,79
PIT + CNN1	$8x10^{-3} \pm 3x10^{-5}$	$7x10^{-4}$	0,77±0,01	0,81
MS + CNN2	3,69±0,5	4,25	2,1±0,2	2,1
PIT + CNN2	$7x10^{-3} \pm 4x10^{-5}$	$7x10^{-4}$	2,02±0,4	2,38
MS + HOG	3,65±0,61	4,14	$1x10^{-2} \pm 2x10^{-4}$	0,01
PIT + HOG	$8x10^{-3} \pm 3x10^{-5}$	$7x10^{-4}$	$1x10^{-2} \pm 8x10^{-3}$	0,01
MS + Haar	3,66±0,63	4,16	0,38±0,07	1,01
PIT + Haar	$8x10^{-3} \pm 3x10^{-5}$	$7x10^{-4}$	0,39±0,08	0,59
MS + LBP	3,73±0,4	4,07	0,4±0,09	1,10
PIT + LBP	$8x10^{-3} \pm 5x10^{-5}$	$7x10^{-4}$	0,42±0,07	0,53

Tabela 24: Desempenho Computacional Total dos métodos executados no Raspberry Pi. MS se refere ao Mapa de Saliências, PIT ao Processamento de Imagens Térmicas, HOG ao classificador HOG+SVM, Haar à Cascata Haar e LBP à Cascata LBP.

	Total		
	Média (s)	Máx (s)	FPS
MS + CNN1	8,81±1,32	11,12	0,09
PIT + CNN1	2,75±0,82	5,34	0,18
MS + CNN2	9,8±2,3	13,23	0,08
PIT + CNN2	6,08±2,28	11,86	0,08
MS + HOG	3,85±0,02	3,87	0,25
PIT + HOG	0,04±0,02	0,07	14
MS + Haar	6,45±0,76	9,37	0,11
PIT + Haar	3,7±0,8	7,97	0,13
MS + LBP	6,5±0,9	9,69	0,11
PIT + LBP	3,81±0,77	7,08	0,14

discutidos nas Seções 4.2.2 e 4.2.3 colocam a eficácia do HOG+SVM em dúvida. Assim, o uso do RaspBerry Pi 2 mostrou poucas vantagens, atingindo bons resultados apenas em um classificador que não tem a capacidade de classificação desejada em comparação as CNNs. Apesar destes resultados deve-se salientar que o código utilizado neste trabalho não foi otimizado para os recursos de *hardware* disponíveis para o Raspberry Pi. Utilizar funcionalidades específicas para o Hardware pode reduzir o tempo de processamento do SRP.

Tabela 25: Desempenho Computacional dos métodos de detecção e classificação métodos executados na BCM. MS se refere ao Mapa de Saliências, PIT ao Processamento de Imagens Térmicas, HOG ao classificador HOG+SVM, Haar à Cascata Haar e LBP à Cascata LBP.

	Detecção		Classificação	
	Média (s)	Máx (s)	Média (s)	Máx (s)
MS + CNN1	0,23±0,01	0,27	0x08±$2x10^{-3}$	0.,08
PIT + CNN1	$6x10^{-5} \pm 3x10^{-5}$	$6x10^{-5}$	0,08±1^{-3}	0,08
MS + CNN2	0,23±0,03	0,22	0,18±$4x10^{-3}$	0,18
PIT + CNN2	$4x10^{-4} \pm 3x10^{-5}$	$6x10^{-5}$	0,17±0,01	0,2
MS + HOG	0,21±0,04	0,23	$7x10^{-4} \pm 10^{-4}$	$9x10^{-4}$
PIT + HOG	$7x10^{-5} \pm 3x10^{-5}$	$6x10^{-5}$	$7x10^{-4} \pm 10^{-4}$	$9x10^{-4}$
MS + Haar	0,22±0,01	0,22	0,03±0,01	0,08
PIT + Haar	$8x10^{-5} \pm 3x10^{-5}$	$6x10^{-5}$	0,04±0,01	0,06
MS + LBP	0,22±$3x10^{-3}$	0,26	0,04±0,02	0,11
PIT + LBP	$8x10^{-5} \pm 5x10^{-5}$	$6x10^{-5}$	0,04±$4x10^{-3}$	0,8

Tabela 26: Desempenho Computacional Total dos métodos executados na BCM. MS se refere ao Mapa de Saliências, PIT ao Processamento de Imagens Térmicas, HOG ao classificador HOG+SVM, Haar à Cascata Haar e LBP à Cascata LBP.

	Total		
	Média (s)	Máx (s)	FPS
MS + CNN1	0,77±0,13	0,74	1,35
PIT + CNN1	0,22±0,08	0,41	2,43
MS + CNN2	1,63±0,32	1,4	0,71
PIT + CNN2	0,61±0,19	0,93	1,08
MS + HOG	0,28±0,02	0,3	3,33
PIT + HOG	0,02±$1x10^{-3}$	0,02	100
MS + Haar	0,42±0,06	0,61	1,64
PIT + Haar	0,13±0,05	0,28	3,57
MS + LBP	0,51±0,08	0,76	1,31
PIT + LBP	0,14±0,04	0,26	3,85

Os resultados na plataforma BCM mostram resultados que atingem o objetivo de processamento em 1 fps. Para a maioria dos métodos, a melhoria no BCM em relação à plataforma embarcada ficou entre 10 e 15 vezes. Neste caso, classificadores que mostraram mais capacidade de reconhecimento para a solução do problema atingiram taxas superiores a 1 fps. Isto se deve a capacidade de processamento do hardware da BCM que é superior à plataforma embarcada. Nos testes do BCM, a vantagem da técnica de Processamento de Imagens Térmicas sob

o Mapa de Saliências é evidente. Observando os valores máximos (piores casos), o Mapa de Saliências consome cerca de 250 ms de processamento. Na maioria dos métodos, 250 ms representa no mínimo 25% do tempo de processamento, enquanto na combinação com HOG + SVM chega a mais de 75%. Por outro lado, a técnica utilizando Processamento de Imagens Térmicas tem praticamente nenhum impacto no tempo de processamento, representando menos de 1% do tempo total. Em situações de grande calor (acima de 30 graus Celsius) se uma pessoa estiver sobre um porção da imagem onde há raios solares, o Processamento de Imagens Térmicas tem dificuldade em identificar o objeto. Por outro lado, o Mapa de Saliências detecta uma quantidade maior de exemplos falsos, o que influência o tempo total de processamento, visto que mais imagens serão processadas pelos classificadores. Apesar do tempo computacional elevado, o Mapa de Saliências tem participação importante no bom desempenho dos SRPs propostos.

Já do ponto de vista dos classificadores, os experimentos com as CNNs mostraram que é possível utilizar esta técnica e atingir valores de pelo menos 1 fps sem o uso de GPUs. Os experimentos mostram a influência do tamanho da rede e a quantidade de mapa de características no tempo de processamento. A CNN2 teve tempos de processamento maiores em comparação à CNN1. A CNN2 ainda tem taxas muito próximas de 1 fps, assim em imagens mais complexas a técnica provavelmente não atingiria o objetivo de 1 fps, visto que teve tempos de processamento de 200 ms para cada objeto. Já a técnica de HOG + SVM mostrou novamente um desempenho melhor neste sentido. No pior caso utilizando Processamento de Imagens Térmicas, o HOG + SVM atingiu uma taxa de 100 fps. Esta taxa é mais do que o suficiente para cumprir o objetivo de pelo menos 1 fps de processamento, além de ser uma taxa muito maior que dos outros classificadores. Os Classificadores em Cascata também mostraram bons resultados com relação ao tempo de processamento, atingindo taxas de mais de 3 fps. Entretanto, o custo benefício do uso da Cascata de classificadores é pequeno, pois a taxa de acerto é menor que a CNN1 que por sua vez, teve um desempenho de tempo não muito distante das Cascatas. É importante atingir a taxa de 1 fps, pois caso a taxa fique abaixo, o SRP pode perder informações e consequentemente não detectar algumas pessoas que possam estar no ambiente.

5.5 AMEAÇAS À VALIDADE DOS RESULTADOS

As Ameaças à Validade dos Resultados especificam até que ponto os resultados empíricos obtidos neste trabalho são aplicáveis a outras situações. Relacionando estas ameaças é possível reconhecer as fraquezas às quais os métodos aqui propostos estão submetidos.

Nos experimentos de reconhecimento dos classificadores, foram apresentadas bases de dados com grande quantidade de exemplos, com vários exemplos de pessoas em diversos

ambientes. Entretanto, a etapa de aumento de dados pode ser questionada. Como o aumento de dados foi desproporcional em relação às bases, é possível que o treinamento tenha beneficiado alguns conjuntos de imagens na etapa de teste. Isto pode impactar no desempenho dos classificadores em outras situações, apesar da grande variedade de exemplos de treinamento. Porém, ao mesmo tempo que a distribuição das bases no aumento de dados é desproporcional, foi necessário que tal abordagem fosse adotada, pois bases com quantidades reduzidas de dados teriam muitos exemplos submetidos a variação de rotação, translação e escala. Essa grande quantidade de imagens geraria imagens muito próximas umas das outras, o que poderia levar os classificadores ao *overfitting*.

Os Testes de oclusão podem ser questionados quanto ao método utilizado para a simulação da oclusão. Ele é pouco efetivo ao simular situações reais, onde as pessoas sofreram oclusão de outros objetos (árvores, carros, edifícios, etc) que podem influenciar na etapa de classificação.

Neste trabalho, as ameaças estão concentradas, principalmente, nos Sistemas de Reconhecimento de Padrões. Apesar das técnicas demonstrarem bom desempenho nos experimentos realizados, elas podem ter um comportamento diferente em outras situações. A utilização de Mapa de Saliências pode sofrer variações em ambientes muito complexos, onde há muita variação de cores, formas ou estão contidos pequenos elementos. Nestes ambientes o Mapa de Saliências pode detectar uma grande quantidade de elementos, o que pode impactar significativamente o desempenho do SRP. Já o Processamento de Imagens Térmicas de baixa resolução sofre forte influência do ambiente. Esta técnica de segmentação e detecção fica limitada a ambientes onde a temperatura é menor que a temperatura corporal humana, pois sem essa diferença não é possível detectar pessoas.

5.6 DISCUSSÃO

Os experimentos mostraram diversos resultados, principalmente quanto à questão dos classificadores. As CNNs obtiveram os melhores desempenhos quanto à qualidade de classificação, justificando o porquê de serem uma das técnicas de aprendizado de máquina que mais recebem atenção nos últimos anos. Os resultados destas redes neurais chegaram muito próximo dos 100% de acerto. Já a combinação do HOG + SVM mostra a eficácia da técnica no reconhecimento de pessoas, mostrando porque a técnica é muito usada no reconhecimento de pessoas. As Cascatas de Classificadores não se mostraram bons candidatos para o problema de reconhecimento de pessoas em imagens aéreas, apresentando muitas dificuldades no reconhecimento. Onde este desempenho das Cascatas da-se pelo fato das características de Haar e LBP serem

descritores de textura, sendo que em imagens aéreas há poucas texturas que descrevem pessoas.

As técnicas de segmentação e detecção mostraram vantagens e desvantagens importantes a serem destacadas. O Mapa de Saliências exigiu tempos de processamento maiores, entretanto apresentou mais precisão ao segmentar os objetos. A segmentação dos objetos se mostrou um ponto importante, principalmente quando se analisou o uso das câmeras térmicas de baixa resolução. Devido à baixa resolução, o uso de imagens térmicas apresenta dificuldades de segmentar os objetos da imagem RGB, gerando variação na translação, o que gerou dificuldades de reconhecimento aos classificadores. A variação de translação também pode envolver a questão de haver um deslocamento entre as câmeras, devido à utilização de um suporte improvisado. Apesar da questão de segmentação, a detecção através do processamento de imagens térmicas se mostrou computacionalmente eficiente possibilitando a detecção de objetos da imagem em um tempo menor.

Os resultados obtidos neste trabalho mostram a eficácia das Redes Neurais Convolucionais. Até onde sabemos, este é o primeiro trabalho da literatura que utiliza as CNNs para o reconhecimento de pessoas em imagens aéreas. Assim, o trabalho contribui mostrando a grande capacidade de reconhecimento das CNNs. A avaliação do desempenho computacional das técnicas utilizadas também é uma importante contribuição, pois trabalhos envolvendo sistemas de reconhecimento de padrões raramente fazem este tipo de avaliação. Foram propostos métodos com resposta em tempo real, utilizando os conceitos de Sistemas de Reconhecimento de Padrões. As técnicas de segmentação e detecção se mostraram determinantes para que o SRP fornecesse resposta em tempo real. Destaca-se a utilização das câmeras térmicas de baixa resolução, que unem um hardware especializado e de baixo custo com baixo custo computacional.

Também se destaca o uso da base de dados utilizada no treinamento. A união de várias bases resultou em uma grande base de imagens, algo determinante, principalmente, para o treinamento eficiente das Redes Neurais Convolucionais.

6 CONCLUSÕES E TRABALHOS FUTUROS

O principal objetivo deste trabalho foi estudar os métodos para a aplicação de reconhecimento de pessoas em imagens aéreas, utilizando técnicas de visão computacional e reconhecimento de padrões. Com o advento dos VANTs esta aplicação tem se tornado cada vez mais comum, sendo utilizada em diversas situações práticas. Isto abre a possibilidade do reconhecimento ser feito de forma autônoma. Esta busca autônoma pode aumentar a eficiência das aplicações com VANT, visto que diminui a intervenção humana.

O trabalho teve como foco principal estudar as técnicas de reconhecimento de padrões que pudessem cumprir a tarefa de reconhecer pessoas em imagens aéreas de forma autônoma e com precisão. Foram estudados algoritmos com diversas abordagens, tendo cada um vantagens e desvantagens na sua utilização. Os algoritmos foram escolhidos por estarem bastante discutidos na literatura, em diversas aplicações. Destaca-se o uso das Redes Neurais Convolucionais que, segundo nossas pesquisas, ainda não foi alvo de estudo para o reconhecimento de pessoas em imagens aéreas.

Métodos de detecção e segmentação também foram estudados, a fim de diminuir o espaço de busca, reduzindo assim o tempo computacional da aplicação, tornando possível que informações em tempo real sejam fornecidas. Assim como as técnicas de reconhecimento, as técnicas de detecção e segmentação também apresentaram vantagens e desvantagens conforme o ambiente e situação quais estão inseridas.

Para a boa condução dos experimentos, foram tomadas diversas medidas na realização do trabalho. As medidas foram concentradas principalmente na construção de uma base de imagens consistente, com o objetivo de fornecer aos classificadores um número maior de exemplos. Assim, na etapa de treinamento os classificadores puderam aprender as mais diversas situações envolvidas no problema, conduzindo seu aprendizado a algo mais próximo de uma situação real.

Os experimentos executados foram: Testes de Reconhecimento dos classificadores, Testes de oclusão e Avaliação do Sistema de Reconhecimento de Padrões. Os Testes de Re-

conhecimento foram executados a fim de avaliar o desempenho de classificação dos métodos utilizados nesta dissertação: Redes Neurais Convolucionais, Cascata Haar, Cascata LBP e HOG + SVM. Já os testes de oclusão buscaram simular situações de oclusão parcial, uma situação comum em imagens com perspectiva aérea. Por fim, a Análise do Sistema de Reconhecimento de Padrões simulou parcialmente um SRP: usando duas plataformas que podem ser aplicadas em uma situação real. Foram utilizadas técnicas de detecção e segmentação juntamente com classificadores a fim de indicar pessoas nas imagens. Na Avaliação do SRP também foi levado em consideração o tempo de processamento de cada técnica, com a finalidade de avaliar o desempenho do sistema e seus componentes em um sistema embarcado (Raspberry Pi 2) e em uma Base de Controle Móvel (Intel Core i5).

Dentre os resultados do reconhecimento de classificadores, destacaram-se as Redes Neurais Convolucionais. As duas arquiteturas utilizadas atingiram resultados com uma acurácia próxima dos 100%. O HOG+SVM teve resultados um pouco menos expressivos, atingindo 92%. Já as duas Cascatas (Haar e LBP) obtiveram baixos resultados de acurácia sendo, respectivamente, 73% e 66%. Destaca-se também a alta taxa de Falsos Positivos obtidos pelas Cascatas.

Nos Testes de Oclusão a CNN2 teve os resultados mais expressivos, alcançando a Sensibilidade média de 0,72. A técnica foi seguida pela HOG + SVM, que obteve Sensibilidade média de 0,5. A CNN1 teve resultados inferiores à CNN2, atingindo apenas 0,38 de Sensibilidade média, praticamente metade da CNN2. As Cascatas Haar e LBP, novamente obtiveram resultados pouco expressivos, atingindo taxas de Sensibilidade média de cerca de 0,20.

Na Análise do SRP, as técnicas seguiram a tendência dos resultados obtidos nos testes de reconhecimento dos classificadores. As CNNs atingiram os melhores resultados, seguidas do HOG+SVM, e por último as Cascatas Haar e LBP. Os experimentos utilizando a câmera térmica, mostraram um problema em relação a segmentação para o reconhecimento. Pela baixa resolução das imagens térmicas, as imagens segmentadas e submetidas a classificação sofrem maior variação na translação. Além de haver o problema relacionado com a utilização do suporte improvisado para as câmeras, o que gera deslocamentos constantes e dificultam a calibração. Assim, alguns classificadores com sensibilidade maior à translação (HOG+SVM e Cascata Haar) tiveram resultados de Sensibilidade piores quando aplicados em conjunto com o Processamento de Imagens Térmicas.

Durante os experimentos com os SRPs, o Mapa de Saliências mostrou um tempo computacional mais elevado, além de submeter mais objetos ao processo de detecção, o que gera mais tempo de processamento. Já o Processamento de Imagens Térmicas teve resultados con-

sideravelmente menores, fornecendo menos objetos para a etapa de classificação. Entretanto, destaca-se a desvantagem de usar o Processamento de Imagens Térmicas, o qual necessita de hardware especializado. Além do hardware, as pessoas presentes na imagem devem ser os objetos mais quentes do ambiente, caso contrário o Processamento de Imagens Térmicas tem dificuldade de detectá-las. Já quanto a etapa de classificação, o HOG+SVM se mostrou o classificador com melhor tempo computacional, seguido pelas Cascatas e depois as CNNs.

Em trabalhos futuros, propõe-se o foco nas Redes Neurais Convolucionais. Esta técnica mostrou grande potencial. Em hardware mais especializado o problema envolvendo tempo computacional pode ser mitigado. Os trabalhos contidos na literatura, com frequência, utilizam placas gráficas (GPUs) afim de melhorar o desempenho computacional da técnica. Dessa forma, hardwares embarcados recentemente lançados no mercado, como a NVIDIA Jeston, podem atender a necessidade de poder computacional da CNN. Há ainda possibilidades de otimização envolvendo o Raspberry Pi 2, através de recursos especiais dos núcleos ARM, tais como instruções SIMD e coprocessador NEON ou até mesmo o uso da GPU contida na plataforma. Além disso, outra alternativa é o uso de FPGA. As FPGAs tem grande poder de processamento, principalmente para problemas com grande quantidade de operações aritméticas. Uma solução simples e eficiente é implantação de CNNs treinadas dentro de FPGAs.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2014.

ANDRILUKA, M.; ROTH, S.; SCHIELE, B. Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE. **IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.** [S.l.], 2009. p. 1014–1021.

ANDRILUKA, M. et al. Vision based victim detection from unmanned aerial vehicles. In: IEEE. **Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on**. [S.l.], 2010. p. 1740–1747.

ANGELOVA, A.; KRIZHEVSKY, A.; VANHOUCKE, V. Pedestrian detection with a Large-Field-Of-View deep network. **2015 IEEE International Conference on Robotics and Automation (ICRA)**, p. 704–711, 2015.

AREL, I.; ROSE, D. C.; KARNOWSKI, T. P. Deep machine learning-a new frontier in artificial intelligence research. **IEEE Computational Intelligence Magazine**, v. 5, n. 4, p. 13–18, 2010.

BAY, H.; TUYTELAARS, T.; Van Gool, L. SURF: Speeded up robust features. In: **European Conference on Computer Vision**. [S.l.]: Springer, 2006. p. 404–417.

BELONGIE, S.; MALIK, J.; PUZICHA, J. Shape matching and object recognition using shape contexts. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 24, n. 4, p. 509–522, 2002.

BLONDEL, P. et al. Fast and viewpoint robust human detection in uncluttered environments. In: **IEEE Visual Communications and Image Processing**. [S.l.: s.n.], 2014. p. 522–525.

BLONDEL, P. et al. Human Detection in Uncluttered Environments: from Ground to UAV View. In: **13th International Conference on Control Automation Robotics & Vision (ICARCV)**. [S.l.: s.n.], 2014. p. 76–81.

BOURDEV, L.; MALIK, J. Poselets: Body part detectors trained using 3d human pose annotations. In: **2009 IEEE 12th International Conference on Computer Vision**. [S.l.: s.n.], 2009. p. 1365–1372.

BRADSKI, G. et al. The OpenCV Library. **Doctor Dobbs Journal**, v. 25, n. 11, p. 120–126, 2000.

BRECKON, T. P. et al. Autonomous real-time vehicle detection from a medium-level UAV. In: **Proc. 24th International Conference on Unmanned Air Vehicle Systems**. [S.l.: s.n.], 2009. p. 21–29.

CAO, X. et al. Vehicle detection and motion analysis in low-altitude airborne video under urban environment. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 21, n. 10, p. 1522–1533, 2011.

CHEN, J. et al. An object recognition strategy base upon foreground detection. In: **Artificial Intelligence and Computational Intelligence**. [S.l.]: Springer, 2011. p. 39–46.

CHEN, L.; JIANG, Z.; FENG, H. Parts-probability-based vehicle detection. **Science China Information Sciences**, Springer, v. 57, n. 11, p. 1–11, 2014.

CHEN, X.; MENG, Q. Vehicle Detection from UAVs by Using SIFT with Implicit Shape Model. In: IEEE. **2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. [S.l.], 2013. p. 3139–3144.

CIRESAN, D.; MEIER, U.; SCHMIDHUBER, J. Multi-column Deep Neural Networks for Image Classification. **International Conference of Pattern Recognition**, n. February, p. 3642–3649, 2012.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

CROW, F. C. Summed-area tables for texture mapping. **ACM SIGGRAPH computer graphics**, ACM, v. 18, n. 3, p. 207–212, 1984.

CRUZ, J. E. C. **Orbitais com o Uso de Abordagens do tipo Descritor-Classificador**. Dissertação (Mestrado) — INPE, 2014.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.** [S.l.: s.n.], 2005. v. 1, p. 886–893.

DAVIS, J. W.; KECK, M. A. A two-stage template approach to person detection in thermal imagery. In: **Seventh IEEE Workshops on Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1.** [S.l.: s.n.], 2005. v. 5, p. 364–369.

DOHERTY, P.; RUDOL, P. A UAV search and rescue scenario with human body detection and geolocalization. In: **AI 2007: Advances in Artificial Intelligence**. [S.l.: s.n.], 2007. p. 1–13.

DOLLÁR, P. et al. Integral Channel Features. In: **British Machine Vision Conference**. [S.l.: s.n.], 2009. v. 2, n. 3, p. 5.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012.

EGAN, J. P. **Signal detection theory and ROC analysis**. New York, NY: Academic Press, 1975. (Series in Cognition and Perception).

FAWCETT, T. An introduction to ROC analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FELZENSZWALB, P. F. et al. Object detection with discriminatively trained part-based models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 32, n. 9, p. 1627–1645, 2010.

FISHER, R. B.; KORYLLOS, K. Interactive textbooks: embedding image processing operator demonstrations in text. **International Journal of Pattern Recognition and Artificial Intelligence**, World Scientific, v. 12, n. 08, p. 1095–1123, 1998.

FLIR. **FLIR LEPTON ® Long Wave Infrared (LWIR) Datasheet**. 2014. 1–50 p.

FLYNN, H.; CAMERON, S. Multi-modal People Detection from Aerial Video. In: SPRINGER. **Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013**. [S.l.], 2013. p. 815–824.

FORSYTH, D. A.; PONCE, J. **Computer vision: a modern approach**. [S.l.]: Prentice Hall Professional Technical Reference, 2002.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: **Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)**. [S.l.: s.n.], 1996. p. 148–156.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997.

GASZCZAK, A.; BRECKON, T. P.; HAN, J. Real-time people and vehicle detection from UAV imagery. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **IS&T/SPIE Electronic Imaging**. [S.l.], 2011. p. 1–13.

GERÔNIMO, D. et al. Survey of pedestrian detection for advanced driver assistance systems. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 7, p. 1239–1258, 2010.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 2–9.

GLEASON, J. et al. Vehicle detection from aerial imagery. In: **2011 IEEE International Conference on Robotics and Automation (ICRA)**. [S.l.: s.n.], 2011. p. 2065–2070.

GRABNER, H. et al. On-line boosting-based car detection from aerial images. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 63, n. 3, p. 382–396, 2008.

GRENZDÖRFFER, G.; ENGEL, A.; TEICHERT, B. The photogrammetric potential of low-cost UAVs in forestry and agriculture. **International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences**, v. 1, p. 1207–1213, 2008.

GUO, T.; KUJIRAI, T.; WATANABE, T. Mapping crop status from an unmanned aerial vehicle for precision agriculture applications. **ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 1, p. 485–490, 2012.

GURURATSAKUL, S. et al. Shark detection using optical image data from a mobile aerial platform. In: **2010 25th International Conference of Image and Vision Computing New Zealand (IVCNZ)**. [S.l.: s.n.], 2010. p. 1–8.

HAAR, A. On the theory of orthogonal function systems. **Math. Ann**, v. 69, n. 3, p. 331–371, 1910.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

HINTON, G. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research (JMLR)**, v. 15, p. 1929–1958, 2014. ISSN 15337928.

HU, M.-K. Visual pattern recognition by moment invariants. **IRE Transactions on Information Theory**, IEEE, v. 8, n. 2, p. 179–187, 1962.

HUNG, C.; BRYSON, M.; SUKKARIEH, S. Multi-class predictive template for tree crown detection. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 68, p. 170–183, 2012.

INTEL. **Intel® Core™ i5-3210M Processor**. 2012. 1 p. Disponível em: <http://ark.intel.com/products/67355/Intel-Core-i5-3210M-Processor-3M-Cache-up-to-3_10-GHz-rPGA>.

ITTI, L. **Models of bottom-up and top-down visual attention**. Tese (Doutorado) — California Institute of Technology, 2000.

ITTI, L.; KOCH, C.; NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, IEEE, n. 11, p. 1254–1259, 1998.

JIA, Y. et al. Caffe: Convolutional architecture for fast feature embedding. In: **Proceedings of the 22nd ACM International Conference on Multimedia**. [S.l.: s.n.], 2014. p. 675–678.

KARPATHY, A. **Convolutional Neural Networks for Visual Recognition**. 2016. Disponível em: <http://cs231n.github.io/convolutional-networks/>.

KHAN, S. M. et al. 3D model based vehicle classification in aerial imagery. In: **2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2010. p. 1681–1687.

KOHAVI, R.; OTHERS. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: **IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2**. [S.l.: s.n.], 1995. v. 14, n. 2, p. 1137–1145.

KOZEMPEL, K.; HAUSBURG, M.; REULKE, R. Airborne vehicle detection using SURF-descriptors and support vector machines. In: **2011 IEEE Forum on Integrated and Sustainable Transportation System (FISTS)**. [S.l.: s.n.], 2011. p. 73–78.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. **Advances In Neural Information Processing Systems**, p. 1–9, 2012.

LAVIGNE, D. A. et al. Unsupervised classification and clustering of image features for vehicle detection in large scale aerial images. In: **2010 13th Conference on Information Fusion (FUSION)**. [S.l.: s.n.], 2010. p. 1–8.

LAWSON, C. L. et al. Basic linear algebra subprograms for fortran usage. **ACM Transactions on Mathematical Software (TOMS)**, ACM, v. 5, n. 3, p. 308–323, 1979.

LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2323, 1998. ISSN 00189219.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. Convolutional networks and applications in vision. In: **Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)**. [S.l.: s.n.], 2010. p. 253–256.

LEVI, G.; HASSNER, T. Age and Gender Classification Using Convolutional Neural Networks. **IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops**, 2015.

LIANG, P. et al. Vehicle detection in wide area aerial surveillance using Temporal Context. In: IEEE. **2013 16th International Conference on Information Fusion (FUSION)**. [S.l.], 2013. p. 181–188.

LIENHART, R.; KURANOV, A.; PISAREVSKY, V. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: **Pattern Recognition**. [S.l.]: Springer, 2003. p. 297–304.

LIN, F. et al. Development of a vision-based ground target detection and tracking system for a small unmanned helicopter. **Science in China Series F: Information Sciences**, Springer, v. 52, n. 11, p. 2201–2215, 2009.

LOWE, D. G. Object recognition from local scale-invariant features. In: **The proceedings of the seventh IEEE international conference on Computer vision, 1999.** [S.l.: s.n.], 1999. v. 2, p. 1150–1157.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, Springer, v. 60, n. 2, p. 91–110, 2004.

LUVIZON, D. C.; NASSU, B. T.; MINETTO, R. Vehicle speed estimation by license plate detection and tracking. In: **2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2014. p. 6563–6567.

MALEK, S. et al. Efficient framework for palm tree detection in UAV images. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 7, p. 1–12, 2014.

MA'SUM, M. A. et al. Simulation of intelligent Unmanned Aerial Vehicle (UAV) For military surveillance. In: IEEE. **2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)**. [S.l.], 2013. p. 161–166.

MATESE, A. et al. Intercomparison of UAV, Aircraft and Satellite Remote Sensing Platforms for Precision Viticulture. **Remote Sensing**, v. 7, n. 3, p. 2971–2990, 2015. ISSN 2072-4292.

MATSON, J. **Resumo de fatos sobre a radiação em Fukushima**. 2016. 1 p. Disponível em: <http://www2.uol.com.br/sciam/artigos/resumo_de_fatos_sobre_a_radiacao_em_fukushima.html>.

MOLINA, P. et al. Searching Lost People With UAVs: the System and Results of the Close-Search Project. **ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, p. 441–446, 2012.

MONTABONE, S.; SOTO, A. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. **Image and Vision Computing**, Elsevier, v. 28, n. 3, p. 391–402, 2010.

MORANDUZZO, T.; MEKHALFI, M. L.; MELGANI, F. Lbp-Based Multiclass Classification Method for UAV Imagery. In: **2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)**. [S.l.: s.n.], 2015. p. 2362–2365.

MORANDUZZO, T.; MELGANI, F. A SIFT-SVM method for detecting cars in UAV images. In: **2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)**. [S.l.: s.n.], 2012. p. 6868–6871.

MORANDUZZO, T.; MELGANI, F. Comparison of different feature detectors and descriptors for car classification in uav images. In: **2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)**. [S.l.: s.n.], 2013. p. 204–207.

MORANDUZZO, T.; MELGANI, F. Automatic car counting method for unmanned aerial vehicle images. **IEEE Transactions on Geoscience and Remote Sensing**, v. 52, n. 3, p. 1635–1647, 2014.

MORANDUZZO, T.; MELGANI, F. Detecting cars in uav images with a catalog-based approach. **IEEE Transactions on Geoscience and Remote Sensing**, v. 52, n. 10, p. 6356–6367, 2014.

MORANDUZZO, T.; MELGANI, F.; DAAMOUCHE, A. An object detection technique for very high resolution remote sensing images. In: **2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA)**. [S.l.: s.n.], 2013. p. 79–83.

MORSE, B. S.; THORNTON, D.; GOODRICH, M. A. Color anomaly detection and suggestion for wilderness search and rescue. In: ACM. **Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction**. [S.l.], 2012. p. 455–462.

NAGENDRAN, A.; HARPER, D.; SHAH, M. New system performs persistent wide-area aerial surveillance. **SPIE Newsroom**, v. 5, p. 20–28, 2010.

OJALA, T.; PIETIKAINEN, M.; HARWOOD, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: **Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing.** [S.l.: s.n.], 1994. p. 582–585.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 24, n. 7, p. 971–987, 2002.

OLIVEIRA, D. C.; WEHRMEISTER, M. A. Towards real-time people recognition on aerial imagery using convolutional neural networks. In: **2016 IEEE 19th International Symposium on Real-Time Distributed Computing (ISORC)**. [S.l.: s.n.], 2016. p. 27–34.

PAPAGEORGIOU, C. P.; OREN, M.; POGGIO, T. A general framework for object detection. In: **Sixth International Conference on Computer Vision, 1998.** [S.l.: s.n.], 1998. p. 555–562.

PERLIN, H. A.; LOPES, H. S. Extracting human attributes using a convolutional neural network approach. **Pattern Recognition Letters**, v. 68, p. 250–259, 2015. ISSN 01678655.

PINGTING, L. et al. Stationary vehicle detection in aerial surveillance with a UAV. In: **2012 8th International Conference on Information Science and Digital Content Technology (ICIDT)**. [S.l.: s.n.], 2012. p. 567–570.

PORTMANN, J. et al. People detection and tracking from aerial thermal views. In: **2014 IEEE International Conference on Robotics and Automation (ICRA)**. [S.l.: s.n.], 2014. p. 1794–1800.

PRATI, R. C.; BATISTA, G.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215–222, 2008.

REILLY, V.; SOLMAZ, B.; SHAH, M. Geometric constraints for human detection in aerial imagery. In: **Computer Vision–ECCV 2010**. [S.l.]: Springer, 2010. p. 252–265.

REILLY, V.; SOLMAZ, B.; SHAH, M. Shadow casting out of plane (SCOOP) candidates for human and vehicle detection in aerial imagery. **International Journal of Computer Vision**, Springer, v. 101, n. 2, p. 350–366, 2013.

REMONDINO, F. et al. UAV photogrammetry for mapping and 3d modeling–current status and future perspectives. **International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 38, n. 1, p. C22, 2011.

RICHARDS, A. **Alien vision: exploring the electromagnetic spectrum with imaging technology**. [S.l.]: SPIE Press, 2001.

ROGALSKI, A. Material considerations for third generation infrared photon detectors. **Infrared physics & technology**, Elsevier, v. 50, n. 2, p. 240–252, 2007.

RUDOL, P.; DOHERTY, P. Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery. In: **2008 IEEE Aerospace Conference**. [S.l.: s.n.], 2008. p. 1–8.

RUIZ-GONZALEZ, R. et al. An SVM-Based Classifier for Estimating the State of Various Rotating Components in Agro-Industrial Machinery with a Vibration Signal Acquired from a Single Point on the Machine Chassis. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 14, n. 11, p. 20713–20735, 2014.

SAIF, A. F. M. S. et al. A Review of Machine Vision based on Moving Objects: Object Detection from UAV Aerial Images. **International Journal of Advancements in Computing Technology**, v. 5, n. 15, p. 57–72, 2013.

SEKMEN, A.; YAO, F.; MALKANI, M. Smart video surveillance for airborne platforms. **Robotica**, v. 27, n. 05, p. 749–761, 2009.

SERRA, J. **Image Analysis and Mathematical Morphology**. Orlando, FL, USA: Academic Press, Inc., 1983. ISBN 0126372403.

SIMÕES, P. R. **O uso de drones em desastres ambientais**. 2016. 1 p. Disponível em: <http://droneng.com.br/o-uso-de-drones-em-desastres-ambientais/>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **CoRR**, abs/1409.1556, 2014.

SU, A. et al. Online cascaded boosting with histogram of orient gradient features for car detection from unmanned aerial vehicle images. **Journal of Applied Remote Sensing**, International Society for Optics and Photonics, v. 9, n. 1, 2015.

SUN, S.; SALVAGGIO, C. Aerial 3d building detection and modeling from airborne lidar point clouds. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, IEEE, v. 6, n. 3, p. 1440–1449, 2013.

SZEGEDY, C. et al. Going deeper with convolutions. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2015. p. 1–9.

SZELISKI, R. **Computer vision: algorithms and applications**. [S.l.]: Springer Science & Business Media, 2010.

TERRA NOTICIAS. **Busca por desaparecidos continua em Mariana**. 2015. 1 p. Disponível em: <http://noticias.terra.com.br/ciencia/busca-por-desaparecidos-continua-em-mariana,be73fbe24aed18b9d06c0fe2216345c9ys51yalu.html>.

TEUTSCH, M.; KRÜGER, W.; HEINZE, N. Detection and classification of moving objects from UAVs with optical sensors. In: **SPIE Defense, Security, and Sensing**. [S.l.: s.n.], 2011. p. 80501J–80501J.

TEUTSCH, M. et al. Low Resolution Person Detection with a Moving Thermal Infrared Camera by Hot Spot Classification. In: **2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2014. p. 209–216.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. [S.l.]: Elsevier Science, 2008. ISBN 9780080949123.

TOKEKAR, P. et al. Sensor planning for a symbiotic UAV and UGV system for precision agriculture. In: IEEE. **2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**. [S.l.], 2013. p. 5321–5326.

TONGPHU, S.; THONGSAK, N.; DAILEY, M. N. Rapid detection of many object instances. In: SPRINGER. **Advanced Concepts for Intelligent Vision Systems**. [S.l.], 2009. p. 434–444.

TUERMER, S. et al. Airborne vehicle detection in dense urban areas using HoG features and disparity maps. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, IEEE, v. 6, n. 6, p. 2327–2337, 2013.

UPTON, E.; HALFACREE, G. **Raspberry Pi user guide**. [S.l.]: John Wiley & Sons, 2014.

VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. In: IEEE. **Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001.** [S.l.], 2001. v. 1, p. I—-511.

WALTHER, D.; KOCH, C. Modeling attention to salient proto-objects. **Neural networks**, Elsevier, v. 19, n. 9, p. 1395–1407, 2006.

WANG, L.; HE, D.-C. Texture classification using texture spectrum. **Pattern Recognition**, Elsevier, v. 23, n. 8, p. 905–910, 1990.

WANG, S. Vehicle detection on aerial images by extracting corner features for rotational invariant shape matching. In: **2011 IEEE 11th International Conference on Computer and Information Technology (CIT)**. [S.l.: s.n.], 2011. p. 171–175.

WU, B.; NEVATIA, R. Cluster boosted tree classifier for multi-view, multi-pose object detection. In: IEEE. **IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007**. [S.l.], 2007. p. 1–8.

XIAO, J. et al. Vehicle and person tracking in aerial videos. In: **Multimodal Technologies for Perception of Humans**. [S.l.]: Springer, 2008. p. 203–214.

YANG, B.; SHARMA, P.; NEVATIA, R. Vehicle detection from low quality aerial LIDAR data. In: **2011 IEEE Workshop on Applications of Computer Vision (WACV)**. [S.l.: s.n.], 2011. p. 541–548.

YANG, Y. et al. Vehicle detection methods from an unmanned aerial vehicle platform. In: **2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES)**. [S.l.: s.n.], 2012. p. 411–415.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. **European Conference on Computer Vision**. [S.l.], 2014. p. 818–833.

ZHANG, C.; MA, Y. **Ensemble machine learning**. [S.l.]: Springer, 2012.

ZHAO, T.; NEVATIA, R. Car detection in low resolution aerial images. **Image and Vision Computing**, Elsevier, v. 21, n. 8, p. 693–703, 2003.

ZHENG, Z. et al. Vehicle detection based on morphology from highway aerial images. In: **2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)**. [S.l.: s.n.], 2012. p. 5997–6000.

ZONGJIAN, L. Uav for mapping—low altitude photogrammetric survey. **International Archives of Photogrammetry and Remote Sensing**, v. 37, p. 1183–1186, 2008.