Veículos Aéreos Não Tripulados para a Vigilância de Áreas Urbanas em Cidades Inteligentes

Mathias A. G. de Menezes[1], **Ricardo Maroquio B.**[2], **Erick M. Moreira**[1], **Hebert Azevedo Sá.**[1], **Paulo F. F. Rosa**[1]

[1]Departamento de Ciência e Tecnologia – Instituto Militar de Engenharia (IME)
Praça Gen. Tibúrcio, 80 – Urca, Rio de Janeiro – RJ, 22290-270

[2]Coordenadoria de Informática
Instituto Federal do Espírito Santo (IFES) – Vitória, ES – Brasil

{mathiasdemenezes,emenezes,rpaulo,azevedo}@ime.eb.br, maroquio@gmail.com

Abstract. *The present work presents a tracking application that integrates object detection with a Region-based Convolutional Neural Network as the object detector and the Discriminatory Correlation Filter with Channel and Spatial Reliability as the tracking algorithm for the proposed tracking method. This approach has the objective and motivation to assist the preventive actions of security systems, used in the context of Smart Cities, in urban structures and regions. The generated model results show an average accuracy of 92% for the object tracker when applied to the video sequences of the image dataset.*

Resumo. *Esta pesquisa apresenta uma aplicação de rastreamento que integra a detecção de objetos com uma Rede Neural Convolucional baseada em Região como detector de alvos de interesse e o Filtro de Correlação Discriminativa com Canal e Confiabilidade Espacial como algoritmo de rastreamento para o método proposto. Esta abordagem tem o objetivo e a motivação de auxiliar as ações preventivas de sistemas de segurança, empregados no contexto de Cidades Inteligentes, em estruturas e regiões urbanas. Os resultados do modelo gerado mostram uma precisão média de 92% para o rastreador de objetos quando aplicado às sequências de vídeo do conjunto de imagens.*

1. Introdução

Os Veículos Aéreos Não Tripulados (VANTs) são comumente usados em operações policiais e militares [Samad et al. 2007], [Semsch et al. 2009], especialmente para monitorar e rastrear entidades móveis suspeitas em áreas urbanas de risco, como favelas, zonas de guerra e regiões conflagradas. Essa abordagem permite aprender e entender o comportamento dessas entidades, fornecendo inteligência e prontidão para se antecipar e agir nos locais e rotas de locomoção [Geng et al. 2018], [Hu et al. 2021]. A vigilância e o rastreamento são particularmente desafiadores nesses ambientes. Isso acontece porque geralmente o monitoramento é realizado por meio de câmeras de vigilância convencionais, que podem ser burladas por dispositivos eletrônicos usados por criminosos [Daikoku et al. 2013]. Muitas pesquisas foram feitas para desenvolver estratégias de detecção e rastreamento para permitir que os VANTs realizem esses tipos de missões. Mas outro grande desafio na coordenação e implantação de VANTs é a quantidade de recursos humanos necessários para essas missões. Atualmente, o controle e a coordenação

de VANTs geralmente exigem dois ou três operadores humanos para manuseá-los. Além disso, segundo [Samant and Chang 2010], desenvolver uma aplicação de rastreamento em um ambiente urbano tem suas complexidades, como em zonas de exclusão aérea e estabelecimento da altura do horizonte da cidade e etc.

Com a motivação de contribuir para a solução deste problema, é proposto um método para detecção e rastreamento de entidades suspeitas, utilizando a Rede Neural Convolucional Baseada em Região (R-CNN) e o Filtro de Correlação Discriminativa com Canal e Confiabilidade Espacial (DCF-CSR), um rastreador implementado em python na biblioteca OpenCV.

O restante do artigo está organizado da seguinte forma: os trabalhos relacionados são apresentados na Seção 2; a formulação do problema é descrita na Seção 3; o método de rastreamento proposto é descrito na Seção 4; a avaliação da abordagem proposta é realizada na Seção 5; e, conclui-se o artigo na Seção 6.

2. Trabalhos Relacionados

A segurança e a proteção de zonas urbanas são garantidas pelo crescimento maciço de Cidades Inteligentes e aplicativos de Internet das Coisas. Um exemplo disso são os dados gerados por câmeras de vigilância em equipamentos de aviação, como VANTs. E a realização de tarefas de detecção, de rastreamento e posicionamento de objetos da perspectiva de um VANT pode efetivamente melhorar a eficiência do monitoramento para uma vigilância urbana inteligente [Baldoni et al. 2017]. Com base neste requisito, [Thakur et al. 2021] apresentou uma solução unindo dois modelos para detecção de objetos, YOLO v4 e DeepSORT. Para isso os autores elaboraram duas equações: uma equação de estado e uma equação de medição, e um filtro de partículas baseado em multímodo interativo para realizar a estimativa de estado dos alvos em trajetória não linear. Os resultados da simulação mostram que o algoritmo proposto pode automaticamente detectar e rastrear veículos em ambientes urbanos. Além disso, o algoritmo de filtro de partículas baseado em um multímodo interativo melhora significativamente o desempenho do VANT em execuções de manobras.

Nas Cidades Inteligentes, sempre haverá emergências imprevistas que devem ser resolvidas para manter a ordem comum. Portanto, é necessário a implementação de um ou mais sistemas inteligentes para detectar ameaças e lidar com elas [Khan et al. 2021]. Em [Wan et al. 2018], foi apresentada uma arquitetura de sistema composta por um agente central e três camadas: uma camada VANT, uma camada multi-robô e uma camada de rede de sensores. Os drones atuaram como sensores móveis. Eles forneceram dados gerais de monitoramento para robôs aéreos e de transporte para um cenário de emergência. Os robôs em terra foram responsáveis por obter dados de monitoramento detalhados e lidar com essas emergências. A rede de sensores continuou monitorando o ambiente e auxiliando os robôs e drones na pista em suas tarefas. De acordo com os resultados, o agente central pode ajustar o sistema de acordo com os requisitos específicos da tarefa.

Para alcançar a urbanização inteligente, o monitoramento contínuo é necessário. A vigilância por vídeo através de câmeras de circuito fechado de televisão tem sido estudada há décadas, mas tem problemas diferentes, como cobertura de área limitada e falta de recursos de compartilhamento e rastreamento de localização. Por outro lado, os sensores de óticos montados em drones são mais escaláveis e flexíveis, com cobertura de vigilância

mais abrangente. Mas, ao mesmo tempo, os drones também enfrentam vários desafios, como recursos limitados de processamento e energia, efeitos de trepidação da câmera em *feeds* de vídeo e interferência de sinais de transmissão [Mao 2021]. E, partindo destas prerrogativas, [Dilshad et al. 2020] se concentra na vigilância por vídeo usando drones na detecção e rastreamento de objetos, sumarização de vídeo, monitoramento persistente do alvo, operação de busca e salvamento em ambiente hostil, gerenciamento de tráfego em cidades inteligentes e gerenciamento de desastres em uma situação apocalíptica.

Como foi mencionado anteriormente, um sistema de vigilância por vídeo é a integração de computadores, redes, comunicações e *codecs* de vídeo. Devido à sua arquitetura distribuída, processamento paralelo de imagens, fácil instalação e expansão, etc., é amplamente utilizado em educação, transporte, indústria e outros campos. No entanto, os aplicativos de vigilância por vídeo em Cidades Inteligentes também enfrentam desafios como eventos de vídeo em grande escala, baixa qualidade de transmissão de dados de vídeo, extensão de tempo e perda de integridade dos dados de vigilância [Utomo et al. 2020]. Tendo em vista estes problemas, [Jin et al. 2020] projetou uma série de algoritmos de otimização e estratégias de escalonamento baseados em enxames de VANTs. Primeiro, foi construída uma rede de cobertura total de dispositivos de enxame de VANTs em um ambiente de comunicação heterogêneo de Cidade Inteligente. Em segundo lugar, formulou-se o problema de escalonamento de enxames de VANT como um problema de empacotamento frágil de dois objetivos e foi projetado um algoritmo de escalonamento ótimo com desempenho de aproximação constante. Os resultados da simulação experimental demonstraram plenamente a eficácia, viabilidade e robustez do esquema proposto em termos de ciclo de vida do sistema, taxa de quadros de vídeo recodificável, relação entre o tempo de voo do VANT e o ciclo de vida do sistema, custo de transmissão e atraso.

A detecção e o rastreamento de múltiplos objetos é um problema na área de visão computacional. A extração de features e o processamento de oclusão são dois elementos-base para detectar e rastrear vários objetos. No entanto, os métodos existentes não funcionam bem nesses aspectos ao detectar vários objetos [Dange and Momin 2019]. De acordo com [Li et al. 2018], as R-CNNs alcançaram grande sucesso na extração de features baseados em região, e o filtro de Peças baseado em Modelo de Peças-deformáveis (MPD) é adequado para detectar objetos em estado de oclusão. E em seu trabalho, foi proposta uma estrutura que integra R-CNN e MDP para detectar vários objetos. Além disso, foi proposto um novo filtro baseado no algoritmo de descoberta de subgrafos densos para refinar as previsões geradas pelo MPD. Ao combinar esses dois modelos, pode-se detectar cada objeto com alta precisão entre todos os objetos da imagem, especialmente se os objetos estiverem próximos um do outro. Ao contrário dos métodos tradicionais, o framework desenvolvido é capaz de detectar vários objetos pertencentes a várias classes, não apenas classes típicas como pessoas ou carros.

Para uma melhor compreensão, na Figura 1 estão apresentadas as diferenças de abordagem dos trabalhos relacionados e as contribuições desta pesquisa, ambos voltados para a solução de problemas que envolvem a detecção e o rastreamento dentro do contexto de Cidades Inteligentes.

	VANTs	Detecção	Rastreamento	R-CNN	DCF-CSR	Monitoramento Persistente	Múltiplas Entidades	Estimativa de Trajetória
Thakur et al. 2021	✓	✓	✓	-	-	✓	✓	✓
Wan et al. 2018	✓	✓	✓	-	-	-	✓	-
Dilshad et al. 2020	✓	✓	✓	-	-	✓	✓	-
Jin et al. 2020	✓	✓	✓	-	-	✓	✓	-
Li et al. 2018	-	✓	-	✓	-	-	✓	-
Esta pesquisa	✓	✓	✓	✓	✓	✓	-	-

Figura 1. Quadro comparativo de trabalhos relacionados.

3. Contextualização do Problema

A motocicleta é um veículo comumente utilizado por vândalos, criminosos e até terroristas. Geralmente, esses indivíduos andam em grupos ou duplas e se misturam com os cidadãos que transitam pelas avenidas e ruas da cidade. Portanto, é difícil identificar um veículo suspeito, seja por uma aeronave de grande porte, como um helicóptero, ou por um pequeno VANT sem uma aplicação embarcada. Assim, é necessário criar o modelo do alvo a ser classificado, preservando seus aspectos para rastreá-lo.

A atuação dos VANTs nesses tipos de ambiente é mostrada na Figura 2: O VANT vai iniciar o voo com a finalidade de estabelecer um perímetro de vigilância ao redor da construção, para evitar que qualquer entidade/objeto suspeito adentre o prédio. E os recursos que o VANT vai usar para resolver esse problema são 2: uma trajetória orbital circular (já que se trata de uma aeronave de asa fixa) e a informação visual captada por uma câmera acoplada ao próprio VANT para detectar e identificar se é uma possível ameaça.

Uma abordagem ideal para resolver este problema é o uso de Redes Neurais Convolucionais Baseadas em Região (R-CNN). Essas redes usam como dados de entrada regiões recortadas das imagens para detectar se um número de objetos de uma determinada categoria está presente, bem como detectar onde cada objeto está localizado na imagem. Este tipo de rede é bastante robusto e pode discernir vários objetos agrupados simultaneamente, independente da oclusão de partes do objeto alvo.

Os VANTs utilizados nesta tarefa vem equipados com uma câmera infravermelha integrada e um link de rádio para comunicação. E geralmente, dois operadores humanos gerenciam o VANT durante a missão: onde um opera a câmera e o outro movimenta a aeronave. Basicamente, a aplicação que é proposta visa automatizar a detecção e o rastreamento de entidades suspeitas e a monitorá-las dentro do campo de visão da câmera,

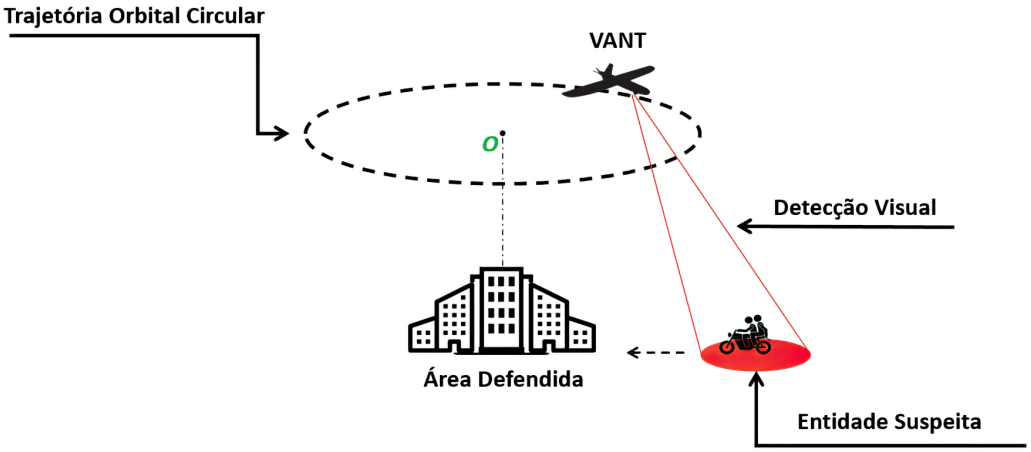

Figura 2. O cenário contém os seguintes atores: um VANT, uma localidade a ser defendida e um objeto suspeito.

contribuindo para a segurança e vigilância no contexto de Cidades Inteligentes.

Conforme está mostrado na Figura 3, o procedimento se dá da seguinte forma: o VANT inicia a missão, a câmera faz a varredura da superfície, tendo como referencia a altitude da aeronave, a angulação da aeronave (o ângulo de guinada e a câmera). Sempre com a finalidade de encontrar a posição do alvo e manter o rastreamento. E enquanto o alvo não for detectado pela câmera, o VANT continuará a trajetória de voo, mas se o alvo for detectado a atuação da aeronave é voltada para manter a trajetória de rastreamento.

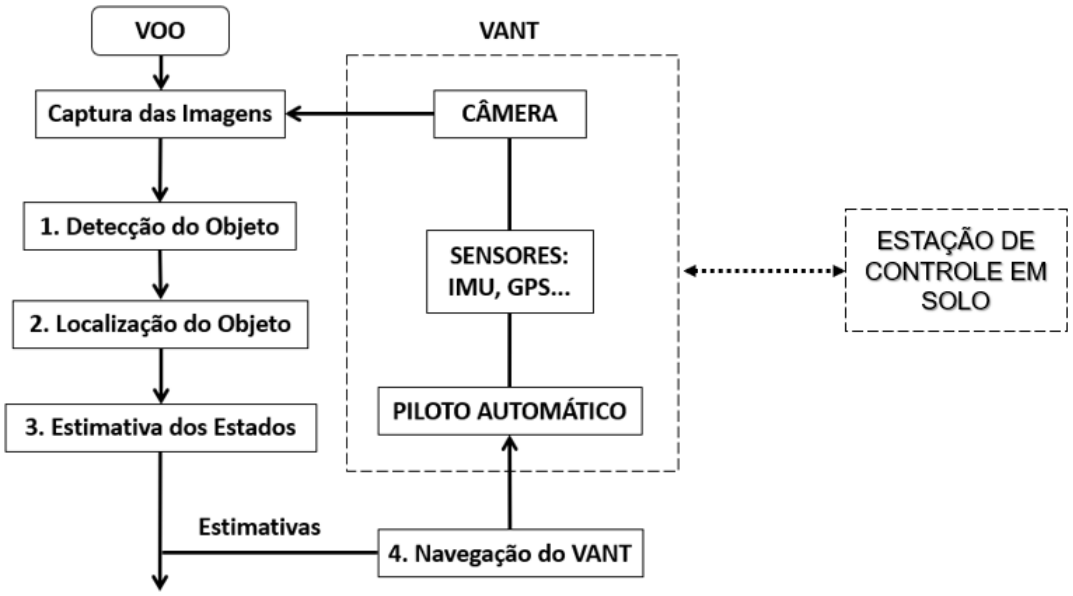

Figura 3. Atuação do VANT.

4. Método Proposto

O método de rastreamento de objetos proposto consiste em duas partes: (i) um modelo de detecção de objetos gerado a partir dos resultados do treinamento e (ii)um rastreador de objetos que usa o (i) modelo como referência. O algoritmo de busca seletiva extrai

regiões de interesse das imagens e as entrega para um modelo de detecção de objetos para prever a sua localização e classificá-los em classes de objetos. As previsões de saída de detecção são fornecidas ao algoritmo de rastreamento para rastrear as *bouding boxes* previstas das classes de objetos. Qualquer outra condição do rastreador que altere a previsão de localização do objeto acionará o classificador de objeto e reiniciará o processo do zero.

O rastreador, chamado Filtro de Correlação Discriminativa com Confiabilidade de Canal e Espacial (DCF-CSR), usa a confiabilidade espacial para definir o suporte de filtros para regiões selecionadas de uma parte do frame para rastreamento. Isso expande e posiciona a área selecionada e rastreia áreas ou objetos não retangulares. Este rastreador tem duas funções padrão, HoGs e Colornames. Além disso, funciona muito bem para frames abaixo de 25 fps [Alan Lukežič and Kristan 2018].

O método se divide nas seguintes etapas mostradas na Figura 4: as imagens do dataset são processadas por uma R-CNN. Onde as regiões de interesse de uma imagem vão ser destacadas com o algoritmo de Busca Seletiva, e a rede neural convolucional utilizada será a VGG-16. (especifica para reconhecimento/detecção de objetos). Desta forma, o modelo do objeto gerado vai alimentar o filtro DCF-CSR, resultando no rastreamento.

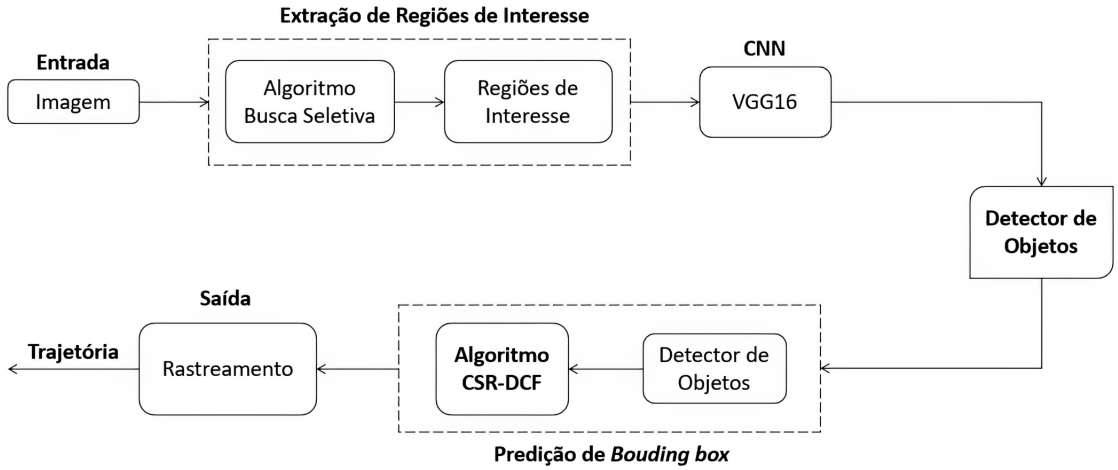

Figura 4. Método proposto.

4.1. R-CNN para Detecção de Objetos

O método R-CNN em [Girshick et al. 2013] é um modelo de aprendizado de máquina que realiza segmentação com base nos resultados de detecção de objetos. O R-CNN inicialmente usa um algoritmo de busca seletiva para extrair um grande número de regiões propostas e, em seguida, calcula os recursos de cada região por meio de uma rede neural convolucional (CNN). Por fim, classifica cada região usando um classificador linear específico, geralmente uma máquina de vetores de suporte (SVM). O R-CNN é capaz de realizar tarefas mais complexas, como detecção de objetos e segmentação grosseira de imagens.

4.2. Regiões Extraídas

Inicialmente, cerca de 2.000 regiões propostas são extraídas usando o algoritmo de Busca Seletiva [Uijlings et al. 2013], que é baseado em técnicas tradicionais simples de visão

computacional. O processo, mostrado na Figura 5 acontece da seguinte forma: primeiro, cada Região de Interesse (RoI) proposta é deformada em uma imagem quadrada de tamanho padrão; segundo, a imagem é alimentada a uma CNN que gera um *array* com 4096 características dimensionais como saída; e, finalmente, um SVM classifica o *array* de recursos produzindo duas saídas: uma classificação e uma indicação de desvio (*offset*) que pode ser usada para ajustar a *bouding box*.

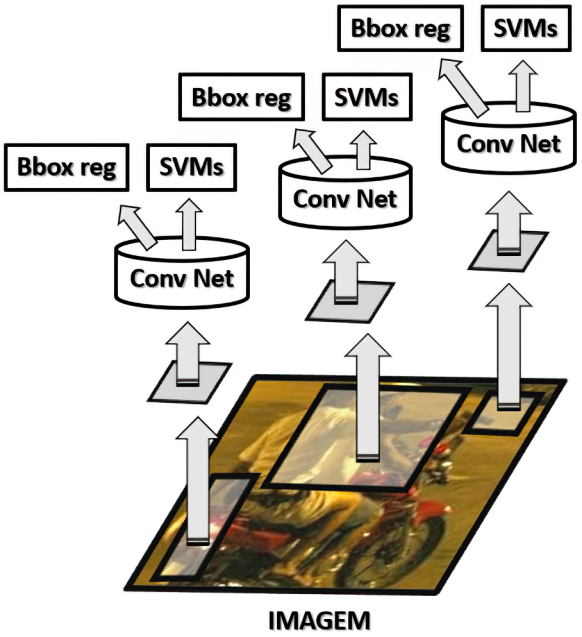

Figura 5. Arquitetura R-CNN: Cada RoI proposto é passado pela CNN para extrair recursos e depois por um classificador SVM.

4.3. Processamento de Características Convolucionais

As regiões propostas extraídas são repassadas à CNN. Para essa tarefa optou-se pela VGG-16 [Simonyan and Zisserman 2014], na Figura 6, que é uma CNN muito robusta e usada para a tarefa de detecção de objetos. Basicamente, a CNN receberá uma região proposta passando por uma série de camadas convolucionais, não lineares, *clustering* e totalmente conectadas para obter duas saídas. Uma saída é uma única classe que melhor descreve a região proposta. A CNN é estruturada em quatro camadas, ou estágios: camada de convolução, camada de agrupamento, camada de normalização e camada totalmente conectada.

4.4. Rastreamento de Objetos

O método de rastreamento proposto se baseia no algoritmo DCF-CSR. Além disso, este algoritmo foi implementado e integrado à biblioteca OpenCV como um módulo de Rede Neural Profunda, em inglês *Deep Neural Network* (DNN). Desta forma é proposto um aplicativo de rastreamento que integra detecção de objetos com R-CNN como detector de objetos e o DCF-CSR como algoritmo de rastreamento para o método de rastreamento.

5. Resultados

O conjunto de imagens foi coletado de repositórios do GitHub. A diferença entre a coleção de imagens desta proposta e outras coleções de imagens é que estas imagens

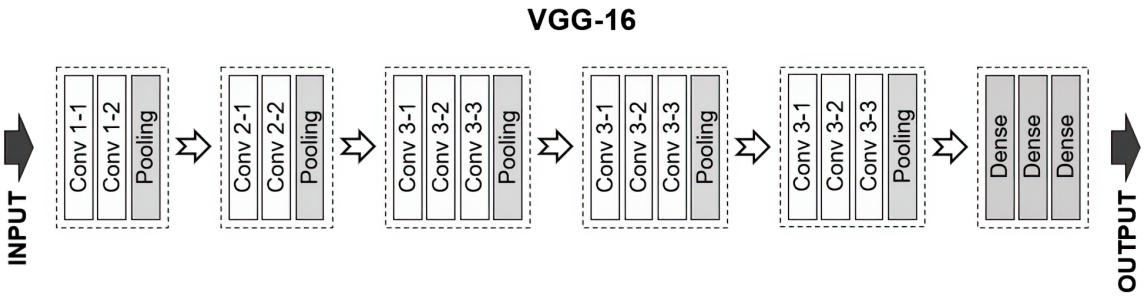

Figura 6. VGG-16: Topologia da CNN.

são captadas por câmeras de drones, câmeras de helicópteros e câmeras de segurança instaladas no ambiente urbano local. Foram coletadas 14226 imagens no total, que foram divididas em 9958 imagens para o processo de treinamento e 4268 imagens para o processo de teste. A saída das classificações tem duas categorias: Entidade Suspeita (ES) e Entidade Não Suspeita (EN).

Depois de submeter a CNN em exatos 12000 ciclos de treinamento com o conjunto de imagens, obteve-se o modelo para classificador de objetos. Em seguida, o mesmo foi validado aplicando-o às imagens de diferentes recursos, atingindo a taxa de precisão de 94%, conforme mostrado na matriz de confusão na Figura 7.

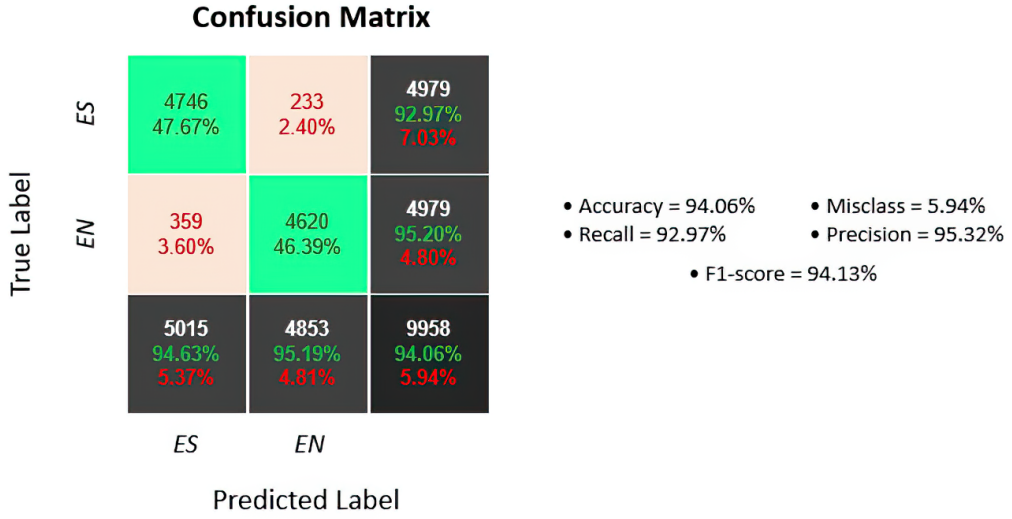

Figura 7. Matriz de Confusão: Estatísticas do classificador treinado.

Na Figura 8 pode-se ver os resultados notáveis do modelo de rastreamento, onde ele rastreia os objetos em cada frame. Durante o rastreamento, principalmente em sequências de vídeo, as diversas posições do objeto, os locais e as formas dificultam o rastreamento. O que significa que uma grande quantidade de imagens no conjunto de dados com diferentes posições e ambientes melhoram os resultados.

O método de rastreamento foi testado em sequências de vídeo feitas por helicóptero e os resultados obtidos foram bastante promissores. Os experimentos mostram que o algoritmo rastreador pode detectar novamente o objeto assim que ele sair do quadro atual. A Figura 9 e a Figura 10 apresentam os resultados qualitativos para as sequências

Figura 8. Rastreamento de Objetos: Resultado do classificador de objetos, imagens do conjunto de dados.

de vídeo feitas por helicópteros, com enquadramentos abaixo de 25 fps. E mesmo que a forma ou a aparência do objeto rastreado mude, o rastreador mantém o objeto ”travado”corretamente. Os resultados experimentais no conjunto de imagens mostram que a proposta alcança melhor desempenho do que o método R-CNN e o filtro DCF-CSR quando aplicados sozinhos para detectar vários objetos.

Figura 9. Sequência de Vídeo 1: Resultados qualitativos do método de rastreamento proposto em quadros extraídos.

Para medir a qualidade do rastreador de objetos, usou-se a métrica *Intersection over Union* (IoU). O método IoU calcula a razão entre a área de sobreposição e a área conjunta entre a caixa delimitadora prevista e a caixa delimitadora de *ground truth*. A IoU é uma métrica de avaliação usada para medir a precisão de um detector/rastreador de objetos em relação a um conjunto de dados específicos. Essa métrica de avaliação é frequentemente usada em desafios de detecção e rastreamento de objetos, como em

Figura 10. Sequência de Vídeo 2: Resultados qualitativos do método de rastreamento proposto em quadros extraídos.

abordagens com R-CNN, Faster R-CNN, YOLO e Deep SORT [Pramanik et al. 2022]. Entretanto, o algoritmo real usado para gerar as previsões não importa. A IoU é simplesmente uma métrica de avaliação. Qualquer algoritmo que forneça caixas delimitadoras previstas como saída pode ser avaliado usando IoU.

Formalmente, para aplicar a IoU para avaliar um detector/rastreador de objetos (arbitrário), precisa-se: (i) das caixas delimitadoras de *ground truth* (ou seja, as caixas delimitadoras rotuladas à mão do conjunto de testes que especificam onde o objeto está na Imagem); e (ii), as caixas delimitadoras previstas do modelo gerado conforme mostram a Figura 11. A precisão média de classificação do método proposto atingiu 92% quando aplicado às sequências de vídeo.

$$\text{IoU} = \frac{S_{Bbox\ Prevista} \cap S_{Bbox\ Ground\ Truth}}{S_{Bbox\ Prevista} \cup S_{Bbox\ Ground\ Truth}}$$

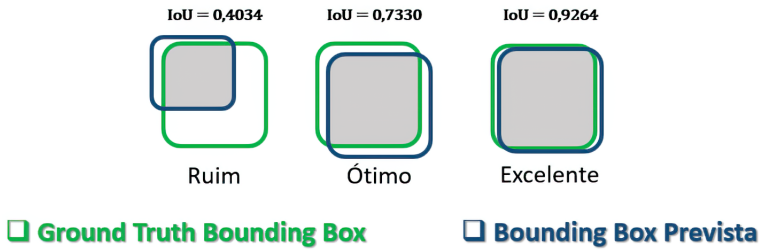

Figura 11. O cálculo da IoU é feito dividindo a área de sobreposição entre as caixas delimitadoras pela área de união.

6. Conclusão

Nesta pesquisa é apresentado um rastreador de objetos para rastrear entidades suspeitas em ambientes urbanos. Com este método, integrou-se o modelo de classificador de objetos baseado em aprendizado profundo com a versão de implementação OpenCV do

algoritmo DCF-CSR suportado por módulos DNN do próprio OpenCV. Os resultados mostraram que o modelo classificador de objetos foi satisfatório após 12000 ciclos de treinamento, com apenas 9958 imagens para treinamento e 4268 imagens para teste. No entanto, quando aplicado à sequências de vídeo em imagens capturadas com baixa resolução por helicópteros, o rastreador teve um desempenho abaixo do esperado. Essas imagens precisam estar em boa resolução, com maior quantidade de angulação, detalhando a posição e a forma das entidades presentes.

Em conclusão, o trabalho teve algumas limitações. É necessário uma maior quantidade de imagens representando a classe de entidades suspeitas, bem como uma melhoria na estrutura da CNN, aumentando suas camadas convolucionais e o número de ciclos de treinamento. Em trabalhos futuros serão implementadas formas de contornar a oclusão dos objetos a fim de manter o monitoramento constante dos objetos e além de estimar a trajetória dos mesmos. Também serão usadas variações do modelo R-CNN, como um Mask R-CNN. Pois, com uma Mask R-CNN é possível construir correlações entre objetos próximos. Dessa forma, essas alterações podem aumentar a precisão do método proposto, possivelmente tornando o rastreador mais robusto e eficaz.

Agradecimentos

Este estudo foi realizado com o apoio do Programa de Cooperação Acadêmica em Defesa Nacional (PROCAD-DEFESA) e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Finanças 001.

Referências

Alan Lukežič, Tomáš Vojíř, L. Z. J. M. and Kristan, M. (2018). Discriminative correlation filter with channel and spatial reliability. *International Journal of Computer Vision*, 126:671–688.

Baldoni, G., Melita, M., Micalizzi, S., Rametta, C., Schembra, G., and Vassallo, A. (2017). A dynamic, plug-and-play and efficient video surveillance platform for smart cities. In *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 611–612.

Daikoku, M., Karungaru, S., and Terada, K. (2013). Automatic detection of suspicious objects using surveillance cameras. In *The SICE Annual Conference 2013*, pages 1162–1167.

Dange, A. D. and Momin, B. F. (2019). The cnn and dpm based approach for multiple object detection in images. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1106–1109.

Dilshad, N., Hwang, J., Song, J., and Sung, N. (2020). Applications and challenges in video surveillance via drone: A brief survey. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 728–732.

Geng, H., Guan, J., Pan, H., and Fu, H. (2018). Multiple vehicle detection with different scales in urban surveillance video. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–4.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587.

Hu, C., Qu, G., Shin, H.-S., and Tsourdos, A. (2021). Distributed synchronous cooperative tracking algorithm for ground moving target in urban by uavs. *International Journal of Systems Science*, 52(4):832–847.

Jin, Y., Qian, Z., and Yang, W. (2020). Uav cluster-based video surveillance system optimization in heterogeneous communication of smart cities. *IEEE Access*, 8:55654–55664.

Khan, S., Teng, Y., and Cui, J. (2021). Pedestrian traffic lights classification using transfer learning in smart city application. In *2021 13th International Conference on Communication Software and Networks (ICCSN)*, pages 352–356.

Li, J., Wong, H.-C., Lo, S.-L., and Xin, Y. (2018). Multiple object detection by a deformable part-based model and an r-cnn. *IEEE Signal Processing Letters*, 25(2):288–292.

Mao, R. (2021). Real-time small-size pixel target perception algorithm based on embedded system for smart city. In *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pages 505–511.

Pramanik, A., Pal, S. K., Maiti, J., and Mitra, P. (2022). Granulated rcnn and multi-class deep sort for multi-object detection and tracking. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):171–181.

Samad, T., Bay, J. S., and Godbole, D. (2007). Network-centric systems for military operations in urban terrain: The role of uavs. *Proceedings of the IEEE*, 95(1):92–107.

Samant, A. and Chang, K. (2010). Image-based tracking and sensor resource management for uavs in an urban environment. *Proceedings of SPIE - The International Society for Optical Engineering*.

Semsch, E., Jakob, M., Pavlicek, D., and Pechoucek, M. (2009). Autonomous uav surveillance in complex urban environments. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 82–85, Milan, Italy.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*.

Thakur, D. N., Nagrath, P., Jain, R., Saini, D., Sharma, N., and D, J. (2021). *Artificial Intelligence Techniques in Smart Cities Surveillance Using UAVs: A Survey*, pages 329–353.

Uijlings, J., Sande, K., Gevers, T., and Smeulders, A. (2013). Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171.

Utomo, W., Bhaskara, P. W., Kurniawan, A., Juniastuti, S., and Yuniarno, E. M. (2020). Traffic congestion detection using fixed-wing unmanned aerial vehicle (uav) video streaming based on deep learning. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pages 234–238.

Wan, S., Lu, J., Fan, P., and Letaief, K. B. (2018). To smart city: Public safety network design for emergency. *IEEE Access*, 6:1451–1460.