
**Detecção e classificação de objetos em imagens
para rastreamento de veículos**

Raphael Montanari

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Raphael Montanari

Detecção e classificação de objetos em imagens para rastreamento de veículos

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação - ICMC-USP, como
parte dos requisitos para obtenção do título de
Mestre em Ciências - Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientadora: Profa. Dra. Roseli Aparecida Francelin
Romero

USP – São Carlos
Outubro de 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

M764d Montanari, Raphael
Detecção e classificação de objetos em imagens
para rastreamento de veículos / Raphael Montanari;
orientadora Roseli Aparecida Francelin Romero. --
São Carlos, 2015.
77 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2015.

1. Robótica. 2. Visão computacional. 3.
Aprendizado de máquina. I. Romero, Roseli Aparecida
Francelin, orient. II. Título.

Raphael Montanari

Detection and classification of objects in images for vehicle tracking

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Roseli Aparecida Francelin Romero

USP – São Carlos
October 2015

*Este trabalho é dedicado à Maria Luiza Cruzera Montanari,
que semeou em mim o sonho de realizar esta pesquisa.*

AGRADECIMENTOS

À professora Roseli pela amizade e confiança e por todo o direcionamento das pesquisas e incentivo para a realização deste trabalho.

À meus pais e padrinhos por estarem sempre comigo dando todo apoio, suporte e amor.

À Deus, por todo amor, força e coragem concedido e por cada momento vivido durante a realização deste trabalho.

Aos integrantes do BIOCOP, LASI e Laboratório de Aprendizado de Robôs: Daniel, Mu-
rillo, Eduardo, Marcelo, Adam, Fernando, Ewerton e Alcides pelas contribuições, colaborações,
sugestões e discussão de ideias.

Aos professores do Centro de Robótica de São Carlos em especial ao coordenador Marco Terra.

Minha gratidão aos amigos e colegas que ganhei durante este período, em especial aos atletas do grupo de corrida Pelotão e aos amigos do Ministério Universidades Renovadas.

Agradecimentos especiais ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro recebido no primeiro semestre e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio por meio do auxílio científico-financeiro (Processo 2012/14725-4).

*“Não sou nada.
Nunca serei nada.
Não posso querer ser nada.
À parte isso, tenho em mim todos os sonhos do mundo.”*
(Fernando Pessoa)

RESUMO

MONTANARI, R.. **Detecção e classificação de objetos em imagens para rastreamento de veículos.** 2015. 77 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A robótica é uma área multidisciplinar que cresce continuamente com a contribuição do avanço científico e aumento frequente do poder computacional do hardware. As pesquisas em robótica estão divididas em diversas linhas de investigação. A visão computacional é uma das linhas de pesquisa de grande interesse devido à farta variedade de métodos e técnicas oferecidas. Um dos maiores desafios para os robôs é descobrir e analisar o ambiente em que estão inseridos. Dentre os principais sensores que podem ser utilizados, as câmeras digitais oferecem um bom benefício: podem ser leves, pequenas e baratas, características fundamentais para alguns robôs. Este trabalho propõe o desenvolvimento e análise de um sistema de visão computacional para rastrear veículos usando sistemas de detecção e classificação de segmentos em imagens. Para atingir os objetivos são investigados métodos de extração de informações das imagens, modelos de atenção visual e modelos de aprendizado bioinspirados para detecção e classificação de veículos. Para a tarefa de atenção visual foram utilizadas as técnicas de geração de mapas de saliência iNVT e VOCUS2, enquanto que para classificação foi empregada a técnicas *bag-of-features* e finalmente, para o rastreamento do veículo especificado, durante seu percurso em uma rodovia, foi adotada a técnica Camshift com filtro de Kalman. O sistema desenvolvido foi implementado com um robô aéreo e testado com imagens reais contendo diferentes veículos em uma rodovia e os resultados de classificação e rastreamento obtidos foram muito satisfatórios.

Palavras-chave: Robótica, Visão computacional, Aprendizado de máquina.

ABSTRACT

MONTANARI, R.. **Detecção e classificação de objetos em imagens para rastreamento de veículos.** 2015. 77 f. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Robotics is a multidisciplinary area that continually grows with the contribution of scientific advancement and frequent increase in computational hardware power. Research in robotics are divided into several lines of investigation. Computer vision is one of the research areas of great interest due to the abundant variety of methods and techniques offered. One of the biggest challenges for the robots is to discover and analyze the environment in which they are inserted. Among the main sensors that can be used, digital cameras offer good benefits: they can be lightweithg, small and cheap, which are fundamental characteristics for some robots. This work undertakes the development and analysis of a computer vision system to track vehicles by detecting and classifying segments in imaging systems. To achieve the objectives, methods on image information extraction, visual attention models and bioinspired learning models were studied for detection and classification of vehicles. For the task of visual attention the INVT and VOCUS2 models were used to generate saliency maps, while for classification was applied the bag-of-features method and finally to track the specified vehicle during its journey on a highway, it was adopted CamShift technique joint with a Kalman filter. The developed system was implemented with an aerial robot and tested with real images containing different vehicles on a highway and the results of classification and tracking obtained were very satisfactory.

Key-words: Robotics, Computer vision, Machine learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Estrutura do modelo de saliência iNVT, adaptada de (ITTI; KOCH, 2001).	30
Figura 2 – Processamento dos canais de características do sistema VOCUS2 (FRINTROP; WERNER; GARCIA, 2015).	35
Figura 3 – Sistema de visão proposto em Benicasa (2013).	38
Figura 4 – Exemplo de correspondência de imagens (LOWE, 2004).	39
Figura 5 – Representação gráfica da função de diferença de Gaussiana (LOWE, 2004).	40
Figura 6 – Construção do descritor do ponto-chave (LOWE, 2004).	41
Figura 7 – Filtros de caixa que aproximam derivadas gaussianas de segunda ordem.	42
Figura 8 – Representação gráfica da vizinhança $3 \times 3 \times 3$.	43
Figura 9 – Filtro wavelet Haar.	44
Figura 10 – Entradas do descritor de uma sub-região (BAY; TUYTELAARS; GOOL, 2006).	44
Figura 11 – Etapas para detecção, classificação e rastreamento de veículos em imagens.	51
Figura 12 – Robô aéreo de asa rotativa utilizado para gravação dos vídeos.	56
Figura 13 – Software Ground Station DJI (INNOVATIONSHOR, 2015 (accessed January 7, 2015)).	56
Figura 14 – Imagem de entrada do sistema.	57
Figura 15 – Imagem de entrada para o detector.	58
Figura 16 – Mapas de conspicuidade de (a) intensidades, (b) cores e (c) orientações e o (d) reconhecimento.	59
Figura 17 – (a) Imagem segmentada e (b) imagem destacando o objeto saliente.	59
Figura 18 – Gráfico dos valores salientes dos objetos que competiam pela atenção.	59
Figura 19 – Exemplo de imagem de entrada e imagem exibindo veículos detectados.	60
Figura 20 – Os três tipos de veículos considerados: (a) carro, (b) caminhão e (c) moto.	60
Figura 21 – Subconjunto de imagens considerados para o treinamento de caminhões.	61
Figura 22 – Sequência de imagens exibindo a posição do veículo em cada quadro.	62

LISTA DE TABELAS

Tabela 1 – Principais características entre os sistemas iNVT e VOCUS2. (FRINTROP; WERNER; GARCIA, 2015)	34
Tabela 2 – A matriz de confusão da classificação.	58
Tabela 3 – A matriz de confusão do classificador.	62

SUMÁRIO

1	INTRODUÇÃO	19
1.1	Contexto e motivação	19
1.2	Objetivos	20
1.3	Organização da dissertação	21
2	TRABALHOS RELACIONADOS	23
2.1	Detecção de objetos em movimento e rastreamento em vídeos	23
2.2	Sistemas computacionais de atenção visual	25
2.3	Sistemas de rastreamento de veículos	27
2.4	Considerações finais	28
3	FUNDAMENTOS TEÓRICOS	29
3.1	Modelo de atenção visual <i>bottom-up</i> - (iNVT)	29
3.2	Detecção visual de objetos com um sistema de atenção computacional (VOCUS2)	34
3.3	Modelo bottom-up e top-down	37
3.4	Extração de características de imagens	37
3.4.1	<i>SIFT</i>	38
3.4.2	<i>SURF</i>	41
3.5	Bag-of-features	45
3.5.1	<i>K-médias</i>	45
3.5.2	<i>Detecção de objetos</i>	46
3.5.3	<i>Support Vector Machines - SVM</i>	46
3.6	<i>Meanshift</i>	47
3.7	<i>Camshift</i>	48
3.8	<i>Kalman</i>	48
3.9	Considerações finais	49
4	MODELO COMPUTACIONAL PROPOSTO PARA O RASTREAMENTO DE VEÍCULOS	51
4.1	Considerações finais	53
5	ANÁLISE DO MODELO PROPOSTO	55
5.1	Coleta de dados	56

5.2	Experimento com modelos de atenção visual para detecção	57
5.3	Experimento para classificação de segmentos com Bag-of-Features	58
5.4	Experimento com o rastreamento	62
5.5	Considerações finais	63
6	CONCLUSÕES E TRABALHOS FUTUROS	65
REFERÊNCIAS		69
Glossário		75
APÊNDICE A	PUBLICAÇÕES DECORRENTES DESTE TRABALHO	77



INTRODUÇÃO

1.1 Contexto e motivação

A robótica é uma ciência que atrai grande atenção da comunidade científica, das indústrias e do público (MATARIC, 2007). Recentemente, robôs móveis autônomos tem-se tornado acessíveis com aplicações na resolução de diversas tarefas em agricultura, exploração de ambientes inóspitos, atividades militares, entre outras. Com o avanço do poder computacional, os robôs conseguem executar tarefas cada vez mais complexas. As diversas oportunidades de aplicação fazem desta tecnologia alvo de um grande número de investigações. A integração de sistemas de visão computacional com a robótica móvel é um campo de grande interesse de pesquisa entre roboticistas.

Uma aplicação em que essa integração vem apresentando sucesso é a criação de Veículos Aéreos Não Tripulados (do inglês *Unmanned aerial vehicle*, UAV). Atualmente, as missões de robôs aéreos concentram-se em áreas de defesa, segurança pública e privada, agricultura e de meio-ambiente. Exemplos destes tipos de missões são encontrados em (SAMAD; BAY; GODBOLE, 2007), (MU *et al.*, 2009), (COSTA *et al.*, 2012) e (MERINO *et al.*, 2006). Os robôs aéreos tornaram-se uma plataforma padrão para o ensino e pesquisa de robótica (KRAJNÍK *et al.*, 2011) e, de acordo com (RYAN *et al.*, 2004), grande parte das pesquisas em robôs aéreos é realizada em três frentes: controle de voo, visão computacional e localização e mapeamento simultâneos (do inglês *Simultaneous localization and mapping*, SLAM) (LEONARD; DURRANT-WHYTE, 1991).

A visão computacional tem significativa contribuição às aplicações de robôs aéreos, pois fornece métodos para a extração de informações do ambiente a partir de imagens digitais relevantes para uma determinada aplicação; tais como, análise de movimento, detecção de objetos, rastreamento e mapeamento e localização simultâneos. Segundo (TRUCCO; VERRI, 1998), a visão computacional é definida como "um conjunto de técnicas computacionais que

visa a estimar ou explicitar as propriedades geométricas e dinâmicas do mundo 3D a partir de imagens digitais".

No Brasil, diversos institutos de pesquisa realizam pesquisas direcionadas à robôs aéreos, principalmente em temas como processamento de imagens (ANDRADE *et al.*, 2010), módulos de missão e piloto automático (BRANCO *et al.*, 2011) e controle (SCHILDIT *et al.*, 2012). Apesar destas iniciativas, quando comparado a outros países, há pouca pesquisa sobre robôs aéreos em realização no Brasil.

Este trabalho está inserido em um projeto do Laboratório de Aprendizado de Robôs (LAR-ICMC/USP) denominado LARVANTAR. O projeto tem como meta a construção de um sistema cooperativo de robôs heterogêneos autônomos de uso geral. O sistema do projeto abrange diversos tipos de robôs e faz alocação de tarefas de acordo com suas habilidades. Os temas em investigação no projeto LARVANTAR são visão computacional, alocação de tarefas e comunicação, planejamento de movimento e gerenciamento de robôs inspirado nos conceitos de robótica de enxames. Alguns robôs que integram o projeto são classificados como veículos aéreos não tripulados de asa rotativa; ou seja, possuem funcionamento similar a helicópteros, diferentemente dos veículos aéreos não tripulados de asa fixa que são similares a aviões.

A motivação do presente trabalho de mestrado é a construção de um módulo de rastreamento de objetos em tempo de execução.

O rastreamento de objetos dá-se através da descoberta e acompanhamento da posição do objeto na imagem. Para encontrar regiões de interesse em uma imagem, sistemas de atenção visual podem ser empregados para reduzir o espaço de busca e facilitar a identificação de objetos na imagem. A classificação de imagens é usada para separar os veículos dos demais objetos encontrados. Métodos de correspondência entre imagens possuem alta precisão, rápida execução e podem ser usados para identificar objetos.

Modelos de atenção visual recentes (BENICASA, 2013) (FRINTROP; WERNER; GARCIA, 2015) são usados para evidenciar diversos objetos contidos em imagens únicas com boas taxas de detecção. Esses modelos podem ser usados para rastrear objetos desde que sejam adaptados para trabalhar com sequencias de imagens.

A importância deste trabalho dá-se pelo fato de que o rastreamento de veículos com imagens capturadas por um robô aéreo móvel ainda é um desafio para os pesquisadores. Esta abordagem, pelas vantagens que proporciona, representa uma ótima oportunidade de investigação.

1.2 Objetivos

Este trabalho estabelece a hipótese de que sistemas de atenção visual podem ser adaptados para detectar objetos em uma sequência de imagens e em conjunto com técnicas de classificação

podem ser aplicados para rastrear veículos a partir de imagens de um robô aéreo.

Assim, os seguintes objetivos são propostos:

- Adaptar o sistema de atenção visual Bottom-Up e Top-Down ([BENICASA, 2013](#)) para detectar segmentos com destaque a partir de uma entrada de vídeo;
- Classificar imagens de veículos dentre os segmentos detectados;
- Rastrear veículos em um vídeo obtido por um robô aéreo;
- Fornecer informações visuais para o módulo de controle do robô aéreo do sistema LAR-VANTAR.

1.3 Organização da dissertação

Este texto está organizado como segue: No capítulo [2](#) estão apresentados os trabalhos relacionados à sistemas computacionais de atenção visual e sistemas de rastreamento de objetos. Descritos no capítulo [3](#) estão os fundamentos teóricos e principais conceitos sobre atenção visual, características de imagens, detecção, classificação e rastreamento de objetos em imagens. No capítulo [4](#) é detalhado o modelo proposto para o rastreamento de veículos. No capítulo [5](#) são explicados os experimentos realizados. As conclusões do trabalho e perspectivas futuras são apresentados no capítulo [6](#).



TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados alguns trabalhos encontrados na literatura sobre detecção de objetos e rastreamento de objetos em movimento. Em seguida, são descritos também alguns trabalhos sobre rastreamento de veículos, que é o foco principal da presente dissertação.

2.1 Detecção de objetos em movimento e rastreamento em vídeos

O rastreamento de objetos em movimento é um campo de pesquisa muito abrangente na área de visão computacional. Análise de conteúdo de vídeo digital é um item importante para a indexação baseada em conteúdo multimídia, bem como, para a recuperação de vídeo baseada em conteúdo e sistemas visuais de vigilância. Processamento de vídeos é uma importante área de pesquisa em ciência de computação na qual estão sendo investigados processos de segmentação, detecção e rastreamento de objetos em vídeos 2D (capturados de câmera fixa ou em movimento) que formam a base para aplicações tais como recuperação de informação, indexação baseada em conteúdo, rastreamento de pessoas, sistemas de vigilância visual e monitoramento de tráfego.

A detecção de um objeto em movimento em um vídeo é uma etapa essencial para aplicações complexas, tais como rastreamento de objetos e recuperação de vídeos. O objetivo da detecção de objetos em movimento é localizar objetos em primeiro plano para extrair informações úteis ao sistema como trajetória, orientação e velocidade, etc. O desafio desta tarefa está em resolver a perda de informação que pode ser causada por planos de fundo complexos ou dinâmicos, mudanças de iluminação, sombras, oclusão, etc. Assim, o rastreamento de objetos em vídeo pode ser definido como o processo de segmentar os objetos de interesse a partir de cenas de vídeo.

Nas últimas décadas, vários métodos foram desenvolvidos para o rastreamento de objetos a partir de vídeos. De acordo com Liu, Yuen e Qiu (2009), esses métodos podem ser categorizados

em três abordagens: baseada em contorno, baseada em orientação e baseada em distribuição. [Huang et al. \(2008\)](#) classificou os métodos de detecção em abordagens tais como baseada em características, baseada em diferenças de quadros e baseada em modelo.

Para aplicações de vigilância visual existem duas situações principais, que podem ser ambientes internos e externos. Segundo [Lei e Xu \(2006\)](#), em ambientes internos o espaço de monitoramento é restrito e a distância entre a câmera e o objeto é relativamente pequena. Essa situação resulta em objetos de tamanho maior com aparência mais bem definida. Portanto, um modelo de aparência ([ZHOU; CHELLAPPA; MOGHADDAM, 2004](#)) pode ser usado para tratar oclusões e rotular objetos. Em ambientes externos, geralmente deve-se considerar cenários complexos e imprevistos, nos quais as formas e tamanhos dos objetos alteram muito e sua posição varia no campo de visão da câmera. As condições climáticas que se alteram durante o dia e as características de movimento de certos objetos provocam mudanças repentinas, nas quais o objeto em movimento pode desaparecer e aparecer em uma dada cena.

Em problemas de Visão Computacional, os algoritmos de detecção e rastreamento de objetos geralmente falham em atuar com tais cenários complexos. Existem uma variedade de métodos dedicados para o rastreamento de objetos no processamento de vídeos, tais como Subtração de plano de fundo ([PICCARDI, 2004](#)), Meanshift ([FUKUNAGA; HOSTETLER, 1975](#)), Camshift ([BRADSKI, 1998](#)) e Fluxo ótico ([HORN; SCHUNCK, 1981](#)), descritas brevemente a seguir.

- Subtração de Plano de Fundo: De acordo com [Huang et al. \(2008\)](#), o método mais simples de detecção baseada em diferença de quadros é a Subtração de plano de fundo. Neste método, os valores de *pixels* individuais são estatisticamente modelados com o tempo. Assim, a detecção de objetos é realizada ao encontrar os *pixels* com valores que desviam do modelo estatístico do plano de fundo de uma certa cena.
- MeanShift:
- CamShift:
- Fluxo Ótico:

A técnica Fluxo ótico é uma abordagem baseada em orientação, visando detectar a direção do movimento com precisão, mas é sensível à mudanças de iluminação.

Também há as abordagens baseadas em características, que consistem em encontrar características correspondentes em uma sucessão de quadros. As características geralmente são cantos, cores e contornos.

2.2 Sistemas computacionais de atenção visual

Rotineiramente, recebemos incontáveis estímulos através de todos os sentidos e temos a impressão de perceber uma rica interpretação do nosso mundo visual. Mudanças em nosso ambiente atraem nossa atenção. Entretanto, apenas uma pequena parte da cena é analisada em detalhes a cada momento. Como o cérebro não consegue processar tamanha quantidade de informação é preciso selecionar o que nos interessa. As pessoas são capazes de automaticamente prestar atenção a regiões de interesse ao seu redor e varrer uma cena mudando rapidamente o foco de atenção. A este processo, dá-se o nome de atenção visual. A ordem na qual uma cena é investigada é determinada por mecanismos de atenção seletiva. Atenção visual define a habilidade mental em selecionar estímulos, respostas, memórias ou pensamentos, que são relevantes entre muitos outros que são irrelevantes.

A base de muitos modelos de atenção remonta ao trabalho de [Treisman e Gelade \(1980\)](#). Eles afirmaram que recursos visuais são importantes e são combinados para dirigir a atenção humana. Em [Koch e Ullman \(1987\)](#) foi proposto um modelo *feed-forward* para combinar esses recursos e introduzir o conceito de mapa de saliência.

Um mapa de saliência é definido como um mapa organizado topograficamente que representa saliências visuais correspondentes de uma cena visual. [Koch e Ullman \(1987\)](#) definiram uma região saliente como "uma certa região determinada primariamente pelo modo como esta região difere de sua vizinhança, considerando cor, orientação, movimento, profundidade, etc.". Eles também introduziram uma rede neural (*winner-take-all*) que seleciona o local mais saliente e emprega o mecanismo de inibição de retorno, permitindo que o foco de atenção mude para o próximo local mais saliente.

Vários sistemas foram criados, implementando os modelos relacionados que processam imagens digitais. A primeira implementação completa e validada do modelo de [Koch e Ullman \(1987\)](#) foi proposta por [Itti, Koch e Niebur \(1998\)](#), sendo aplicado a cenas sintéticas e reais. Desde então, tem havido um crescente interesse nessa área. Várias abordagens com diferentes pressupostos para a modelagem da atenção têm sido propostos e têm sido avaliadas em relação a diferentes conjuntos de dados.

Baseado nessa arquitetura, [Clark e Ferrier \(1988\)](#) propuseram uma das primeiras implementações de um sistema de atenção. Os cálculos são realizados usando o recurso de operações por filtro. Implementadas através de sistema de processamento de imagem para fins especiais, os modelos são baseados em filtros.

Outro modelo pioneiro de atenção baseado em filtro foi introduzido por [Milanese \(1993\)](#), que incluiu informação *top-down* de um sistema de reconhecimento de objeto que é percebido por memórias associativas distribuídas (DAMs). Apesar da introdução de conceitos sobre mapas de conspicuidade e cálculos de recursos baseados em mecanismos centro-vizinhanças (chamados de "operadores de conspicuidade"), o sistema não define os valores de referência para várias

técnicas que são usadas em modelos computacionais de atenção. Um dos modelos mais antigos de atenção, que é amplamente conhecido é o ajuste seletivo de Tsotsos – modelo ST de atenção visual ([TSOTSOS, 1990](#)). Trata-se de um modelo conectivo de arquitetura piramidal com um feixe inibidor. Também é possível considerar a informação *top-down* por sinais específicos do alvo por meio da inibição de todas as regiões com características diferentes das características do alvo ou através de regiões de uma localização especificada. O modelo foi implementado com várias características tais como, luminância, orientação, cores oponentes, movimento e profundidade estereoscópica. Originalmente, cada versão do modelo ST processava apenas uma dimensão característica, mas, posteriormente, foi estendido para executar recursos de interação. Uma adaptação incomum do modelo de Tsotsos é fornecida por [Ramström e Christensen \(2002\)](#) na qual o controle de distribuição do sistema de atenção é realizado usando conceitos da teoria dos jogos. Metaforicamente, os cantos da pirâmide estão sujeitos à negociação em mercado, as características são os bens, os bens raros são caros (as características são salientes), e os resultados da negociação representam a saliência. Um dos sistemas de atenção atualmente mais conhecido, baseado no modelo de Koch-Ullman, é o modelo de visão neuromórfica (iNVT) que é mantido e atualizado pelo grupo de Itti ([ITTI; KOCH; NIEBUR, 1998](#)). Esse modelo, bem como a sua aplicação, serve de base para diversos grupos de pesquisa. Uma razão para isso é a boa documentação e a disponibilidade *on-line* do código fonte.

Considerando que o iNVT é um dos sistemas mais conhecidos e distribuídos, muitos grupos o testaram e várias melhorias foram implementadas. Por exemplo, [Draper e Lionelle \(2005\)](#) desenvolveram o sistema de atenção seletiva para a etapa inicial do processo (SAFE) que, contudo, apresenta várias diferenças. Por exemplo, não há combinação de mapas de características em toda a escala, mas sim de manutenção, que resulta em uma pirâmide de mapas de saliência. Eles mostraram que essa abordagem é mais estável no que diz respeito às transformações geométricas, tais como, translações, rotações e reflexões. Além disso, [Frintrop \(2006\)](#) sugeriu separar os cálculos de intensidade em cálculos de liga-desliga, em vez de combiná-los em um único mapa. Certos efeitos de determinação imediata de regiões, que só diferem em uma característica, são detectados por esta separação. O mesmo se aplica para a separação de vermelho e verde, bem como o azul e o amarelo.

O sistema de atenção de Hamker estabeleceu a imitação dos processos neurais no córtex visual humano ([HAMKER, 2005](#)). A saliência *bottom-up* é semelhante ao modelo iNVT de Itti. O sistema é capaz de aprender um alvo e lembrar-se dos valores da característica de um estímulo apresentado. Um ponto interessante é que o sistema de Hamker pode realizar uma espécie de reconhecimento de objetos muito difícil, através das chamadas unidades de detecção de jogo. Uma abordagem para seleção hierárquica baseada em regiões de interesse é apresentada por [Sun e Fisher \(2003\)](#). As regiões de interesse são calculadas em diferentes escalas; primeiro, numa escala grosseira e, em seguida, se a região for suficientemente interessante, investiga-se uma escala mais fina. Backer apresentou um modelo interessante de atenção com dois estágios de seleção ([BACKER; MERTSCHING; BOLLMANN, 2001](#)). A primeira fase assemelha-se

às arquiteturas padrão (KOCH; ULLMAN, 1987), no qual o resultado não é um único foco, mas um número pequeno, geralmente 4, de locais importantes. Na segunda fase, um desses locais é selecionado e gera-se um único foco de atenção. O modelo investiga alguns dos dados experimentais ignorados em outros trabalhos de rastreamentos de objetos e inibição baseada em objeto de retorno. O sistema VOCUS, de (FRINTROP, 2006), tem vários aspectos que o tornam adequado para aplicações em visão computacional e robótica. A parte *top-down* permite uma pesquisa simples para os objetos de destino.

2.3 Sistemas de rastreamento de veículos

Sistemas de rastreamento de veículos

Em (SIVARAMAN; TRIVEDI, 2013), é apresentado um levantamento das pesquisas sobre detecção, rastreamento e comportamento de veículos, baseados em visão, analisando o progresso das técnicas de visão computacional para percepção de ambiente. Os autores descrevem as pesquisas que colocam detecção de veículos no contexto de análise de ambiente, baseado em sensores, e detalham avanços nos algoritmos discutindo visão monocular, visão estéreo e fusão de sensores. O rastreamento de veículos é discutido nos domínios da visão monocular e estéreo, filtragem, estimativa e modelos dinâmicos. Ainda, são mencionadas as pesquisas recentes em veículos inteligentes, concentrados em utilizar medidas espaço-temporais, trajetórias e outras características para descrever o comportamento dos veículos.

Coifman *et al.* (1998) descreveram as questões associadas com rastreamento baseado em características e apresenta uma implementação em tempo real de um protótipo, avaliando o desempenho do sistema em um grande conjunto de dados.

Em (KOLLER; WEBER; MALIK, 1994), é tratado o problema da vigilância de tráfego em um sistema avançado de gerenciamento de transporte e foi proposta uma abordagem para detecção e rastreamento de veículos em cenas de tráfego em vias que alcança um grau de acurácia e confiança superior aos sistemas anteriores. Isto, devido ao tratamento de oclusões e ao emprego de rastreio de contorno baseado na intensidade e movimento das bordas.

No trabalho proposto por (LIPTON; FUJIYOSHI; PATIL, 1998), foi descrito um método de extração de alvos, em movimento, a partir de um fluxo de vídeo em tempo real, classificando os alvos em categorias predefinidas e, então, faz o rastreamento deles. Alvos móveis são detectados pela diferença de *pixels* entre quadros consecutivos. Os alvos são aplicados à classificação para dividí-los em três categorias, pessoas, veículos e itens do plano de fundo. Depois de classificados, os alvos são rastreados por uma combinação de diferenciação temporal com correspondência de padrões.

(GUPTE *et al.*, 2002) apresentaram algoritmos para detecção baseados em visão e classificação de veículos em sequencia de imagens monoculares de cenas de tráfego gravados

por uma câmera estacionária. O processamento é realizado em três níveis: imagens brutas, região de interesse e região do veículo. Veículos são modelados como segmentos retangulares com comportamentos dinâmicos. O método proposto é baseado na correspondência entre regiões e veículos, enquanto os veículos se movem em uma sequência de vídeo.

[Sullivan et al. \(1997\)](#) demonstraram um modelo para localizar, rastrear e classificar veículos. Métodos, baseados em modelo para rastreamento visual de veículos para uso em sistemas de tempo real, foram descritos, com objetivos de prover monitoramento contínuo e classificação de tráfego a partir de uma câmera fixa em uma rodovia movimentada.

[Kim e Malik \(2003\)](#) introduziram uma abordagem de rastreamento baseada em modelo para detecção veículos em 3D. O algoritmo de detecção e descrição de veículos é baseado no agrupamento de características probabilísticas.

Nesta dissertação, um sistema de visão monocular será desenvolvido para rastreamento e detecção de veículos. O sistema de atenção visual, proposto por Itti, será utilizado para a tarefa de detecção de objetos. Os objetos detectados são classificados, usando a técnica *bag-of-features* e os objetos de interesse, após a classificação, são enviados para sistema de rastreamento. O rastreamento será realizado usando o método Camshift em conjunto com filtro de Kalman.

A diferença para outros sistemas de rastreamento está no escopo da aplicação que é usada para rastreamento a partir de imagens filmadas por um robô aéreo, tendo uma câmera dinâmica e o uso de um sistema de atenção visual para detectar objetos salientes em imagens, além de classificar objetos, usando características locais em imagens. Assim, pretende-se contribuir com um método robusto para rastrear veículos em ambientes no qual não é possível o uso de câmeras fixas.

2.4 Considerações finais

Neste capítulo, foi introduzida uma visão geral dos processos de detecção e rastreamento de objetos e de sistemas de atenção visual. Alguns trabalhos de detecção e rastreamento de veículos foram também apresentados, por ser este o foco do presente trabalho. No próximo capítulo, serão apresentados os principais métodos que formam a base do sistema a ser desenvolvido.



FUNDAMENTOS TEÓRICOS

O rastreio de objetos é a tarefa de estimação de movimento em sequências de imagens. O rastreamento possibilita acompanhar o percurso de um objeto em tempo real, isto é, rastrear um objeto em movimento em frente a uma câmera por meio do fluxo de cada pixel, ou com base na semelhança e na combinação de amostras. A principal etapa para realizar o rastreio de um objeto é detectar a presença deste na cena. Para isso existem diversas técnicas de detecção de conteúdo em imagens.

Os tópicos de interesse serão explicados neste capítulo que está estruturado da seguinte forma: nas Seções 2.2, 3.1, 3.2 e 3.3 são descritos um resumo da base teórica e três sistemas computacionais de atenção visual; na Seção 3.4 são apresentados as definições de características em imagens e dois dos principais métodos existentes para extração das mesmas o SIFT 3.4.1 e o SURF 3.4.2; na Seção 3.5 é descrito o modelo para classificação de imagens *Bag-of-Features* e os algoritmos de aprendizado de máquina usados, k-Means 3.5.1 e SVM 3.5.3; nas Seções 3.6, 3.7 e 3.8 são explicados os algoritmos que podem ser usados para rastreamento de objetos Meanshift, Camshift e filtro de Kalman.

3.1 Modelo de atenção visual *bottom-up* - (iNVT)

O iNVT (*iLab Neuromorphic Vision C++ Toolkit*) é uma biblioteca de classes para desenvolvimento de modelos neuromórficos de atenção visual. Os modelos neuromórficos são inspirados nas funções biológicas do cérebro humano. De acordo com Itti e Koch (2001) existem cinco pontos importantes que devem ser considerados sobre modelos computacionais para atenção visual.

- A percepção de qualquer estímulo de entrada é totalmente dependente do contexto ao seu redor;

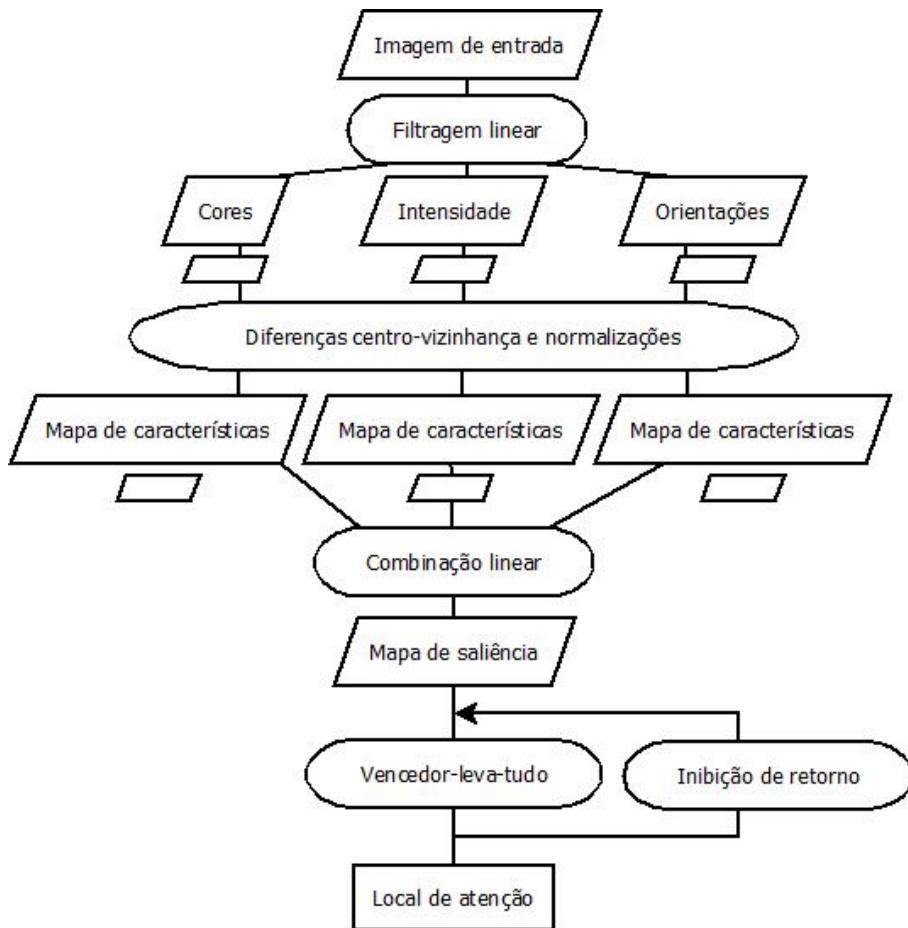


Figura 1 – Estrutura do modelo de saliência iNVT, adaptada de ([ITTI; KOCH, 2001](#)).

- Uma estratégia bastante eficiente de controle *bottom-up* dá-se na utilização de um único mapa de saliência, que codifica topograficamente os estímulos;
- Para o desenvolvimento da atenção, o processo de inibição é muito importante para evitar que uma região focada anteriormente seja novamente focada;
- A interação rígida entre atenção e movimentos oculares insere desafios computacionais ao sistema de coordenadas utilizado para controlar a atenção;
- A escolha dos locais de atenção é fortemente determinado pela compreensão de cena e o reconhecimento de objetos.

O mapa de saliência é amplamente utilizado por modelos *bottom-up* e é formado pela composição de vários mapas com características visuais primitivas da imagem como, por exemplo, cor, intensidade e orientação. Qualquer que seja a dimensão característica, esta composição produz uma saliência independente. A cena visual com suas diversas regiões que disputam essa medida o fazem por muitas escalas especiais. As mais salientes são oriundas da região vencedora. A Figura 1 ilustra o modelo proposto por [Itti e Koch \(2001\)](#) para compreensão de como o mapa de saliência é gerado.

Extração de Características Visuais Primitivas

A decomposição de uma imagem em um conjunto de canais distintos é realizada através do modelo proposto por Koch e Ullman (1987). Nesse modelo é estabelecida a extração de características visuais primitivas ou características de baixo nível. Essas características são extraídas da imagem original em várias escalas espaciais, através de filtragens lineares. Qualquer modelo de atenção *bottom-up* tem como primeiro estágio de processamento a computação de características visuais primitivas. Há inspiração biológica considerando que a análise das características da forma pré-atenção são paralelas a todo o campo visual (ITTI; KOCH, 2001).

De um modo geral, a alimentação inicial do modelo por uma imagem estática dá-se por extração das seguintes características visuais: cor, intensidade e orientação. r , g e b são os canais vermelho, verde e azul da imagem de entrada. A imagem de intensidades é obtida por $I = (r + g + b)/3$. I será aplicado para gerar a pirâmide Gaussiana $I\sigma$, discutida a seguir. Para fins de normalização, os canais r , g e b são normalizados a partir de I . O objetivo desta normalização é inibir regiões que apresentem baixos valores de luminosidade (não salientes). Neste caso, r , g e b são normalizados somente quando I for maior do que 1/10 de seu valor máximo sobre toda imagem. Em seguida, quatro canais de cores são criados: $R = r - (g + b)/2$ para o vermelho, $G = g - (r + b)/2$ para o verde, $B = b - (r + g)/2$ para o azul e $Y = (r + g)/2 - |r - g|/2 - b$ para o amarelo. O valor zero é atribuído para sinais negativos (ITTI; KOCH; NIEBUR, 1998). A obtenção de imagens sem ruídos e detalhes indesejados com realce para as características importantes dá-se pela geração de uma pirâmide Gaussiana que é composta de nove níveis para cada canal, sendo: $I\sigma$, $R\sigma$, $G\sigma$, $B\sigma$ e $Y\sigma$, onde $\sigma \in [0..8]$. A pirâmide Gaussiana é gerada de acordo com o algoritmo proposto em (BURT; ADELSON, 1983), descrito na seção seguinte.

Pirâmide Gaussiana

Operações progressivas de filtragem passa-baixas e subamostragem são estabelecidas pela pirâmide Gaussiana de Burt e Adelson (1983). Um filtro Gaussiano com dimensão de 5x5 pixels foi usado no modelo de saliência de Itti, Koch e Niebur (1998).

A representação em forma de pirâmide é utilizada com o objetivo de destacar características salientes e inibir demais regiões da imagem. Para a sua geração, um filtro Gaussiano é aplicado a cada nível da pirâmide previamente à geração do nível seguinte. Considerando uma imagem de entrada representada inicialmente por uma matriz G_0 , o nível zero da pirâmide, composta por linhas e colunas (x, y) , onde cada coordenada (*pixel*) representa um valor correspondente da imagem relacionada a cada característica. O nível 1 contém a imagem G_1 , que é a redução ou versão convolvida de G_0 . De forma similar aos canais considerados, uma pirâmide Gaussiana G_σ pode ser definida recursivamente como segue:

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G(x, y), \quad \text{para } \sigma = 0 \quad (3.1)$$

$$G_\sigma(x, y) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m+2, n+2) G_{\sigma-1}(2x+m, 2y+n), \text{ para } 0 < \sigma \leq 8, \quad (3.2)$$

onde $w(m, n)$ são os pesos gerados a partir de uma função Gaussiana, empregados para gerar os níveis da pirâmide para todos os canais.

Pirâmide direcional

As informações sobre as orientações locais do modelo de Itti, Koch e Niebur (1998) são importantes no desenvolvimento da atenção visual.

De acordo com Greenspan *et al.* (1994), a extração destas características pode ser obtida através da aplicação de filtros direcionais sobre a imagem. O perfil de sensibilidade do campo receptivo dos neurônios é aproximado ao de orientação seletiva presente no córtex visual primário.

Os mapas de orientações $O_\sigma(\theta)$ são criados através da convolução do mapa de intensidades I_σ , com filtros direcionais de Gabor para quatro orientações $\theta \in 0^\circ, 45^\circ, 90^\circ, 135^\circ$. Os filtros são usados para identificar as barras ou bordas em uma determinada direção.

Diferenças centro-vizinhança

A diferença centro-vizinhança é implementada como a diferença entre escalas, ou seja, o centro é um pixel da imagem na escala $c \in 2, 3, 4$ e a vizinhança é o pixel correspondente em outra escala $s = c + \delta$, com $\delta \in 3, 4$ da representação piramidal.

O contraste de intensidades é usado para construir o primeiro conjunto de mapas, definido como segue:

$$I(c, s) = |I(c) \ominus I(s)|. \quad (3.3)$$

O segundo conjunto de mapas é construído a partir dos canais de cores, definidos como:

$$RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \quad (3.4)$$

$$BY(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|, \quad (3.5)$$

O terceiro conjunto de mapas é gerado a partir de informações de orientação local, de acordo com as seguintes equações:

$$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|, \quad (3.6)$$

onde $\theta \in 0^\circ, 45^\circ, 90^\circ, 135^\circ$.

Saliência

Uma miríade de modelos de atenção bottom-up segue a hipótese de Koch e Ullman (1987). Vários mapas de características compõem um mapa de saliência. O mapa de saliência é um mapa escalar bidimensional de atividade representado topograficamente pela conspicuidade ou saliência visual (ITTI; KOCH, 2000). Para a construção de um único mapa de saliência, os mapas de características são individualmente somados nas diversas escalas. Geram-se então três mapas de conspicuidade: I para intensidade, C para cor e O para orientação. Itti, Koch e Niebur (1998) propuseram a normalização dos mapas de características antes de formar o mapa de saliência para aumentar o contraste das regiões com níveis mais salientes, fazendo com que regiões salientes onde há pouco contraste sejam inibidas.

Uma vez realizado o processo de normalização, os mapas de características são combinados em três mapas de conspicuidade.

De acordo com Itti, Koch e Niebur (1998), a motivação para a criação dos três canais é a hipótese de que as características similares competem pela saliência, enquanto modalidades diferentes contribuem independentemente para o mapa de saliência.

Por fim, os três mapas de conspicuidade são normalizados e somados. O resultado é uma entrada final para o mapa de saliência S , como segue:

$$S = \frac{1}{3}(N(I) + N(C) + N(O)) \quad (3.7)$$

Seleção da atenção e inibição de retorno

Para o desenvolvimento da seleção visual entre as regiões mais salientes do mapa de saliência, Itti, Koch e Niebur (1998) empregam uma rede neural composta por neurônios do tipo Integra e Dispara, para representar o mapa de saliência S e fazer com que o estímulo de cada neurônio seja o valor de saliência dos pontos no mapa de saliência. A rede de neurônios Integra e Dispara alimenta uma rede neural do tipo WTA (*Winner-Takes-All*) (KOCH; ULLMAN, 1987) (TSOTSOS *et al.*, 1995). A rede é usada para localizar a região mais saliente no mapa de saliências, indicada pelo neurônio vencedor (ITTI; KOCH, 2000). Os neurônios da rede Integra e Dispara são empregados, neste caso, como integradores dos valores de S . Na rede, todos os neurônios recebem ativação de forma independente, até que o neurônio vencedor alcance o limite e dispare. Há, então, desencadeamento simultâneo de três mecanismos: primeiro, o foco da atenção é direcionado para a localização do neurônio vencedor; segundo, o inibidor global é acionado e todos os demais neurônios são inibidos; e por último, a região sob o foco da atenção é temporariamente inibida na rede de neurônios, permitindo que a próxima região saliente seja destacada. Assim o foco da atenção não é mais redirecionado para a região anterior que é caracterizada por um mecanismo de inibição de retorno (ITTI; KOCH; NIEBUR, 1998).

3.2 Detecção visual de objetos com um sistema de atenção computacional (VOCUS2)

Através do sistema de saliências VOCUS2 ([FRINTROP; WERNER; GARCIA, 2015](#)) calcula-se um mapa de saliência a partir de uma entrada de imagem ou vídeo. Sua estrutura é baseada na tradicional abordagem proposta por ([ITTI; KOCH; NIEBUR, 1998](#)) na qual são calculadas características paralelamente e o contraste no centro-vizinhança é calculado por diferenças de Gaussianas. Este algoritmo tem desempenho que está no estado da arte das tarefas de segmentação de objetos salientes. O sistema é rápido, tem uma estrutura simples, coerente e produz mapas de saliência bem detalhados.

O VOCUS2 ([FRINTROP; WERNER; GARCIA, 2015](#)) é um sistema de obtenção de saliências cujos fundamentos são os mesmos encontrados no modelo de Itti-Koch ([ITTI; KOCH; NIEBUR, 1998](#)): canais de características são computados em paralelo; as pirâmides computadas permitem processamento multi-escala e contrastes são obtidos através da Diferença de Gaussianas.

Uma das estruturas mais importantes é a de escala de espaço (que usa uma pirâmide gêmea), e sua taxa de centro-vizinhança, que se mostrou o parâmetro pivotal de sistemas de saliências. Como na abordagem de ([ITTI; KOCH; NIEBUR, 1998](#)), o sistema resultante tem uma estrutura baseada em conceitos da percepção humana. Mas ao invés de produzir mapas de saliência baseados em segmentos, este sistema gera mapas baseado em *pixels*. Entretanto, para algumas tarefas, mapas de saliência baseados segmentos são melhores quando da determinação dos limites precisos do objeto. Para obter tais limites, o VOCUS2 integra um *framework* de geração de propostas de objeto.

O sistema VOCUS2 funciona como mostrado na Figura 2. A imagem de entrada é convertida em um espaço de cores-opONENTES com canais para a intensidade, verde-vermelho e azul-amarelo. Para cada canal, são computadas duas pirâmides de imagem (uma para o centro e uma para a vizinhança) nas quais o contraste entre o centro-vizinhança é calculado. A Tabela 1 indica as diferenças entre o sistema VOCUS2 e o iNVT.

Tabela 1 – Principais características entre os sistemas iNVT e VOCUS2. ([FRINTROP; WERNER; GARCIA, 2015](#)).

	iNVT	VOCUS2
Características	intensidade (I), cor (C), orientação (O)	intensidade (I), cor (C)
Estrutura piramidal	uma pirâmide	pirâmides gêmeas (diferença principal)
Fusão de características	uma escala por camada sub-amostragem ponderação pela unicidade prioridade para canal de cor e depois intensidade	múltiplas escalas por camada interpolação média aritmética igualdade entre 3 canais

A maior parte do processamento ocorre pela extração de características de intensidade e de cor. O canal da característica orientação não é utilizado, pois atribui altos valores de saliência aos cantos do objeto e torna o método menos eficiente para segmentação. Para calcular as

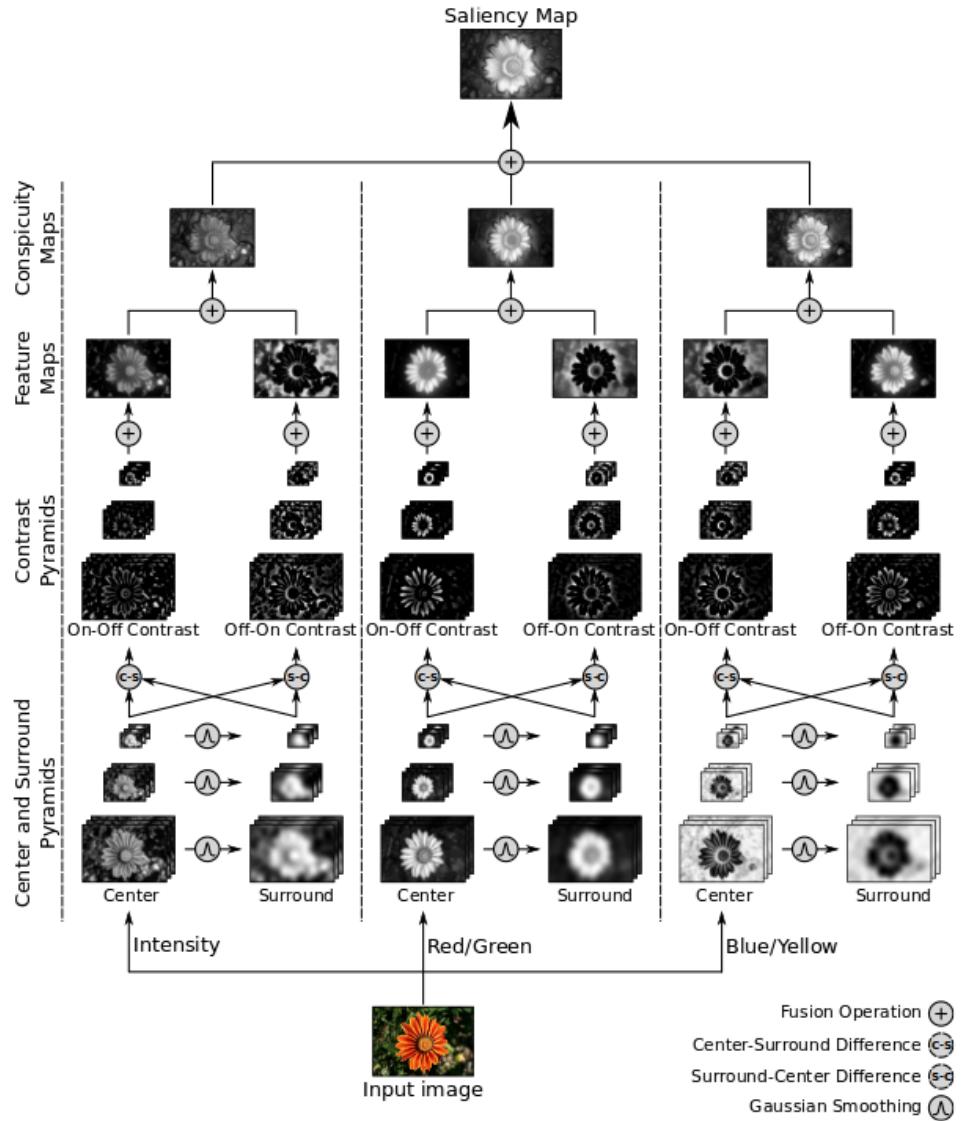


Figura 2 – Processamento dos canais de características do sistema VOCUS2 (FRINTROP; WERNER; GARCIA, 2015).

características cores, é utilizado o espaço de cores RGB. Os canais de característica intensidade (I) e características cores-opONENTES vermelho-verde (RG) e azul-amarelo (BY) são obtidos pelas Equações 3.8.

$$\begin{aligned}
 I &= \left(\frac{R + G + B}{3} \right) \\
 RG &= R - G \\
 BY &= B - \frac{R + G}{2}
 \end{aligned} \tag{3.8}$$

Espaços de escala com pirâmides gêmeas

Duas diferenças entre o VOCUS2 e o iNVT no espaço de escala são que ao invés de utilizar uma pirâmide Gaussiana simples, é usado um espaço com escalas e oitavas, e o uso

de pirâmides gêmeas ao invés da subtração de camadas da pirâmide. Estas pirâmides gêmeas possuem uma pirâmide central $C = (C_0, \dots, C_k)$ e uma pirâmide de vizinhança $S = (S_0, \dots, S_k)$. Cada imagem central C_i tem uma imagem de vizinhança correspondente S_i , que é a imagem C_i suavizada por um fator sigma σ_x que corresponde à taxa de centro-vizinhança desejada. O fator sigma pode ser obtido através da Equação 3.9. O valor de σ_c é o utilizado para obter a imagem central C_i e σ_s é o fator de suavização efetivo para a imagem de vizinhança S_i .

$$\sigma_x = \sqrt{\sigma_s^2 - \sigma_c^2} \quad (3.9)$$

A vantagem desta abordagem é não estar limitada pelas taxas centro-vizinhança dadas pela pirâmide, podendo variar o valor de forma flexível (FRINTROP; WERNER; GARCIA, 2015).

Apesar de uma granularidade mais fina poder ser obtida utilizando um espaço de escala com várias escalas por camada, os mapas de escala usados para subtração ainda tem que ser escolhidos a partir do conjunto disponível da pirâmide e taxas centro-vizinhança não podem ser escolhidas arbitrariamente.

Contraste centro-vizinhança

Os contrastes de cor e de intensidade podem ser computados subtraindo o mapa do centro e o mapa de vizinhança. Para distinguir objetos com alto brilho em um ambiente escuro de objetos escuros em ambientes com alto brilho, o cálculo é separado em contrastes *off-on* e *on-off*, correspondendo às células da visão humana que respondem a somente um destes contrastes.

Isto oferece dois mapas de contraste para cada camada i das pirâmides: $X_i^f = C_i^f - S_i^f$ (para contrastes *on-off*) e $Y_i^f = S_i^f - C_i^f$ (para contrastes *off-on*), com $f \in I, RG, BY$. Em ambos os contrastes, valores abaixo de zero são truncados em zero. As pirâmides resultantes são chamadas de pirâmides de contraste.

Fusão dos canais de características

Para realizar a fusão dos canais de características, o algoritmo soma as pirâmides entre escalas para obter os mapas de características (F_1^f e F_2^f), como visto na Equação 3.10.

$$\begin{aligned} F_1^f &= \bigoplus_i X_i, \text{ with } i \in \{1, \dots, k\} \\ F_2^f &= \bigoplus_i Y_i, \text{ with } i \in \{1, \dots, k\} \end{aligned} \quad (3.10)$$

Diferentemente do iNVT, esta soma entre escalas (\oplus) interpola à escala mais fina, não a mais granulosa, antes de somar os mapas. Os dois mapas de características de cada canal são fundidos em mapas de conspicuidade (Equação 3.11) e os mesmos são combinados em um único

mapa de saliência S (Equação 3.12).

$$C^f = f(F_1^f, F_2^f), \text{ com } f \in \{I, RG, BY\}, \quad (3.11)$$

$$S = g(C^I, C^{RG}, C^{BY}), \quad (3.12)$$

3.3 Modelo bottom-up e top-down

Muitos trabalhos baseiam-se na busca por características a partir da cena. Benicasa (2013) considera o sistema visual dos primatas que seleciona as características baseado em objetos. Esse mecanismo, chamado *top-down*, modula o sistema através do enviesamento da atenção de acordo não apenas com características primitivas (cor, orientação, intensidade, etc.), mas também com informações de memória, objetos pré-conhecidos e importantes para a realização da tarefa. Para correlacionar características dos objetos, neurônios específicos (chamados osciladores) representam um objeto, caso estejam sincronizados. Assim cada conjunto de osciladores representa um dos objetos da cena, ficando em diferentes sincronias, tornando possível agrupá-los por similaridade, proximidade, etc.

A partir dessa correlação, foi proposta a utilização do modelo de rede neural *Locally Excitatory Globally Inhibitory Oscillator Network* (LEGION), o que possibilita a segmentação dos diferentes objetos da cena seguindo o modelo *top-down*, guiando a atenção, além do enviesamento *bottom-up* de busca por características primitivas, pelos objetos de interesse (figuras geométricas planas, no caso de (BENICASA, 2013)). Todos os objetos competem pela atenção, fazendo uso de redes Integra & Dispara (I&D) e da inibição de retorno, para que o mesmo objeto não seja classificado mais de uma vez. Após a segmentação, o objeto é classificado por uma rede Perceptron multicamadas (MLP) e, quando um objeto é reconhecido, um peso maior é aplicado aos neurônios que o representam, aumentando a possibilidade de ser o neurônio vencedor ao passar pelo mapa auto-organizável (SOM), que define o foco de atenção levando em consideração as características primitivas e os objetos reconhecidos. Esse modelo é particularmente interessante por conseguir representar objetos em cenas reais e foi utilizado para reconhecer figuras geométricas em sala de aula e também placas de sinalização de trânsito. A Figura 3 apresenta uma visão geral desse sistema.

3.4 Extração de características de imagens

Uma característica é uma propriedade que pode representar uma imagem ou parte dela. Pode ser um pixel, um círculo, uma linha, uma região com textura média dos níveis de cinza, etc. Apesar de não existir uma definição formal, características podem ser definidas como partes detectáveis da imagem com algum significado.

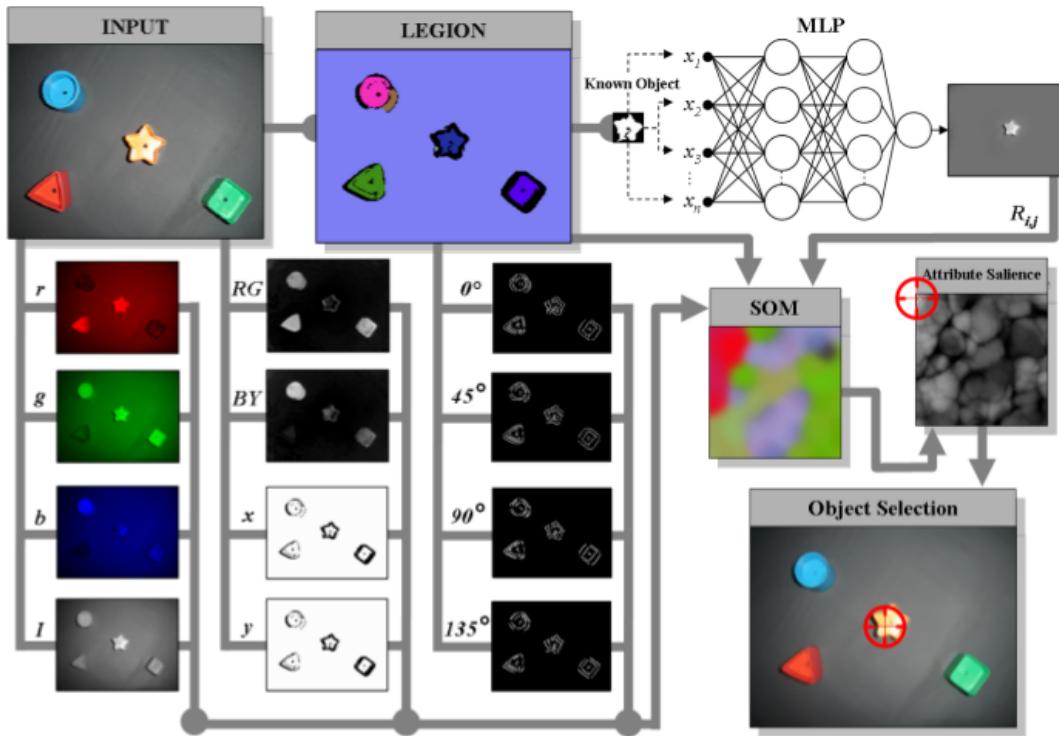


Figura 3 – Sistema de visão proposto em Benicasa (2013).

As características podem ser extraídas por diversos algoritmos, por exemplo, o SIFT (Scale-Invariant Feature Transform) (LOWE, 2004), o SURF (Speeded-Up Robust Features) (BAY; TUYTELAARS; GOOL, 2006), o FAST (Features from Accelerated Segment Test) (ROSTEN; DRUMMOND, 2006) ou o ORB (Oriented FAST and Rotated BRIEF) (RUBLEE *et al.*, 2011).

3.4.1 SIFT

O SIFT é um algoritmo de visão computacional composto por duas partes distintas: o detector, usado para detectar os extremos da imagem e fazer a localização de pontos-chave, e o descritor, empregado na definição da orientação e descrição dos pontos-chave. O detector é baseado em cálculos de diferenças de Gaussianas e o descritor utiliza histogramas de gradientes orientados para descrever a vizinhança local dos pontos de interesse. O SIFT detecta a orientação do gradiente dominante em sua posição e registra o histograma do gradiente local resultante em relação a essa orientação. Assim, as características do SIFT são relativamente bem comportadas em pequenas transformações.

O algoritmo SIFT para extração de características de imagens transforma uma imagem em uma "grande coleção de vetores de características locais" (LOWE, 2004). Cada um desses vetores de características é invariante à escala, rotação e translação da imagem. Em uma imagem são detectados e construídos vários pontos-chave e seus descritores. Esse conjunto de descritores pode ser usado para fazer correspondências entre imagens como exemplificado na Figura 4.



Figura 4 – Exemplo de correspondência de imagens (LOWE, 2004).

Detector

O detector busca pontos invariantes à mudança de escala, possibilitando a detecção de pontos-chave em vários níveis de aproximação do objeto de interesse. Para isso, o detector procura por características estáveis em relação à escala aplicando uma função Gaussiana. Assim, o espaço escalar de uma imagem é dado pela operação de convolução entre a Gaussiana de escala-variável com uma imagem $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (3.13)$$

sendo x e y as coordenadas do centro do quadro, σ a escala aplicada na imagem e G a função Gaussiana definida como:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (3.14)$$

A detecção estável de pontos-chave é feita usando o espaço escalar extremo na função de Diferença de Gaussiana (DoG) que pode ser calculada pela diferença de duas escalas próximas (Figura 5).

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (3.15)$$

Em que k é o fator constante de separação entre o espaço de escalas.

Os extremos são dados pelos máximos ou mínimos locais para cada DoG, que podem ser obtidos comparando a intensidade de cada ponto com a intensidade de seus vizinhos. A Figura 5 representa o modelo de detecção de pontos-chave em uma imagem. Para cada escala, a imagem é convolucionada com gaussianas para produzir um conjunto de imagens escaladas. As gaussianas de imagens adjacentes são subtraídas para produzir a diferença de gaussianas.

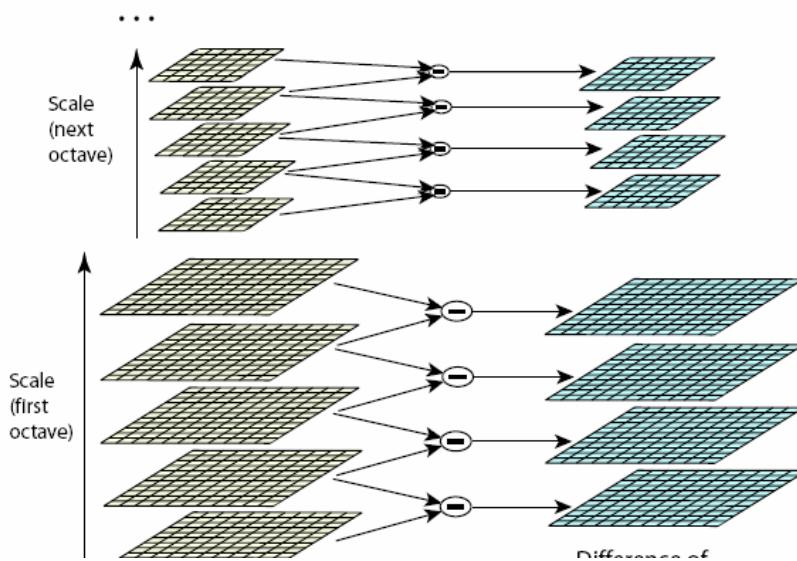


Figura 5 – Representação gráfica da função de diferença de Gaussiana (LOWE, 2004).

Descriptor

Uma orientação é atribuída para cada ponto-chave para construir os descritores. Monta-se um histograma das orientações para uma região vizinha ao redor do ponto-chave. Cada ponto na vizinhança do ponto-chave é adicionado ao histograma com um determinado peso. Os picos no histograma de orientações correspondem às direções dominantes dos gradientes locais e são utilizados para definir a orientação do ponto-chave. Assim, cada ponto-chave tem quatro dimensões: sua posição x, y, magnitude e orientação. O descritor do ponto-chave é criado computando-se as magnitudes e orientações dos gradientes amostrados (Figura 6) ao redor da localização do ponto-chave.

Os vetores resultantes são chamados de características SIFT e, em conjunto com a técnica K-NN (k-nearest neighbours) (FIX; HODGES, 1951), podem ser usados para identificar possíveis objetos em uma imagem. Quando existe um conjunto de características coincidentes em duas imagens, é altamente provável que as características se refiram a um mesmo objeto.

Devido ao grande número de características SIFT em uma imagem de um objeto, a técnica é capaz de reconhecer objetos mesmo que exista um considerável nível de oclusão do objeto na imagem.

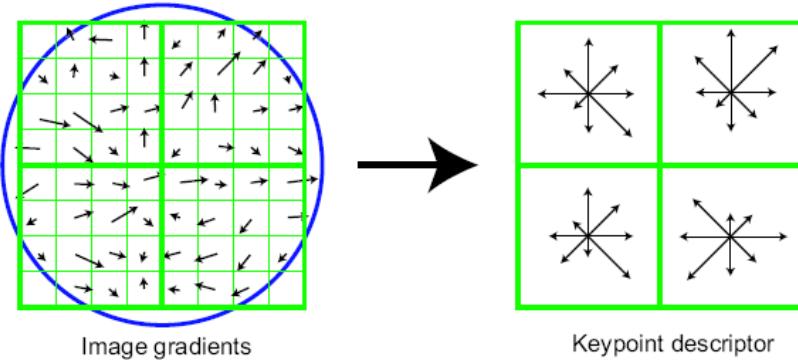


Figura 6 – Construção do descriptor do ponto-chave (LOWE, 2004).

3.4.2 SURF

O SURF (*Speeded-Up Robust Features*) ((BAY; TUYTELAARS; GOOL, 2006)) é um detector e um descriptor de pontos de interesse invariante às operações de escala e rotação. O detector SURF utiliza como base a matriz Hessiana, mas, no entanto, ao invés de utilizar uma medida para selecionar a localização e a escala, tal como é feito pelo detector Hessian-Laplace, (BAY; TUYTELAARS; GOOL, 2006) utiliza o determinante do Hessiano para ambos. Já o descriptor SURF representa a distribuição das respostas da *wavelet Haar* na vizinhança do ponto de interesse.

Detector

Para a tarefa de detectar um objeto em uma imagem, ao invés de procurar o objeto como um todo, apenas os pontos de interesse do objeto são utilizados para identificação. Este tipo de abordagem é escolhida por várias razões, das quais as principais são: o custo computacional de procurar em dados com grande dimensionalidade, como armazenadas em imagens; e o alto nível de redundância incorporada, porque *pixels* não se movem de forma independente e possuem um elevado grau de correlação. Existem vários métodos para definir e detectar pontos de interesse, que vão desde os que consideram cantos como pontos de interesse para os que consideram um *blob* de cada vez.

Detector rápido de Hessian

O detector de Hesse rápido detecta características blob-like. Ele baseia-se na matriz de Hesse, que a escala é definida como se segue:

$$\mathcal{H}(x, y, \sigma) = \begin{bmatrix} \frac{\delta^2}{\delta x^2} G(\sigma) * I(x, y) & \frac{\delta}{\delta x} \frac{\delta}{\delta y} G(\sigma) * I(x, y) \\ \frac{\delta}{\delta x} \frac{\delta}{\delta y} G(\sigma) * I(x, y) & \frac{\delta^2}{\delta y^2} G(\sigma) * I(x, y) \end{bmatrix} \quad (3.16)$$

Sabe-se que, no caso contínuo, gaussianas são adequadas para a análise do espaço de escala (KOENDERINK, 1984; LINDEBERG, 1990). No entanto, todos os detectores baseados em Hessian possuem um ponto fraco: quando se trabalha com imagens distintas, a gaussiana precisa ser também discretizada, como consequência, há uma perda de repetibilidade sob as rotações da imagem ao redor dos múltiplos ímpares de $\pi/4$. No entanto, a matriz de Hessian é escolhida porque a taxa de repetição é ainda muito alta em qualquer ângulo de rotação (BAY *et al.*, 2008). Como os filtros de gaussiana discretizados já não são ideal e as circunvoluções ainda são custosos, eles são aproximados por filtros de caixa (Figura 7). Desta forma, quando usado em conjunto com imagens integrais, os cálculos podem ser realizados em tempo constante. Embora esta seja uma aproximação ainda mais forte, o desempenho é comparável ou até melhor do que com o discreto e a gaussiana cortada (BAY *et al.*, 2008).

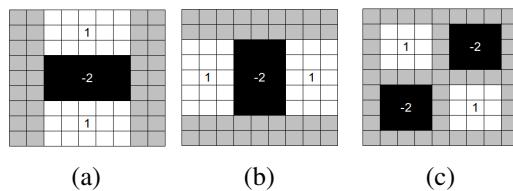


Figura 7 – Filtros de caixa que aproximam derivadas gaussianas de segunda ordem.

Para determinar quão "forte" é o ponto atual para classificá-lo como de interesse ou não é necessário calcular a matriz aproximada de Hessian utilizando o filtro de caixa.

Invariância de escala

Os cálculos anteriores são realizadas em diferentes escalas porque os pontos de interesse podem ser comparados entre as imagens, onde elas são vistas em diferentes escalas. O espaço de escala é implementada como uma imagem piramidal. Sem filtros de caixa, geralmente, a imagem piramidal é construída por sucessivos processos de suavização da imagem como uma subamostra da gaussiana e, em seguida, a fim de atingir um nível mais elevado da pirâmide. Com os filtros de caixa e as imagens integrais, não há a necessidade de filtrar a imagem iterativamente e subamostrá-lo. Em vez disso, é o filtro que é o sobredimensionado e aplicada exatamente a mesma velocidade sobre a imagem original. Os níveis mais elevados da pirâmide são atingidos através da aplicação de filtros gradualmente maiores. O espaço escala é dividido em oitavas. Cada oitava é uma série de mapas de resposta do filtro, obtido por convolução da mesma imagem de entrada, com um filtro de tamanho crescente. Devido à natureza discreta dos filtros de caixa, o seu tamanho deve ser aumentado para um mínimo de dois *pixels*, a fim de manter a presença do *pixel* central.

Classificação do ponto de interesse

Para classificar um ponto de interesse de uma supressão não-máxima em um $3 \times 3 \times 3$ é aplicada em escala e imagem do espaço (Figura 8). Cada amostra é comparada com seus 8 vizinhos na imagem atual e os 18 vizinhos de escala superior e inferior no espaço de escalas.

Se ele tem a maior pontuação (determinante da matriz Hessian) de seus vizinhos, então ele é considerado um ponto de interesse, caso contrário, ele é descartado.

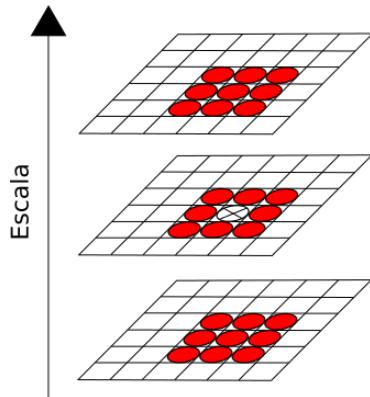


Figura 8 – Representação gráfica da vizinhança $3 \times 3 \times 3$.

Uma vez que é classificado como um ponto de interesse, a localização é refinada para a precisão de *subpixel*, ajustando uma parábola para o ponto de amostragem e seus vizinhos imediatos (BROWN; LOWE, 2002).

Descriptor

Após a detecção dos pontos de interesse, é necessário atribuir uma descrição de cada um, a fim de identificar e distingui-los um do outro e combiná-los entre as imagens. Idealmente, um bom descriptor irá fornecer uma descrição para cada ponto que é único, mas idêntico para todos os pontos de vista possíveis do mesmo ponto.

Este caso ideal é difícil de se alcançar, se não impossível. No entanto, os descriptores locais utilizam as informações sobre a textura dos pontos de interesse vizinhos para distingui-los, tanto quanto possível um do outro. Especificamente, o descriptor SURF, tende a funcionar muito bem nesta tarefa. O descriptor SURF é construído em três etapas, que agora são descritas.

Orientação do ponto do atributo de interesse

Para atingir invariância de rotação, o primeiro passo consiste em atribuir uma orientação para o ponto, de modo que, quando é visto a partir de outra perspectiva, pode ser adequada. Para esta finalidade, e para tomar vantagem da utilização de imagens integrais, filtros *wavelet* Haar são utilizados (Figura 9). Sendo s a escala em que foi detectado o ponto de interesse, filtros de $4s$ de tamanho são usados e respostas *wavelet* em direções x e y com um passo de amostragem de s , são calculados em torno de um vizinho circular de raio $6s$.

Vetor descriptor

Embora haja algumas variações, o descriptor SURF padrão consiste de um vetor com 64 entradas. Para construir este vetor, o primeiro passo é o de construir um quadrado $20s$ tamanho da região de f , centrado no ponto de interesse e com a orientação selecionada na etapa anterior. Esta região é dividida em 4×4 sub-regiões quadradas. Usando o filtro de Haar com tamanho

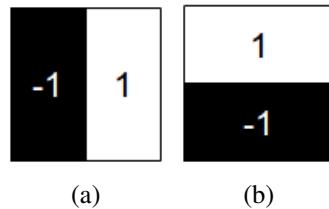


Figura 9 – Filtro wavelet Haar.

$2S$, as respostas de filtro em 5×5 pontos de amostragem igualmente espaçadas são calculados na direção x e y . Note-se que estas instruções são definidas em relação à orientação da região quadrada. Mas, em vez de rotacionar a imagem em si, as respostas do filtro são calculadas na imagem sem rotação e então interpoladas. Após ponderação das respostas de filtro usando uma gaussiana com $\sigma = 3.3$ centrada no ponto de interesse, quatro somas em cada sub-região são calculadas: as somas de dx e dy e, para ter informações sobre a polaridade das mudanças de intensidade (Figura 10), o somas de $|dx|$ e $|dy|$.

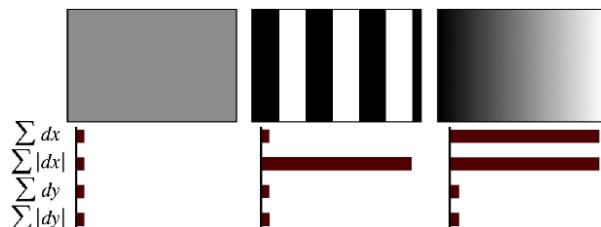


Figura 10 – Entradas do descritor de uma sub-região (BAY; TUYTELAARS; GOOL, 2006).

Assim como o SIFT, o descritor SURF também consiste em transformar uma região em torno da característica em um vetor de componentes de frequência (BOTTERILL; MILLS; GREEN, 2008). De acordo com Bay, Tuytelaars e Gool (2006), embora o SURF possa ser semelhante em conceito com o SIFT, o SURF é menos sensível ao ruído e supera o SIFT. Isso se dá por conta da integração global de informações obtidas a partir de gradiente da sub-região, em vez de gradientes individuais, como no caso do SIFT. Outra vantagem em relação ao SIFT é seu reduzido tempo de processamento para detecção e descrição dos pontos de interesse em imagens. Isso se dá pelo fato do descritor SURF utilizar apenas 64 dimensões, reduzindo assim o esforço computacional dos operadores diferenciais no espaço de escala.

Vetor descritor

Um ponto de interesse em uma imagem, é considerado igual a outro ponto de interesse em outra imagem se eles estão próximos o suficiente no sentido do vizinho mais próximo. O algoritmo mais utilizado para determinar o vizinho mais próximo é o kd-tree (MOORE, 1991).

3.5 Bag-of-features

Bag-of-Features (BoF) é uma abordagem popular para classificação visual de objetos cujo interesse é devido à seu poder e simplicidade. Sua origem deriva do modelo *bag-of-words* proposto por (SALTON; MCGILL, 1983). Esta abordagem é usada para diversas tarefas de visão computacional, tais como classificação de imagens, localização de robôs e reconhecimento de texturas. Os métodos que aplicam este modelo baseiam-se em coleções não ordenadas de descritores de imagens. Eles possuem a característica de descartar informações espaciais e são conceitualmente mais simples que os métodos alternativos. A ideia principal por trás desta abordagem é calcular características em uma imagem e, a partir de combinações com um conjunto de características, classificar a imagem. No modelo BoF, as características extraídas são agrupadas e as partições geradas são usadas para montar um dicionário de palavras visuais. Após quantizar as características usando o dicionário visual, as imagens são representadas pela frequência das palavras visuais.

Em um dicionário visual, cada imagem dentro do BoF representa um grupo encontrado pelo algoritmo de agrupamento. Se dada uma imagem, as características quantizadas forem similares às características encontradas no dicionário, pode se classificar a imagem usando o rótulo da imagem que está no dicionário.

Para formar o dicionário visual do modelo, é necessário um algoritmo agrupador para separar as características em conjuntos semelhantes que possam ser comparadas futuramente por um classificador. Os algoritmos mais utilizados para o agrupamento de características é a técnica K-médias e o classificador SVM que serão descritos a seguir.

3.5.1 K-médias

K-médias (LLOYD, 1982) é um método que segue o paradigma de aprendizado não supervisionado para fazer extração de conhecimento sem utilizar informações sobre as classes dos exemplos. Assim busca organizar um conjunto de objetos em grupos de acordo com alguma medida de similaridade ou dissimilaridade. O k-médias é um algoritmo agrupador que dada uma função de dissimilaridade retorna uma partição do conjunto de objetos. É o algoritmo mais conhecido para agrupamento de dados. Ele busca particionar o conjunto de objetos em k grupos, sendo cada objeto associado ao grupo mais próximo.

Seu funcionamento é simples, inicialmente seleciona-se k pontos aleatórios sobre o espaço, esses pontos são chamados de centroides. Para cada objeto computa-se o centroide mais próximo e o rotula como pertencente ao centroide. Em seguida recalcula-se a posição dos centroides com base na posição de seus objetos associados. Para recalcular a posição dos centroides, considera-se a distância média de seus objetos relacionados. Assim, a cada passo do algoritmo os centroides são movidos em direção a seus objetos associados e o algoritmo é interrompido quando não houver mais variações nos seus centroides. No término da execução,

se tem as coordenadas dos centroides que partitionam o espaço.

3.5.2 Detecção de objetos

O primeiro passo do modelo é extrair características de imagens do objeto que se deseja detectar. Cada descritor de característica é composto por um vetor e representa uma palavra que será usada para montar o dicionário visual. Em seguida, a matriz de descritores é submetida ao método k-médias, com número de grupos variando de 1 até o número de características. O próximo passo do algoritmo é gerar os histogramas das imagens de treinamento e validação. Para cada característica, localiza-se o centroide mais próximo e soma-se a quantidade de ocorrências para formar o histograma. Após formar o histograma da imagem, este é normalizado no intervalo de grupos. Depois de construídos os histogramas de todas as imagens, é construído um histograma médio para cada classe. Com o histograma de cada classe construído, pode-se fazer a classificação de novas imagens comparando o histograma de uma imagem desconhecida com os histogramas de classes conhecidas. O histograma que apresenta a menor dissimilaridade representa a classe da qual a imagem desconhecida será rotulada.

Qualquer algoritmo classificador pode ser usado para classificar as características. Os mais usados são o *Naive Bayes* e o *Support Vector Machines*.

3.5.3 Support Vector Machines - SVM

SVM é um classificador binário baseado na teoria da aprendizagem estatística ([CORTES; VAPNIK, 1995](#)). A principal vantagem deste algoritmo, quando comparado com outros algoritmos de aprendizagem tais como Redes Bayesianas, Discriminante Linear de Fisher e Redes Neurais Artificiais tradicionais, é a sua resolução através de otimização por Programação quadrática, onde o problema dos mínimos locais é ausente. Este classificador tem sido cada vez mais utilizado em aplicações como classificação de padrões, reconhecimento de imagens, seleção de genes, classificação de textos, entre outras.

A teoria a qual se baseia o SVM estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa capacidade de generalização. Tenta-se prever corretamente a classe a qual determinado dado de entrada pertence baseando-se nos seus próprios dados e nos dados do mesmo domínio em que o aprendizado ocorreu ([LORENA A. C. E CARVALHO, 2003](#)). Logo, diante de uma amostragem de duas classes distintas, a classificação SVM permite a obtenção de diversas funções de um separador das mesmas. Diante de diversas possibilidades para obtenção de uma função separadora de classes, deve-se avaliar qual é a função que faz a melhor distinção entre duas categorias diferentes de dados.

3.6 Meanshift

O *Meanshift* é um algoritmo que desloca de forma interativa um grupo de pontos para a média do grupo de pontos na sua vizinhança. Possui aplicações em agrupamento (no qual sua metodologia é muito similar), estimativa de densidade de probabilidade e rastreamento.

Seja um conjunto S de n pontos de dados x_i pertencente a $d - D$ no espaço Euclidiano X . $K(x)$ é a função de núcleo que indica o quanto x contribuiu para a estimativa da média. Então, a média da amostra m de x com núcleo K é dado por:

$$m(x) = \frac{\sum_{i=1}^n K(x - x_i)x_i}{\sum_{i=1}^n K(x - x_i)} \quad (3.17)$$

A diferença $m(x) - x$ é chamada *mean shift* (média de deslocamento). O algoritmo *Meanshift* move interativamente o ponto de dado para sua respectiva média. Em cada interação, $x \leftarrow m(x)$. O algoritmo para quando $m(x) = x$.

A sequência $x, m(x), m(m(x)), \dots$ é chamada de trajetória de x . Se as médias amostrais são computadas em vários pontos, em seguida, a cada iteração, a atualização é feita simultaneamente para todos esses pontos.

Kernel

Tipicamente, *kernel* K é uma função de $\|x\|^2$:

$$K(x) = k(\|x\|^2) \quad (3.18)$$

k é chamado de perfil de K e suas propriedades são: k é não negativo, k é não incremental: $k(x) \geq k(y)$ se $x < y$. k é uma função contínua definida por partes e

$$\int_0^\infty k(x)dx < \infty \quad (3.19)$$

Estimativa de densidade

Para um conjunto de n pontos de dados x_i contidos em um espaço $d - D$, a estimativa de densidade de *kernel* com *kernel* $K(x)$ (profile $k(x)$) e raio h é:

$$\tilde{f}_K(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - xi}{h}\right) \quad (3.20)$$

A qualidade do estimador de densidade de *kernel* é medida pelo erro quadrático médio entre a densidade real e a estimativa.

3.7 Camshift

O algoritmo de rastreamento por *Meanshift* utiliza de distribuição fixada de cores. Em algumas aplicações a distribuição de cores pode mudar, por exemplo, pela variação de profundidade. *Continuous Adaptive Mean Shift* (CAMSHIFT), é um algoritmo que trata dinamicamente dessa mudança, pois utiliza de uma janela de busca de tamanho dado e calcula a distribuição de cores nesta janela. A localização da janela de busca W dentro da imagem I é determinada da seguinte forma:

Calcule o momento (média) zero dentro de W

$$M_{00} = \sum_{(x,y) \in W} I(x,y). \quad (3.21)$$

Calcule o primeiro momento para x e y

$$M_{10} = \sum_{(x,y) \in W} xI(x,y), \quad M_{01} = \sum_{(x,y) \in W} yI(x,y). \quad (3.22)$$

A localização da janela de busca é configurada na

$$x_c = \frac{M_{10}}{M_{00}}, \quad y_c = \frac{M_{01}}{M_{00}}. \quad (3.23)$$

Algoritmo 1: Algoritmo CAMSHIFT

Input: Imagem

- 1 Seleciona posição inicial da janela de busca.
 - 2 Executa o rastreamento Meanshift com método revisado de atribuição de posição da janela de busca.
 - 3 Armazenar momento zero.
 - 4 Atribuir tamanho da janela de busca em função do momento zero.
 - 5 Repita as instruções 2 e 4 até a convergência.
-

3.8 Kalman

Os filtros de Kalman são utilizados para estimar o estado de um sistema linear no qual o estado é assumido ser uma distribuição Gausiana. Estes filtros são compostos por duas fases, predição e correção. A etapa de previsão usa o modelo de estado para prever o novo estado das variáveis:

$$\begin{aligned} \bar{X}^t &= \mathbf{D}X^{t-1} + W, \\ \bar{\Sigma}^t &= \mathbf{D}\Sigma^{t-1}\mathbf{D}^t + Q, \end{aligned}$$

no qual \bar{X}^t e $\bar{\Sigma}^t$ são os estados e a covariância de predições no tempo t . \mathbf{D} é o estado de transição da matriz que define a relação entre as variáveis de estado no tempo t e $t - 1$. \mathbf{Q} é covariância do ruído \mathbf{Q} . Da mesma forma, o passo de correção utilizado das observações Z^t para atualizar o estado do objeto:

$$\begin{aligned} K^t &= \bar{\Sigma}^t \mathbf{M}^t [\mathbf{M}^t \bar{\Sigma}^t \mathbf{M}^T + R^t]^{-1}, \\ X^t &= \bar{X}^t + K^t v, \\ v &= [Z^t - \mathbf{M}^t \bar{X}^t], \\ \Sigma^t &= \bar{\Sigma}^t - K^t \mathbf{M}^t \bar{\Sigma}^t, \end{aligned} \tag{3.24}$$

no qual, \mathbf{X} é chamado de inovação, \mathbf{M} é a matriz de medição K é o ganho de Kalman, que é a equação de Riccati 3.24 usada para propagação dos estados modelos. Nota-se que o estado atualizado, X^t ainda é uma distribuição Gaussiana. No caso de funções f^t e h^t serem não-linear, elas podem ser linearizadas por expansão de séries de Taylor obtendo um filtro de Kalman estendido (BAR-SHALOM, 1987). De forma similar ao filtro de Kalman, o filtro estendido de Kalman assume que o estado é distribuído de forma Gaussiana. O filtro de Kalman tem sido amplamente utilizada pela comunidade de visão computacional para o rastreamento. Broida e Chellappa (1986) utilizaram o filtro de Kalman para rastrear pontos em imagens com ruído. Em rastreamento de objetos com câmeras estéreo, Beymer e Konolige (1999) utilizaram o filtro de Kalman para predição da posição e velocidade do objeto em $x - z$ dimensões. Já Rosales e Sclaroff (1999) usaram o filtro de Kalman estendido para estimar trajetórias 3D de um objeto de movimento 2D.

3.9 Considerações finais

Neste capítulo, apresentou-se os principais fundamentos teóricos de rastreamento de objetos em imagens. Três abordagens para detecção de objetos foram apresentadas. A abordagem por atenção visual é recente mas possui resultados promissores. A estratégia *bag-of-features* descrita é amplamente reconhecida como um bom classificador de imagens.

Conhecendo-se os sistemas mais importantes para as tarefas de detecção de objetos em imagens, classificação de segmentos em imagens e rastreamento de objetos é possível descrever o modelo proposto que está descrita no capítulo a seguir.

CAPÍTULO
4

MODELO COMPUTACIONAL PROPOSTO PARA O RASTREAMENTO DE VEÍCULOS

Neste capítulo está descrito o sistema proposto para o cumprimento dos objetivos. No capítulo 3 vimos algumas técnicas de atenção visual, extração de características de imagens, classificação e rastreamento que são usadas para o rastreamento de veículos.

A Figura 11 exibe os processos e dados produzidos durante a operação do sistema e que serão descritos nas seções deste capítulo.

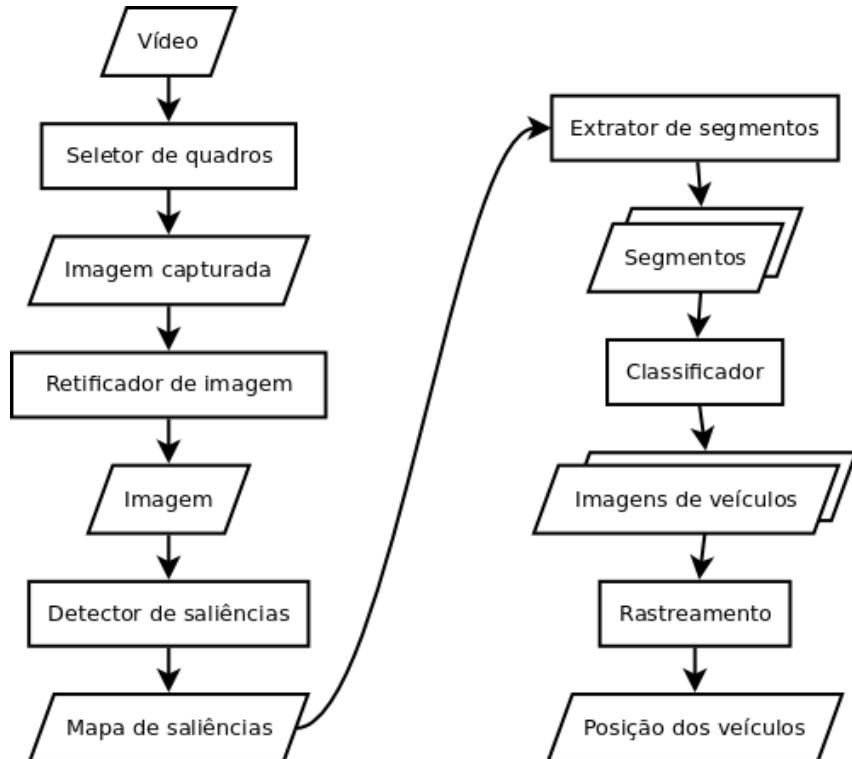


Figura 11 – Etapas para detecção, classificação e rastreamento de veículos em imagens.

O processo de detecção e classificação de segmentos possui um elevado custo computa-

cional. Para o sistema funcionar em tempo de execução não é possível processar todos os 60 quadros por segundo do vídeo gerado pela câmera de captura. Assim, nem todos os quadros precisam ser processados para não sobrecarregar o sistema, mas, para isso, é necessário escolher qual quadro possui mais informações relevantes dentro de um conjunto de quadros. Este quadro mais relevante é chamado quadro-chave. No sistema proposto, a seleção de quadros-chave é feita pela detecção de movimento na cena. O método utilizado para detectar o movimento na cena é calcular a quantidade de *pixels* alterados entre pares subsequentes de quadros. Se a quantidade ultrapassar um limite o novo quadro é selecionado como quadro-chave e é utilizado para detectar e classificar novos objetos. Caso contrário, o quadro é utilizado apenas para atualizar a posição dos objetos que já estão sob rastreio. Neste modelo, a taxa de captura de quadros-chave depende do vídeo sendo processado. Vídeos com muito movimento dos objetos extraem mais quadros-chave. Vídeos com pouco movimento podem utilizar um quadro-chave a cada determinado intervalo de tempo.

A maioria das câmeras está sujeita a distorções e para reduzir os efeitos negativos de tais distorções é necessário um processo de calibração e retificação. A calibração da câmera é o processo de corrigir e remover a distorção das lentes da câmera o que é feito apenas uma vez. Nesta etapa, alguns dos quadros de vídeo captados são usados para detectar um padrão de tabuleiro de xadrez e determinar a matriz de distorção e a matriz de câmara. Após o processo de calibração as imagens são retificadas para obter uma imagem com menos distorção.

Nesse processo é gerado um mapa de saliência de cada quadro-chave. O quadro-chave é inserido como parâmetro de entrada no detector de saliência e é processado para gerar o mapa de saliência. Dois detectores de saliência foram usados a fim de comparar as técnicas, os modelos Bottom-Up e Top-Down (3.3) e o VOCUS2 (3.2).

A função do módulo extrator de segmentos é separar os diversos segmentos presentes no mapa de saliência em imagens recortadas dos segmentos. Assim, a região ao redor do ponto de maior saliência do mapa é recortada para uma nova imagem e em seguida removida do mapa de saliência. Este processo ocorre até que todas as áreas salientes do mapa sejam removidas.

Cada segmento recortado do mapa de saliências é então classificado pelo modelo Bag-of-Features treinado com um vocabulário visual que contem imagens dos objetos de interesse. Para treinar o classificador do Bag-of-Features, são extraídas características locais SURF (3.4.2) de um banco de imagens criado durante os experimentos deste projeto. O algoritmo k-Médias é executado para agrupar as características em centroides, representando, assim, características que são similares. Após o agrupamento, são extraídas características de imagens rotuladas dos objetos de interesse e são comparados aos centroides dos grupos. Os grupos recebem o rótulo de acordo com as características mais similares. Para cada segmento do mapa de saliências, são detectadas e extraídas as características e são comparadas aos grupos. Sendo o segmento classificado de acordo com o grupo que tiver mais características da imagem. Os segmentos classificados como objetos de interesse são enviados para o módulo de rastreamento enquanto os

demais segmentos são descartados.

No módulo de Rastreamento, os objetos segmentados são incluídos em uma lista e os objetos sob rastreio são atualizados com a nova localização. O processo de rastreamento ocorre em todo quadro, não apenas no quadro-chave, e para cada objeto são aplicados o Camshift [3.7](#) e um filtro de Kalman [3.8](#) para estimar e suavizar a trajetória do veículo no fluxo de imagens.

As posições detectadas dos veículos são enviadas para o módulo de controle do LAR-VANTAR que envia comandos para o robô aéreo se aproximar ou se afastar de algum objetivo específico.

4.1 Considerações finais

Neste Capítulo foi descrita a proposta para o rastreamento de veículos. Os módulos do sistema foram apresentados assim como as técnicas utilizadas em cada módulo. No Capítulo [5](#) são detalhados o processo de construção de tal modelo, as decisões de projeto tomadas e os resultados alcançados.



ANÁLISE DO MODELO PROPOSTO

No presente capítulo são apresentados os experimentos realizados explicitando os equipamentos e ferramentas que foram utilizados, bem como os resultados obtidos durante a fase de implementação do modelo proposto no capítulo 4, utilizando os métodos e conceitos apresentados no capítulo 3. Na seção 5.1 é explanado o processo de coleta de dados. Na seção 5.2 é exposta a implementação do detector de objetos salientes usando a técnica VOCUS2 e o modelo usando as abordagens *bottom-up* e *top-down*. Na seção 5.3 explicam-se o desenvolvimento e os resultados da classificação em conjunto com o *Bag-of-Features*. Na seção 5.4 são descritos os testes usando as técnicas Camshift e filtro de Kalman para rastrear os objetos.

Este trabalho aplicou a metodologia descrita no capítulo 4 para detectar e rastrear veículos durante o voo de um robô aéreo sobrevoando uma rodovia. Todos os experimentos foram realizados em um computador com processador Intel Core i5, 3.40GHz e 6GB de memória RAM. As bibliotecas de softwares livres OpenCV, VOCUS2 e iNVT foram usadas para implementar algumas técnicas do sistema. Os equipamentos utilizados nos experimentos foram uma câmera de vídeo e um robô aéreo do tipo asa rotativa (Fig. 12).

O robô aéreo é equipado com sistema de controle NAZA V2, da fabricante DJI (([INNOVATIONSHOR, 2015 \(accessed January 7, 2015\)](#))). O sistema de controle é constituído de uma unidade de medida inercial IMU, sistema de posicionamento global GPS, gerenciador de energia, e link de dados. Este sistema mantém a posição estável do robô e comunica com uma base terrestre pelo link de dados (receptor/robô). A estação terrestre comunica-se com um *tablet* via conexão *bluetooth* no qual está instalado o software de controle. O software fornece diversas funcionalidades e configurações, tais como seguir uma rota pré-planejada, decolagem e pouso autônomos, controle por *joystick* virtual, iniciar, trocar e interromper missões, definir a altitude, velocidade, tempo de voo e outras configurações. A Figura 13 apresenta a tela do software DJI da estação terrestre.

Uma câmera GoPro de baixo custo e alta resolução foi instalada no robô aéreo para



Figura 12 – Robô aéreo de asa rotativa utilizado para gravação dos vídeos.

aquisição dos vídeos. Esta câmera possui correção de ruído, caixa de proteção contra impactos e conexão Wi-Fi.

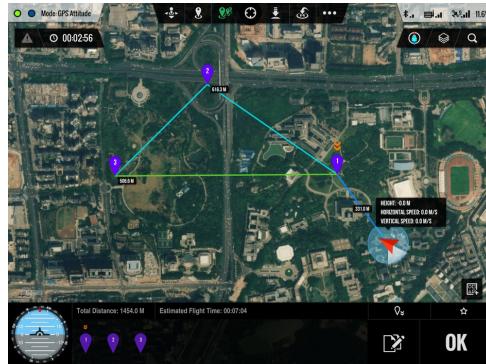


Figura 13 – Software Ground Station DJI ([INNOVATIONSHOR, 2015 \(accessed January 7, 2015\)](#)).

5.1 Coleta de dados

Durante a revisão bibliográfica, com o melhor de nosso conhecimento, não foi encontrado um banco de dados com vídeos ou imagens de veículos a partir de ângulos aéreos; apenas alguns vídeos filmados a partir de helicópteros em voo que não puderam ser utilizados devido à baixa qualidade das filmagens.

Os dados utilizados pelo sistema foram vídeos adquiridos pela câmera de robô aéreo sobrevoando uma rodovia. Foram filmados 30 minutos de vídeo em alta resolução com o robô aéreo em quatro diferentes altitudes: 15, 20, 25 e 30 metros do solo.

Para o treinamento dos classificadores são necessárias diversas imagens do objeto que se deseja rastrear. Para compor o banco de imagens para treinamento, 287 imagens foram

manualmente selecionadas e rotuladas a partir dos vídeos gravados. Essas imagens contém um ou mais veículos na mesma cena. A Figura 14 é um exemplo das imagens que são consideradas como entrada para o sistema.



Figura 14 – Imagem de entrada do sistema.

Durante o processamento, são criados descritores das características das imagens usados para o rastreio de objetos que são temporariamente armazenados.

Embora a câmera capture o vídeo em 60 quadros por segundo, o sistema não tem capacidade para processar toda essa quantidade de informação. Para contornar esta dificuldade, são selecionados quadros-chave com base no movimento dos objetos da cena para o processamento. Um novo quadro-chave é selecionado a cada segundo ou quando uma quantidade de *pixels* diferentes entre dois quadros exceder um limite.

5.2 Experimento com modelos de atenção visual para detecção

Na tabela 2, pode ver-se a matriz de confusão obtida para a classificação das cinco classes de veículos consideradas. Houve um erro de classificação maior entre caminhonetes e caminhões e entre caminhões e ônibus. Como esperado, motos têm o mínimo erro de classificação devido ao seu tamanho, enquanto outras classes têm mais semelhança entre si. Na Figura 16d observa-se o resultado para a imagem apresentada na Figura 15.

Nas figuras 16a a 16c, podem ser vistas as intensidades de cores e características de orientação extraídos pelo processo *bottom-up* aplicado ao processamento da imagem original.

Na figura 16d é mostrada a posição do objeto de maior atenção. Na figura 17a, é mostrada a imagem correspondente à segmentação pela rede LEGION e na figura 17b. O objeto saliente de maior atenção aparece na cor vermelha. Finalmente, na figura 18 é mostrado um gráfico que contém os valores da saliência dos objetos que competiram pela atenção. Podemos ver que a

região do asfalto foi a de maior valor saliência, mas considerando o enviesamento *top-down* de objetos conhecidos pelo MLP, o objeto com valor igual a 0,35 de saliência é o preferido e destacado pelo sistema visual.

O sistema de atenção visual detectou os veículos em cenas reais com taxa de acerto de 83 %. Isto é devido à necessidade de lidar com diferentes condições de luz (não consideradas nas imagens utilizadas para o treinamento) e também devido ao ruído nas imagens durante o processo de captura de imagens.

Tabela 2 – A matriz de confusão da classificação.

Classe real	Carros	Caminhões	Motos	Camionete	Ônibus
Carros	22	3	2	3	0
Caminhões	1	20	0	2	7
Motos	3	0	27	0	0
Camionete	8	5	0	17	0
Ônibus	0	8	0	1	21



Figura 15 – Imagem de entrada para o detector.

5.3 Experimento para classificação de segmentos com Bag-of-Features

O sistema proposto foi testado para detectar veículos de três classes diferentes: carros, caminhões e motos, como mostrado na Figura 20.

Para reconhecer os veículos como modelo *bag-of-features* foram usadas múltiplas SVMs com a estratégia *one-against-all* para classificar os veículos nas classes correspondentes. Três SVM foram treinadas para executar o reconhecimento, uma por classe e para cada classificador. Como amostras positivas foram usadas as imagens do conjunto de dados que correspondem à classe específica, e como amostras negativas foram usadas as imagens das outras classes

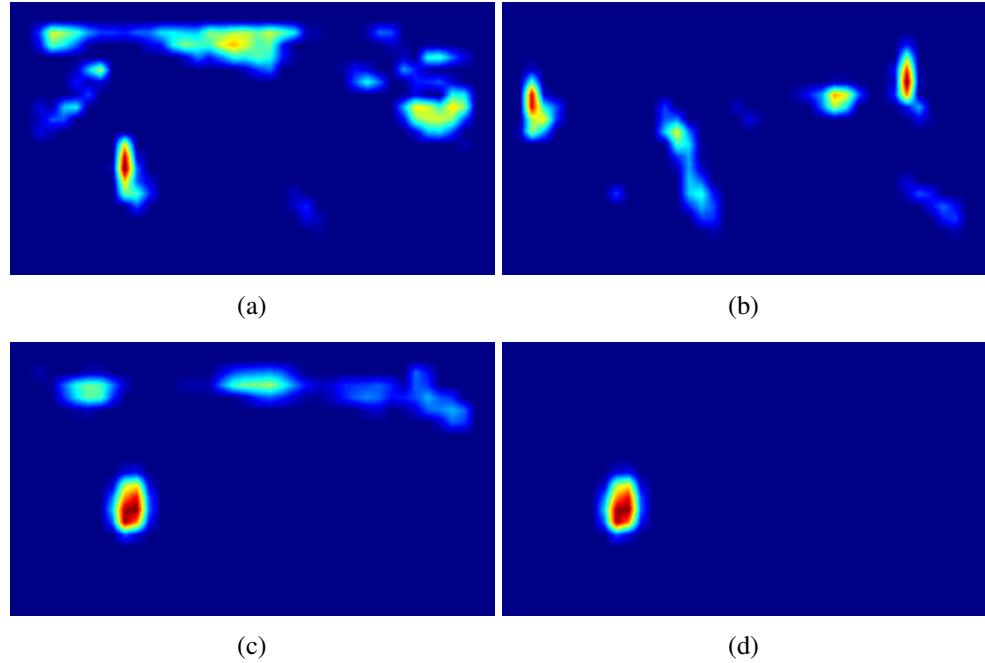


Figura 16 – Mapas de conspicuidade de (a) intensidades, (b) cores e (c) orientações e o (d) reconhecimento.

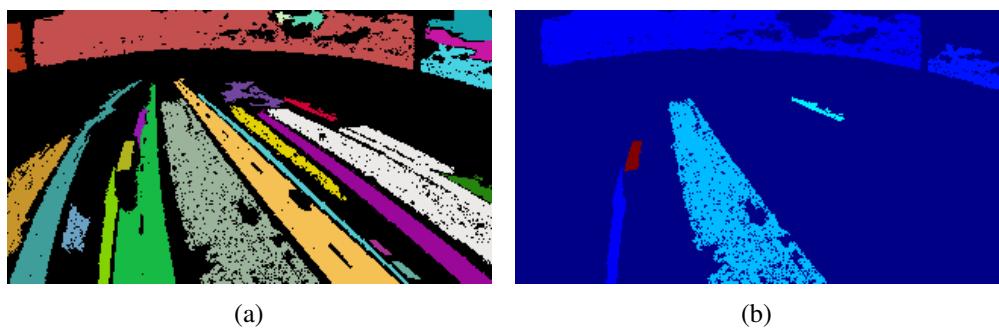


Figura 17 – (a) Imagem segmentada e (b) imagem destacando o objeto saliente.

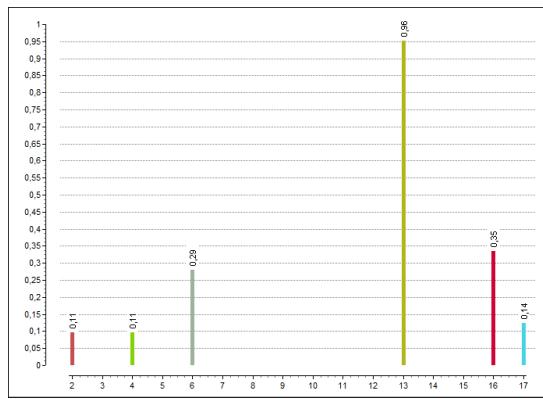
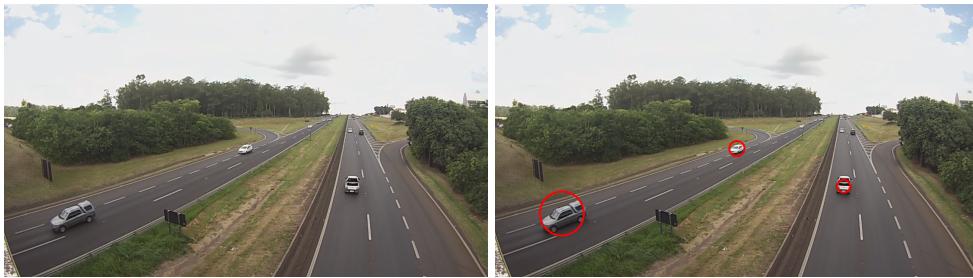


Figura 18 – Gráfico dos valores salientes dos objetos que competiam pela atenção.

do conjunto. Por exemplo, as imagens de caminhões foram usadas como amostras positivas e imagens das classes carro, moto e desconhecido foram usadas como amostras negativas para o treinamento da SVM que classificaria caminhões. Assim, o segmento dado ao modelo BoF pelo

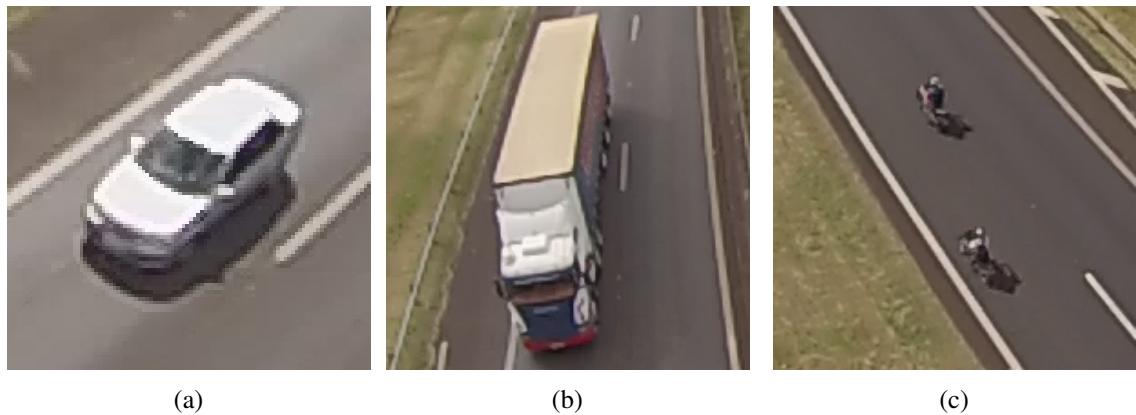


(a) Imagem de entrada do sistema.

(b) Veículos detectados na cena.

Figura 19 – Exemplo de imagem de entrada e imagem exibindo veículos detectados.

sistema de atenção visual é classificado em ordem pelos SVMs. Em outras palavras, o segmento dado ao BoF pelo sistema de atenção visual é classificado primeiro pela SVM de uma classe. Em caso positivo, o sistema considera o segmento com o rótulo do classificador atual. Caso contrário, o segmento é passado aos classificadores SVM restantes, seguindo a mesma regra. Se o segmento não for classificado por qualquer classificador, ele é considerado um objeto sem interesse para o sistema.



(a)

(b)

(c)

Figura 20 – Os três tipos de veículos considerados: (a) carro, (b) caminhão e (c) moto.

Foram consideradas 40 imagens contendo veículos de cada classe como amostras positivas e 40 imagens sem esses veículos como amostras negativas para o treinamento das SVMs. Na Tabela 3 é mostrada a matriz de confusão gerada pela classificação das SVMs da fase de teste. Cada linha representa um classificador treinado para predizer apenas uma classe. Os valores das colunas são as previsões dos respectivos classificadores para cada classe considerada. As amostras classificadas como desconhecidas são as amostras que nenhum dos 3 classificadores conseguiu reconhecer. Por exemplo, o classificador de carro - representado na primeira linha da matriz - rotulou corretamente 29 carros, confundiu 9 carros classificados como caminhão e um carro classificado como moto. Na coluna desconhecido são todos os objetos que esse classificador não reconheceu como carro, podendo ser ele qualquer outro exemplo negativo como caminhão, moto, árvores ou qualquer outro segmento de não interesse. Portanto, a acurácia do classificador de carros é feita pelas classificações corretas (20 carros como carros) pelo total de imagens



Figura 21 – Subconjunto de imagens considerados para o treinamento de caminhões.

reconhecidos e classificados por esse classificador (39 imagens nesse caso). Para cada classe de interesse - carro, caminhão e moto - os classificadores foram testados individualmente. Já para a linha de desconhecidos na Tabela 3 foram utilizados árvores, placas de trânsito, fragmentos da estrada e outros objetos de não interesse e nenhum dos classificadores os reconheceu como um exemplo positivo para sua classe. Dessa forma, eles foram ignorados e corretamente classificados como objetos desconhecidos. Finalmente, a última linha da tabela mostra os resultados dos classificadores combinados. Isto é, a imagem passava em um classificador por vez e caso fosse classificada como positiva em algum dos classificadores era rotulada como daquela classe e a próxima imagem era então testada. A ordem dos classificadores seguia a ordem mostrada nas linhas da tabela. Primeiro, o classificador de carro, depois o de caminhão e, finalmente, o de moto. Foram selecionadas 114 imagens contendo diferentes tipos de veículos. O resultado obtido foi de uma predição correta de 29 dos 39 carros, 39 dos 40 caminhões e 23 das 35 motos contidas no banco de imagens varridas no teste. Apenas um carro não foi classificado e passou pelos classificadores sendo considerado desconhecido. A acurácia do sistema foi de 79,82%. Com relação ao sistema de atenção visual, os segmentos foram detectados em tempo satisfatório. As vezes o sistema não reconhecia segmentos válidos nas primeiras tentativas devido à variação de luminosidade do ambiente e ao ruído das imagens de vídeo, oriundos da trepidação do robô aéreo. Porém, tais segmentos eram ignorados com sucesso pelos classificadores como mostram os resultados.

Tabela 3 – A matriz de confusão do classificador.

Classes	carro	caminhão	moto	desconhecido	#Amostras	Acurácia (%)
carro	29	9	1	40	79	87.34
caminhão	3	35	0	31	78	84.61
moto	9	3	22	39	73	83.56
desconhecido	0	0	0	26	26	100
classificadores	29/39	39/40	23/35	1/0	114	79.82

5.4 Experimento com o rastreamento

Após os passos de detecção e classificação, segmentos são enviados para serem rastreados nos demais quadros não chave. Os objetos rastreados são mantidos em uma lista. Para adicionar novos segmentos à lista faz-se duas comparações entre o novo segmento e os segmentos já existentes na lista: comparação de histograma e número de correspondências de características. Se algum objeto da lista corresponder ao segmento detectado, a posição do objeto é atualizada de acordo com a posição do segmento na imagem. Se não existir qualquer objeto correspondente ao segmento, então o segmento é inserido como novo objeto a ser rastreado. Objetos são removidos da lista se não forem detectados em dez quadros consecutivos. Isto ocorre quando o objeto sai dos limites da cena ou então quando aproxima-se da linha do horizonte, tornando-se pequeno demais para ser detectado.

Para cada objeto da lista é computado o histograma do segmento e a posição inicial na imagem. Para cada quadro seguinte é executado o Camshift, no qual o objeto é buscado nos arredores de sua última localização com base em seu histograma. Além disso, as posições são filtradas pelo filtro de Kalman para reduzir ruídos de possíveis medições do rastreamento Camshift.



Figura 22 – Sequência de imagens exibindo a posição do veículo em cada quadro.

5.5 Considerações finais

Neste capítulo foram descritos os equipamentos e ferramentas usadas, os experimentos realizados e os resultados obtidos com o modelo proposto. Conforme resultados apresentados nas seções 5.2, 5.3 e 5.4 pode-se verificar que a abordagem utilizada foi eficaz nas tarefas desejadas.



CONCLUSÕES E TRABALHOS FUTUROS

A atenção visual é um mecanismo biologicamente inspirado. A fim de receber informações do ambiente, órgãos dos sentidos como os olhos, ouvidos, nariz recebem informações sensoriais e transmitem a informação sensorial para o cérebro. A capacidade de selecionar e processar apenas as regiões mais relevantes de uma cena visual é parte fundamental neste processo. Isso significa que para reconhecer e localizar objetos em uma determinada cena, o sistema visual seleciona-os, um a cada vez. Portanto, o principal objetivo é estimar as probabilidades de identificar características visuais dos objetos diretamente no local da imagem. Dentre as diferentes áreas de aplicação da atenção visual, o foco da pesquisa desenvolvida durante este mestrado foi o desenvolvimento de um sistema de atenção visual para detecção de objetos e classificação de veículos para a tarefa de rastreamento de veículos em imagens aéreas. As características essenciais da atenção visual *bottom-up* e *top-down* foram abordadas. Dessa forma, fenômenos de descontinuidade de intensidade em diferentes escalas e orientações, bem como a busca por regiões de interesse a partir de características de alto nível, especificadas através conhecimento prévio na forma ou modelos sobre o que se está buscando na cena são computadas diretamente na imagem, sem qualquer informação contextual. Ou seja, investigar a possibilidade de criar um mecanismo de atenção visual que poderia ser otimizado para destacar quaisquer regiões significativas da classe desejada do objeto. Uma revisão bibliográfica foi realizada e os principais aspectos da atenção visual foram abordados. Avanços tecnológicos constantes foram observados e vários sistemas novos implementados. O aspecto fundamentalmente desenvolvido neste trabalho evidencia, claramente, a inserção de informações de alto nível em sistemas de atenção visual *bottom-up* pode ser realizada com sucesso. A arquitetura do sistema foi apresentada, amplamente discutida e aplicada ao rastreamento de veículos através de sistemas de detecção e classificação de segmentos em imagens. Técnicas robustas de classificação foram empregadas com sucesso no rastreamento de veículos a partir de imagens de um robô aéreo. A adaptação do sistema de atenção visual *bottom-up* e *top-down* para detectar segmentos usando entrada de vídeo foi realizada. Em seu contexto global, esta dissertação contribui com dados e

informações úteis para o módulo de controle do robô do sistema LARVANTAR. Como resultado deste trabalho e por força de sua inserção, dois trabalhos foram submetidos (Anexo A) e três trabalhos em colaboração foram publicados além de várias atividades correlatas que serviram de suporte e benefício sustentados em seu conhecimento. As contribuições específicas deste trabalho podem ser assim identificadas:

- (a) identificação de regiões mais significativas de uma imagem: melhor forma de encontrar regiões de interesse na imagem para analisar partes específicas ao invés da imagem por inteiro, acelerando o processo de detecção e classificação de diversos tipos de veículos nas cenas;
- (b) processamento da imagem ampliado e melhor caracterizado através de seleção de quadros-chave, rastreio em todos os quadros;
- (c) identificação e aplicação de algoritmos robustos que geraram modelos realistas e passíveis de melhora de atenção;
- (d) os experimentos podem ser facilmente otimizados e aplicados em situações reais através do uso de robôs aéreos.

Em geral, as seguintes questões foram abordadas: objetivo da visão em encontrar o objeto, reconhecimento do objeto de forma segmentada, integração dos sistemas *bottom-up* e *top-down* consistente com a seleção visual por identificação das principais características, a redução da atenção espacial reduz a incerteza sobre a localização e melhora o reconhecimento do objeto.

Entre as limitações que foram encontradas durante a execução deste trabalho estão:

- **Capacidade computacional** - Os algoritmos de visão computacional são complexos e demandam alta capacidade de processamento.
- **Grande quantidade de dados** - A quantidade de dados utilizada deve ser limitada para evitar estouro de memória.
- **Qualidade dos sensores** - Sensores possuem ruídos que podem interferir no funcionamento do sistema. É necessário tratar eventuais ruídos.

Como continuidade, pretende-se estudar e analisar melhorias e realizar experimentos para incrementar a quantidade de características primitivas utilizadas; ampliar a análise dos pesos e otimizar o sistema de classificação por emprego de algoritmos mistos; embarcar o sistema em plataforma Raspberry Pi e Arduino Uno; aplicar redes neurais convolucionais para classificação dps segmentos para melhorar acurácia do classificador.

A principal aplicação do sistema desenvolvido nesta dissertação é útil para o sistema LARVANTAR para o controle do robô aéreo e nele será implementado.

As estratégias *top-down* e *bottom-up* ajudam no processamento da informação e no foco de conhecimento mais relevantes. Partindo-se de macroescala com redução para segmentos menores torna-o um modelo realisticamente validável. A percepção fica evidenciada e o ambiente melhor descrito.

REFERÊNCIAS

- ANDRADE, R. M.; JOSÉ, C. G.; CASTRO, A. P. A.; SHIGUEMORI, E. H. Acompanhamento de marcos em sequências de imagens aéreas para navegação autônoma de veículos aéreos não tripulados. In: **SIBGRAPI 2010, Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images**. Porto Alegre: IEEE Computer Society, 2010. ISBN 978-1-4244-8420-1. Citado na página 20.
- AROCA, R. V.; AGUIAR, F. G.; AIHARA, C.; TONIDANDEL, F.; MONTANARI, R.; FRACCAROLI, E. S.; SILVA, M. O. da; ROMERO, R. A. F. Olimpíada brasileira de robótica: relatos da primeira regional em são carlos-sp. In: . [S.l.: s.n.], 2014. p. 35–41. Citado na página 77.
- BACKER, G.; MERTSCHING, B.; BOLLMANN, M. Data-and model-driven gaze control for an active-vision system. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 23, n. 12, p. 1415–1429, 2001. Citado na página 26.
- BAR-SHALOM, Y. **Tracking and data association**. [S.l.]: Academic Press Professional, Inc., 1987. Citado na página 49.
- BATISTA, M. R.; JORGE, L. L. R.; CALVO, R.; FRACCAROLI, E. S.; DOREA, Y. M.; SILVA, M. O. da; MONTANARI, R.; ROMERO, R. A. F. A solution to swarm robot escorting using the probabilistic lloyd method. In: . [S.l.: s.n.], 2013. Citado na página 77.
- BAY, H.; ESS, A.; TUYTELAARS, T.; GOOL, L. van. Speeded-up robust features (surf). **Computer Vision and Image Understanding (CVIU)**, v. 110, n. 3, p. 346–359, June 2008. Citado na página 42.
- BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: **In ECCV**. [S.l.: s.n.], 2006. p. 404–417. Citado 4 vezes nas páginas 13, 38, 41 e 44.
- BENICASA, A. X. **Sistemas computacionais para atenção visual Top-Down e Bottom-up usando redes neurais artificiais**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo, São Carlos, 9 2013. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29042014-162209/>>. Citado 5 vezes nas páginas 13, 20, 21, 37 e 38.
- BEYMER, D.; KONOLIGE, K. Real-time tracking of multiple people using continuous detection. In: **IEEE Frame Rate Workshop**. [S.l.: s.n.], 1999. Citado na página 49.
- BOTTERILL, T.; MILLS, S.; GREEN, R. Speeded-up Bag-of-Words algorithm for robot localisation through scene recognition. In: **Image and Vision Computing New Zealand**. [S.l.: s.n.], 2008. p. 1–6. Citado na página 44.
- BRADSKI, G. R. Computer vision face tracking for use in a perceptual user interface. Citeseer, 1998. Citado na página 24.

BRANCO, K.; PELIZZONI, J.; NERIS, L. O.; TRINDADE, O.; OSORIO, F.; WOLF, D. Tiriba - a new approach of uav based on model driven development and multiprocessors. In: **Robotics and Automation (ICRA), 2011 IEEE International Conference on**. [S.l.: s.n.], 2011. p. 1–4. ISSN 1050-4729. Citado na página [20](#).

BROIDA, T. J.; CHELLAPPA, R. Estimation of object motion parameters from noisy images. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, n. 1, p. 90–99, 1986. Citado na página [49](#).

BROWN, M.; LOWE, D. G. Invariant features from interest point groups. In: **BMVC**. [S.l.: s.n.], 2002. Citado na página [43](#).

BURT, P. J.; ADELSON, E. H. The laplacian pyramid as a compact image code. **Communications, IEEE Transactions on**, IEEE, v. 31, n. 4, p. 532–540, 1983. Citado na página [31](#).

CLARK, J. J.; FERRIER, N. J. Modal control of an attentive vision system. In: **ICCV**. [S.l.: s.n.], 1988. p. 514–523. Citado na página [25](#).

COIFMAN, B.; BEYMER, D.; MCLAUCHLAN, P.; MALIK, J. A real-time computer vision system for vehicle tracking and traffic surveillance. **Transportation Research Part C: Emerging Technologies**, Elsevier, v. 6, n. 4, p. 271–288, 1998. Citado na página [27](#).

CORTES, C.; VAPNIK, V. Support-vector networks. **Mach. Learn.**, Kluwer Academic Publishers, Hingham, MA, USA, v. 20, n. 3, p. 273–297, set. 1995. ISSN 0885-6125. Citado na página [46](#).

COSTA, F. G.; UNEYAMA, J.; BRAUN, T.; PESSIN, G.; OSORIO, F. S.; VARGAS, P. A. The use of unmanned aerial vehicles and wireless sensor network in agricultural applications. In: **Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International**. [S.l.: s.n.], 2012. p. 5045 –5048. ISSN 2153-6996. Citado na página [19](#).

DRAPER, B. A.; LIONELLE, A. Evaluation of selective attention under similarity transformations. **Computer vision and image understanding**, Elsevier, v. 100, n. 1, p. 152–171, 2005. Citado na página [26](#).

FIX, E.; HODGES, J. **Discriminatory analysis. Nonparametric discrimination: Consistency properties**. [S.l.], 1951. Citado na página [40](#).

FRACCAROLI, E. S.; SILVA, M. O. da; BATISTA, M. R.; MONTANARI, R.; ROMERO, R. A. F. Controlador fuzzy para impressora 3d de baixo custo. In: . João Pessoa: [s.n.]. p. 364–375. Citado na página [77](#).

FRINTROP, S. **VOCUS: A visual attention system for object detection and goal-directed search**. [S.l.]: Springer, 2006. v. 3899. Citado 2 vezes nas páginas [26](#) e [27](#).

FRINTROP, S.; WERNER, T.; GARCIA, G. M. Traditional saliency reloaded: A good old model in new shape. In: . [S.l.: s.n.], 2015. p. 82–90. Citado 6 vezes nas páginas [13](#), [15](#), [20](#), [34](#), [35](#) e [36](#).

FUKUNAGA, K.; HOSTETLER, L. D. The estimation of the gradient of a density function, with applications in pattern recognition. **Information Theory, IEEE Transactions on**, IEEE, v. 21, n. 1, p. 32–40, 1975. Citado na página [24](#).

- GREENSPAN, H.; BELONGIE, S.; GOODMAN, R.; PERONA, P.; RAKSHIT, S.; ANDERSON, C. H. Overcomplete steerable pyramid filters and rotation invariance. In: IEEE. **Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on.** [S.I.], 1994. p. 222–228. Citado na página 32.
- GUPTE, S.; MASOUD, O.; MARTIN, R. F.; PAPANIKOLOPOULOS, N. P. Detection and classification of vehicles. **Intelligent Transportation Systems, IEEE Transactions on**, IEEE, v. 3, n. 1, p. 37–47, 2002. Citado na página 27.
- HAMKER, F. H. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. **Computer Vision and Image Understanding**, Elsevier, v. 100, n. 1, p. 64–106, 2005. Citado na página 26.
- HORN, B. K.; SCHUNCK, B. G. Determining optical flow. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **1981 Technical symposium east.** [S.I.], 1981. p. 319–331. Citado na página 24.
- HUANG, K.; WANG, L.; TAN, T.; MAYBANK, S. A real-time object detecting and tracking system for outdoor night surveillance. **Pattern Recognition**, Elsevier, v. 41, n. 1, p. 432–444, 2008. Citado na página 24.
- INNOVATIONSHOR, N. D. **Ipad Ground Station.** 2015 (accessed January 7, 2015). <<http://http://www.dji.com/product/ipad-ground-station>>. Citado 3 vezes nas páginas 13, 55 e 56.
- ITTI, L.; KOCH, C. A saliency-based search mechanism for overt and covert shifts of visual attention. **Vision research**, Elsevier, v. 40, n. 10, p. 1489–1506, 2000. Citado na página 33.
- _____. Computational modelling of visual attention. **Nature reviews neuroscience**, Nature Publishing Group, v. 2, n. 3, p. 194–203, 2001. Citado 4 vezes nas páginas 13, 29, 30 e 31.
- ITTI, L.; KOCH, C.; NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE Computer Society, v. 20, n. 11, p. 1254–1259, 1998. Citado 6 vezes nas páginas 25, 26, 31, 32, 33 e 34.
- KIM, Z.; MALIK, J. Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking. In: IEEE. **Computer Vision, 2003. Proceedings. Ninth International Conference on.** [S.I.], 2003. p. 524–531. Citado na página 28.
- KOCH, C.; ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. In: **Matters of intelligence.** [S.I.]: Springer, 1987. p. 115–141. Citado 4 vezes nas páginas 25, 27, 31 e 33.
- KOENDERINK, J. J. The structure of images. **Biological cybernetics**, Springer, v. 50, n. 5, p. 363–370, 1984. Citado na página 42.
- KOLLER, D.; WEBER, J.; MALIK, J. **Robust multiple car tracking with occlusion reasoning.** [S.I.]: Springer, 1994. Citado na página 27.
- KRAJNÍK, T.; VONÁSEK, V.; FIŠER, D.; FAIGL, J. Ar-drone as a robotic platform for research and education. In: **International Conference on Research and Education in Robotics.** Heidelberg: Springer, 2011. ISSN 1865-0929. Citado na página 19.

LEI, B.; XU, L.-Q. Real-time outdoor video surveillance with robust foreground extraction and object tracking via multi-state transition management. **Pattern Recognition Letters**, Elsevier, v. 27, n. 15, p. 1816–1825, 2006. Citado na página 24.

LEONARD, J.; DURRANT-WHYTE, H. Simultaneous map building and localization for an autonomous mobile robot. In: **Intelligent Robots and Systems '91. 'Intelligence for Mechanical Systems, Proceedings IROS '91. IEEE/RSJ International Workshop on**. [S.l.: s.n.], 1991. p. 1442 –1447 vol.3. Citado na página 19.

LINDEBERG, T. Scale-space for discrete signals. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, IEEE, v. 12, n. 3, p. 234–254, 1990. Citado na página 42.

LIPTON, A. J.; FUJIYOSHI, H.; PATIL, R. S. Moving target classification and tracking from real-time video. In: **IEEE. Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on**. [S.l.], 1998. p. 8–14. Citado na página 27.

LIU, C.; YUEN, P. C.; QIU, G. Object motion detection using information theoretic spatio-temporal saliency. **Pattern Recognition**, Elsevier, v. 42, n. 11, p. 2897–2906, 2009. Citado na página 23.

LLOYD, S. P. Least squares quantization in pcm. **Information Theory, IEEE Transactions on**, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 45.

LORENA A. C. E CARVALHO, A. C. P. L. F. Uma introdução às support vector machines. **Relatório Técnico - ICMC**, n. 192, p. 43–67, 2003. ISSN 0885-6125. Citado na página 46.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **Int. J. Comput. Vision**, Kluwer Academic Publishers, Hingham, MA, USA, v. 60, n. 2, p. 91–110, nov. 2004. ISSN 0920-5691. Disponível em: <<http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>>. Citado 5 vezes nas páginas 13, 38, 39, 40 e 41.

MATARIC, M. J. **The Robotics Primer**. [S.l.]: MIT Press, 2007. ISBN 978-0-262-63354-3. Citado na página 19.

MERINO, L.; CABALLERO, F.; DIOS, J. R. M. de; FERRUZ, J.; OLLERO, A. A cooperative perception system for multiple uavs: Application to automatic detection of forest fires. **J. Field Robotics**, p. 165–184, 2006. Citado na página 19.

MILANESE, R. Detecting salient regions in an image: from biological evidence to computer implementation. 1993. Citado na página 25.

MONTANARI, R.; TOZADORE, D. C.; FRACCAROLI, E. S.; BENICASA, A. X.; ROMERO, R. A. F. A visual attention approach for the tracking of vehicles through uav. XI Workshop de Visão Computacional (WVC2015). 2015. Citado na página 77.

MOORE, A. **A tutorial on kd-trees**. [S.l.], 1991. Citado na página 44.

MU, C.; YAN, Q.; FENG, Y.; CAI, J.; YU, J. **Overview of powerlines extraction and surveillance using remote sensing technology**. 2009. 74981M-74981M-8 p. Disponível em: <<http://dx.doi.org/10.1117/12.833688>>. Citado na página 19.

PICCARDI, M. Background subtraction techniques: a review. In: **IEEE. Systems, man and cybernetics, 2004 IEEE international conference on**. [S.l.], 2004. v. 4, p. 3099–3104. Citado na página 24.

- RAMSTRÖM, O.; CHRISTENSEN, H. I. Visual attention using game theory. In: SPRINGER. **Biologically motivated computer vision**. [S.l.], 2002. p. 462–471. Citado na página 26.
- ROSALES, R.; SCLAROFF, S. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In: IEEE. **Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on**. [S.l.], 1999. v. 2. Citado na página 49.
- ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. In: **European Conference on Computer Vision**. [s.n.], 2006. v. 1, p. 430–443. Disponível em: <http://edwardrosten.com/work/rostens_2006_machine.pdf>. Citado na página 38.
- RUBLEE, E.; RABAUD, V.; KONOLIGE, K.; BRADSKI, G. Orb: An efficient alternative to sift or surf. In: **International Conference on Computer Vision**. Barcelona: [s.n.], 2011. Citado na página 38.
- RYAN, A.; ZENNARO, M.; HOWELL, A.; SENGUPTA, R.; HEDRICK, J. K. An overview of emerging results in cooperative uav control. In: **In Proceedings of 43rd IEEE Conference on Decision and Control**. [S.l.: s.n.], 2004. p. respectively. Citado na página 19.
- SALTON, G.; MCGILL, M. **Introduction to modern information retrieval**. McGraw-Hill, 1983. ISBN 9780070544840. Disponível em: <<http://books.google.com.br/books?id=7f5TAAAAMAAJ>>. Citado na página 45.
- SAMAD, T.; BAY, J. S.; GODBOLE, D. Network-Centric Systems for Military Operations in Urban Terrain: The Role of UAVs. **Proceedings of the IEEE**, IEEE, v. 95, n. 1, p. 92–107, jan. 2007. ISSN 0018-9219. Disponível em: <<http://dx.doi.org/10.1109/JPROC.2006.887327>>. Citado na página 19.
- SCHILDIT, A. N.; SANCA, A. S.; AES, J. ao P. F. G.; DEUS, M. S. de; ALSINA, P. J. Hardware and telemetry architectures to an unmanned aerial vehicle of quadrotor type. In: **5th Workshop on Applied Robotics and Automation - ROBOCONTROL 2012**. [S.l.: s.n.], 2012. Citado na página 20.
- SIVARAMAN, S.; TRIVEDI, M. M. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. **Intelligent Transportation Systems, IEEE Transactions on**, IEEE, v. 14, n. 4, p. 1773–1795, 2013. Citado na página 27.
- SULLIVAN, G. D.; BAKER, K. D.; WORRALL, A. D.; ATTWOOD, C.; REMAGNINO, P. Model-based vehicle detection and classification using orthographic approximations. **Image and Vision Computing**, Elsevier, v. 15, n. 8, p. 649–654, 1997. Citado na página 28.
- SUN, Y.; FISHER, R. Object-based visual attention for computer vision. **Artificial Intelligence**, Elsevier, v. 146, n. 1, p. 77–123, 2003. Citado na página 26.
- TREISMAN, A. M.; GELADE, G. A feature-integration theory of attention. **Cognitive psychology**, Elsevier, v. 12, n. 1, p. 97–136, 1980. Citado na página 25.
- TRUCCO, E.; VERRI, A. **Introductory Techniques for 3-D Computer Vision**. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132611082. Citado na página 19.
- TSOTSOS, J. K. Analyzing vision at the complexity level. **Behavioral and brain sciences**, Cambridge Univ Press, v. 13, n. 03, p. 423–445, 1990. Citado na página 26.

TSOTSOS, J. K.; CULHANE, S. M.; WAI, W. Y. K.; LAI, Y.; DAVIS, N.; NUFLO, F. Modeling visual attention via selective tuning. **Artificial intelligence**, Elsevier, v. 78, n. 1, p. 507–545, 1995. Citado na página [33](#).

ZHOU, S. K.; CHELLAPPA, R.; MOGHADDAM, B. Visual tracking and recognition using appearance-adaptive models in particle filters. **Image Processing, IEEE Transactions on**, IEEE, v. 13, n. 11, p. 1491–1506, 2004. Citado na página [24](#).

GLOSSÁRIO

Atenção Bottom-up: Processo de atenção exógena no qual o foco se desloca para elementos da cena que se destacam dos demais por alguma característica destoante..

Atenção Top-Down: Processo de atenção endógena no qual informações do córtex cerebral alteram o foco da atenção..

BoF: Sigla de *Bag-of-Features*. Modelo computacional que faz busca em um conjunto de características para classificar imagens..

Camshift: Sigla de *Continuously Adaptive Meanshift*. Técnica para rastreamento visual de objetos..

Característica visual: Qualquer parte destacável de uma imagem..

Classificação: Ação de categorizar um elemento dentro de um conjunto possível de classes..

Detecção: Ação de encontrar um elementos específico dentro de um conjunto de elementos..

FAST: Sigla de *Features from Accelerated Segment Test*. Algoritmo de detecção de características locais visuais em imagens..

Filtro de Kalman: Método de previsão e correção de estados de variáveis para redução de ruídos e incertezas de um sistema..

iNVT: Sigla de *iLab Neuromorphic Vision C++ Toolkit*. Biblioteca de código livre para o desenvolvimento de modelos neuromórficos de visão..

k-Means: Algoritmo de aprendizado de máquina usado para agrupamento de dados..

k-NN: Sigla de *k-Nearest Neighbors*. Algoritmo de aprendizado de máquina usado para classificação de padrões..

LEGION: Sigla de *Locally Excitatory Globally Inhibitory Oscillator Network*. Rede neural artificial composta de osciladores neurais com excitação de osciladores acoplados e inibição global que pode ser usada para segmentação de imagens, seleção de objetos e segregação de discurso..

Mapa de saliência: Mapa topologicamente organizado que representa saliências visuais de uma cena correspondente..

Meanshift: Técnica para localização ponto de máximo de uma função de densidade..

MLP: Sigla de *Multi Layer Perceptron*. Rede neural multicamadas usado para classificação ou regressão de dados..

OpenCV: Sigla de *Open Source Computer Vision Library*. Biblioteca de código livre que contém uma coleção de algoritmos de visão computacional..

ORB: Sigla de *Oriented FAST and Rotated BRIEF*. Algoritmo de detecção e descrição de características visuais em imagens..

Quadro-chave: Imagem de um fluxo de vídeo selecionada por se destacar devido à alguma característica..

Rastreamento: Ação de seguir a trajetória de um objeto durante um período de tempo..

Saliência visual: Qualidade perceptual subjetiva distinta, que faz com que alguns itens no mundo destaquem-se de seus vizinhos e imediatamente prendem o foco de atenção..

SIFT: Sigla de *Scale-invariant feature transform*. É um algoritmo de visão computacional para detectar e descrever características locais em imagens.

SOM: Sigla de *Self-Organizing Map*. Rede neural artificial auto organizável para aprendizado não supervisionado..

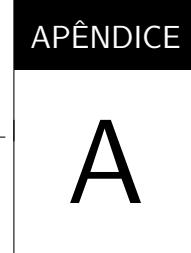
SURF: Sigla de *Speeded Up Robust Features*. É um detector e descritor de características locais em imagens, parcialmente inspirado no SIFT, que pode ser usado para tarefas como reconhecimento de objetos e reconstrução 3D..

SVM: Sigla de *Support Vector Machine*. Algoritmo de aprendizado de máquina usado para classificação de padrões..

UAV: Sigla de *Unmanned Aerial Vehicle*. Veículo aéreo não tripulado..

VOCUS: Sigla de *Visual Object detection with a Computational attention System*. Sistema computacional de atenção visual para geração de mapas de saliências..

Winner-Take-All: Mecanismo de competição no qual o elemento vencedor é selecionado e os demais elementos são inibidos..



PUBLICAÇÕES DECORRENTES DESTE TRABALHO

(MONTANARI *et al.*, 2015) MONTANARI, R. ; TOZADORE, D. ; FRACCAROLI, E. S. ; ROMERO, R. A. F. . Ground vehicle detection and classification by an unmanned aerial vehicle. In: 12th Latin American Robotics Symposium (LARS2015), Uberlândia, 2015.

(MONTANARI *et al.*, 2015) MONTANARI, R. ; TOZADORE, D. ; FRACCAROLI, E. S. ; BENICASA, A. X. ; ROMERO, R. A. F. . A visual attention approach for the tracking of vehicles through UAV. In: XI Workshop de Visão Computacional (WVC2015), São Carlos, 2015.

Colaboração:

(FRACCAROLI *et al.*,) FRACCAROLI, E. S. ; SILVA, M. O. ; BATISTA, M. R. ; MONTANARI, R. ; ROMERO, R. A. F. . Controlador fuzzy para impressora 3D de baixo custo. In: III Congresso Brasileiro de Sistemas Fuzzy, 2014, João Pessoa. Controlador fuzzy para impressora 3D de baixo custo, 2014. p. 364-375.

(AROCA *et al.*, 2014) AROCA, R. V. ; AGUIAR, F. G. ; AIHARA, C. ; TONIDANDEL, F. ; MONTANARI, R. ; FRACCAROLI, E. S. ; SILVA, M. O. ; ROMERO, R. A. F. . Olimpíada Brasileira de Robótica: relatos da primeira regional em São Carlos-SP. In: Workshop of Robotics in Education, 2014, São Carlos. Workshop of Robotics in Education, 2014. p. 35-41.

(BATISTA *et al.*, 2013) BATISTA, M. R. ; JORGE, L. L. R. ; CALVO, R. ; FRACCAROLI, E. S. ; DOREA, Y. M. ; SILVA, M. O. ; MONTANARI, R. ; ROMERO, R. A. F. . A Solution to Swarm Robot Escorting using the Probabilistic Lloyd Method. In: Simpósio Brasileiro de Automação Inteligente, 2013, Fortaleza. A Solution to Swarm Robot Escorting using the Probabilistic Lloyd Method, 2013.