UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE ENGENHARIA ELÉTRICA
BACHARELADO EM ENGENHARIA ELÉTRICA

ADRIANE ALESSI
JOÃO EDUARDO HOFFMANN

SISTEMA ADAPTATIVO DE IDENTIFICAÇÃO E RASTREAMENTO PARA CLASSIFICAÇÃO E LOCALIZAÇÃO DE OBJETOS EM TEMPO REAL BASEADO EM CÂMERAS PARA VEÍCULOS AUTÔNOMOS

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA
2020

ADRIANE ALESSI
JOÃO EDUARDO HOFFMANN

SISTEMA ADAPTATIVO DE IDENTIFICAÇÃO E RASTREAMENTO PARA CLASSIFICAÇÃO E LOCALIZAÇÃO DE OBJETOS EM TEMPO REAL BASEADO EM CÂMERAS PARA VEÍCULOS AUTÔNOMOS

Trabalho de Conclusão de Curso apresentado(a) como requisito parcial à obtenção do título de Bacharel(a) em Engenharia Elétrica, do Departamento Acadêmico de Engenharia Elétrica, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Max Mauro Dias Santos

PONTA GROSSA
2020

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEP. ACADEMICO DE ELETRONICA-PG

TERMO DE APROVAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO - TCC

SISTEMA ADAPTATIVO DE IDENTIFICAÇÃO E RASTREAMENTO PARA CLASSIFICAÇÃO E LOCALIZAÇÃO DE OBJETOS EM TEMPO REAL BASEADO EM CÂMERAS PARA VEÍCULOS AUTÔNOMOS

Por

Adriane Alessi e Joao Eduardo Hoffmann

Monografia apresentada às 14 horas do dia 11 de dezembro de 2020 como requisito parcial, para conclusão do Curso de Engenharia Elétrica da Universidade Tecnológica Federal do Paraná, Câmpus Ponta Grossa. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação e conferidas, bem como achadas conforme, as alterações indicadas pela Banca Examinadora, o trabalho de conclusão de curso foi considerado APROVADO.

Banca examinadora:

Prof. Cristhiane Gonçalves	Membro
Prof. Fernanda Cristina Côrrea	Membro
Prof. Max Mauro Dias Santos	Orientador
Prof. Josmar Ivanqui	Professor responsáv

Documento assinado eletronicamente por (Document electronically signed by) **MAX MAURO DIAS SANTOS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 11/12/2020, às 15:08, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por (Document electronically signed by) **FERNANDA CRISTINA CORREA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 11/12/2020, às 15:11, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por (Document electronically signed by) **CRISTHIANE GONCALVES**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em (at) 11/12/2020, às 15:11, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por (Document electronically signed by) **JOSMAR IVANQUI**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em (at) 12/12/2020, às 08:41, conforme horário oficial de Brasília (according to official Brasilia-Brazil time), com fundamento no (with legal based on) art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site (The authenticity of this document can be checked on the website) https://sei.utfpr.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador (informing the verification code) **1787051** e o código CRC (and the CRC code) **061D1452**.

Referência: Processo nº 23064.021957/2020-01 SEI nº 1787051

Eu, Adriane, dedico este trabalho ao meu pai Amilton e minha mãe Cleci por tudo que fizeram por mim, me apoiando de todas as formas para chegar onde estou. Também às minhas irmãs por serem meu porto seguro, dando conselhos em todas as minhas dificuldades.

Eu, João, dedico este trabalho ao meu pai Jarbas e minha avó Anna, que sempre me apoiaram e incentivaram durante toda esta trajetória e em todas as minhas decisões.

Dedicamos este trabalho aos nossos amigos e as empresas Klabin e DAF, pelo suporte e aprendizado.

AGRADECIMENTOS

Agradecemos, primeiramente a Deus pelas experiências e oportunidades vividas em nosso período de graduação.

Este trabalho não poderia ser finalizado sem a ajuda de diversas pessoas e instituições às quais prestamos nossas homenagens. Dentre eles nossos pais, avós e amigos, que por todo esse tempo nos deram suporte para chegar até aqui, somos gratos por tê-los em nossa jornada.

Ao Prof. Orient. Max Mauro Dias Santos, que nos mostrou os caminhos a serem seguidos e pela confiança depositada.

Aos professores e colegas do departamento, que ajudaram de forma direta e indireta na conclusão deste trabalho.

Enfim, a todos os que de alguma forma contribuíram para a realização deste trabalho.

RESUMO

ALESSI, Adriane; HOFFMANN, João Eduardo. **Sistema Adaptativo de Identificação e Rastreamento para Classificação e Localização de Objetos em Tempo Real Baseado em Câmeras para Veículos Autônomos**. 2020. 56 f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Elétrica) – Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2020.

As conquistas recentes de visão computacional mostram um ganho proporcional na precisão de tarefas de localização e classificação de objetos e na complexidade das estruturas neurais. O foco contínuo na precisão exige equilíbrio de processamento para expandir o número de aplicações e usos de um determinado sistema e garantir sua versatilidade. Para esse propósito, duas abordagens de visão computacional para localização, detecção e rastreamento de objetos, são descritas e associadas. Com as atuais abordagens de visão computacional e redes neurais convolucionais consideradas, um algoritmo de vários estágios é implementado adotando o modelo neural YOLOv3, para detecção de objetos, e a abordagem hierárquica HART, para rastreamento de objetos. Além disso, é realizada a proposição de um sistema adaptável de mudança de estágio pós-detecção com base no número de objetos detectados nas imagens de sequência e no tempo de processamento relativo ao estágio. O sistema adaptável compõe um modelo capaz de apresentar ótimos resultados de desempenho com recursos de detecção e rastreamento de objetos. No final, apresenta-se uma avaliação comparativa de desempenho, utilizando o conjunto de dados KITTI, com a intenção de analisar e enfatizar o desempenho de cada estágio para sequências relacionadas a ambientes de tráfego. A avaliação, consistindo em métricas e definições de precisão e tempo de processamento, é realizada para configurações multiestágios de localização e rastreamento pós-localização com mudança adaptável de estágios.

Palavras-chave: Detecção. Rastreamento. YOLO. HART.

ABSTRACT

ALESSI, Adriane; HOFFMANN, João Eduardo. **Adaptive Identification and Tracking System for Classifying and Locating Real-Time Objects Based on Cameras for Autonomous Vehicles**. 2020. 56 p. Final Coursework (Bachelor's Degree in Electrical Engineering) – Federal University of Technology – Paraná. Ponta Grossa, 2020.

The achievements of recent computer vision proposals show a proportional gain of object classification and localization accuracy to the complexity of neural network structures. The continuous focus on precision demands balance in processing to expand the number of applications and uses of a given system and guarantee its versatility. For this purpose, two computer vision approaches for object localization, detection and tracking, are described and associated. With current computer vision approaches and convolutional neural networks considered, a multistage algorithm is implemented adopting the YOLOv3 neural model for object detection and the HART approach for object tracking. In addition, we propose a post-detection adaptable stage change system based on the number of detected objects in sequence images and stage-related processing time. The adaptable system composes a model capable of presenting optimal performance results with detection and tracking features. In the end, we present a comparative performance evaluation, using the KITTI dataset, with the intention of analyzing and emphasizing the performance of each stage for sequences related to traffic environments. The evaluation, consisting of average precision and processing time metrics and definitions, is performed for multistage configurations of localization and post-localization tracking with adaptive stage changes.

Keywords: Detection. Tracking. YOLO. HART.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS E ACRÔNIMOS

SIGLAS

ALFD	Descritor de Fluxo Local Agregado, do inglês *Aggregated Local Flow Descriptor*
ANN	Rede Neural Artificial, do inglês *Artificial Neural Network*
CNN	Rede Neural Convolucional, do inglês *Convolutional Neural Network*
CPU	Unidade Central de Processamento, do inglês *Central Processing Unit*
DFN	Rede de Fluxo Dorsal, do inglês *Dorsal Flow Network*
DL	Aprendizado Profundo, do inglês *Deep Learning*
DNN	Redes Neurais Profundas, do inglês *Deep Neural Network*
DPM	Modelo de Parte Deformável, do inglês *Deformable Part Model*
FFT	Transformada Rápida de Fourier, do inglês *Fast Fourier Transform*
FPS	Quadros por Segundo, do inglês *Frames per Second*
GOTURN	Rastreamento de Objeto Genérico Utilizando Redes de Regressão , do inglês *Generic Object Tracking Using Regression Networks.*
GPU	Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit*
HART	Rastreamento Recorrente, Atentivo e Hierárquico, do inglês *Hierarchical Attentive Recurrent Tracking*
HOG	Histograma de Gradientes Orientados, do inglês *Histogram of Oriented Gradients*
IoU	Intersecção sobre União, do inglês *Intersection over Union*
IPT	Trajetória de Ponto de Interesse, do inglês *Interest Point Trajectory*
KCF	Filtro de Correlação Kernelizado, do inglês *Kernelized Correlation Filter*
LSTM	Memória Longa de Curto Prazo, do inglês *Long Short-term Memory*
mAP	Precisão Média, do inglês *Mean Average Precision*
ML	Aprendizado de Máquina, do inglês *Machine Learning*
MLP	Perceptron Multicamadas, do inglês *Multilayer Perceptron*
MOT	Rastreamento de Múltiplos Objetos, do inglês *Multiple Object Tracking*
NMS	Não-Máxima Supressão, do inglês *Non-maximum Suppression*
NOMT	Rastreamento Aproximadamente Online de Múltiplos Alvos, do inglês *Near-Online Multi-target Tracking*
RATM	Modelo de Rastreamento Atentivo Recorrente, do inglês *Recurrent Attentive Tracking Model*
RGB	Vermelho, Verde e Azul, do inglês *Red, Green and Blue*
RoI	Regiões de Interesse, do inglês *Regions of Interest*
RPN	Rede de Proposta de Região, do inglês *Region Proposal Network*
SORT	Rastreamento Simples, *Online*, e em Tempo Real do inglês *Simple Online and Realtime Tracking*

SVM	Máquina de Vetores de Suporte, do inglês *Support Vector Machine*
SSD	Detector de Disparo Único, do inglês *Single Shot Detector*
UKF	Filtro de Kalman Desodorizado, do inglês *Unscented Kalman Filter*
VOC	Classes de Objetos Visuais, do inglês *Visual Object Classes*
YOLO	Você Só Olha Uma Vez, do inglês *You Only Look Once*

ACRÔNIMOS

COCO	Objetos Comuns em Contexto do inglês *Common Objects in Context*
ConvNet	Rede Neural Convolucional, do inglês *Convolutional Neural Network*
ImageNet	Conjunto de Dados de Imagem em Grande Escala, do inglês *Large Scale Image Dataset*
RADAR	Detecção e Alcance por Rádio, do inglês *Radio Detection and Ranging*
R-CNN	Rede Neural Convolucional Baseada em Região, do inglês *Region Based Convolutional Neural Network*
ResNet	Rede Neural Residual, do inglês *Residual Neural Network*
Soft-NMS	Não-Máxima Supressão Suave, do inglês *Soft Non-maximum Suppression*
KITTI	*Instituto Tecnológico de Carlsrue e Instituto Tecnológico da Toyota*, do inglês *Karlsruhe Institute of Technology and Toyota Technological Institute*
LiDAR	Detecção e Alcance por Laser, do inglês *Laser Detection and Ranging*

SUMÁRIO

1 INTRODUÇÃO

Os sistemas autônomos de percepção veicular envolvem respostas visuais e sensoriais interconectadas ao meio ambiente, exigindo uma complexa tomada de decisão em termos de segurança humana nas mais diversas situações em tempo real. Neste contexto, tarefas de visão computacional, como classificação e localização de objetos, tornaram-se essenciais para processos autônomos, tendo em vista que o conhecimento de objetos e suas interações suportam previsões dinâmicas e servem como entrada para abordagens sensoriais integradas.

Nas últimas décadas, várias abordagens de visão computacional foram desenvolvidas para realizar o reconhecimento de objetos (REDMON; FARHADI, 2018; REN et al., 2017; BEWLEY et al., 2016; KOSIOREK; BEWLEY; POSNER, 2017; KRIZHEVSKY; SUTSKEVER; HINTON, 2012; GEIGER et al., 2013; EVERINGHAM et al., 2010; LIN et al., 2014). Isso inclui a detecção de objetos, um subconjunto de reconhecimento de objetos, onde o problema se estende à classificação e localização de objetos, permitindo que vários objetos sejam identificados e localizados na mesma imagem. Recentemente, as abordagens de localização de objetos, com base nas informações das sequências anteriores, mostraram-se consistentes e com previsões rápidas, garantindo o uso de um único modelo e entradas de imagem de sequência, com várias entradas de localização, para realizar o rastreamento de vários objetos.

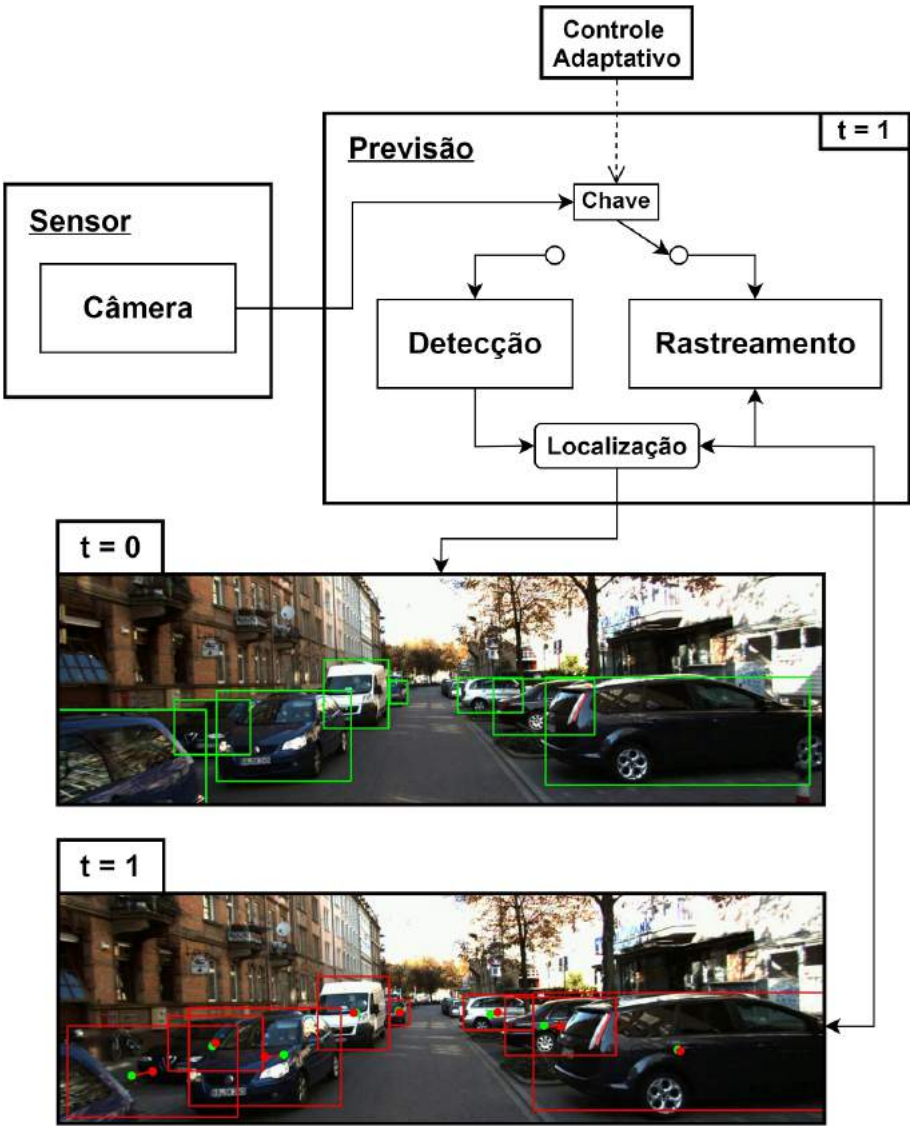
O problema de visão computacional do rastreamento de objetos está associado à presença de movimento de objetos em uma sequência de imagens. Um modelo de rastreamento ideal localiza um objeto, dadas suas características gerais definidas em uma região delimitada, lidando com desafios relacionados à localização dinâmica de objetos, como desaparecimento, oclusão, deformidades, variações de escala e alterações de iluminação.

Os avanços nas aplicações de visão computacional aumentaram o impacto de novas tecnologias no meio ambiente, levando à busca de propostas cada vez mais precisas para garantir melhor confiabilidade nos resultados da percepção ambiental. Considerando o impacto potencial dessas tecnologias, a busca por processos mais rápidos se torna essencial para aplicações em tempo real. Consequentemente, com base em investigações recentes sobre abordagens de detecção e rastreamento de objetos, e sua adequação a sistemas em tempo real, propomos um modelo multiestágios para classificação e localização de objetos. O modelo contém um mecanismo de mudança adaptativa de estágios para garantir precisão mais próxima de qualquer sistema de detecção avançado com tempo de processamento apropriado para atuação em tempo real.

O desempenho desejado é obtido aproveitando os estágios de rastreamento

em situações em que o número de objetos localizados é menor que um determinado limite de objetos localizados. Os resultados são demonstrados usando uma avaliação comparativa de desempenho com configurações contrastantes do modelo. Na Figura 1 é apresentado o modelo desempenhando a localização e o rastreamento pós-localização de múltiplos objetos utilizando uma abordagem de vários estágios com opção de configuração adaptativa para fins de tempo real. Para o quadro com os resultados de localização, $t = 0$, as caixas verdes contêm os objetos localizados por detecção ou rastreamento. Para $t = 1$, as caixas delimitadoras vermelhas contêm objetos rastreados com informação temporal, círculos verdes indicam centroides de quadro de sequência anterior e círculos vermelhos indicam centroides de objetos rastreados.

Figura 1 – Modelo de classificação e localização de objetos proposto.



Fonte: Adaptado de SANTOS et al. (2020).

1.1 MOTIVAÇÃO

Com o aumento do impacto de novas tecnologias no meio ambiente, modelos recorrentes de previsão são desenvolvidos com grande enfoque na precisão de resultados para dado sistema, tornando maior o tempo de processamento na entrega de previsões. Torna-se, portanto, significativo o desenvolvimento de estratégias e abordagens com enfoque no balanço de recursos computacionais e versatilidade de aplicações.

1.2 OBJETIVO GERAL

O objetivo desse trabalho é estudar os ganhos apresentados por um modelo multiestágios adaptativo de classificação e localização em tempo real em comparação com um modelo recorrente de detecção.

1.3 OBJETIVO ESPECIFICO

O objetivo específico desse trabalho é apresentar o desenvolvimento e implementação de um modelo multiestágios adaptativo de classificação e localização de objetos em tempo real para veículos autônomos.

1.4 JUSTIFICATIVA

Para que exista confiabilidade em um sistema autônomo com propósitos em tempo real são necessários precisão e baixo tempo de processamento na realização de interações. Um modelo multiestágios versátil proporciona escolhas variadas de configuração para garantir precisão e tempo de processamento adequado aos mais diversos sistemas e aplicações.

1.5 ORGANIZAÇÃO DO TRABALHO

O restante deste manuscrito está estruturado da seguinte forma. A Seção 2 discute investigações nas tarefas de detecção e rastreamento, quanto à classificação e localização de objetos. A Seção 3 apresenta uma visão geral das características, arquitetura e do sistema de mudança adaptativa de estágios do modelo proposto. A Seção 4 destaca a definição de métricas, apresentação e análise da avaliação comparativa experimental. Finalmente, a Seção 5 resume a relevância da metodologia desenvolvida neste trabalho.

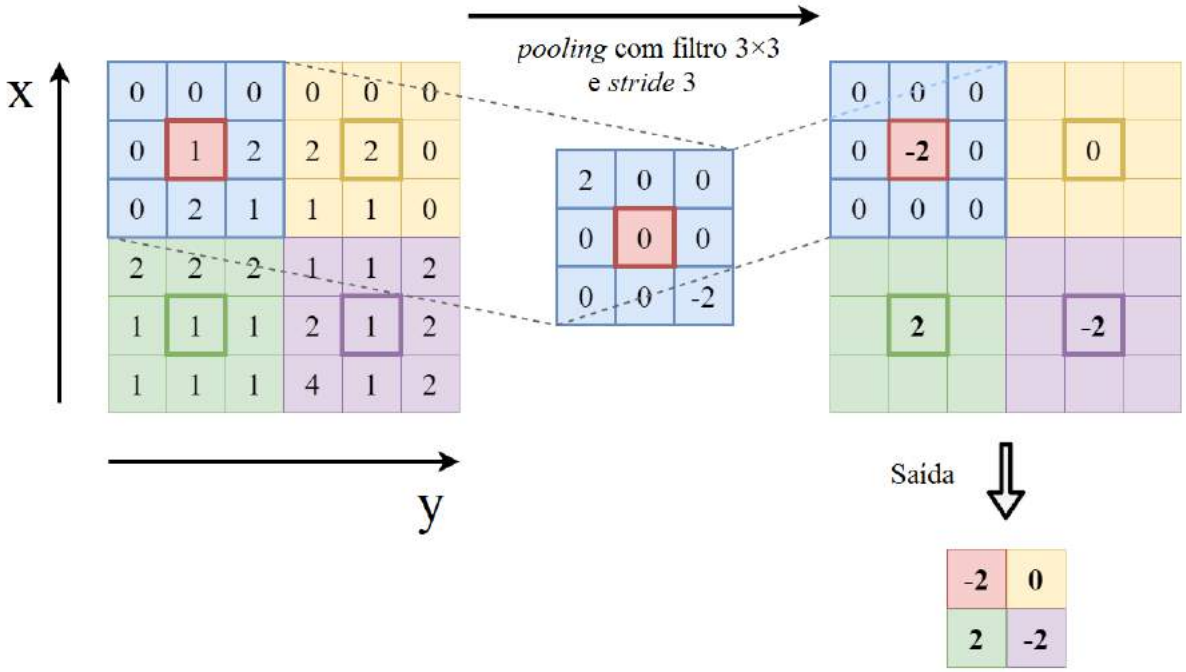
2 REVISÃO DA LITERATURA

2.1 REDES NEURAIS CONVOLUCIONAIS

Considerando a importância da migração de Redes Neurais Artificiais (ANNs) para Redes Neurais Profundas (DNNs) quanto a capacidade de manipulação de grandes quantidades de dados, proporcional ao aumento progressivo do número de camadas e consequentemente à complexidade de arquiteturas para tarefas de previsão. Surge a necessidade de reformular a estrutura geral de aprendizado de forma a otimizar o processo de previsão para matrizes de entrada, utilizando operações de convolução para cálculos derivados do Aprendizado Profundo (DL) clássico e o devido redimensionamento matricial em cada etapa do processo. As Redes Neurais Convolucionais (CNNs) tornam-se, portanto, fundamentais na resolução de problemas recorrentes no contexto de Aprendizado de Máquina (ML) (ALBAWI, 2012).

A entrada de uma CNN é representada por uma matriz de convolução em que os dados são constituídos pelos *pixels* brutos da imagem. As regiões desejadas da matriz de entrada realizam operações com outras matrizes, também chamadas de filtros, estas regiões de ligação entre matrizes são conhecidas como camadas de conexão. As ligações podem ser realizadas entre regiões de cada matriz ou de forma completa, conectando todos os pontos de entrada a pontos de saída. Os filtros e as conexões realizadas entre as matrizes são de escolha do desenvolvedor da arquitetura de detecção e classificação de objetos. Dentre as operações, *pooling* é responsável por apresentar matrizes de resposta com dimensão inferior à entrada retendo as informações de importância, tendo em vista a execução da operação *stride*, que utiliza o conceito de janela deslizante, eliminando informações desnecessárias que apresentem características semelhantes de nós vizinhos (ALBAWI, 2012). A Figura 2 demonstra o funcionamento das operações de convolução, *pooling* e *stride*. As regiões de *stride* são representadas pelas cores azul, amarela, verde e roxa, e o *pixel* resultante de uma operação está representado em vermelho.

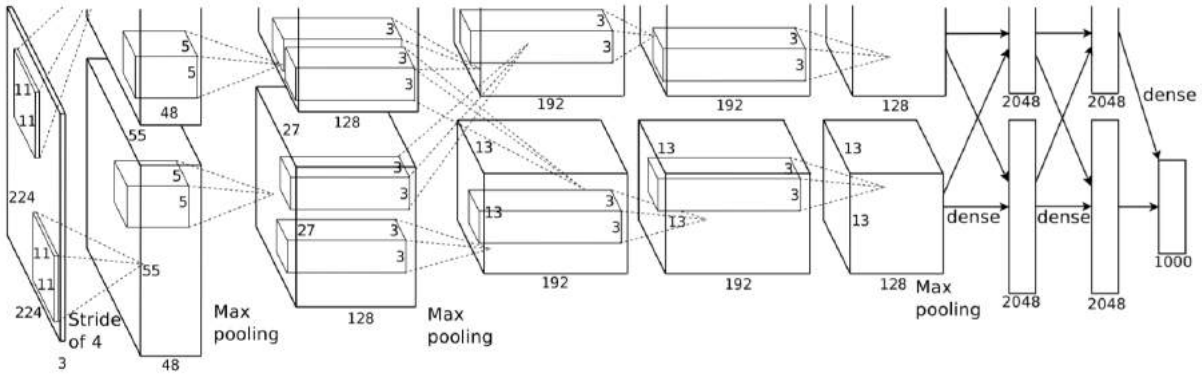
Figura 2 – Operações de convolução, *pooling* e *stride*.



Fonte: Autoria própria.

As camadas de conexão total, do inglês *fully-connected layer*, tem a função de organizar todos os nós de entrada na forma vetorial, em que cada componente apresenta uma informação. A camada final de conexão total é um vetor que contém informações de classe do objeto, pontuações e probabilidades (ALBAWI, 2012). A Figura 3 demonstra todas as etapas em uma arquitetura de CNN, o processamento é feito de forma compartilhada entre duas unidades de processamento gráfico (GPUs).

Figura 3 – Arquitetura de rede neural de convolução.

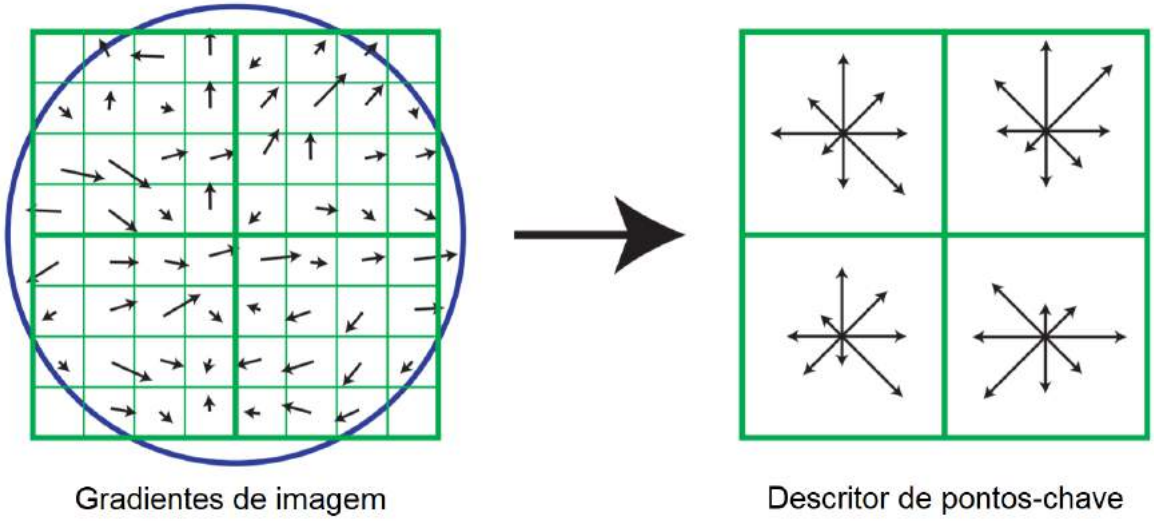


Fonte: KRIZHEVSKY, SUTSKEVER e HINTON (2012).

2.2 DETECÇÃO DE OBJETOS

Antes do advento das Redes Neurais Convolucionais (CNNs), a detecção de objetos esteve relacionada a descritores e processos focados na extração de recursos. Os Histogramas de Gradientes Orientados (HOGs) são uma reminiscência de histogramas de orientação de arestas, descritores de pontos-chave (LOWE, 2004) e contextos de forma (DALAL; TRIGGS, 2005). Com a adição de estágios iniciais, relacionados à normalização do contraste da imagem local e ao denso cálculo de gradiente celular, os descritores HOG tornaram inteligível a relevância das características das bordas, regiões com mudanças bruscas de intensidade, explícitas na magnitude dos gradientes, para a extração de informações úteis. Na Figura 4 é apresentado um descritor de pontos-chave, criado ao calcular a magnitude e a orientação do gradiente em cada ponto da imagem de amostra.

Figura 4 – Descritor de pontos-chave.

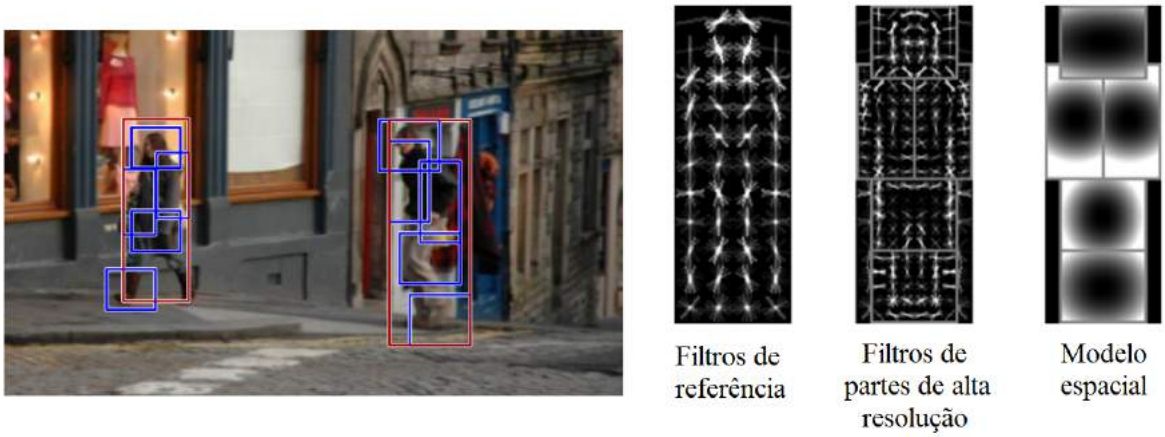


Fonte: Adaptado de LOWE (2004).

Na última década, várias abordagens foram propostas e demonstradas por diferentes pesquisadores na detecção de objetos em primeiro plano, o principal desafio na detecção de objetos em movimento é estimar a sua região de delimitação com mais precisão (REDMON; FARHADI, 2018). Considerando a existência de múltiplos objetos em imagens, FELZENSZWALB e HUTTENLOCHER (2004) propuseram um algoritmo com a capacidade de segmentar uma imagem em regiões potenciais usando uma representação gráfica. A dissimilaridade entre regiões foi associada a um conjunto de arestas ponderadas e conexão de vértices em uma imagem, apresentando uma quantificação de dissimilaridade relacionada aos atributos locais dos *pixels*. Os componentes são então atribuídos como conectados, definindo regiões por similaridade de *pixels*.

Após o desenvolvimento de estratégias de segmentação da informação inicial em regiões em potencial, os sistemas passaram a redirecionar classificadores para executar a detecção, estes classificadores são adicionados ao objeto desejado, o qual passa por uma avaliação, considerando vários locais e escalas, em uma imagem de entrada (HE et al., 2016). Sistemas como os Modelos de Parte Deformável (DPMs), utilizam de uma abordagem de janelas deslizantes em que o classificador é executado em locais espaçados uniformemente ao longo de toda a imagem analisada (HE et al., 2016). Considerando a existência de múltiplos DPMs, o modelo de HE et al. (2016), se destaca por apresentar classificadores de objetos em uma imagem com diferentes resoluções e uma previsão de objetos com diferentes formas, possível, em virtude do uso de modelos de referência, treinados com o uso de múltiplas imagens teste (FELZENSZWALB et al., 2010). Os modelos, são caracterizados por representarem o objeto em sua totalidade, os filtros de partes de alta resolução, são relativos ao objeto e possuem partes delimitadas no modelo, enquanto o modelo espacial apresenta o grau de mobilidade de cada parte (FELZENSZWALB et al., 2010). Os filtros de referência utilizados no DPM de HE et al. (2016), apresentam uma estrutura obtida na forma matricial no padrão de HOGs (FELZENSZWALB et al., 2010). A Figura 5 apresenta os modelos do sistema DPM.

Figura 5 – Modelos (DPM) utilizados na detecção de pessoas em imagens.

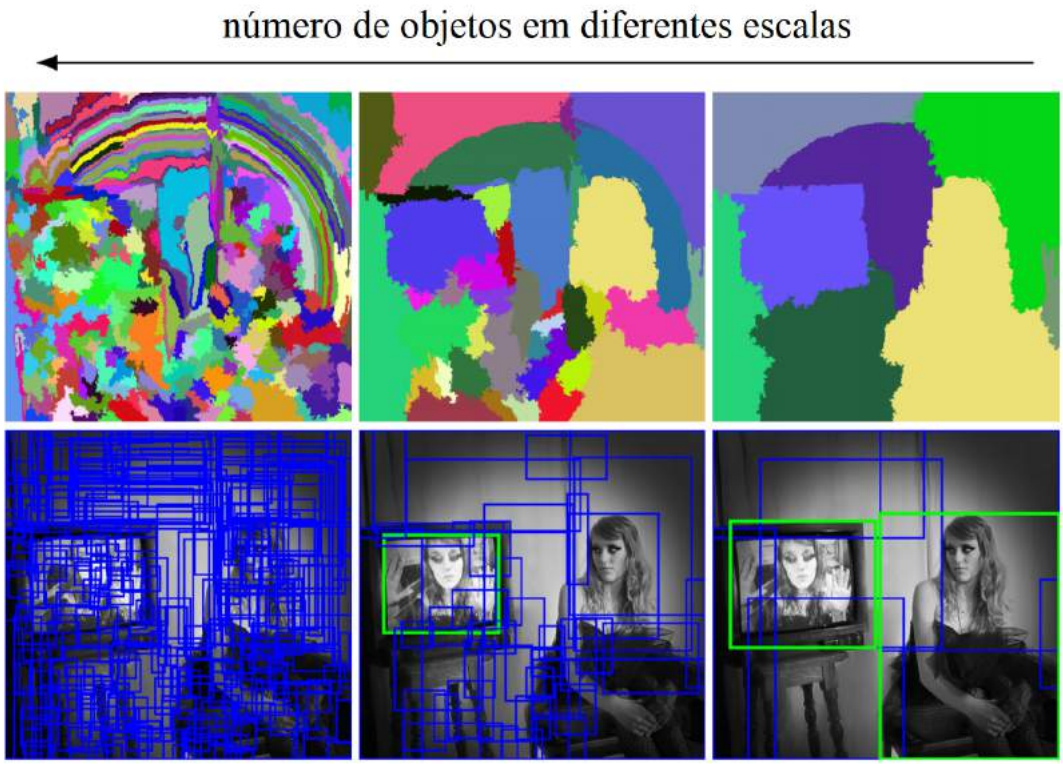


Fonte: Adaptado de FELZENSZWALB et al. (2010).

Construída com base nas contribuições anteriores, a Pesquisa Seletiva (UIJLINGS et al., 2013) utiliza um método de segmentação (FELZENSZWALB; HUTTENLOCHER, 2004), considerando várias escalas para objetos e fazendo uso de um algoritmo hierárquico, para criar regiões iniciais e realizar o agrupamento iterativo contínuo de regiões, ao calcular, repetidamente, similaridades entre regiões próximas, guiando à obtenção de uma única região. O objetivo do sistema é alcançado com a geração

de um conjunto independente de classes para locais de objetos. A Figura 6 demonstra a necessidade de considerar um grande número de regiões e escalas para delimitar objetos com diferentes dimensões.

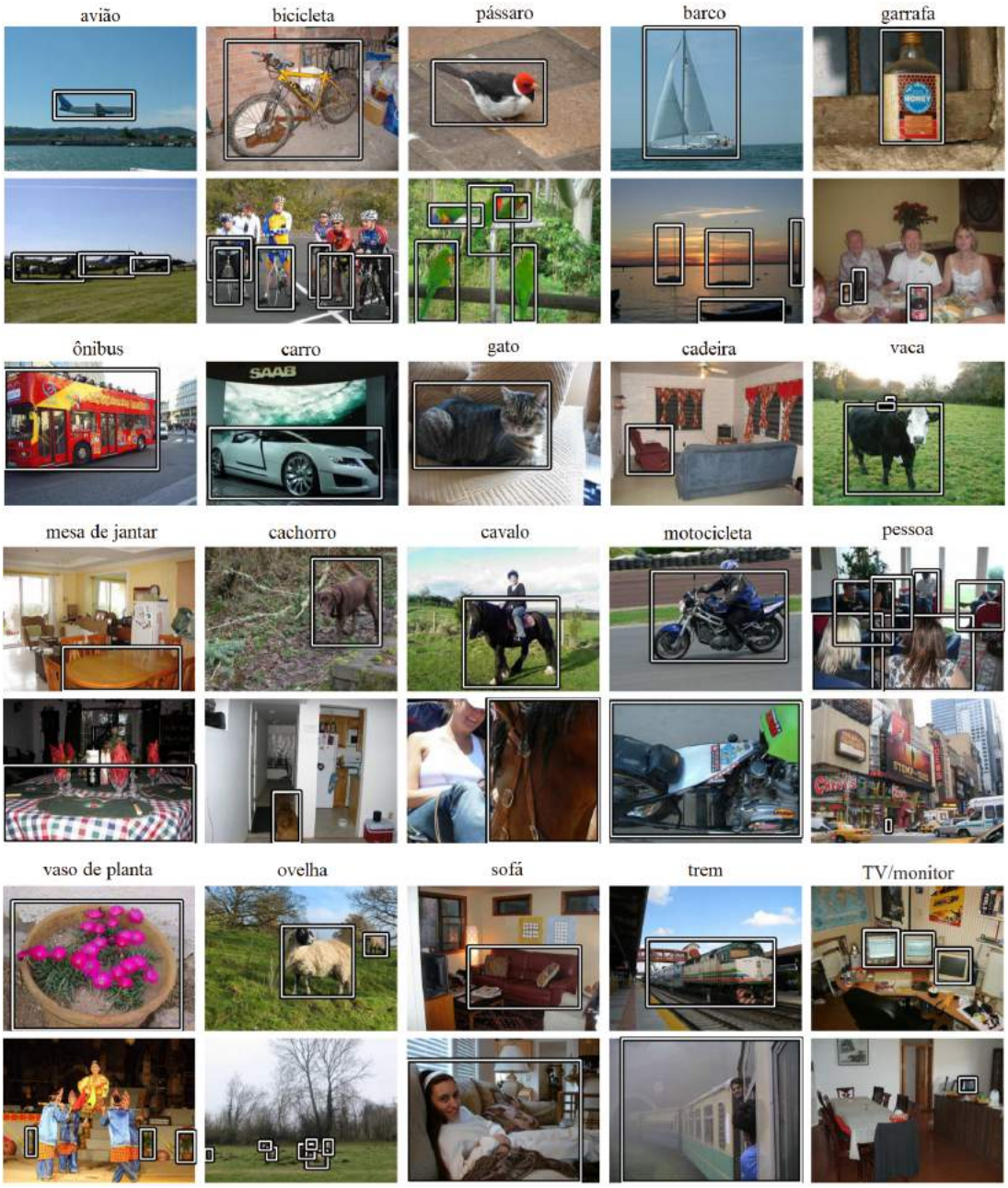
Figura 6 – Diferenciação em escalas presente na pesquisa seletiva.



Fonte: Adaptado de UIJLINGS et al. (2013).

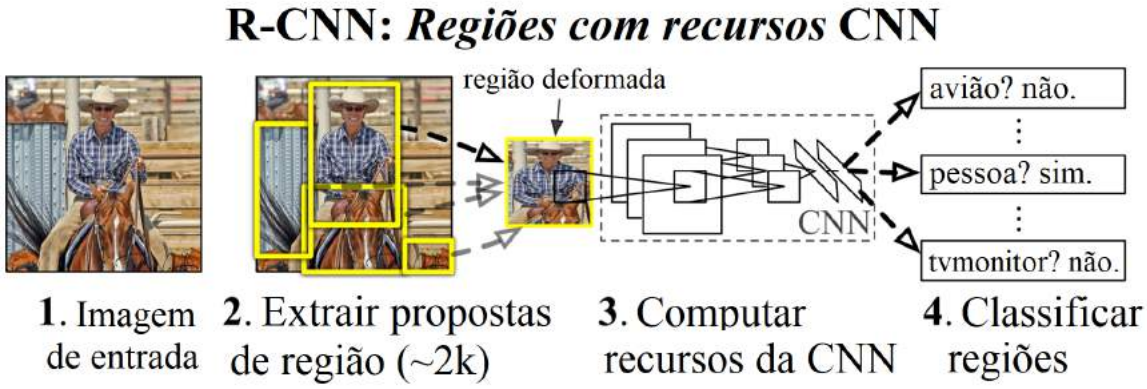
Com o recente desenvolvimento do Aprendizado Profundo (DL), e apresentação dos ganhos relativos na classificação de imagens com precisão substancialmente mais alta (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), relacionados ao treinamento de CNNs em grandes bases de dados, GIRSHICK et al. (2014) desenvolveram o modelo R-CNN, baseado em regiões de interesse e estruturado com a combinação da pesquisa seletiva por propostas de região e extração de recursos da CNN para classificar as regiões independentes. O algoritmo desenvolvido apresentou uma melhoria relativa de 30% em relação aos melhores resultados da avaliação PASCAL VOC (EVERINGHAM et al., 2010). A Figura 7 apresenta imagens rotuladas da base de dados de treinamento PASCAL VOC, na qual a abordagem R-CNN foi treinada. A Figura 8 apresenta o modelo R-CNN, o sistema (1) recebe uma imagem de entrada, (2) extrai cerca de 2000 propostas de região, (3) calcula recursos para cada proposta, fazendo uso de uma rede neural convolucional profunda (CNN), e então (4) classifica cada região, utilizando SVMs específicas à classe.

Figura 7 – Imagens rotuladas de exemplo na base de dados VOC2007.



Fonte: Adaptado de EVERINGHAM et al. (2010).

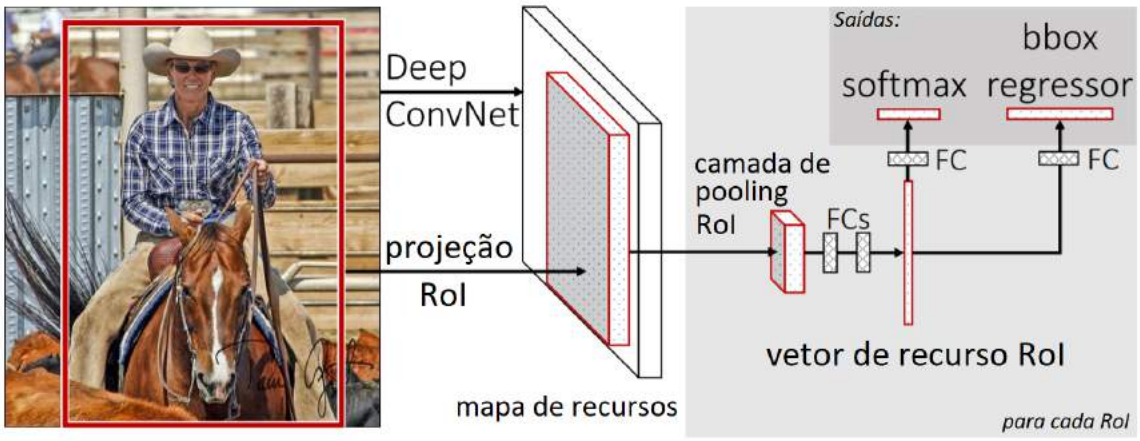
Figura 8 – Modelo R-CNN.



Fonte: Adaptado de GIRSHICK et al. (2014).

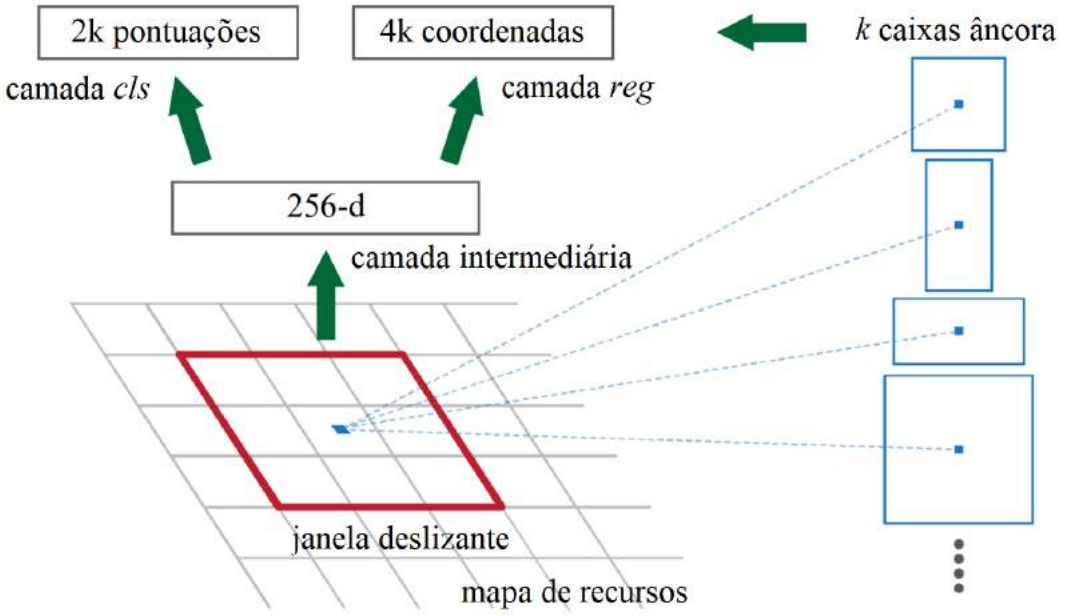
Girshick et al. (GIRSHICK, 2015) aprimorou o treinamento e os testes gerando regiões de interesse (RoI) como resultado de um processo da CNN. Uma camada de *pooling* das RoI é aplicada em todas as regiões para remodelá-las como novas entradas, corretamente dimensonadas para servirem como entrada de redes neurais totalmente conectadas. Ren et al. (REN et al., 2017) implementaram melhorias adicionais no processo, definindo um mapa de recursos, resultado de um processo neural inicial, como entrada da camada de *pooling* das RoI e uma rede de proposta de região (RPN), resultando em novas propostas para objetos e suas pontuações relativas em cada posição, como mostrado na Figura 9 e a Figura 10.

Figura 9 – Arquitetura do modelo *Fast* R-CNN.



Fonte: Adaptado de GIRSHICK (2015).

Figura 10 – Rede de proposta de região (RPN).



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

Após o desenvolvimento e aprimoramento dos detectores de dois estágios, o foco em sistemas em tempo real induziu a busca de novas arquiteturas de detecção de objetos, capazes de apresentar resultados precisos em menor tempo. REDMON et al. (2016) desenvolveu o modelo "*You Only Look Once*" (YOLO), uma arquitetura unificada para previsões de caixas delimitadoras e pontuações de classes relativas. A estrutura unificada é capaz de detectar objetos ao dividir a imagem de entrada em uma célula de grade, representando individualmente um possível centro para objetos e executando etapas de reconhecimento por meio de uma única inferência de rede.

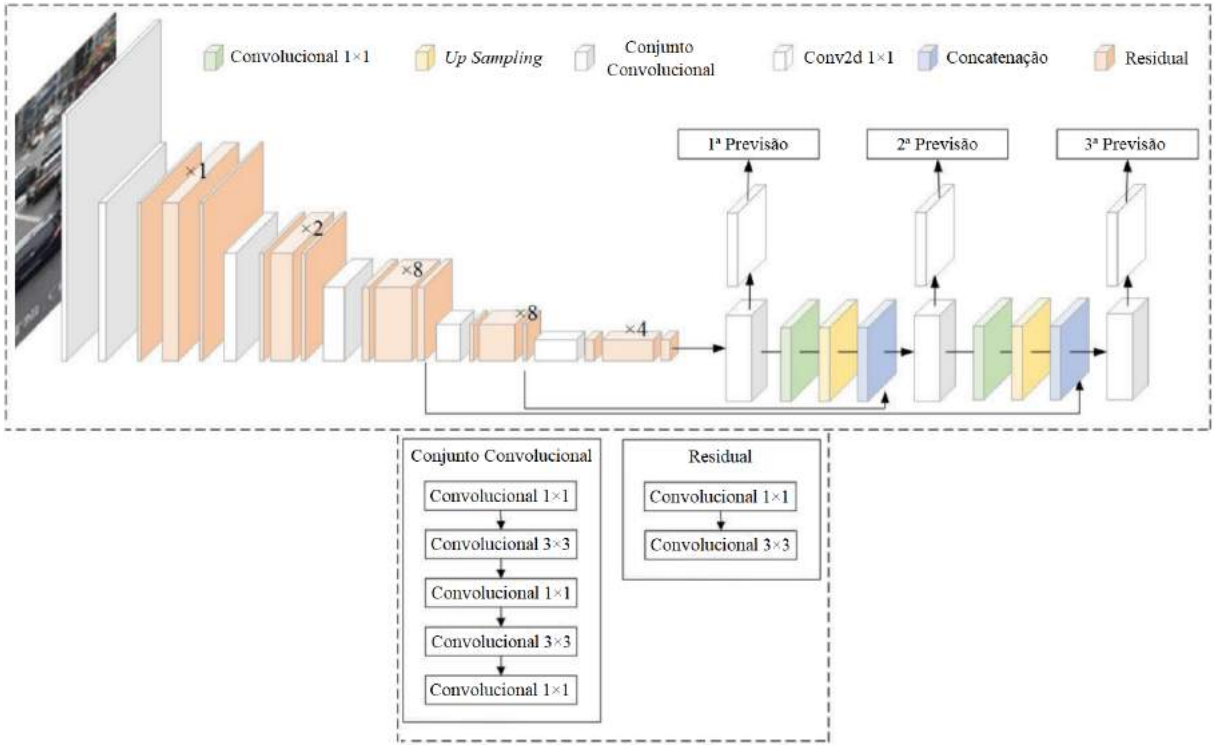
Inferências de rede mais rápidas foram obtidas com o *Single Shot MultiBox Detector* (SSD) (LIU et al., 2016), desenvolvido como uma abordagem simples de previsão de caixas delimitadoras, contando com deslocamentos de forma e confidências de classe. O equilíbrio na precisão foi alcançado na aplicação de filtros, para diferentes detecções de proporção, caracterizando mapas de entrada para fases posteriores da rede, com o objetivo de aumentar a precisão para previsões de objetos de pequena dimensão, dificuldade proveniente da baixa capacidade de generalização de objetos por classe, consequência da menor otimização de pesos no treinamento de redes com menos camadas.

Posteriormente, Redmon e Farhadi (REDMON; FARHADI, 2017) propuseram melhorias ao modelo, adicionando normalização em lote, *batch normalization*, a todas as camadas convolucionais e melhorias na convergência, desempenhando, adicionalmente, um pré-treinamento na rede de classificação ImageNet por dez *epochs*, visando melhorias de confiabilidade às previsões. Além disso, o modelo YOLOv2 utiliza

de camadas convolucionais para prever a localização de caixas âncora, em contraste ao uso único de camadas totalmente conectadas para prever caixas delimitadoras nos modelos anteriores, resultando no aumento da capacidade de revocação do modelo.

Outras melhorias, voltadas à previsão de pontuações de confiança para objetos, foram impostas ao modelo (REDMON; FARHADI, 2018), como o uso de regressão logística, classificadores logísticos independentes e a substituição da camada Softmax, responsável por extrapolar a última camada da rede em valores de confiança. As mudanças estruturais resultaram em uma nova arquitetura para extração de recursos, chamada Darknet-53 (REDMON; FARHADI, 2018), contando com a adição de blocos residuais inspirados nas contribuições da proposta ResNet (HE et al., 2016), facilitando o treinamento de Redes Neurais Profundas (DNNs). Além disso, uma previsão em três escalas inspirada nas pirâmides de recursos é utilizada para aumentar o número de previsões para cada candidato a objeto. O modelo conclui por generalizar representações de objetos em um único processo de detecção, exibindo desempenho de detecção semelhante aos modelos R-CNNs com velocidades de inferência apropriadas para aplicações em tempo real. A Figura 11 apresenta o modelo YOLOv3, fazendo uso da arquitetura de rede Darknet-53 e três previsões em escala.

Figura 11 – Estrutura do modelo YOLOv3.



Fonte: Adaptado de MAO et al. (2019).

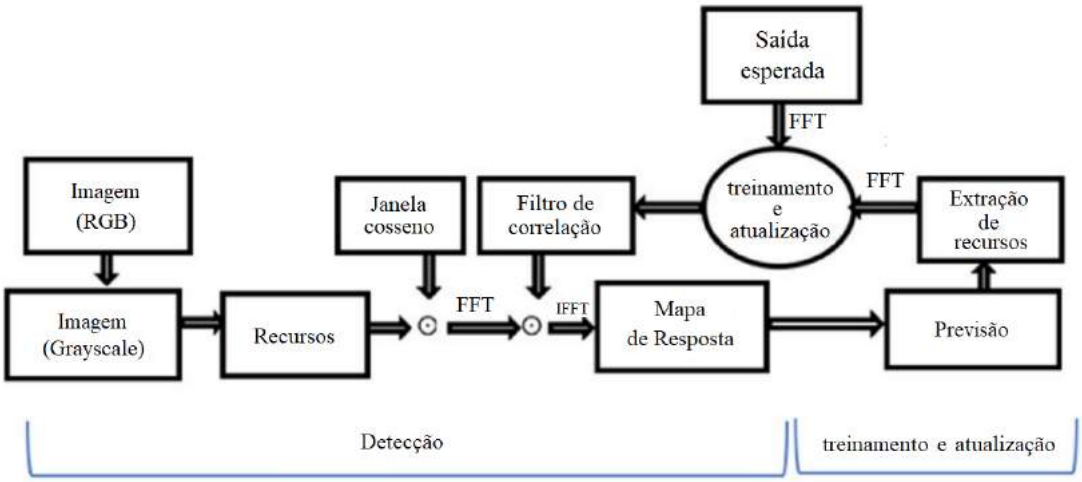
2.3 RASTREAMENTO DE OBJETOS

Inicialmente, as propostas e abordagens, voltadas ao desempenho do rastreamento de objetos, foram baseadas no movimento de objetos em sequências de imagens. O algoritmo de rastreamento de centroides foi uma das primeiras propostas matemáticas ao rastreamento e foi estruturado como um processo de várias etapas, caracterizado pela atribuição de caixas delimitadoras e identificadores para calcular o centroide de determinado objeto. Nos quadros de sequência posteriores, os identificadores são atribuídos com base na distância entre os centroides que apresentam o mesmo identificador. Melhorias adicionais foram alcançadas com a implementação do filtro de Kalman (KALMAN, 1960), que permitiu o rastreamento de movimento baseado na distância euclidiana para previsões de posição e velocidade, melhorando o desempenho geral do modelo.

Um algoritmo amplamente utilizado no rastreamento robusto e em tempo real é o Filtro de Correlação Kernelizado (KCF) (HENRIQUES et al., 2015), algoritmo desenvolvido para fins de tempo real. Foi amplamente utilizado no campo da visão computacional, graças à sua alta velocidade na previsão e robustez de processamento. O algoritmo é capaz de apresentar resultados adequados e em alta velocidade devido à filtragem de correlação, que veio do campo de processamento de sinais, e ao uso da Transformada Rápida de Fourier (FFT) no domínio da frequência, o que melhora notavelmente a velocidade de computação (YADAV; PAYANDEH, 2018). Uma imagem RGB é convertida em escala de cinza e, em seguida, o KCF emprega HOGs para calcular o recurso desejado nos conteúdos extraídos anteriormente, instituindo valores de peso.

A correlação de *kernels* é usada para determinar o novo local e o cálculo do mapa de resposta da correlação e conteúdo de destino estimado são obtidos com o uso de uma FFT. Para obter o mapa de resposta no domínio espacial, é feito uso da Transformada Inversa e Rápida de Fourier, concluindo com a extração de recursos e atualização dos filtros de correlação, para, então, atualizar o resultado de rastreamento e aparência do destino no novo local previsto. Apesar de sua velocidade e precisão, o modelo KCF demonstra algumas limitações. Devido a uma estrutura sem recursos de aprendizado, os resultados podem apresentar delimitações inadequadas de um objeto sendo rastreado em ambientes dinâmicos (YADAV; PAYANDEH, 2018), impossibilitando aplicações exigentes em tempo real, como pode ser visto na Figura 12.

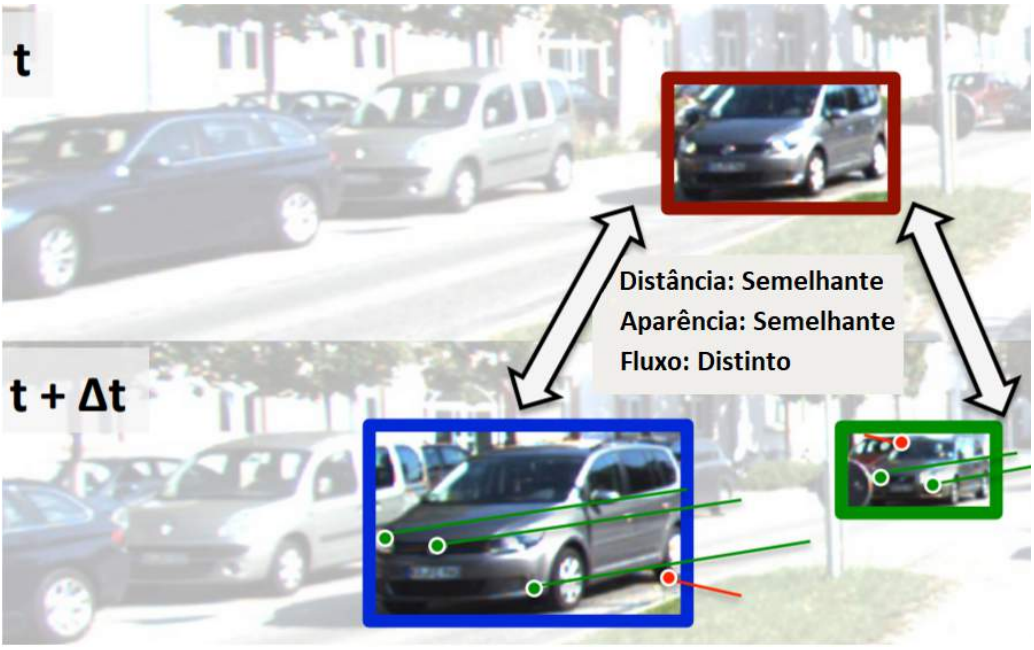
Figura 12 – Diagrama de blocos do algoritmo KCF.



Fonte: Adaptado de YADAV e PAYANDEH (2018).

Com o surgimento de algoritmos *online* e globais, capazes de rastrear objetos com considerações de período atual e através associações de informações de longo prazo, respectivamente, o Rastreamento Aproximadamente *Online* de Múltiplos Alvos (NOMT) (CHOI, 2015) foi proposto com apoio em ambas as frentes de rastreamento para prever confianças de identidade de objetos, temporalmente distantes, ao adicionar dados aos quadros de tempo iterativamente em cada quadro processado com informações temporais de forma bidirecional. O Descritor de Fluxo Local Agregado (ALFD) age para relacionar duas caixas delimitadoras temporariamente distantes com o auxílio de Trajetórias de Pontos de Interesse (IPTs), relativas à trajetórias de fluxo óptico. A afinidade entre duas caixas é calculada pelo produto entre um parâmetro treinado em sobreposições de caixas temporalmente distantes e o ALFD. A implementação do método evidenciou a capacidade, a partir de sistemas de discriminação de identidade, de corrigir erros simples de associação, comuns nas primeiras iterações, ao realizar um acréscimo de associações globais, identificando e relacionando novos alvos no processo. A Figura 13 apresenta duas possíveis associações temporais com a abordagem NOMT, entre t e $t + \Delta t$, e os parâmetros de discriminação de identidade.

Figura 13 – Associação temporal da abordagem NOMT.

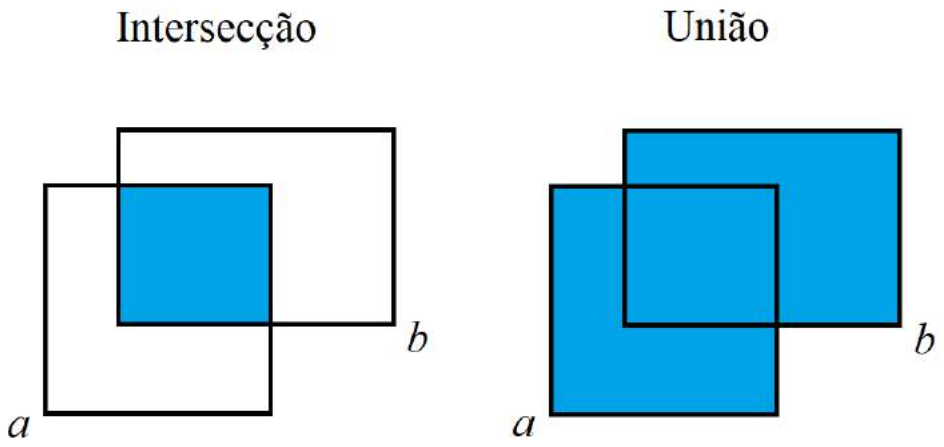


Fonte: Adaptado de CHOI (2015).

O Rastreamento Simples, *Online* e em Tempo Real (SORT) (BEWLEY et al., 2016) é uma abordagem de rastreamento de vários objetos para aplicações em tempo real. A arquitetura de rastreamento foi desenvolvida como uma combinação de técnicas de rastreamento de movimento, como Filtros de Kalman, para previsão de movimento, e o Algoritmo Húngaro (KUHN, 2010), empregado para lidar com associações de identificadores usando distâncias de Intersecção sobre União (IoU) entre caixas delimitadoras previstas e de referência. A abordagem alcança resultados precisos quando aplicada a objetos com baixa incerteza de estado e alta generalização, em comparação com os dados usados para o treinamento. A falta de recursos de atenção para discriminar o estado dos objetos gera deficiências no rastreamento por meio de oclusões. Com base nas deficiências apresentadas pela abordagem original, melhorias foram implementados adicionando, uma CNN para discriminação de aparência profunda, mudanças na estrutura de Kalman para filtrar através de $\mathbf{k}$ quadros de rastreamento e uma nova maneira de incorporar a distância entre os estados previstos e as novas medições. Com as otimizações, *Deep* SORT (WOJKE; BEWLEY; PAULUS, 2017) apresenta uma melhoria na redução de trocas de identidade, mantém a exatidão e precisão da previsão, e reduz o tempo de execução da abordagem como resultado de uma implementação do descritor de atenção profunda. A Equação 1 e a Figura 14, relativas ao cálculo de IoU, considerando caixas delimitadoras previstas $\boldsymbol{a}$ e de referência $\boldsymbol{b}$, são apresentadas a seguir.

$$\mathrm{IoU}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cap \boldsymbol{b}}{\boldsymbol{a} \cup \boldsymbol{b}} = \frac{\text{área de sobreposição}}{\text{área de união}} \tag{1}$$

Figura 14 – Representação de Intersecção e União.



Fonte: Autoria própria.

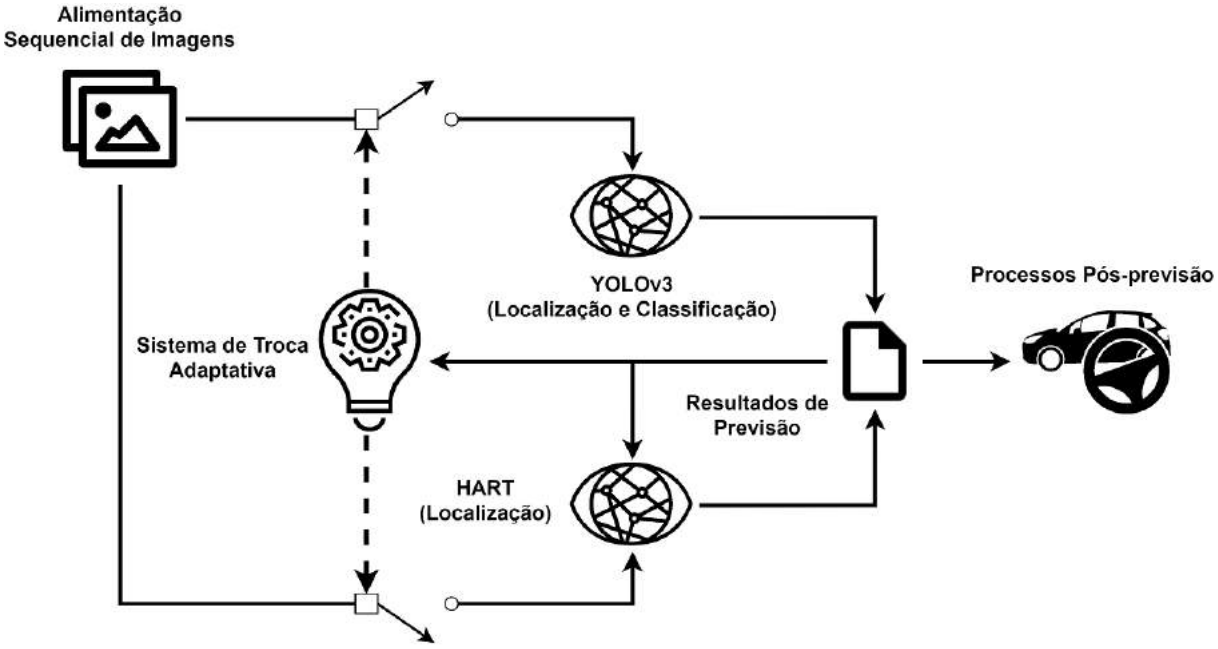
Com os avanços das pesquisas e implementações no campo da computação visual, HELD, THRUN e SAVARESE (2016) propuseram o método de Rastreamento de Objeto Genérico Utilizando Redes de Regressão (GOTURN) para o rastreamento individual de objetos, utilizando uma CNN para o aprendizado de padrões de objetos. O método foi definido por uma abordagem baseada em regressão, que permitiu a localização do objeto em uma única execução de *feed-forward* de rede, realizada facilitando o aprendizado de uma relação genérica entre aparência e movimento de um objeto em potencial. A rede foi treinada com dados de quadro sequenciais, representadas por regiões de interesse em quadros anteriores, contendo o objeto alvo, e regiões de pesquisa no quadro atual para localização.

A abordagem de Rastreamento Recorrente, Atentivo e Hierárquico (HART) (KOSIOREK; BEWLEY; POSNER, 2017) foi desenvolvida com recursos de arquitetura de vários estágios, inspirados na percepção visual humana e nos sistemas de cognição. O modelo proposto explora a propriedade dos mecanismos de atenção visual para criar laços de *feedback* para agrupamento de recursos, diferenciação do plano de imagem e preenchimento perceptivo na previsão de um vislumbre inicial da localização, mapa de localização e localização final (KOSIOREK; BEWLEY; POSNER, 2017; STOLLENGA et al., 2014). A abordagem apresenta custos computacionais desejáveis em tempo real e interpretabilidade devido a considerações de recursos do ambiente, contidos na imagem, nos estágios do mapeamento de localização, aumentando o potencial de rastreamento de objetos em situações desafiadoras e dinâmicas (KOSIOREK; BEWLEY; POSNER, 2017),

3 MATERIAL E MÉTODOS

O modelo desenvolvido dispõe da estrutura neural de detecção YOLOv3 (REDMON; FARHADI, 2018) e da abordagem de rastreamento HART (KOSIOREK; BEWLEY; POSNER, 2017) para executar tarefas de classificação e localização de objetos em uma abordagem multiestágio para fins em tempo real. Além disso, um sistema adaptativo é implementado com o intuito de reduzir cargas computacionais ao realizar trocas de configuração com base na quantidade de objetos em cena. O modelo proposto para classificar e localizar objetos em tempo real é ilustrado na Figura 15.

Figura 15 – Estágio de previsão proposto.



Fonte: Adaptado de SANTOS et al. (2020).

3.1 MODELO MULTIESTÁGIO DE DETECÇÃO E RASTREAMENTO

Quanto à dependência de informações temporais, o rastreamento *offline* de objetos depende da localização de um determinado objeto em quadros anteriores enquanto a detecção de objetos é independente, resultando na classificação e localização de objetos com base em dados de quadro único. Portanto, a concordância na execução de múltiplos estágios de localização foi obtida ao desempenhar a função de rastreamento após estágios de localização nos dados de sequência anteriores.

As estratégias e abordagens de rastreamento e detecção foram definidas com base na capacidade de previsibilidade em tempo real. A abordagem HART, implemen-

tada no estágio de rastreamento, foi configurada com entradas de sequência atual e anterior, e, devido à execução do rastreamento unitário de objetos, o estágio atribui inferências encadeadas para cada objeto localizado anteriormente na sequência de imagens. O estágio de detecção apresenta a execução do modelo YOLOv3 para quadros correntes de sequência. No Apêndice A, as Figuras 31, 32 e 33 apresentam a estrutura do algoritmo desenvolvido e implementado em Python.

3.2 ARQUITETURAS NEURAIS E IMPLEMENTAÇÃO

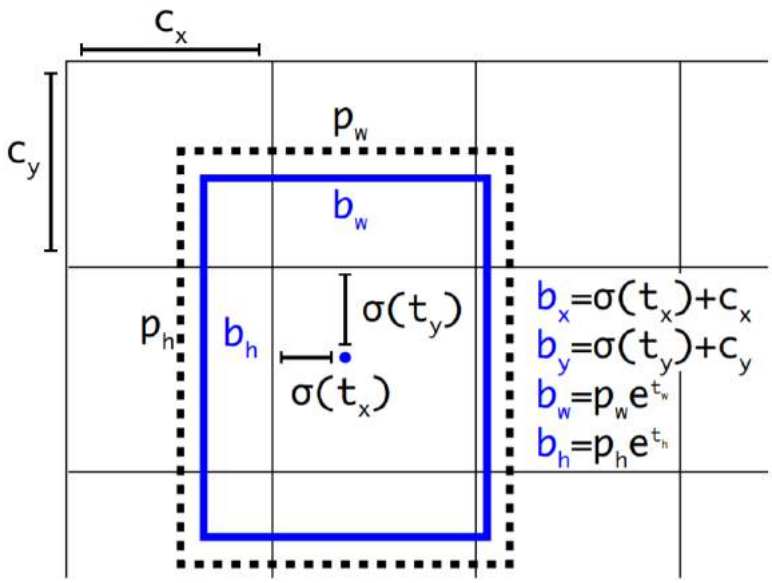
Esta seção apresenta as arquiteturas e a implementação de ambas as abordagens do modelo multiestágio, destacando a estrutura geral, os componentes individuais do sistema e a escolha dos parâmetros de inferência, relevando a aplicabilidade em tempo real.

3.2.1 Modelo YOLOv3

Com o propósito de realizar a previsão dos dados de classe e local de múltiplos objetos em um único estágio, o modelo YOLOv3 faz uso da rede Darknet-53, cuja estrutura é uma abordagem híbrida da rede Darknet-19 (REDMON; FARHADI, 2017) com recursos de blocos residuais profundos (HE et al., 2016), reunidos com o intuito de apresentar ganhos significativos de precisão, a partir de uma estrutura profunda, e compensar o treinamento com uma rápida convergência, resultado da adição de blocos residuais.

A inferência da rede YOLOv3 efetua a previsão de caixas delimitadoras em 3 escalas diferentes para obter coordenadas, pontuação relativa do objeto e 80 possíveis confidências de classe, a partir do conjunto de dados para treinamento *Common Objects in Context* (COCO) (LIN et al., 2014). As caixas delimitadoras são previstas utilizando caixas âncora relacionadas à classe. A rede prevê quatro coordenadas para cada caixa t_x, t_y, t_w, t_h. As coordenadas do centroide b_x, b_y estão relacionadas a deslocamentos previstos no canto superior esquerdo da imagem c_x, c_y e valores normalizados para o local da célula que representa a localização prevista para o objeto $\sigma\left(t_x\right)$, $\sigma\left(t_y\right)$, utilizando uma função de normalização *Sigmoid*. A previsão da altura e largura b_w, b_h é realizada como uma compensação exponencial e^{t_w}, e^{t_h} da dimensão de três caixas âncora distintas p_w, p_h para cada escala, obtidas com uma operação de *k-means clustering* que define as nove caixas âncora que melhor generalizam a dimensão dos objetos na base de dados para treinamento (REDMON; FARHADI, 2018). A Figura 16 apresenta visualmente as coordenadas previstas.

Figura 16 – Coordenadas de caixa, previstas com o modelo YOLOv3.

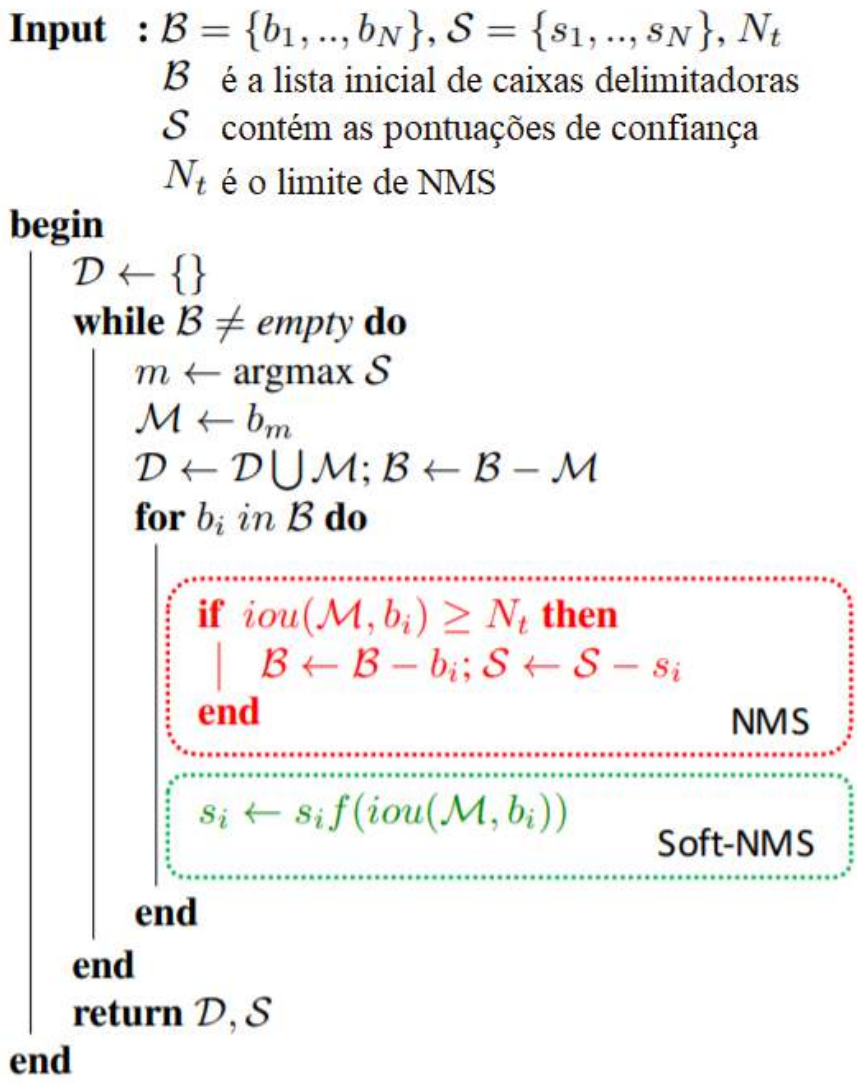


Fonte: REDMON e FARHADI (2017).

A previsão da pontuação do objeto é realizada por meio de regressão logística, apresentando valores de 0 a 1 e definindo a probabilidade de que uma região, relacionada a uma caixa delimitadora, contenha um objeto. As confidências de classe são previstas usando classificadores logísticos independentes para lidar melhor com a classificação de múltiplos rótulos, em oposição ao uso de uma camada Softmax (REDMON; FARHADI, 2018).

A rede YOLOv3, pré-treinada (REDMON; FARHADI, 2018) no conjunto de dados COCO (LIN et al., 2014) para uma entrada de dimensão de imagem de 416×416, foi implementada no estágio de detecção. Considerando essa dimensão de entrada e as previsões de saída em três escalas diferentes, a rede gera várias previsões de caixas delimitadoras para um objeto em potencial, o que é indesejável. Para filtrar caixas delimitadoras de um possível objeto, foi implementado um limite de confiança para eliminar caixas delimitadoras com baixa pontuação de confiança e um algoritmo de Não-Máxima Supressão Suave (Soft-NMS) (BODLA et al., 2017). A implementação do algoritmo permite a filtragem de caixas delimitadoras de objetos, selecionando caixas de proposta com altas pontuações de confiança, criando uma lista de propostas e eliminando os componentes com IoU maior que o limite definido. O algoritmo Soft-NMS é repetido até que não existam caixas delimitadoras a serem selecionadas para o processo. A Figura 17 representa as etapas de processo do algoritmo Soft-NMS, o conteúdo em verde substitui o conteúdo em vermelho, realizando um processo de revisão das pontuações de detecção ao escalar cada uma como função linear ou gaussiano da sobreposição.

Figura 17 – Representação do algoritmo Soft-NMS.



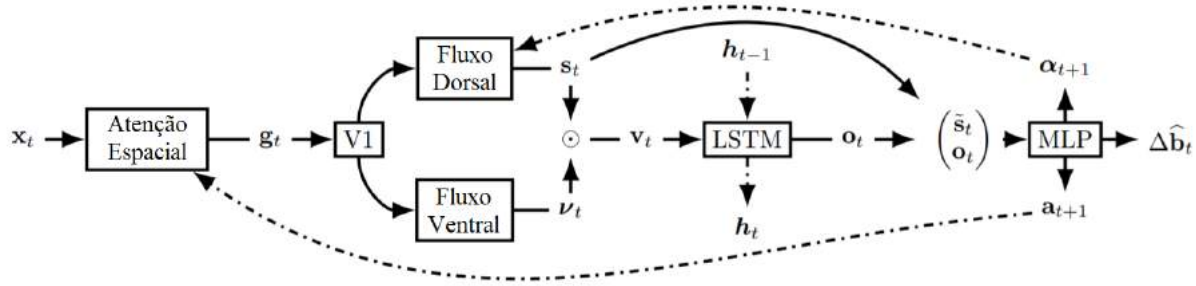
Fonte: Adaptado de BODLA et al. (2017).

3.2.2 Abordagem HART

A abordagem HART é baseada no Modelo de Rastreamento Atentivo Recorrente (RATM) (KAHOU et al., 2017), valendo-se da adição de blocos estratégicos e novas funções de perda, projetadas para a regressão de valores de previsibilidade das caixas delimitadoras e mecanismos de atenção, em uma estratégia unificada (KOSIOREK; BEWLEY; POSNER, 2017). Baseado nas contribuições no processamento de sequências de NING et al. (2017), o método HART insere dados de imagem processados em uma rede de Memória Longa de Curto Prazo (LSTM), capaz de aprender a trocar informações espaço-temporais e de aparência, resultando em previsões mais robustas ao enfrentar desafios de rastreamento, como oclusão e deformidade de objetos (KOSIOREK; BEWLEY; POSNER, 2017). As Figuras 18 e 19 apresentam a

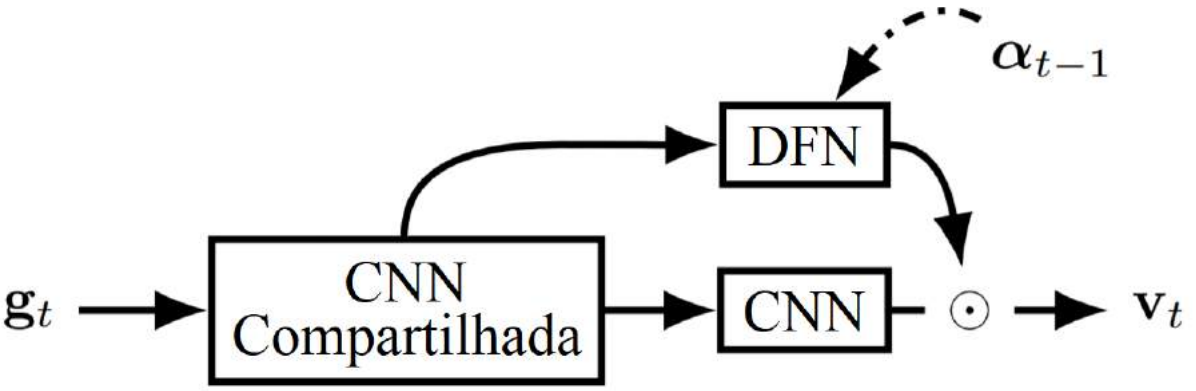
estrutura da abordagem HART. O bloco de atenção espacial extrai um vislumbre $\mathbf{g}_t$ da imagem de entrada $\mathbf{x}_t$. V1 é implementado como uma CNN compartilhada entre o bloco de fluxo dorsal (DFN) e ventral (CNN), e em conjunto com o fluxo ventral são responsáveis por extrair recursos de aparência. O fluxo dorsal calcula a segmentação do primeiro e segundo plano do vislumbre de atenção. Os recursos mascarados $\mathbf{v}_t$ contribuem para a memória de trabalho $\mathbf{h}_t$. A saída da rede LSTM é então utilizada no cálculo de atenção a_{t+1}, aparência α_{t+1} e correção das caixas delimitadoras $\Delta\widehat{\mathbf{b}}_t$. O símbolo $\odot$ representa o produto de Hadamard e implementa o mascaramento de recursos visuais por um mapa de localização. Linhas tracejadas correspondem à conexões temporais enquanto linhas sólidas descrevem o fluxo de dados em um intervalo de tempo.

Figura 18 – Estrutura da abordagem HART.



Fonte: Adaptado de KOSIOREK, BEWLEY e POSNER (2017).

Figura 19 – Arquitetura do bloco de atenção à aparência da abordagem HART.



Fonte: Adaptado de KOSIOREK, BEWLEY e POSNER (2017).

O modelo apresenta um estágio de atenção espacial para extrair um vislumbre inicial do objeto $\mathbf{g}_t$ na imagem de entrada $\mathbf{x}_t \in \mathbb{R}^{H\times W}$, utilizando duas matrizes para a largura $\mathbf{A}_t^x \in \mathbb{R}^{w\times W}$ e altura $\mathbf{A}_t^y \in \mathbb{R}^{h\times H}$ da imagem, contendo um Gaussiano por linha da matriz. Uma soma cumulativa de atualizações de atenção, de t = 1 a t =

T, é empregada para gerar uma atenção espacial por vez t (KOSIOREK; BEWLEY; POSNER, 2017).

$$\mathbf{g}_t = \mathbf{A}_t^y \mathbf{x}_t \left(\mathbf{A}_t^x\right)^\top \tag{2}$$

Um estágio para a extração de recursos com base na aparência é então executado, compreendendo informações espaciais e de aparência em um vetor $\mathbf{v}_t$ e adaptando a entrada com considerações de camada convolucional e de *pooling* máximo em V1. Esta etapa apresenta estágios de fluxo ventral e dorsal, contendo CNNs responsáveis pela extração adicional de características com base na aparência e segmentação do plano da imagem, respectivamente. Enquanto o fluxo ventral trata os recursos visuais e produz mapas de recursos $\boldsymbol{\nu}_t$, filtros convolucionais $\psi_t^{h_t \times b_i \times c_i}$ são computados no fluxo dorsal, como saída de um Perceptron Multicamadas (MLP), calculando a aparência α_{t+1}, para K camadas convolucionais aplicadas à saída do estágio V1. O resultado é transformado em uma distribuição espacial de Bernoulli s_t por uma função de ativação Sigmoid (KOSIOREK; BEWLEY; POSNER, 2017).

$$\Psi_t = \left\{\psi_t^{h_i \times b_i \times c_i}\right\}_{i=1}^{K} = \mathrm{MLP}(\alpha_t) \tag{3}$$

Os resultados dos estágios paralelos são então mesclados fazendo uso do produto de Hadamard $\odot$, resultando no vetor $\mathbf{v}_t$ com dimensão equivalente às matrizes de entrada e referência na atualização do estado oculto $\boldsymbol{h}_t$ das LSTMs. A matriz do produto Hadamard é então definida como uma entrada para a rede LSTM, resultando em uma saída $\mathbf{o}_t$ de calculo de atenção $\Delta\mathbf{a}_{t+1}$, aparência α_{t+1} e correção da caixa delimitadora $\Delta\widehat{\mathbf{b}}_t$ para a etapa relativa ao processo atual (KOSIOREK; BEWLEY; POSNER, 2017). As operações são detalhadas nas equações (4) a (6).

$$\mathbf{v}_t = \mathrm{MLP}(\mathrm{vec}(\boldsymbol{\nu}_t \odot \mathbf{s}_t)) \tag{4}$$

$$\mathbf{o}_t, \boldsymbol{h}_t = \mathrm{LSTM}(\mathbf{v}_t, \boldsymbol{h}_{t-1}) \tag{5}$$

$$\boldsymbol{\alpha}_{t+1}, \Delta\mathbf{a}_{t+1}, \Delta\widehat{\mathbf{b}}_t = \mathrm{MLP}(\mathbf{o}_t, \mathrm{vec}(\mathbf{s}_t)) \tag{6}$$

A previsão da caixa delimitadora $\widehat{\mathbf{b}}_t$ é desenvolvida progressivamente à medida que a junção de atenção $\mathbf{a}_t$ atualiza, contendo um termo de atualização com um pequeno parâmetro programável $\boldsymbol{c}$ que garante os limites de atualização no treinamento inicial e correções nos valores da caixa delimitadora $\Delta\widehat{\mathbf{b}}_t$, saída de uma rede MLP final (KOSIOREK; BEWLEY; POSNER, 2017). Esta operação está detalhada nas equações (7) e (8).

$$\mathbf{a}_{t+1} = \mathbf{a}_t + \tanh(\boldsymbol{c})\Delta\mathbf{a}_{t+1} \tag{7}$$

$$\widehat{\mathbf{b}}_t = \mathbf{a}_t + \Delta\widehat{\mathbf{b}}_t \tag{8}$$

O treinamento do modelo, realizado em 21 sequências de treinamento KITTI MOT, obtidas com a configuração veicular apresentada na Figura 20 e divididas na proporção 80/20 para os conjuntos de treinamento e testes, é executado para pequenos lotes, *minibatches*, de conjuntos de dados $\{x^i_{1:T}, b^i_{1:T}\}^M_{i=1}$ com um tamanho de M. A perda do sistema é definida como a soma de um termo de perda de rastreamento $\mathcal{L}_{\mathrm{t}}(\widehat{y}, y)$, termos de perda espacial e de aparência $(\mathcal{L}_{\mathrm{s}}(\widehat{y}, y) + \mathcal{L}_{\mathrm{a}}(\widehat{y}, y))$ e termos de regularização $(R(\boldsymbol{\lambda}) + \beta R(\widehat{y}, y))$. O aprendizado por função de perda ponderada $\boldsymbol{\lambda} = \{\lambda_t, \lambda_s, \lambda_{\mathrm{a}}\}$ e considerações relativas à regularização também são aplicados para limitar o número de hiper parâmetros do sistema (KOSIOREK; BEWLEY; POSNER, 2017; GEIGER et al., 2013).

Figura 20 – Configuração veicular para obtenção dos dados do conjunto KITTI.

Fonte: Adaptado de BERNINI et al. (2014).

$$\boldsymbol{\lambda}\mathcal{L}(\widehat{y}, y) = \lambda_t\mathcal{L}_t(\widehat{y}, y) + \lambda_s\mathcal{L}_s(\widehat{y}, y) + \lambda_a\mathcal{L}_a(\widehat{y}, y) \tag{9}$$

$$\mathcal{L}_{\mathsf{HART}}(\widehat{y}, y) = \boldsymbol{\lambda}\mathcal{L}(\widehat{y}, y) + R(\boldsymbol{\lambda}) + \beta R(\widehat{y}, y) \tag{10}$$

O termo de perda para o rastreamento foi definido com base no trabalho de YU et al. (2016), demonstrando que a relação da regressão para o termo de IoU, invariante ao objeto e escala da imagem em previsões de caixa delimitadora, leva a convergência mais rápida e ganhos consideráveis de precisão quando comparada às previsões de coordenadas independentes de caixa delimitadora. As entradas da IoU foram de-

finidas como caixas delimitadoras rotuladas $\mathbf{b}_t$ e previstas $\widehat{\mathbf{b}}_t$ (KOSIOREK; BEWLEY; POSNER, 2017).

$$\mathcal{L}_{\mathrm{t}}(\widehat{y}, y) = \mathbb{E}_{p\left(\hat{\mathbf{b}}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1\right)}\left[-\log \mathrm{IoU}\left(\widehat{\mathbf{b}}_t, \mathbf{b}_t\right)\right] \tag{11}$$

Os termos de perda para a atenção espacial foram atribuídos considerando a necessidade de destacar o objeto rastreado da imagem e minimizar os impactos negativos de erros presentes na previsão do vislumbre de atenção. Como o aumento de erros de localização de objetos no estágio de previsão do vislumbre é proporcional à diminuição do vislumbre em dimensão, o primeiro termo da perda restringe a atenção prevista para cobrir a caixa delimitadora e o segundo termo garante a dimensão apropriada do vislumbre da atenção como aplicação de um ajuste (KOSIOREK; BEWLEY; POSNER, 2017).

$$\mathcal{L}_{\mathrm{s}}(\widehat{y}, y) = \mathbb{E}_{p(\mathbf{a}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}\left[-\log\left(\frac{\mathbf{a}_t \cap \mathbf{b}_t}{\mathrm{rea}\left(\mathbf{b}_t\right)}\right) - \log\left(1 - \mathrm{IoU}\left(\mathbf{a}_t, \mathbf{x}_t\right)\right)\right] \tag{12}$$

O termo de perda para a atenção da aparência foi desenvolvido com a intenção de suprimir os distratores do objeto rastreado. A entropia cruzada, foi atribuída, com a função binária de destino $\tau(\mathbf{a}_t, \mathbf{b}_t) : \mathbb{R}^4 \times \mathbb{R}^4 \rightarrow \{0,1\}^{h_v \times w_v}$ em escala com a saída V1, para gerar uma máscara binária, correspondente ao vislumbre g, com *bits* definidos como um na ocorrência de uma sobreposição entre o vislumbre e a caixa delimitadora e zero caso contrário (KOSIOREK; BEWLEY; POSNER, 2017). A equação de entropia cruzada e o termo de perda de atenção para a aparência estão detalhadas abaixo.

$$H(p, q) = -\sum_z p(z) \log q(z) \tag{13}$$

$$\mathcal{L}_{\mathrm{a}}(\widehat{y}, y) = \mathbb{E}_{p(\mathbf{a}_{1:T},\mathbf{s}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}\left[H\left(\tau\left(\mathbf{a}_t, \mathbf{b}_t\right), \mathbf{s}_t\right)\right] \tag{14}$$

A regularização L2 e os filtros convolucionais de múltiplas camadas Ψ_t são aplicados para regressão do modelo, envolvendo os parâmetros $\widehat{y}$, e como uma prevenção contra o sobreajuste na CNN. Além dos termos de regularização L2, um termo de regularização de pesos β é aplicado aos pesos definidos para termos de perda (KOSIOREK; BEWLEY; POSNER, 2017).

$$R_{\mathsf{HART}} = R(\boldsymbol{\lambda}) + \beta R(\widehat{y}, y) \tag{15}$$

$$R(\boldsymbol{\lambda}) = -\sum_i \log\left(\boldsymbol{\lambda}_i^{-1}\right) \tag{16}$$

$$R(\widehat{y}, y) = \frac{1}{2}\|\theta\|_2^2 + \frac{1}{2}\left\|\mathbb{E}_{p(\boldsymbol{\alpha}_{1:T}|\mathbf{x}_{1:T},\mathbf{b}_1)}\left[\boldsymbol{\Psi}_t|\boldsymbol{\alpha}_t\right]\right\|_2^2 \quad (17)$$

A CNN da abordagem HART foi implementada como as três primeiras camadas de uma rede AlexNet modificada (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) no estágio pós-atenção V1. As modificações consistiram na implementação de um *stride* inicial de 1 (um), substituindo o *stride* original de 4 (quatro), e remoção de uma operação de *pooling* máximo para obter melhor resolução e evitar operações de *up sampling*, considerando como entrada um vislumbre de pequena dimensão em comparação ao tamanho de entrada original da AlexNet de 227×227. Quanto à inferência, imagens de sequência do conjunto de dados KITTI (GEIGER et al., 2013), contendo imagens com dimensões 375×1242, são redimensionadas para 187×621. Dessa forma, no estágio inicial de atenção espacial, o sistema gera uma amostra de tamanho 56×56 e, como resultado das modificações iniciais da rede, a CNN contida no estágio V1 pode gerar um mapa de características de tamanho $14 \times 14 \times 384$. O fluxo ventral possui uma única camada convolucional com um *kernel* 1×1 e o fluxo dorsal apresenta duas camadas dinâmicas de filtro com tamanho 1×1 e 3×3, respectivamente. Ambos os fluxos apresentam 5 mapas de recursos para cada camada destacada (KOSIOREK; BEWLEY; POSNER, 2017).

3.3 SISTEMA ADAPTATIVO DE TROCA DE ESTÁGIOS

Considerando que o modelo proposto apresenta diferentes estratégias para cada etapa, individualizadas de acordo com a necessidade de execuções de inferência, única ou cíclica, e dependências em fatores externos, implementou-se um sistema adaptativo que aproveita a dessemelhança no tempo de processamento entre os estágios de detecção e rastreamento, ocasionada pela variação dos ciclos internos de rastreamento conforme o número de objetos localizados no estágio anterior muda. Dado que o tempo de processamento para a fase de detecção é aproximadamente constante, localizando uma ampla e invariante gama de objetos em cada execução, e o tempo de processamento do estágio de rastreamento varia conforme a quantidade de objetos localizados no processo anterior, a proposição adaptativa é definida para que o uso apropriado dos estágios de rastreamento confiram a redução geral do tempo de processamento, impactando positivamente no consumo computacional, em situações de fluxo de tráfego variável, e evitando a execução de estágios de rastreamento em ambientes superpovoados, como situações de tráfego intenso.

A mudança adaptativa é realizada ao comparar o número de objetos localizados, no final de uma fase de previsão, com um limite, representando o ponto em que as variações conferem vantagens em termos de redução do tempo de processamento

para uma determinada configuração. Para obter o limite de objetos localizados, foi realizada uma comparação de períodos de processamento entre os estágios de detecção e rastreamento para uma variedade prática de objetos localizados. A partir da qual, definiu-se o limite como o valor de objetos localizados que apresentou o tempo de processamento mais aproximado entre os estágios, com um tempo menor para os estágios de rastreamento, dada a adaptabilidade do sistema para configurações de baixo consumo computacional. Para realizar a análise e verificar o limite de objetos localizados, dados foram coletados em 15 sequências de dados brutos KITTI (GEIGER et al., 2013), 5 (cinco) de cada categoria relacionada ao tráfego (cidade, residencial e rodoviária), considerando o número de quadros, objetos localizados na sequência e tempo médio de processamento para ambos os estágios. As sequências de cada categoria foram definidas pelo equilíbrio entre tráfego leve, médio e pesado. As Figuras 21, 22 e 23 apresentam quadros presentes nas sequências de dados brutos selecionadas.

Figura 21 – Quadro de sequência da categoria Cidade.

Fonte: GEIGER et al. (2013).

Figura 22 – Quadro de sequência da categoria Residencial.

Fonte: GEIGER et al. (2013).

Figura 23 – Quadro de sequência da categoria Rodoviária.

Fonte: GEIGER et al. (2013).

Todos os dados, para a definição do limiar e consequente avaliação do modelo, foram coletados ao cronometrar as passagens por estágios na execução do algoritmo proposto. O tratamento de dados, anterior e posterior às inferências de rede, foi realizado em uma CPU Intel Core i7-7700HQ 2,80GHz, e inferências de rede em uma GPU Nvidia GeForce GTX 1060 6GBs.

Tabela 1 – Tempo médio de processamento para objetos localizados na sequencia amostral.

Número de objetos	Número de quadros		Tempo médio de processamento (s)	
	Detecção	Rastreamento	Detecção	Rastreamento
0	290	291	0.0923	-
1	387	389	0.0930	0.0246
2	375	379	0.0933	0.0396
3	374	371	0.0934	0.0475
4	367	366	0.0940	0.0628
5	334	338	0.0940	0.0835
6	239	239	**0.0941**	**0.0892**
7	151	153	0.0944	0.1137
8	58	58	0.0951	0.1194
9	20	20	0.0966	0.1330
10	6	6	0.0950	0.1575
11	7	7	0.0954	0.1563

Os valores presentes na Tabela 1 revelam que para 6 (seis) objetos localizados o tempo médio de processamento do estágio de rastreamento é o menor e mais próximo do tempo apresentado pelo estágio de detecção, resultado que define o valor do limite proposto. A definição do limite habilita a capacidade adaptativa de executar a classificação e localização de objetos, com foco na adequação do sistema aos ganhos de velocidade e precisão.

3.4 ESTRUTURA DE AVALIAÇÃO

Modelos recentes de visão computacional foram desenvolvidos como uma estrutura única para executar a classificação e localização, detectando ou rastreando

vários objetos em imagens ou sequências (REDMON; FARHADI, 2018; REN et al., 2017; BEWLEY et al., 2016; KOSIOREK; BEWLEY; POSNER, 2017). Portanto, a estrutura de avaliação geralmente é competitiva, desafiando novos modelos a alcançar resultados de desempenho de ponta em conjuntos de dados rotulados (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; GEIGER et al., 2013; EVERINGHAM et al., 2010; LIN et al., 2014).

Dada a propriedade de seleção de opções para configurações de múltiplos estágios, avaliadas pela adição de uma proposta de mudança de estágio adaptável, uma avaliação comparativa robusta baseia-se no desempenho das configurações adaptativas e cíclicas mais relevantes para o modelo. A avaliação foi estruturada com métricas de precisão média (mAP), e tempo de processamento, adaptadas para serem executadas nas sequências de amostras.

A escolha das configurações do modelo foi baseada em estratégias, bem situadas no campo da visão computacional, que enfatizam os atributos do rastreamento em tempo real, destacando as propriedades da configuração adaptativa proposta. Uma configuração de detecção encadeada de objetos foi escolhida como referência para a avaliação de precisão média e três configurações cíclicas de vários estágios, variando na proporção do estágio de rastreamento, foram definidas como uma referência para a avaliação do tempo de processamento e para contribuir na análise da variação progressiva do estágio de rastreamento em termos de desempenho geral.

3.4.1 Conjuntos de Dados para Avaliação

Para a avaliação da precisão média, foram seguidos os conceitos de treinamento e teste apresentados por KOSIOREK, BEWLEY e POSNER (2017), definindo 20% dos dados de treinamento das sequências KITTI MOT (GEIGER et al., 2013) para avaliação, consistindo em 1602 quadros com dados rotulados para 8 classes. A avaliação do tempo de processamento foi realizada na sequência KITTI MOT para validação e testes (GEIGER et al., 2013) que apresentou a maior relação entre o número de quadros e quantidade média de objetos por quadro, uma vez que a variação na condição de tráfego é essencial para que alterações adaptáveis ocorram. Depois de verificar o número de quadros em cada sequência e medir o número médio de objetos localizados por quadro, utilizando previsões do modelo YOLOv3 (REDMON; FARHADI, 2018) como a abordagem mais precisa de localização, selecionou-se a sequência, apresentada nas Figuras 24, 25 e 26, com a relação de 1176 quadros e a média de 2,11 objetos localizados por quadro.

Figura 24 – Quadro da sequência de avaliação de tempo de processamento para um período de tráfego pesado.

Fonte: GEIGER et al. (2013).

Figura 25 – Quadro da sequência de avaliação de tempo de processamento para um período de tráfego médio.

Fonte: GEIGER et al. (2013).

Figura 26 – Quadro da sequência de avaliação de tempo de processamento para um período de tráfego leve.

Fonte: GEIGER et al. (2013).

4 RESULTADOS E DISCUSSÃO

Este capítulo visa apresentar os resultados e discussão das avaliações comparativas de precisão média e tempo de processamento para as configurações de maior relevância do modelo multiestágio desenvolvido.

4.1 AVALIAÇÃO COMPARATIVA DE PRECISÃO MÉDIA

A avaliação comparativa de precisão média foi definida considerando métricas competitivas de avaliação para modelos de classificação e localização de objetos em situações de tráfego veicular (EVERINGHAM et al., 2010; GEIGER et al., 2013). A distinção de previsões como verdadeiras ou falsas com base em métricas e limites bem conceituados leva à apresentação de resultados válidos e seguros para condições variadas de tráfego em tempo real.

As métricas definidas compreendem curvas de precisão $\times$ revocação e precisão média, como a área sob a curva, calculada na interpolação para todos os pontos do nível de revocação, para classes de automóveis e pedestres, dada a presença máxima de ambas em ambientes de tráfego e a generalização à classe de CNNs treinadas em dados de tráfego. Precisão e revocação são definidas pela sobreposição entre caixas delimitadoras de referência B_{gt} e caixas delimitadoras previstas B_p com uma definição de limite mínimo de IoU. Os resultados condicionais são divididos em verdadeiros positivos TP, falsos positivos FP ou falsos negativos FN, decompondo as previsões como corretas, erradas ou não detectadas, respectivamente. Os conceitos de IoU e os conceitos do modelo são dados pelas equações 18 a 21.

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{18}$$

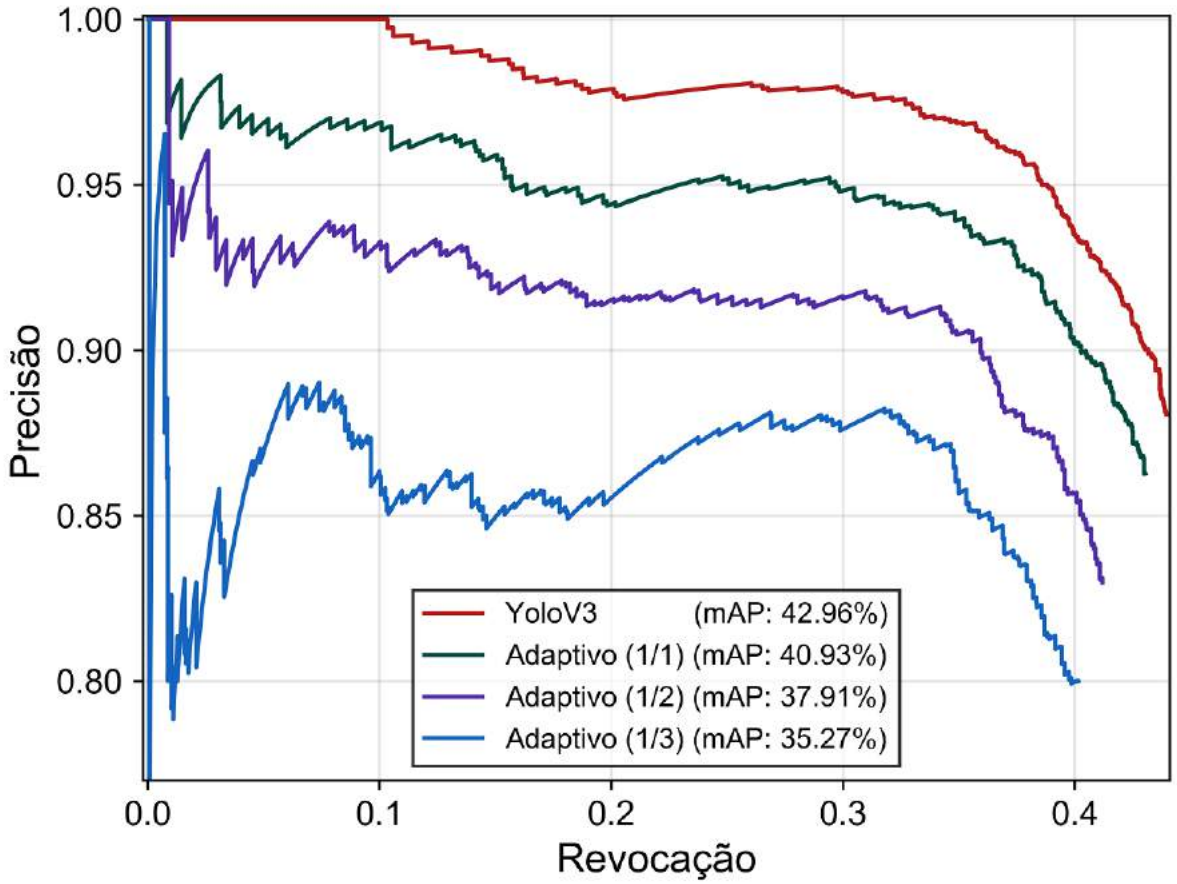
$$\textit{Previsão} = \begin{cases} TP & \text{se IoU} >= \textit{limite} \\ FP & \text{se IoU} < \textit{limite} \end{cases} \tag{19}$$

$$\textit{Precisão} = \frac{TP}{TP + FP} \tag{20}$$

$$\textit{Revocação} = \frac{TP}{TP + FN} \tag{21}$$

O valor de 0,5 foi definido para os limites de confiança e sobreposição de objetos previstos, como um valor médio, competitivo e consistente na classificação de falsos positivos para avaliação (GEIGER et al., 2013). As curvas 27 e 28 apresentam os resultados para as classes Carro e Pedestre.

Figura 27 – Curva de precisão × revocação para a classe Carro

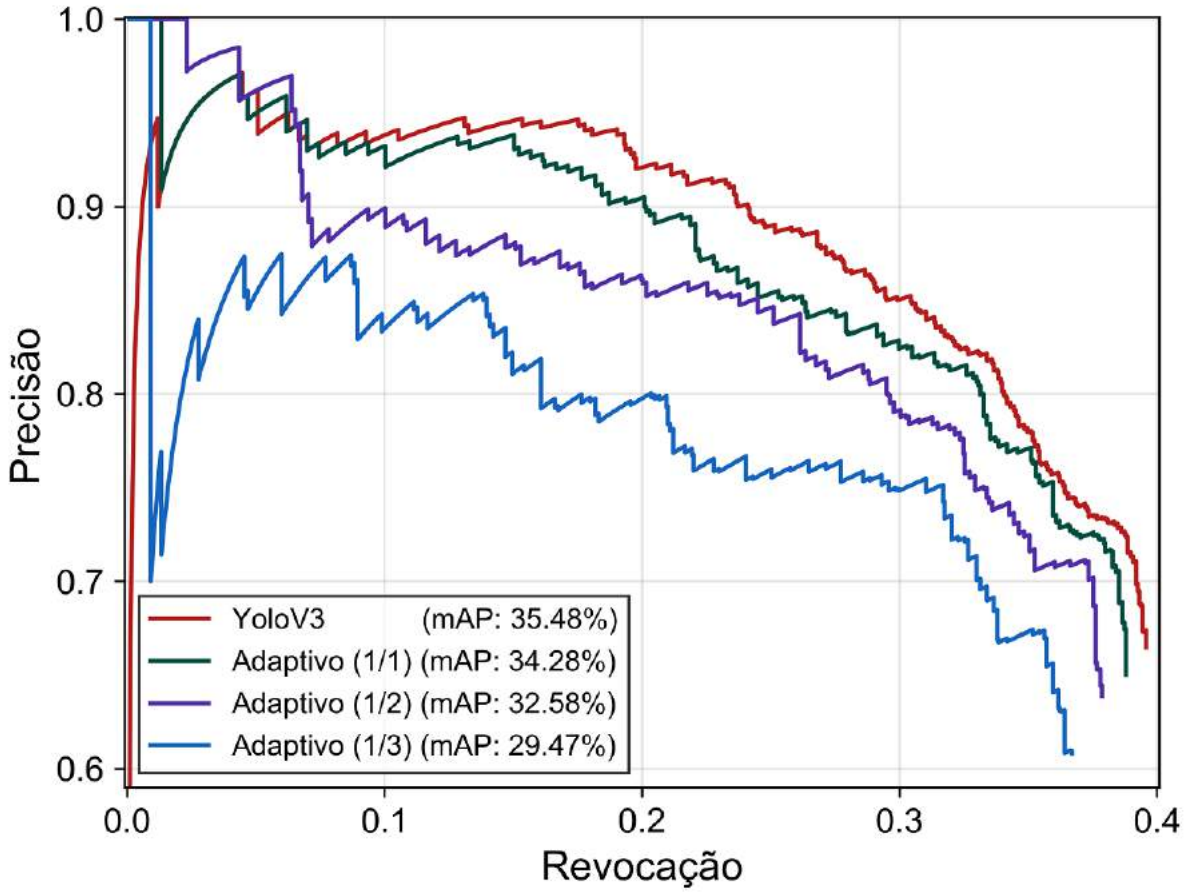


Fonte: Autoria própria.

As Figuras 27 e 28 mostram que a abordagem YOLOv3 (REDMON; FARHADI, 2018) apresenta a maior quantidade de verdadeiros positivos de alta confiança devido ao aprofundamento da rede, incluindo uma profunda estrutura de camadas e parâmetros treinados, levando à uma elevada otimização de pesos. Em contraste, o modelo proposto apresentou uma perda inicial em precisão média devido à troca adaptativa entre estágios e relativas abordagens de classificação e localização, observado, adicionalmente, para estágios de rastreamento encadeados, na configuração Adaptivo (1/2) e de forma agravada em Adaptativo (1/3).

A perda de precisão de sobreposição de objetos, agravada por estágios de rastreamento, é explícita em como a curva apresenta maior invariabilidade inicial para a abordagem de detecção YOLOv3 (REDMON; FARHADI, 2018), e aumenta a instabilidade para as três configurações adaptativas do modelo proposto, de forma intensificada a medida que a quantidade de inferências encadeadas aumentam. A ocorrência é causada pelo deslocamento e perda de atenção na previsão de objetos com movimento variado ao utilizar informações temporais provenientes de objetos parentes de alta confiança, previstos em estágios anteriores de detecção ou rastreamento. As perdas em precisão são menos impactantes para classes com baixa variância de mo-

Figura 28 – Curva de precisão × revocação para a classe Pedestre



Fonte: Autoria própria.

vimento, como pedestres, onde a abordagem adaptativa de múltiplos estágios prova ser mais consistente, apresentando resultados de alta confiança próximos ao modelo YOLOv3.

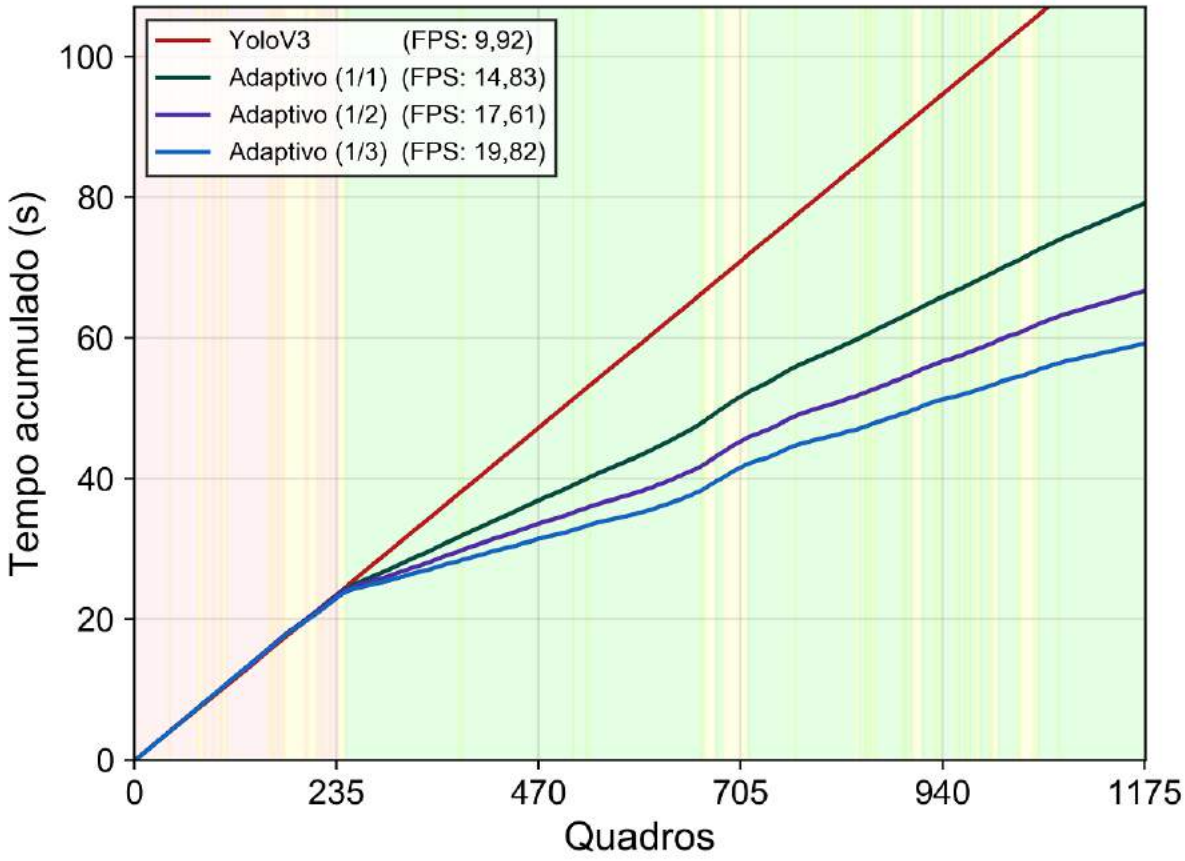
O aumento de falsos positivos, a medida que a revocação, de previsões sequenciadas de forma decrescente quanto à confiança, diminui, é consequência de inconsistências na classificação e localização com ausência de recursos temporais de aparência e movimento, presentes em inferências de rastreamento, para execuções de estágios de detecção. A capacidade do modelo de manter valores consistentes de falsos positivos, próximos do modelo YOLOv3 (REDMON; FARHADI, 2018), conforme observado nas configurações do modelo adaptativo, é causada pela imposição da localização de objetos para cada caixa parente anterior, de mesma classe, nas execuções de estágios de rastreamento, independente das pontuações de sobreposição e confiança anteriores, contrastando a inexistência de associações temporais em abordagens de detecção, levando à variação de identidade e perdas quanto à aparência para previsões sequenciais.

4.2 AVALIAÇÃO COMPARATIVA DE TEMPO DE PROCESSAMENTO

As métricas para a avaliação do tempo de processamento foram definidas de acordo com a capacidade de realização de tarefas de visão computacional sobre fluxos e ambientes variáveis de tráfego, permitindo a visualização do comportamento de diferentes configurações de vários estágios, mudanças adaptativas, com base no número de objetos na cena, e variações de desempenho, induzidas por mudanças entre condições de tráfego, leve, médio e pesado. Para visualizar o desempenho de cada modelo para todas as condições de tráfego na sequência de avaliação, duas curvas de tempo acumulado $\times$ quadros, com padrões coloridos para cada estado de tráfego, foram adotadas, para todos os quadros de sequência, contendo uma grande quantidade de intervalos de tráfego médio e leve, e períodos de tráfego pesado, presentes nos 237 quadros iniciais da sequência. Adicionalmente, a taxa de quadros por segundo (FPS), foi calculada, como a razão entre a quantidade total de quadros da sequência e o tempo de execução, para cada configuração, com o intuito de verificar uma estatística geral de desempenho.

Todos os valores dos limites foram mantidos a partir da avaliação de precisão média, considerando que seu uso discrimina corretamente o número de objetos localizados para diferentes condições de tráfego. Carimbos de tempo foram coletados para todos os estágios de detecção e rastreamento como um único processo de, pré-tratamento dimensional para imagens de entrada, inferência(s) de rede e correções de sobreposições de identidade (NMS). As Figuras 29 e 30 apresentam as curvas de tempo acumulado $\times$ quadros para intervalos de interesse. As cores verde, amarela e vermelha foram adotadas para representar sequências de tráfego leve, médio e pesado, respectivamente. A legenda da Figura 29 apresenta, adicionalmente, a taxa de quadros por segundo (FPS) de cada modelo ou configuração adotados.

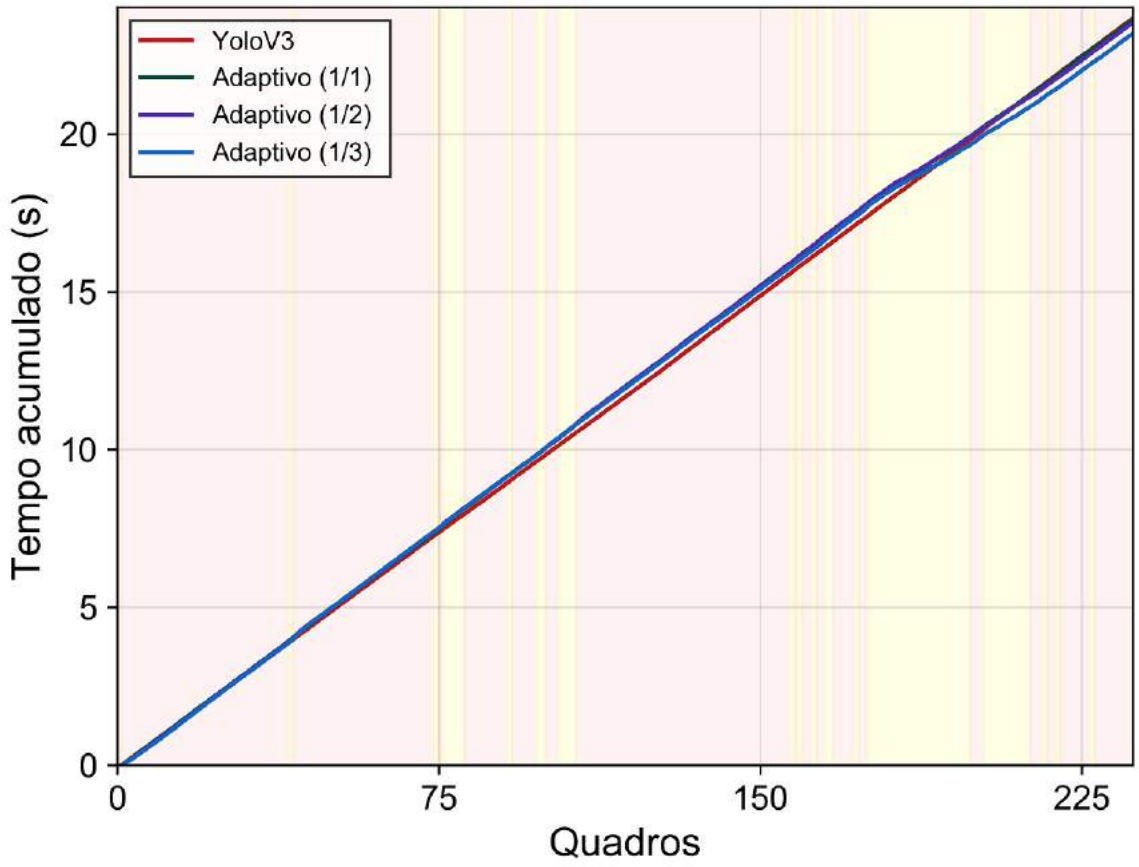
Figura 29 – Curva de tempo acumulado × quadros para todos os quadros da sequência.



Fonte: Autoria própria.

As Figuras 29 e 30 contém as curvas de avaliação de tempo de processamento, exibindo a projeção temporal linear do modelo de detecção a medida que o número de saídas de previsão permanece aproximadamente o mesmo para todos os quadros da sequência. O uso da técnica de não-máxima supressão suave Soft-NMS aumenta ligeiramente os intervalos de tempo de processamento, de acordo com o número de entradas, conforme pode ser visto na Tabela 1. O comportamento do tempo de execução para as configurações do modelo adaptativo nos períodos de tráfego intenso é aproximado do apresentado pela abordagem YOLOv3 (REDMON; FARHADI, 2018), dado que, temporalmente, para um grande número de objetos localizados nosso modelo realiza inferências de detecção em cadeia. Em contraste, todas as configurações do modelo adaptativo apresentam comportamentos variáveis de tempo de execução, explorando as vantagens do desempenho de estágios de rastreamento em cadeia para mudanças nas condições de tráfego pesado a médio, enquanto aproveita o uso de poucas ou nenhuma estratégia de localização para intervalos de estado de tráfego leve, resultando em um ganho geral na taxa de quadros de 49,5%, 77,5% e 99,8% sobre o modelo YOLOv3, para as configurações Adaptativo(1/1), Adaptativo(1/2) e Adaptativo(1/3), respectivamente.

Figura 30 – Curva de tempo acumulado × quadros para o intervalo de tráfego médio a pesado.



Fonte: Autoria própria.

As configurações adaptativas alcançam resultados competitivos ao realocar ganhos de precisão para cenas populadas e tomar vantagem de ganhos de velocidade de processamento em ambientes com condições médias e leves de tráfego, aproximando os intervalos de tempo para previsões de novas identidades, fator crucial em situações em que objetos apresentam alta variabilidade de movimento.

5 CONCLUSÕES E PERSPECTIVAS

O trabalho desenvolvido envolve a proposição de um modelo composto por uma estrutura de múltiplos estágios, passivos de definições estratégicas na realização das tarefas de classificação e localização de objetos. A capacidade de anexar estágios de rastreamento permite a alternância entre os ganhos de precisão de previsibilidade e tempo de processamento, permitindo que sistemas com limitações possam se beneficiar da troca e executem tarefas de visão computacional na configuração mais adequada. O uso adicional do sistema adaptativo de mudança de estágio induz melhorias adicionais no consumo de tempo de processamento, apresentando os resultados de desempenho mais desejáveis dentre as configurações avaliadas ao adotar uma estratégia encadeada de detecção de objetos para condições de tráfego médio a pesado, acima do limiar adaptativo, e uma estratégia cíclica de detecção e rastreamento para condições de tráfego leve a médio, abaixo do limiar adaptativo. Quanto ao tempo de processamento na estratégia proposta, as mudanças de estágio unem o comportamento linear de abordagens de detecção e a quantidade encadeada mais apropriada de inferências de rastreamento para alcançar ganhos na redução do consumo de recursos computacionais e garantir versatilidade de aplicações ao modelo proposto.

REFERÊNCIAS

ALBAWI, Saad. Understanding of a convolutional neural network. Istanbul Kemerburgaz University Istanbul, Turkey, 2012. Citado 2 vezes nas páginas 16 e 17.

BERNINI, N. et al. Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey. In: **17th International IEEE Conference on Intelligent Transportation Systems (ITSC)**. [S.l.: s.n.], 2014. p. 873–878. Citado na página 35.

BEWLEY, A. et al. Simple online and realtime tracking. In: **2016 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2016. p. 3464–3468. Citado 3 vezes nas páginas 13, 27 e 40.

BODLA, Navaneeth et al. Soft-nms – improving object detection with one line of code. In: **The IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 31 e 32.

CHOI, W. Near-online multi-target tracking with aggregated local flow descriptor. In: **2015 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2015. p. 3029–3037. Citado 2 vezes nas páginas 26 e 27.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)**. San Diego, CA, USA: [s.n.], 2005. p. 886–893. Citado na página 18.

EVERINGHAM, M. et al. The pascal visual object classes (voc) challenge. **International Journal of Computer Vision**, v. 88, p. 303–338, 2010. Citado 5 vezes nas páginas 13, 20, 21, 40 e 42.

FELZENSZWALB, P. F. et al. Object detection with discriminatively trained part-based models. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 9, p. 1627–1645, 2010. Citado na página 19.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient graph-based image segmentation. **International Journal of Computer Vision**, v. 59, p. 167–181, 2004. Citado 2 vezes nas páginas 18 e 19.

GEIGER, A. et al. Vision meets robotics: the kitti dataset. **The International Journal of Robotics Research**, v. 32, n. 11, p. 1231–1237, 2013. Citado 8 vezes nas páginas 13, 35, 37, 38, 39, 40, 41 e 42.

GIRSHICK, R. Fast r-cnn. In: **2015 IEEE International Conference on Computer Vision (ICCV)**. [S.l.: s.n.], 2015. p. 1440–1448. Citado na página 22.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **2014 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2014. p. 580–587. Citado 2 vezes nas páginas 20 e 22.

HE, K. et al. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778. Citado 3 vezes nas páginas 19, 24 e 30.

HELD, D.; THRUN, S.; SAVARESE, S. Learning to track at 100 fps with deep regression networks. In: **ECCV**. [S.l.: s.n.], 2016. Citado na página 28.

HENRIQUES, J. F. et al. High-speed tracking with kernelized correlation filters. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 37, n. 3, p. 583–596, 2015. Citado na página 25.

KAHOU, S. E. et al. Ratm: Recurrent attentive tracking model. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.: s.n.], 2017. p. 1613–1622. Citado na página 32.

KALMAN, R. A new approach to linear filtering and prediction problems. **Transaction of the ASME - Journal of Basic Engineering**, v. 82, n. 1, p. 35–45, 1960. Citado na página 25.

KOSIOREK, A.; BEWLEY, A.; POSNER, I. Hierarchical attentive recurrent tracking. In: **31st Conference on Neural Information Processing Systems (NIPS 2017)**. Long Beach, CA, USA: [s.n.], 2017. p. 1–9. Citado 10 vezes nas páginas 13, 28, 29, 32, 33, 34, 35, 36, 37 e 40.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey. Imagenet classification with deep convolutional neural networks. **Neural Information Processing Systems**, v. 25, 01 2012. Citado 5 vezes nas páginas 13, 17, 20, 37 e 40.

KUHN, H. W. The hungarian method for the assignment problem. In: **50 Years of Integer Progremming 1958-2008**. Berlin, Heidelberg: Springer-Verlag, 2010. p. 29–47. Citado na página 27.

LIN, T. et al. Microsoft coco: Common objects in context. **Computer Vision – ECCV**, p. 740–755, 2014. Citado 4 vezes nas páginas 13, 30, 31 e 40.

LIU, Wei et al. Ssd: Single shot multibox detector. In: **Computer Vision – ECCV 2016**. [S.l.: s.n.], 2016. p. 21–37. Citado na página 23.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, v. 60, p. 91–110, 2004. Citado na página 18.

MAO, Q. et al. Mini-yolov3: Real-time object detector for embedded applications. **IEEE Access**, v. 7, p. 133529–133538, 2019. Citado na página 24.

NING, G. et al. Spatially supervised recurrent convolutional neural networks for visual object tracking. In: **2017 IEEE International Symposium on Circuits and Systems (ISCAS)**. [S.l.: s.n.], 2017. p. 1–4. Citado na página 32.

REDMON, J. et al. You only look once: Unified, real-time object detection. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 779–788. Citado na página 23.

REDMON, J.; FARHADI, A. Yolo9000: Better, faster, stronger. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 6517–6525. Citado 3 vezes nas páginas 23, 30 e 31.

REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. In: . arXiv preprint arXiv:1804.02767: [s.n.], 2018. Citado 10 vezes nas páginas 13, 18, 24, 29, 30, 31, 40, 43, 44 e 46.

REN, S. et al. Faster r-cnn: towards real-time object detection with region proposal networks. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 39, n. 6, p. 1137–1149, 2017. Citado 4 vezes nas páginas 13, 22, 23 e 40.

SANTOS, M. M. D. et al. Real-time adaptive object localization and tracking for autonomous vehicles. **IEEE Transactions on Intelligent Vehicles**, p. 1–1, 2020. Citado 2 vezes nas páginas 14 e 29.

STOLLENGA, M. F. et al. Deep networks with internal selective attention through feedback connections. **NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems**, v. 2, p. 3545–3553, 2014. Citado na página 28.

UIJLINGS, J. et al. Selective search for object recognition. **International Journal of Computer Vision**, v. 4, p. 154–171, 2013. Citado 2 vezes nas páginas 19 e 20.

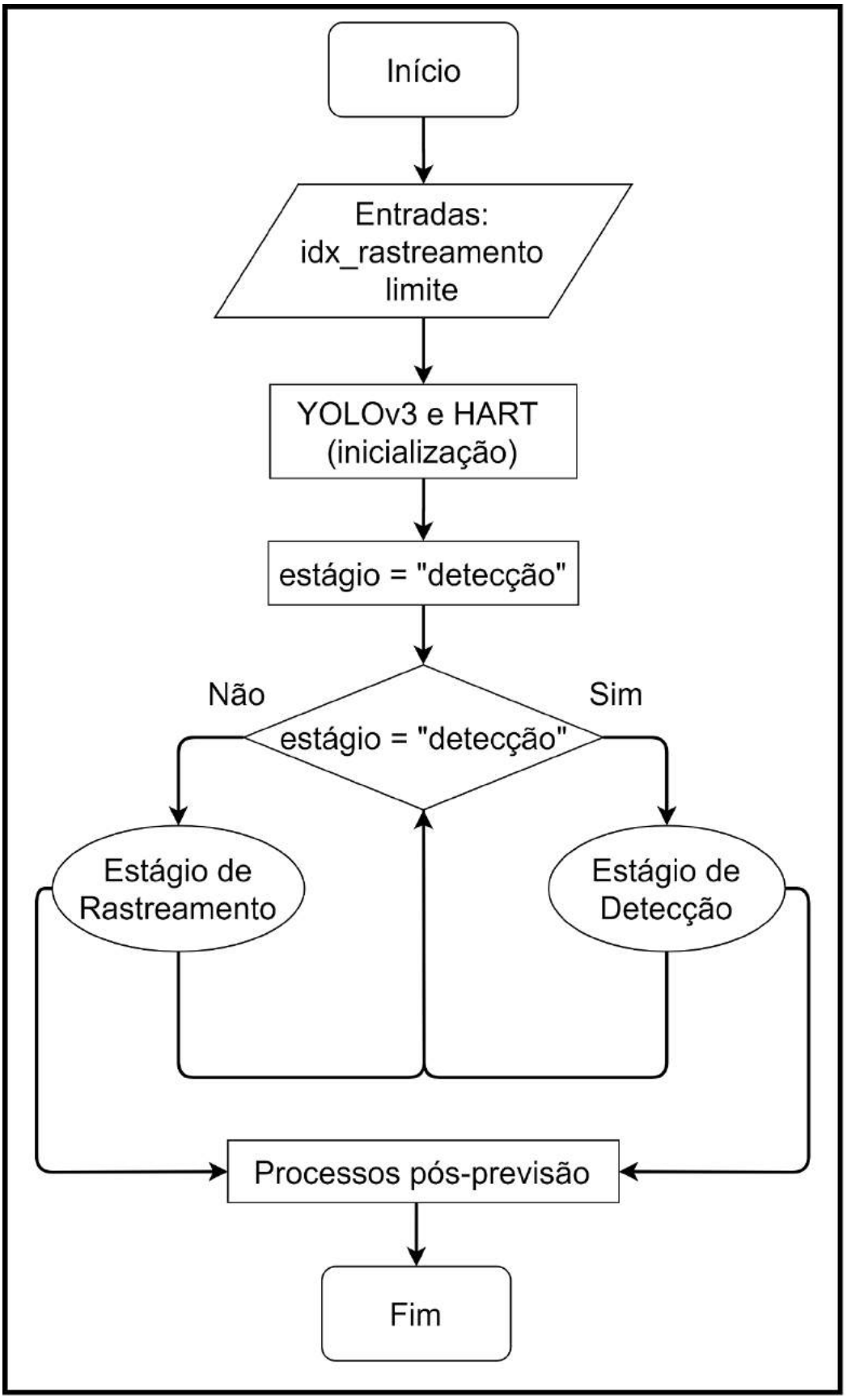
WOJKE, N.; BEWLEY, A.; PAULUS, D. Simple online and realtime tracking with a deep association metric. In: **2017 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2017. p. 3645–3649. Citado na página 27.

YADAV, S.; PAYANDEH, S. Understanding tracking methodology of kernelized correlation filter. In: **2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)**. [S.l.: s.n.], 2018. p. 1330–1336. Citado 2 vezes nas páginas 25 e 26.

YU, Jiahui et al. Unitbox: An advanced object detection network. **MM '16: Proceedings of the 24th ACM international conference on Multimedia**, 2016. Citado na página 35.

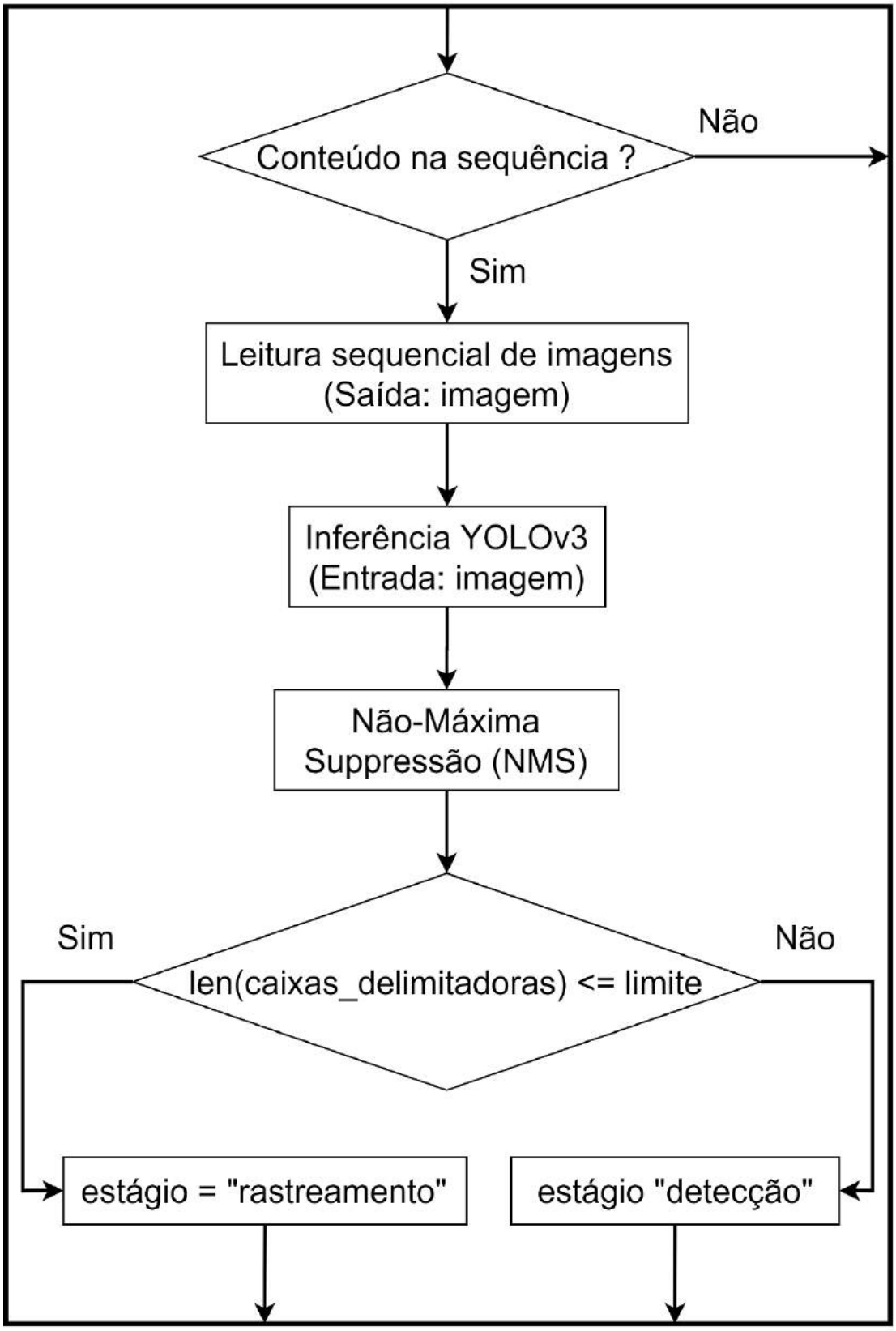
APÊNDICE A – FLUXOGRAMAS DO ALGORITMO

Figura 31 – Fluxograma geral do algoritmo desenvolvido.



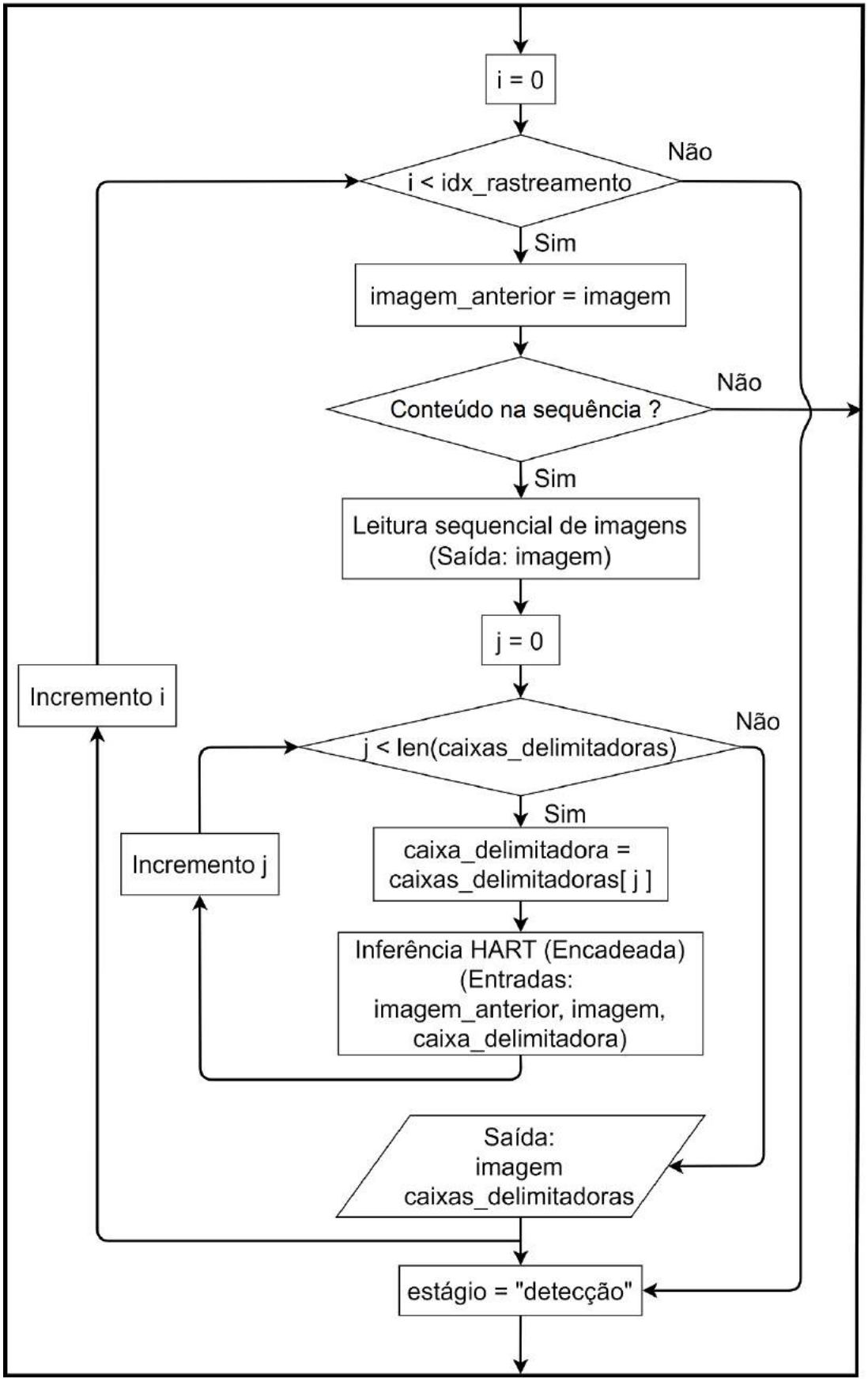
Fonte: Autoria própria.

Figura 32 – Fluxograma do estágio de detecção.



Fonte: Autoria própria.

Figura 33 – Fluxograma do estágio de rastreamento.



Fonte: Autoria própria.