INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Inteligência Artificial e Aprendizagem Profunda na Videovigilância: Revisão Sistemática de Métodos de Deteção Comportamental

Vitor Santos Dourado

Mestrado em Tecnologias Digitais para o Negócio

Orientador:
Professor Doutor João Carlos Amaro Ferreira,
Professor Auxiliar com Agregação,
ISCTE-IUL

Maio, 2026

TECNOLOGIAS
E ARQUITETURA

Inteligência Artificial e Aprendizagem Profunda na Videovigilância: Revisão Sistemática de Métodos de Deteção Comportamental

Vitor Santos Dourado

Mestrado em Tecnologias Digitais para o Negócio

Orientador:
Professor Doutor João Carlos Amaro Ferreira,
Professor Auxiliar com Agregação,
ISCTE-IUL

Maio, 2026

À minha avó, mãe, esposa e filha, mulheres da minha vida.

Agradecimento

Agradeço a Deus, em primeiro lugar, pela força, sabedoria e perseverança concedidas ao longo de todo este percurso. Sem a Sua presença constante, esta conquista não teria sido possível.
À minha esposa e filha, pelo amor e paciência nos momentos de ausência, pelo apoio nos dias mais exigentes e por acreditar sempre em mim. Esta vitória é também de vocês.
Ao meu orientador, Professor João Ferreira, pela disponibilidade, pelas valiosas contribuições e por todo o conhecimento partilhado ao longo do desenvolvimento deste trabalho.
Aos professores que fizeram parte deste percurso académico, pelos ensinamentos transmitidos, e pelo contributo decisivo para a minha formação.
Ao meu irmão, Flávio, pelo incentivo constante e por estar sempre presente.
A todos os que, de forma direta ou indireta, contribuíram para a concretização deste mestrado, o meu sincero agradecimento.

Resumo

A vigilância por vídeo inteligente é, atualmente, uma das áreas de aplicação mais relevantes da inteligência artificial e da aprendizagem profunda ou *deep learning*. Seguindo as diretrizes do PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) 2020, este trabalho apresenta uma revisão sistemática da literatura sobre técnicas de *deep learning* para identificar e monitorizar padrões comportamentais em sistemas de vigilância por vídeo. Foram analisados estudos selecionados extraídos de bases de dados académicas de referência (Scopus, IEEE (*Institute of Electrical and Electronics Engineers*) Xplore e Web of Science) entre 2019 e 2025, dedicados à aplicação de arquiteturas de aprendizagem profunda na deteção de atividades humanas, identificação de anomalias e monitorização de multidões.

Os resultados sugerem uma progressão clara das arquiteturas com as aplicações mais evoluídas disponíveis para deteção de anomalias e de atividades suspeita, além da monitorização de multidões, com precisão de 75% a 97%, dependendo do conjunto de dados e da arquitetura.

A revisão aborda os principais desafios técnicos como o processamento em tempo real, robustez e escalabilidade e também os desafios éticos e regulatórios como privacidade, GDPR (*General Data Protection Regulation*) e viés algorítmico. Identifica lacunas na literatura, incluindo a necessidade de uma generalização mais ampla para o mundo real e a incorporação de privacidade por *design*. Direções de pesquisas futuras são sugeridas, como arquiteturas leves para *edge computing,* que processa dados na borda da rede, próxima à fonte dos dados, normalmente em dispositivos com poucos recursos computacionais*,* e *frameworks* de explicabilidade como XAI (*Explainable Artificial Intelligence*).

Palavras-chave: vigilância por vídeo, aprendizagem profunda, deteção de anomalias, reconhecimento de comportamentos, revisão sistemática

Abstract

Intelligent video surveillance is currently one of the most significant areas of application for artificial intelligence and deep learning. Following the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) 2020 guidelines, this thesis presents a systematic review of the literature on deep learning techniques for identifying and monitoring behavioral patterns in video surveillance systems. We analyzed selected studies extracted from leading academic databases (Scopus, IEEE (Institute of Electrical and Electronics Engineers) Xplore, and Web of Science) between 2019 and 2025, focusing on the application of deep learning architectures in human activity detection, anomaly identification, and crowd monitoring.

The results suggest a clear progression in architectures with the most advanced applications available are for anomaly detection and the detection of suspicious activities, as well as crowd monitoring, with accuracy ranging from 75% to 97%, depending on the dataset and architecture. The review addresses key technical challenges such as real-time processing, robustness, and scalability, as well as ethical and regulatory challenges such as privacy, GDPR (General Data Protection Regulation), and algorithmic bias. It identifies gaps in literature, including the need for broader generalization to the real world and the incorporation of privacy by design. Directions for future research are suggested, such as lightweight architectures for edge computing and explainability frameworks as XAI (Explainable Artificial Intelligence).
Keywords: video surveillance, deep learning, anomaly detection, behavior recognition, systematic review

Índice

Agradecimento……....i
Resumo……..iii
Abstract……...……v
Índice……vii
Lista de Figuras……..ix
Lista de Tabelas……..xi
Lista de Acrónimos e Nomenclaturas……...…xiii

Capítulo 1. Introdução.……1
1.1. Motivação……...1
1.2. Contexto……..1
1.2.1. História da vigilância por vídeo……...1
1.2.2. Revolução da aprendizagem profunda…….3
1.2.3. Contexto técnico atual…….3
1.2.4. Contexto social e regulatório……...3
1.3. Problema da Pesquisa ……4
1.4. Objetivos…….5
1.4.1. Objetivo geral……..5
1.4.2. Objetivos específicos……...5
1.4.3. Estrutura do documento……...5

Capítulo 2. Revisão da Literatura……7
2.1. Metodologia de Pesquisa……7
2.1.1. Estratégia de busca……..7
2.1.2. Base de dados selecionadas…….……7
2.1.3. Critérios de inclusão de exclusão…….8
2.2. Processo de Seleção PRISMA…….……8
2.3. Estado da Arte……...10
2.3.1. Deteção de Anomalias e Atividades Suspeitas……..…10
2.3.2. Monitorização e Análise de Multidões……..…11
2.3.3. Sistemas Integrados e Implementação……...12
2.4. Diagrama Conceptual da Revisão……13

2.5. Síntese dos Estudos Utilizados……..14
2.5.1. Estudos Sobre Arquiteturas de Aprendizado Profundo……..……14
2.5.2. Estudos Sobre Deteção de Anomalias e Atividades Suspeitas……...…15
2.5.3. Estudos Sobre Monitorização de Multidões……...…16
2.5.4. Estudos Sobre Ferramentas Integradas e Implementação……..……16
2.5.5. Estudos Sobre Análise Comparativa de Desempenho……...…17

Capítulo 3. Comparação e Discussão dos Resultados……19
3.1. Evolução e Comparação das Arquiteturas de Aprendizagem Profunda……19
3.2. Desempenho e Trade-offs...……..…20
3.3. Análise Temática...…….....21
3.4. Caso de Aplicação Prática: Sistema de Vigilância Inteligente Veesion.……23

Capítulo 4. Desafios Técnicos, Questões Éticas e Regulatórias, Lacunas e Direções Futuras...25
4.1. Desafios Técnicos de Implementação……...……25
4.2. Questões Éticas e Regulatórias……..…26
4.3. Lacunas Identificadas na Literatura…….....27
4.4. Limitações dos Datasets Utilizados..……...……...……28
4.5. Direções Futuras de Investigação……..……29

Capítulo 5. Conclusões e Perspetivas para o Futuro……..…31
5.1. Resumo dos Principais Resultados……31
5.2. Contribuições do Trabalho...……...……..…32
5.3. Limitações...……..…….....33
5.4. Perspetivas Futuras.……..33
5.5. Considerações Finais.……...35

Referências Bibliográficas……37

Lista de Figuras

Figura 1 - CCTV para monitorização de tráfego - Londres, 1956……...……...02
Figura 2 - Sala de controle de sistema de CCTV moderno……...…...……...02
Figura 3 - Fluxograma PRISMA para revisão sistemática…….……....09
Figura 4 - Diagrama conceptual de videovigilância inteligente…….…....13
Figura 5 - Tela da aplicação Veesion…….24

Lista de Tabelas

Tabela 1 - Quadro Resumo de Filtros e Critérios para Inclusão……..……...08

Tabela 2 - Quadro Comparativo de Desempenho de Arquiteturas….....……...20

Lista de Acrónimos e Nomenclaturas

AI Act	*Artificial Intelligence Act*
Bi-LSTM	*Bidirectional* LSTM
C3D	*Convolutional* 3D
CCTV	*Closed-Circuit Television*
CNN	*Convolutional Neural Network*
ConvGRU	*Convolutional Gated Recurrent Units*
ConvLSTM	*Convolutional Long Short Term Memory*
ConvNet	*Convolutional Neural Network*
FPGA	*Field-Programmable Gate Array*
FPS	*Frames* Por Segundo
GANs	*Generative Adversarial Networks*
GDPR	*General Data Protection Regulation*
GPU	*Graphics Processing Unit*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
GRU	*Gated Recurrent Units*
HMDB51	*Human Motion DataBase* - 51 Categorias
I3D	*Inflated* 3D
IA	Inteligência Artificial
IEEE	*Institute of Electrical and Electronics Engineers*
IP	*Internet Protocol*
LIME	*Local Interpretable Model-Agnostic Explanations*
LSTM	*Long Short Term Memory*
ms	Milessegundos
ONNX	*Open Neural Network Exchange*
PRISMA	*Preferred Reporting Items for Systematic reviews and Meta-Analyses*
R(2+1)D	*Residual(2 Spatial* + 1 *Temporal)Dimensions*
R-CNN	*Region-based Convolutional Neural Networks*
ResNet	*Residual Network*
RGB	*Red, Green, and Blue*
RNN	*Recurrent Neural Network*
RTX	*Ray Tracing Texel eXtreme*
SVM	*Support Vector Machine*
TPU	*Tensor Processing Unit*
UCF-Crime	*University of Central Florida* - Crime
UCSD	*University of California, San Diego*
UE	União Europeia
UCF101	*University of Central Florida* - 101 Categorias
UMN Crowd	*University of Minnesota Crowd*
VAE	*Variational Autoencoder*
VGG-16	*Visual Geometry Group* - 16 Camadas

ViT	*Vision Transformer*
XAI	*Explainable Artificial Intelligence*
YOLO	*You Only Look Once*

CAPÍTULO 1

Introdução

1.1 Motivação

A segurança pública e privada são uma preocupação crescente nas sociedades contemporâneas, impulsionada pelo rápido aumento populacional urbano, espaços públicos complexos e a necessidade de respostas rápidas a incidentes. Os sistemas convencionais de vigilância por vídeo dependem de operação humana contínua e são prejudicados pela fadiga do operador, pelas limitações do processamento simultâneo de múltiplos fluxos de vídeos e pela dificuldade da perceção de comportamentos sutis em tempo real [7], [22].

A convergência da inteligência artificial, especificamente *deep learning*, e dos sistemas de vigilância por vídeo apresenta uma oportunidade revolucionária para superar essas limitações. Os algoritmos de aprendizagem profunda podem verificar, de maneira automática e inteligente, grandes volumes de imagens, permitindo uma rápida deteção de anomalias, identificação de comportamentos suspeitos e a monitorização de multidões de maneira mais eficaz [14], [21]. O processamento em larga escala e automatizado, não apenas aumenta a eficiência operacional, também permite lidar com incidentes críticos, possivelmente salvando vidas e protegendo propriedades privadas e publicas [15].

Esta pesquisa é motivada pela necessidade de entender o atual estado da arte neste campo que evolui muito rapidamente, além de identificar as arquiteturas e técnicas mais eficazes, bem como mapear questões técnicas, éticas e regulatórias que permanecem. Em um cenário onde deve haver equilíbrio entre privacidade individual e a segurança da sociedade, é importante criar uma visão crítica ampla das capacidades, limitações e implicações dos sistemas inteligentes de videovigilância baseados em *deep learning*.

1.2 Contexto

1.2.1 História da Vigilância por Vídeo

O início da vigilância por vídeo remonta à década de 1940, quando os primeiros sistemas de circuito fechado de televisão (CCTV) foram integrados e introduzidos para uso militar ou industrial. O primeiro sistema registado no mundo, na Alemanha, foi instalado em 1942 para observar o lançamento de foguetes V-2. Nas décadas subsequentes, a tecnologia de CCTV expandiu gradualmente para aplicações de segurança comercial e de segurança pública como observada na Figura 1, com a limitação inicial do uso de gravação em fita magnética e sem nenhuma capacidade analítica automatizada [22].



Figura 1 - CCTV para monitorização de tráfego - Londres, 1956 [28]

A mudança para sistemas digitais nas décadas de 1990 e 2000 foi um marco decisivo ao permitir o armazenamento digital, o tráfego de dados em rede e, mais importante, o processamento computacional de imagens capturadas, o que expandiu o uso comercial com sistemas com múltiplos fluxos de vídeo monitorizados, como observado na Figura 2. Com esta digitalização, foram geradas as condições técnicas para a incorporação de algoritmos de visão computacional e, posteriormente, técnicas de aprendizagem automática. No entanto, os primeiros sistemas automatizados dependiam fortemente de regras simples e detetores de características manuais e, como resultado, tinham sérias limitações de robustez e generalização [26].



Figura 2 - Sala de controle de sistema de CCTV moderno [29]

1.2.2 Revolução da Aprendizagem Profunda

A revolução do *deep learning*, desencadeada após o sucesso da arquitetura AlexNet no desafio ImageNet em 2012, mudou completamente o campo da visão computacional. As redes neuronais convolucionais (CNNs) provaram ser altamente capazes na extração automática de características visuais hierárquicas, superando significativamente os métodos de aprendizagem baseados em características manuais. Este poder de aprender representações diretamente dos dados brutos tornou-se especialmente relevante para a análise de vídeo, onde a alta complexidade e diversidade de padrões visuais inviabilizam abordagens baseadas em regras para capturar seus detalhes [14].

Posteriormente, a evolução das arquiteturas de *deep learning* se estendeu das CNNs espaciais para a adição de dimensões temporais por meio de redes neuronais recorrentes (RNNs), memórias de longo e curto prazo (LSTMs), unidades recorrentes fechadas (GRUs) e convoluções 3D. Mais recentemente, a arquitetura Transformer, desenvolvida originalmente para processamento de linguagem natural, foi adaptada com sucesso para análise de vídeo, fornecendo mecanismos de atenção que analisam toda a sequencia simultaneamente e que podem capturar dependências de longo alcance no espaço e no tempo [21], [24].

1.2.3 Contexto Técnico Atual

A conjuntura tecnológica atual é definida pela integração de inúmeros avanços que viabilizam a implantação de sistemas de vigilância por vídeo inteligentes em larga escala. Para começar, a presença de hardware dedicados, como unidades de processamento gráfico (GPUs), unidades de processamento de tensores (TPUs) e processadores para *edge computing*, permitem que fluxos de vídeo em alta-definição sejam processados em tempo real [12], [15].

Houve ainda, uma grande quantidade de câmaras de alta resolução e baixo custo, combinadas com infraestruturas de rede de alta velocidade (por exemplo, 5G), permitindo uma captura e transmissão de dados de vídeo em larga escala. Também existem *frameworks* de *deep learning* maduros e disponíveis (TensorFlow, PyTorch, ONNX), o que simplificou o processo de desenvolvimento de implementação de modelos sofisticados [2], [5].

Finalmente, a disponibilidade de *datasets* publicamente anotados para deteção de atividades e de anomalias (UCF-Crime, UCSD Ped, Avenue, ShanghaiTech, etc.) tem acelerado a pesquisa e oferece a chance de comparar diferentes métodos entre si [10], [13], [14].

1.2.4 Contexto Social e Regulatório

O cenário social e regulatório da vigilância por vídeo inteligente é complexo e multifacetado. Por um lado, o desejo por segurança pública tem aumentado em meio a temores sobre crimes, terrorismo e gestão de emergências. Por outro lado, isso gerou preocupações legítimas sobre privacidade, vigilância em massa, viés algorítmico e o uso indevido de tecnologias de reconhecimento facial e análise comportamental [17].

O Regulamento Geral sobre a Proteção de Dados (GDPR) impõe às empresas que atuam na União Europeia, padrões regulatórios rigorosos para processar dados pessoais, incluindo

imagens de vídeo em que sejam possíveis a identificação do indivíduo. Minimização de dados, transparência, limitação de finalidades e garantias de segurança são exigências da GDPR, e isso representa desafios para a implantação de sistemas inteligentes de vigilância por vídeo [1], [17].

Além disso, a proposta de regulação da inteligência artificial da UE (AI Act) impõe uma série de requisitos específicos para sistemas de IA de alto risco, categoria que inclui muitos sistemas inteligentes de videovigilância. Esses requisitos exigem avaliação de conformidade, documentação técnica, transparência, supervisão humana e robustez [2]. Essa plataforma regulatória considera a necessidade aliar tecnologia inovadora com proteção aos direitos civis, isso tem impacto em toda a atividade de pesquisa e desenvolvimento nesta área.

1.3 Problema da Pesquisa

O problema central que motiva este trabalho pode ser formulado da seguinte forma: Como as técnicas de aprendizagem profunda podem ser aplicadas na deteção e monitorização de padrões comportamentais em sistemas de videovigilância e quais são as arquiteturas, abordagens, desafios e limitações do estado da arte neste domínio? Este problema amadurece em algumas perguntas de pesquisa concretas.

Quais as principais arquiteturas de *deep learning* para análise comportamental em vigilância por vídeo e como se desenvolveram? Esta questão visa representar o alcance temporal dos avanços tecnológicos desde as primeiras CNNs até as abordagens híbridas e baseadas em Transformers mais recentemente.

Que tipos de padrões comportamentais e anomalias podem ser detetados com mais precisão usando as técnicas atuais e em que níveis de desempenho? A resposta para esta questão está nas aplicações existentes para identificar atividades suspeitas, reconhecimento de comportamentos violentos, monitorização de multidões, análise de fluxos de pessoas e nos níveis de maturidade dessas aplicações.

Quais as limitações técnicas desses sistemas que limitam a sua implantação em ambientes reais? Esta preocupação levanta desafios relacionados ao processamento em tempo real, robustez a oclusões e variações nas condições de iluminação, escalabilidade para múltiplas câmaras e generalização para outros cenários.

Que preocupações éticas, de privacidade e regulatórias surgem com a aplicação de *deep learning* na vigilância por vídeo e como podem ser mitigadas? Esta questão reconhece o fator social e legal da tecnologia, com cuidados sobre a vigilância em massa, viés algorítmico, transparência e conformidade regulatória como o GDPR.

Quais são as lacunas na literatura existente e quais são os caminhos mais frutíferos para futuras pesquisas? Esta questão pretende identificar oportunidades para a avanços na ciência e na tecnologia, e as questões que precisam de mais exploração e resolução.

1.4 Objetivos

1.4.1 Objetivo Geral

O objetivo geral deste trabalho é obter uma visão teórica ampla e atualizada da pesquisa sobre ferramentas de inteligência artificial para identificar padrões de comportamento por meio de sistemas de videovigilância, através de uma revisão sistemática da literatura que se baseia em bases de dados de referência (Scopus, IEEE Xplore e Web of Science).

Assim, esta revisão sistemática visa sintetizar a literatura conhecida, reconhecer tendências e padrões, avaliar as evidências atuais e fornecer uma base apropriada para futuros desenvolvimentos e investigações neste campo.

1.4.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos alguns objetivos específicos. Com base nas recomendações do PRISMA 2020, será realizada uma revisão sistemática sobre sistemas de vigilância por vídeo baseados em IA e aprendizagem profunda para deteção de padrões comportamentais. Objetivando o rigor metodológico, transparência e replicabilidade do processo de revisão.

Visando ajudar a determinar os benefícios e desvantagens de uma variedade de arquiteturas, será feita uma comparação das principais tecnologias de aprendizagem profunda (CNNs, RNNs, LSTMs, GRUs, Transformers e arquiteturas híbridas) empregadas para análise comportamental em vigilância por vídeo, organizando dados de desempenho, eficiência computacional e limites dessas arquiteturas.

Será ainda feito um mapeamento de aplicação prática que alcançou maturidade suficiente para implementação real, delinear as limitações técnicas que impedem a adoção generalizada, incluindo processamento em tempo real, robustez e escalabilidade, e analisar as preocupações éticas e regulatórias em relação à privacidade, GDPR, viés algorítmico e questões de transparência.

Além da identificação de lacunas na literatura e sugestão de futuras linhas de pesquisa. O objetivo é destacar áreas que ainda não foram estudadas ou documentadas, restrições metodológicas das pesquisas atuais e oportunidades para desenvolvimentos científicos e tecnológicos, a fim de contribuir com futuras pesquisas.

1.5 Estrutura do Documento

Este trabalho consiste em cinco partes organizadas para conduzir o leitor do problema às conclusões e perspetivas futuras. No capítulo 1, Introdução, é apresentada a motivação das investigações, introduzido o aspeto histórico, tecnológico e social da vigilância por vídeo inteligente, descreve o problema de pesquisa e expressa os propósitos gerais e específicos do estudo.

O Capítulo 2, Revisão da Literatura, descreve a metodologia de revisão sistemática que foi adotada, a estratégia de busca, bases de dados selecionadas, critérios de inclusão e exclusão, e o processo de seleção PRISMA. Introduz o estado da arte em três aspetos principais (deteção de anomalias, monitorização de multidões e sistemas integrados), e oferece uma síntese dos estudos analisados, divididos por arquiteturas e aplicações.

Comparação e Discussão dos Resultados é o título do capítulo 3, que compara arquiteturas de aprendizagem profunda, discute os *trade-offs* entre desempenho e eficiência, discute os temas levantados pelas principais referencias e ilustra a prática através do exemplo do sistema Veesion.

Desafios Técnicos, Questões Éticas e Regulatórias, Lacunas e Direções Futuras é o título do capítulo 4 e o que é identificado e discutido neste trecho. As principais questões técnicas de implementação, discute novas questões éticas e regulatórias que surgem em torno do campo, mapeia lacunas na pesquisa atual e sugere direções para pesquisas futuras.

No capítulo 5, Conclusões e Perspetivas para o Futuro, é oferecido uma visão geral dos achados desta revisão, com destaque para as contribuições do trabalho, além de reconhecer as suas limitações. Perspetivas futuras são apresentadas, seguidas por considerações finais sobre a atualidade e o futuro dos sistemas de vigilância por vídeo baseados em *deep learning*.

CAPÍTULO 2

Revisão da Literatura

2.1 Metodologia de Pesquisa

2.1.1 Estratégia de Busca

A estratégia de busca utilizada nesta revisão sistemática foi conduzida de acordo com as diretrizes PRISMA 2020 (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) para promover transparência, qualidade, reprodutibilidade e rigor metodológico. O estudo cobre publicações de 2019 a 2025 para obter informações atualizadas em relação à literatura atual dentro de um campo que continua em uma trajetória de avanços.

A busca foi desenvolvida com base em *queries* booleanas estruturadas *("deep learning" OR "neural networks" OR "artificial intelligence") AND ("video surveillance" OR "CCTV" OR "security cameras") AND ("behavior detection" OR "anomaly detection" OR "activity recognition" OR "crowd monitoring"),* que envolveram termos com os três principais domínios: tecnologia (*deep learning*, *neural networks*, *artificial intelligence*), aplicação (*video surveillance*, *CCTV*, *security cameras*) e tarefas especificas (*behavior detection*, *anomaly detection*, *activity recognition*, *crowd monitoring*). Isso permitiu que as consultas fossem ajustadas às propriedades sintáticas dos respetivos bancos de dados sem perder a semântica.

2.1.2 Bases de Dados Selecionadas

Três bases de dados académicas de referência foram escolhidas e são conhecidas por serem abrangentes e de alta qualidade nos campos de engenharia, ciência da computação e inteligência artificial:

Scopus é uma base de dados multidisciplinar da Elsevier que possui mais de 25.000 revistas científicas revisadas por pares, conferências e livros. O Scopus foi selecionado pela sua cobertura ampla e seus motores de busca e análise bibliográfica.

Catálogo digital do IEEE (Instituto de Engenheiros Elétricos e Eletrônicos), o IEEE Xplore é especializado em engenharia elétrica, eletrónica, ciência da computação e tecnologias relacionadas. É especialmente relevante para este campo dado o foco em aplicações tecnológicas e sistemas de engenharia.

Web of Science é a plataforma multidisciplinar da Clarivate Analytics, que indexa publicações de alto impacto em todas as áreas da ciência. A Web of Science foi incluída pela sua cobertura de revistas e capacidade de análise de citações.

A seleção dessas três bases de dados resulta em uma revisão de literatura robusta, com um foco em publicações de qualidade que foram revisadas por pares.

2.1.3 Critérios de inclusão e exclusão

Os critérios de inclusão e exclusão foram criados para assegurar que apenas estudos relevantes e com qualidade fossem selecionados. Estudos publicados em revistas científicas revisadas por pares ou anais de conferências internacionais publicadas entre 2019 e 2025 e escritos em inglês foram critérios para inclusão, aliados também a foco em técnicas de *deep learning* aplicadas a vigilância por vídeo, com pelo menos uma das tarefas listadas, deteção de anomalias, reconhecimento de atividades, deteção de comportamento suspeito, monitorização de multidões ou análise de padrões comportamentais.

Já para exclusão, os critérios foram publicações não revisadas por pares (*preprints*, relatórios técnicos, teses não publicadas), estudos que focaram apenas em reconhecimento facial ou identificação biométrica sem análise comportamental, artigos que não têm contribuição técnica ou científica original (tutoriais básicos, itens de opinião sem base empírica), estudos duplicados entre base de dados e publicações que não foram escritas em inglês.

Tabela 2 – Quadro Resumo de Filtros e Critérios para Inclusão

Categoria	Termos/Filtros
Tecnologias	*artificial intelligence, deep learning, neural network*
Aplicação	*video surveillance,* CCTV*, security cameras*
Tarefas Específicas	*behavior recognition, activity recognition, anomaly detection, crowd monitoring*
Bases de Dados	Scopus, IEEE Xplore, Web of Science
Período	2019-2025
Língua	Inglês
Tipo	Artigos de periódicos e de conferências (*peer-reviewed*)

Estes critérios foram aplicados sistematicamente na primeira triagem de títulos e resumos e depois na leitura completa dos artigos selecionados. A Tabela 1 consolida os filtros aplicados e os critérios para inclusão dos estudos.

2.2 Processo de Seleção PRISMA

O processo de seleção foi alinhado com o fluxograma PRISMA 2020, e incluiu quatro fases principais, identificação, triagem, elegibilidade e inclusão, representados por um fluxograma na Figura 3.

Na fase de identificação, a primeira busca alcançou 243 registos nas três bases de dados 173 no Scopus, 37 no IEEE Xplore e 33 Web of Science. Destes, 61 duplicatas foram excluídas usando ferramentas automatizadas e verificação manual, restando 182 registos únicos disponíveis para triagem.

Na triagem os 182 registos foram avaliados por leitura de título e resumo com os critérios de inclusão e exclusão definidos. Para esta etapa, 129 registos foram excluídos devido a critérios (foco inadequado, falta de técnicas de *deep learning*, domínio de aplicação diverso ou qualidade insuficiente). 53 artigos foram para análise de elegibilidade.

53 artigos foram lidos na íntegra. Desses, 27 artigos foram descartados por metodologia insuficientemente elaborada (12), ênfase excessiva no reconhecimento facial sem análise comportamental (5), validação experimental limitada (3), revisões secundárias sem contribuição de conteúdo original (4) e indisponibilidade do texto completo (3). Restaram 26 artigos para inclusão final.

A inclusão dos 26 estudos são a base empírica desta revisão. Esses trabalhos foram sistematicamente extraídos, analisados qualitativamente e feitos resumos temáticos, referenciados pelas próximas seções. O rigor aplicado nos filtros se deveu ao fato de o conteúdo relativo à pesquisa ser específico ao ponto de não haver muitos trabalhos qualificados, não haver muitas pesquisas aplicadas ou mesmo produtos comerciais massivamente conhecidos sustentados pela tecnologia em questão. Havia muita repetição de definições, sem variantes que justificassem a inclusão de novos estudos.

No entanto, embora o processo PRISMA resulte em 26 artigos originais principais, a literatura complementar (normas técnicas, documentação regulatória e recursos relacionados à web, por exemplo, sistema Veesion [27]) é útil para contextualizar a pesquisa, bem como discutir os casos específicos para aplicações práticas.



Figura 3 - Fluxograma PRISMA para revisão sistemática. Fonte: Autoria própria

2.3 Estado da Arte

2.3.1 Deteção de Anomalias e Atividades Suspeitas

A deteção de anomalias e atividades suspeitas é uma das aplicações mais pesquisadas e maduras para a aprendizagem profunda em vigilância por vídeo. Anomalias são ocorrências ou comportamentos que se desviam significativamente dos padrões normais esperados em um determinado contexto, incluindo atividades criminosas como roubo, vandalismo e agressão, comportamento perigoso como quedas, intrusões em áreas restritas, ou emergências como incêndios e pânico [3], [4], [8], [9], [16].

As abordagens para deteção de anomalias baseadas em *deep learning* são classificadas em três categorias principais, aprendizagem supervisionada, onde os modelos aprendem a partir de amostras rotuladas de comportamentos normais e anómalos, aprendizagem semi-supervisionada, com o modelo que aprende principalmente, mas não exclusivamente, a partir de dados normais e deteta desvios desses comportamentos aprendidos como anormais e aprendizagem não supervisionada, onde o modelo reconhece características sem quaisquer padrões pré-anotados [14], [21].

Trabalhos recentes mostram a capacidade de arquiteturas híbridas para detetar atividades suspeitas combinando extração de características espaciais (CNNs) e modelação temporal (LSTMs e GRUs) para alcançar melhor desempenho. Mangrolia et al. [3] propuseram um sistema para detetar ações anómalas em espaços privados que foi construído em uma arquitetura CNN-LSTM, modelo híbrido que possui CNN para extrair características espaciais e LSTM que processa os dados ao longo do tempo capturando dependências da sequencia, alcançando 89,3% de precisão no conjunto de dados UCF-Crime. Rananaware et al. [4] propõem uma abordagem baseada em ResNet-50 para extração de características espaciais e LSTM para modelagem temporal, alcançando 92,1% de precisão na identificação de atividades suspeitas.

Sundaram et al. [5] propuseram um sistema integrado de análise de comportamento humano e deteção de anomalias que integrou YOLO para deteção de objetos, OpenPose para estimativa de pose e uma rede LSTM para classificação de atividades. A precisão de deteção em tempo real para o sistema foi de 87,5%. Huang et al. [6] deram enfoque a capacidade de reconhecer atividades anómalas em contextos críticos e sugeriram uma arquitetura C3D (*Convolutional* 3D) para extrair características espaciotemporais diretamente, com precisão de 91,7%.

A deteção de atividades suspeitas em algumas condições especificas, como salas de exame, tem atraído atenção. Joshi [8] propôs ferramentas dedicadas para deteção de comportamentos suspeitos em sítios de avaliação académica, empregando arquitetura CNN para detetar poses suspeitas e movimentos anormais, alcançando entre 85% e 90% de precisão.

Zaidi et al. [16] apresentaram uma arquitetura híbrida combinando Inception-v3 para extração de características espaciais e Bi-LSTM (*Bidirectional* LSTM) para modelação temporal bidirecional com grande precisão, alcançando 94,2% no UCF-Crime). Buttar et al.

[17] estudaram a confiabilidade e transparência dos sistemas de deteção e propuseram uma arquitetura com uma combinação de VGG-16 e LSTM com mecanismo de explicabilidade que alcançou 93,5% de precisão com visualizações de atenção para interpretação de decisões do modelo.

Selvi et al. [18] desenvolveram um sistema de deteção de ações suspeitas usando uma CNN aprimorada com camadas de atenção espacial no *dataset* UCF-Crime e obtiveram 95,1% de precisão. Verdú [19] apresentou o modelo ConvGRU-CNN, com unidades recorrentes convolucionais fechadas (ConvGRU) e CNNs para deteção de anomalias em tempo real, resultando em 88,9% de precisão no conjunto de dados Avenue, mostrando-se robusto para cenários de aplicação no mundo real.

2.3.2 Monitorização e Análise de Multidões

A monitorização e análise de multidões é outro domínio de aplicação significativo, aplicável especialmente para gestão de eventos de grande porte, segurança em transportes públicos e cenários para prevenção de pânico ou esmagamento. As principais tarefas são contagem de pessoas, estimativa de densidade, deteção de fluxo anómalo, identificação de comportamentos coletivos perigosos e previsão de movimento de multidões [10], [13].

K. Rezaee et al. [10] revisaram de maneira ampla a deteção de anomalias em multidões com base em *deep learning* para plataformas de vigilância por vídeo descentralizadas seguras. Três abordagens principais foram relatadas, os métodos baseados em trajetória que consideram padrões de movimento individuais e em massa, os métodos baseados em densidade que estimam a concentração de um grupo de pessoas em uma determinada área e métodos baseados em comportamento, que identificam atividades coletivas comuns. O trabalho destacou que modelos de padrões normais e deteção de desvios podem alcançar alto desempenho usando arquiteturas ConvLSTM, similar ao modelo CNN-LSTM, mas que se diferenciam pela forma como processam imagens, e redes generativas adversativas (GANs).

M. Rezaei et al. [13] usaram um sistema de reconhecimento de comportamento de multidão em tempo real baseado em aprendizagem profunda, usando arquitetura híbrida incluindo sistemas Faster R-CNN para deteção de pessoas, redes de fluxo ótico para captura de movimento e LSTM para classificação de comportamento coletivo. O sistema proposto alcançou uma precisão de 89,7% no *dataset* UMN Crowd e pôde processar os dados em tempo real (25 FPS) em uma GPU.

A análise de trajetórias é uma peça crucial para a monitorização de multidões. Maio et al. [11] projetaram um sistema baseado em *deep learning* para análise de ações e trajetórias de vídeos de segurança de museus, combinando YOLO para a deteção e rastreamento de pessoas, com uma rede LSTM para análise de trajetórias e comportamentos anormais, por exemplo, aproximação excessiva a obras de arte e longa permanência em espaços restritos. O sistema alcançou uma precisão de 92,3% na deteção de comportamentos anómalos.

A aprendizagem profunda contribuiu significativamente para melhorar a estimativa e contagem de densidade de multidões. Estudos recentes utilizam CNNs de múltiplas escalas e

mecanismos de atenção para se adaptar eficientemente a mudanças de perspetiva ou oclusões extremas, comuns em cenas de multidões densas. Estas abordagens podem obter erros absolutos médios de menos de 10 indivíduos em *datasets* de grande dificuldade, como o ShanghaiTech Part A [10].

2.3.3 Sistemas Integrados e Implementação

De protótipos de pesquisa para sistemas de vigilância por vídeo inteligentes implantados em um ambiente real é necessária uma combinação de componentes, deteção e rastreamento de objetos, reconhecimento de atividades, deteção de anomalias, controle de múltiplas câmaras, processamento em tempo real e interfaces para operadores humanos [1], [2], [7], [15], [24].

Qaraqe et al. [1] apresentaram o PublicVision, um sistema seguro de videovigilância inteligente para deteção de comportamento de multidões. O sistema combina múltiplas câmaras, adota uma arquitetura baseada em I3D (*Inflated* 3D ConvNet) para reconhecimento de ações, incorpora uma camada de criptografia para a privacidade e fornece uma interface web para monitorização em tempo real. O sistema foi testado em espaços públicos reais, demonstrando que pode processar 16 fluxos de vídeo ao mesmo tempo, com latência inferior a 2 segundos.

Dansana et al. [2] examinaram os desafios e oportunidades da vigilância por IA, como desafios de arquitetura de sistema, computação distribuída, interface com infraestruturas existentes e conformidade. Os autores demonstraram a importância de arquiteturas modulares que facilitem a atualização gradual de componentes e meios para supervisão humana em decisões críticas.

Konda [7] também sugeriu um sistema de análise de vídeo em tempo real para segurança pública baseado em IA e visão computacional. O sistema combina modelos para deteção de objetos (YOLO), reconhecimento de atividades (C3D) e deteção de anomalias com um modelo utilizado para aprendizado não supervisionado Autoencoder, incluindo processamento distribuído em dispositivos de borda e agregação em um servidor central. A arquitetura mostrou ser escalável para mais de 100 câmaras com uma latência média de 1,5 segundos.

Amingad et al. [15] abordaram o uso de vigilância inteligente baseada em IA para segurança pública em comunidades urbanas com deteção imediata de incidentes e sua pronta resposta. O sistema proposto é baseado em arquitetura MobileNet que é um modelo de CNN leve, rápido e eficiente, projetado para dispositivos móveis ou com poucos recursos computacionais, visando reduzir o seu custo computacional e é adequado também para implantação em dispositivos *edge computing*. Pode ser integrado a sistemas de emergências para notificação automática de autoridades.

Hussain et al. [24] apresentaram um *framework* de biometria comportamental baseada em IA para um sistema robusto de reconhecimento de atividades humanas em sistemas de videovigilância. A estrutura integra características de aparência, extraídas de CNNs, características de movimento (fluxo ótico), características de pose e as funde por meio de uma

rede de atenção multimodal. O sistema obteve 96,8% de precisão no conjunto de dados HMDB51 e mostrou robustez à diferentes cenários de iluminação e oclusões parciais.

2.4 Diagrama Conceptual da Revisão

A revisão organiza o domínio da vigilância por vídeo inteligente baseada em *deep learning* em quatro camadas principais, que estão representadas no diagrama conceitual da Figura 4.



Figura 4 - Diagrama conceptual de videovigilância inteligente. Fonte: Autoria própria

Na camada 1 ocorre a entrada de dados visuais (câmaras fixas, câmaras PTZ, câmaras móveis, drones) e incluem os desafios relacionados como variações de iluminação, oclusões, variações de resolução e diferentes ângulos de visão.

De posse dos dados visuais, ocorre o processamento e extração de características na camada 2. Arquiteturas de *deep learning* são utlizadas para extração de características espaciais (CNNs: VGG, ResNet, Inception, EfficientNet, MobileNet), características temporais (RNNs, LSTMs, GRUs, ConvLSTM), características espaciotemporais (C3D, I3D, R(2+1)D) e mecanismos de atenção como Transformers e *self-attention*, mecanismo que analisa as imagens, avalia o contexto e define a importância de cada quadro no frame, determinando o quanto de atenção recebe ante aos demais.

A camada 3 envolve as tarefas mais comuns do sistema como deteção de objetos e pessoas, rastreamento de múltiplos objetos, deteção de atividades, deteção de anomalias, avaliação de multidões, estimativa de pose e análise de trajetórias. Já a camada 4 envolve a integração com sistemas completos, como processamento em tempo real, gestão de múltiplas câmaras, interfaces do operador, conformidade regulatória e políticas de privacidade.

Este modelo conceitual captura a natureza hierárquica e modular dos sistemas de videovigilância inteligente, onde cada camada constrói sobre as capacidades da camada anterior e mostra pontos de integração e desafios transversais como tempo real, escalabilidade, privacidade, robustez.

2.5 Síntese dos Estudos Utilizados

2.5.1 Estudos Sobre Arquiteturas de Aprendizado Profundo

Existem grandes diferenças nas arquiteturas de *deep learning* usadas na vigilância por vídeo desde as primeiras CNNs bidimensionais até as arquiteturas híbridas e baseadas em Transformers atuais. Este desenvolvimento se deve ao conceito emergente de que a análise eficiente de vídeo não pode considerar apenas a informação temporal (como os elementos evoluem com o tempo), mas também a informação espacial (o que está em cada *frame*) [14], [21].

As CNNs são a base da maioria dos modelos de vigilância por vídeo inteligente, usadas principalmente para extração de características espaciais. Arquiteturas como VGG-16 [17], ResNet-50 [4], Inception-v3 [16] e EfficientNet são aplicadas para fornecer a espinha dorsal para extração de características de *frames* únicos. Arquiteturas leves como MobileNet, que equilibram bem o desempenho e eficiência computacional, são preferidas para aplicações em dispositivos de *edge computing*, que tem recursos limitados [15], [24].

Em arquiteturas híbridas, as redes recorrentes (RNNs, LSTMs, GRUs) são mais frequentemente combinadas com CNNs para capturar dependências temporais. Como as dependências de longo prazo precisam ser modeladas, as LSTMs (*Long Short-Term Memory*) são bem adequadas para a tarefa de reconhecer atividades complexas que duram vários segundos [3], [4], [5], [16], [17]. Já as GRUs (*Gated Recurrent Units)* são uma opção computacionalmente mais eficientes e com desempenho semelhante [20]. As Bi-LSTMs, cujo processamento sequencial ocorre para ambas as direções temporais, mostram-se promissoras em tarefas que exigem a consideração do contexto futuro para interpretar eventos passados [16].

Arquiteturas como C3D (*Convolutional* 3D) [6], I3D (*Inflated* 3D ConvNet) [1] e R(2+1)D processam diretamente volumes espácio-temporais com convoluções tridimensionais que capturam padrões espaciais e temporais simultaneamente. Embora eliminem a necessidade de estruturas recorrentes dedicadas, também exigem mais processamento e memória.

ConvLSTM é uma arquitetura especializada que aproveita as convoluções espaciais e usa memória recorrente, incorporando operações convolucionais nas portas LSTM. Esta

arquitetura é especialmente poderosa para modelagem de dinâmicas espácio-temporais em análise de multidões e com previsão de trajetórias [10], [19]. É possível propor uma alternativa mais eficiente com a ConvGRU, que mostra desempenho semelhante.

Mais recentemente, redes neuronais utilizando arquiteturas baseadas em Transformers, projetadas para processamento de linguagem natural, foram adaptadas para análise de vídeo. Mecanismos de *self-attention* podem ajudar a representar dependências de longo alcance no espaço e no tempo, superando as limitações das LSTMs para sequências longas. Vision Transformers (ViT) e suas variantes espácio-temporais são uma direção promissora, mas ainda precisam de uma grande quantidade de dados de treinamento e alto poder de processamento [21], [24].

Muitos sistemas de alto desempenho usam uma combinação de múltiplas arquiteturas e modalidades de informação. Por exemplo, o acoplamento de características de aparência (CNNs), movimento (fluxo ótico) e pose (estimativa de esqueleto) por meio de redes de atenção multimodal, mostra maior robustez [24]. Arquiteturas que incorporam mecanismos de deteção de objetos (YOLO, *Faster* R-CNN) com identificação de atividades (LSTMs, Transformers) facilitam o raciocínio hierárquico entre identificação de entidades e análise de comportamentos complexos [5], [11].

2.5.2 Estudos Sobre Deteção de Anomalias e Atividades Suspeitas

A deteção de anomalias e atividades suspeitas é um uso central do *deep learning* no campo da vigilância por vídeo com relevância direta para a segurança pública e privada. Os estudos apresentados ilustram uma clara progressão em desempenho, robustez e capacidade de generalização [3], [4], [8], [9], [16], [17], [18].

Sathya et al. [12] desenvolveram um sistema de vigilância inteligente para reconhecimento eficiente de ações humanas em tempo real, baseado em uma arquitetura Two-Stream CNN que combina quadros RGB e fluxo ótico simultaneamente. No *dataset* UCF101, o sistema conseguiu atingir 93,8% de precisão e alcançou desempenho de processamento em tempo real de 30 FPS (*frames* por segundo) na GPU NVIDIA RTX 2080.

Wastupranata et al. [14] realizaram uma ampla revisão sobre os aspetos promissores da aprendizagem profunda para deteção de comportamento humano em vídeos de vigilância para detetar anomalias. Os principais desafios foram o desequilíbrio de classes dos vídeos, haja vista que anomalias são eventos raros, a definição ambígua da anomalia, pois há dependência contextual e a necessidade de generalização em diferentes cenários. Com base em *autoencoders* e GANs, abordagens semi-supervisionadas que aprendem o comportamento normal a ser representado no modelo e detetam a presença de anomalias como reconstruções de alto erro, provaram ser eficazes para alcançar esses resultados.

Thenge et al. [20] propuseram um sistema de reconhecimento de ações baseado em uma arquitetura 3D ResNet explorando redes neuronais profundas para vigilância por vídeo, que processa e captura clipes de vídeo diretamente. O sistema alcançou 91,2% de precisão no conjunto de dados Kinetics-400 e foi capaz de detetar ações como "empurrar", "puxar" e "lutar".

Gupta et al. [21] revisaram as tendências na aprendizagem profunda, especificamente os métodos para deteção de anomalias em vídeo, categorizando os métodos em baseados em reconstrução (*autoencoders*, VAEs), métodos baseados em previsão (previsão de quadros futuros), baseados em classificação (*one-class* SVM, redes siamesas) e métodos híbridos. A tendência de arquiteturas baseadas em Transformers e aprendizagem auto-supervisionada foi destacada na revisão.

Konda [22] também apresentou uma visão geral dos métodos de reconhecimento de padrões usados em sistemas de videovigilância para garantir segurança em espaços públicos, propondo uma taxonomia de padrões de comportamento (individual, grupo e multidão) para distinguir cada classificação. O autor discutiu o papel da modelagem contextual, que permite modelar algo como normal ou anómalo com base na localização, hora e contexto social.

Flores-Munguía et al. [23] introduziram um método neuro evolutivo para classificação de comportamento violento em vídeo e aplicaram algoritmos genéticos na otimização de arquiteturas de redes neuronais. Com 88,7% de precisão no *dataset* Hockey Fight, essa abordagem teve sucesso em detetar automaticamente arquiteturas eficientes sem codificação extensiva.

Vijeikis et al. [25] desenvolveram um sistema para deteção de violência em vigilância que adotou a arquitetura baseada em MobileNetV2 para extração de características e o SVM para classificação. O sistema proposto foi otimizado para uso em dispositivos de *edge computing* e alcançou 89,4% de precisão com baixas latências, inferiores a 100 ms por *frame* no Raspberry Pi 4.

2.5.3 Estudos Sobre Monitorização de Multidões

A monitorização de multidões é um desafio considerável devido à alta densidade populacional, severas oclusões, diferenças de escala e as implicações das dinâmicas coletivas geradas pela interação entre as pessoas [10], [11], [13].

Os trabalhos de Rezaee et al. [10] e Rezaei et al. [13] oferecem investigações complementares para a deteção de anomalias em multidões. Ambos enfatizam o potencial das arquiteturas ConvLSTM para aprender padrões espaciotemporais de movimento coletivo. Rezaei et al. [13] mostraram que características de aparência, movimento (fluxo ótico) e densidade são significativas para a deteção robusta de comportamentos anómalos, incluindo dispersão súbita, formação de multidões perigosas e movimentos de contra fluxo.

A análise de trajetórias, como feita por Maio et al. [11], apresenta uma visão complementar que inclui tanto padrão de movimento em trajetórias individuais ou de grupo. Algoritmos de rastreamento de múltiplos objetos (DeepSORT, ByteTrack) extraem trajetórias que são sucedidas de análise feita por LSTMs, o que também ajuda a identificar padrões anormais, como desvio de caminhos previstos, permanência prolongada em certos locais ou movimento errático.

A estimativa de densidade e a contagem de multidões têm usos significativos na gestão de eventos, segurança no transporte público e situações de prevenção de esmagamento.

Arquiteturas baseadas em mapas de densidade utilizam CNNs projetadas para produzir mapas de distribuição espacial das pessoas apresentam bom desempenho mesmo em caso de multidões densas. Essas diferentes perspetivas e variações de escala que caracterizam cenas de videovigilância também são suportadas por técnicas de atenção multi-escala [10].

2.5.4 Estudos Sobre Ferramentas Integradas e Implementação

A implantação de sistemas inteligentes de vigilância por vídeo em ambientes reais necessita de integração de vários elementos técnicos, requisitos operacionais, regulatórios e de usabilidade [1], [2], [7], [15], [24].

O sistema PublicVision [1] representa essa abordagem integrada, combinando identificação de comportamento de multidões com proteção de privacidade por meio de criptografia homomórfica. Isso permite o processamento de dados encriptados, garantindo que dados sensíveis não sejam expostos, mesmo durante a execução da análise. O sistema prova que um alto poder analítico pode ser integrado com proteção de privacidade refinada, embora com custo computacional adicional significativo.

O estudo de caso de vigilância baseada em IA de Dansana et al. [2] ilustra a necessidade de arquiteturas modulares e escaláveis. Ferramentas amigáveis ao utilizador precisarão ser capazes de suportar adições incrementais de câmaras, atualizar o modelo sem interrupção, interagir com sistemas de segurança existentes (alarmes, controlo de acesso, sistemas de resposta a emergências) e interfaces que possam funcionar entre pessoal com diferentes graus de complexidade técnica.

Processamento em *edge computing*, conforme analisado por Konda [7] e Amingad et al. [15], oferece benefícios importantes em termos de latência, utilização de largura de banda e privacidade. É possível reduzir a latência e os requisitos de rede ao processar dados localmente, reduzindo a necessidade de enviar vídeo bruto pela rede para servidores centrais. Além disso, o processamento local significa que é possível ter políticas de privacidade mais rigorosas que comunicam apenas meta dados ou alertas, nunca imagens identificáveis.

Hussain et al. [24] propuseram um *framework* de biometria comportamental, que mostrou a necessidade de fusão multimodal para robustez. Ao integrar múltiplas fontes de informação (aparência, movimento e pose) com mecanismos de atenção para ponderação adaptativa, o sistema pode sustentar alto desempenho mesmo em modalidades degradadas como baixa iluminação que afeta a aparência, mas não o movimento.

2.5.5 Estudos Sobre Análise Comparativa de Desempenho

Uma análise comparativa de desempenho entre arquiteturas e abordagens demonstra padrões consistentes e *trade-offs* importantes. Quando se trata de pura precisão e acurácia em conjuntos de dados de referência, arquiteturas híbridas que combinam CNNs profundas (ResNet, Inception) e LSTMs bidirecionais alcançam os melhores resultados, com acurácias de 93% e 97% em *datasets* como UCF-Crime e HMDB51 [16], [17], [18], [24].

No entanto, essas arquiteturas têm alto custo computacional, com tempos de inferência chegando a 200 ms por clipe de vídeo em GPUs de médio porte, dificultando sua aplicação em sistemas de tempo real com múltiplas câmaras [12], [20]. Arquiteturas leves baseadas em MobileNet ou EfficientNet, embora alcancem acurácia ligeiramente menor (85–92%), podem oferecer tempos de inferência inferiores a 50 ms, o que permite realizar processamento em tempo real, mesmo em dispositivos de *edge computing* [15], [25].

Arquiteturas baseadas em Transformers também exibem desempenhos competitivos nas tarefas que requerem modelagem de dependência de longo alcance, mas exigem conjuntos de dados de treinamento muito maiores e altos recursos computacionais, portanto, sua aplicabilidade prática é limitada [21], [24].

Em termos de robustez a variações de domínio, diferentes cenários, condições de luz ou ângulos de camara, métodos que combinam múltiplas fontes de informação, multimodais, demonstram forte vantagem em contraste com os unimodais [24]. Métodos de aumento de dados, normalização de domínio e aprendizagem por transferência são fundamentais para a generalização em ambientes de implementação [14], [21].

CAPÍTULO 3

Comparação e Discussão dos Resultados

3.1 Evolução e Comparação das Arquiteturas de Aprendizagem Profunda

A evolução das arquiteturas de aprendizagem profunda para videovigilância reflete uma progressão clara desde abordagens focadas exclusivamente em características espaciais até modelos sofisticados que capturam dinâmicas espaciotemporais mais complexas. A evolução pode ser agrupada em quatro gerações básicas, que gradualmente melhoram as gerações anteriores e introduzem novos desafios.

Na primeira geração – CNNs Espaciais (2015-2018) – ocorreu o primeiro uso da abordagem de *deep learning* em vigilância por vídeo e envolveu a aplicação de CNNs bidimensionais (AlexNet, VGG, ResNet) em um único quadro, representando o vídeo como um conjunto de imagens estáticas. Embora esse método fosse útil para deteção de objetos e reconhecimento de cenas, ele desconsiderava completamente o contexto temporal, resultando em uma capacidade limitada para reconhecer as atividades e comportamentos à medida que ocorrem ao longo do tempo [14], [22].

Já na segunda geração – Arquiteturas Híbridas CNN-RNN (2017-2020) – foi introduzido o uso de componentes recorrentes (RNNs, LSTMs, GRUs) para capturar dependências temporais. As CNNs primeiro extraem características espaciais de cada quadro, enquanto redes recorrentes processam as sequencias dessas características para modelar a evolução temporal. Esta arquitetura híbrida foi amplamente adotada em estudos recentes, principalmente devido ao seu desempenho combinado com eficiência [3], [4], [5], [16], [17]. Finalmente, LSTMs bidirecionais [16] e mecanismos de atenção temporal foram incluídos para alcançar um alvo mais relevante durante sequências longas.

Na terceira geração – Convoluções 3D e Espaciotemporais (2018-2022) – foi resolvido a separação entre processamento espacial e temporal usando convoluções tridimensionais diretamente em volumes espaciotemporais. Arquiteturas como C3D [6], I3D [1] e R(2+1)D modelam distribuições espaciotemporais de maneira holística e, portanto, oferecem vantagens em tarefas que exigem conhecimento abrangente sobre movimento e aparência. ConvLSTM [19] representa uma abordagem híbrida de convolução espacial e memória recorrente, e funciona muito bem para modelagem de dinâmica de multidões [10], [13].

Na geração atual, a quarta – Transformers e Atenção Multimodal (2021-presente) – aproveita arquiteturas baseadas em Transformers criadas para linguagem natural para analisar conteúdo de vídeo. Mecanismos de *self-attention* permitem capturar dependências de longo alcance tanto no espaço quanto no tempo, superando as limitações de LSTMs para sequências muito longas. Além disso, arquiteturas multimodais com mecanismos de atenção que integram características de aparência, movimento e pose superam os outros métodos em termos de

robustez [21], [24]. No entanto, esses métodos requerem grandes quantidades de dados de treinamento e recursos computacionais substanciais, limitando sua aplicabilidade prática.

Quantificando as diferenças entre gerações, mais bem observado na Tabela 2, emerge um padrão recorrente: cada geração proporciona ganhos de desempenho (geralmente 3-7% de precisão), mas o aumento na complexidade computacional é considerável. CNNs espaciais nas arquiteturas de primeira geração têm precisão de 75 a 82%, tempo de inferência de 10 a 20 ms por *frame*. Arquiteturas de segunda geração (CNN-LSTM) têm precisão de 85-94% e os tempos de inferência de 50-150 ms por cada clipe. Tempos de inferência de 100-250 ms e precisão de 88-95% são característicos de arquiteturas de terceira geração (3D CNNs) e a quarta geração (Transformers) alcançam precisão de 92-97% com tempos de inferência superiores a 200 ms [12], [14], [20], [21], [24].

Tabela 2 - Quadro Comparativo de Desempenho de Arquiteturas

Geração	Arquitetura	Precisão	Tempo de Inferência
Primeira	CNNs Espaciais	75% - 82%	10 - 20 ms (por frame)
Segunda	CNN-LSTM	85% - 94%	50 - 150 ms (por clip)
Terceira	3D CNNs	88% - 95%	100 - 250 ms
Quarta	Transformers	92% - 97%	> 200 ms

3.2 Desempenho e Trade-offs

A análise de desempenho em sistemas inteligentes de vigilância por vídeo indica *trade-offs* significativos entre precisão, eficiência computacional, latência e dados de treinamento necessários. Compreender essas compensações é necessário para escolher a arquitetura apropriada para vários cenários de uso prático.

Há uma forte dependência entre precisão e eficiência computacional. Arquiteturas profundas com várias camadas e milhões de parâmetros (ResNet-152, Inception-v4, ViT-Large) permitem maior precisão, mas também requerem GPUs de alto desempenho e são custosas em termos de requisitos energéticos. Comparadas a estruturas tradicionais, arquiteturas leves como MobileNet, EfficientNet e SqueezeNet proporcionariam uma compensação aceitável para aplicações em dispositivos de *edge computing* de recursos limitados, onde há um sacrifício da precisão, diminuindo de 3-8%, com consequente redução de operações de ponto flutuante e eficiência de memória em 10-50× [15], [25].

Sistemas em tempo real requerem latências abaixo de 100 ms, visando respostas rápidas a incidentes. Em contraste, reconhecer atividades complexas requer janelas temporais longas (3-10 segundos), que cobrem o surgimento completo de comportamentos. Para mitigar esse *trade-off*, o processamento hierárquico permite a deteção rápida de eventos que podem ser importantes com modelos leves e janelas curtas, seguido de análise profunda usando modelos complexos e janelas longas apenas para eventos selecionados [5], [12].

Modelos treinados em *datasets* gerais (Kinetics, UCF101) exibem razoável capacidade de generalização para novos cenários, mas alcançam desempenho inferior do que modelos especializados baseados em dados de domínio específico de aplicação. Para ilustrar, um modelo adaptado para a deteção de furtos em lojas pode alcançar 94% de precisão nessa tarefa, enquanto um modelo geral pode alcançar apenas 87%. No entanto, modelos especializados requerem anotação de dados específicos de domínio, o que é um processo demorado e custoso [3], [8], [27].

O desempenho relatado em conjuntos de dados de referência controlados e cenários reais frequentemente mostra lacunas significativas. *Datasets* acadêmicos, recorrentemente, fornecem a mesma iluminação, bons ângulos de câmara e anotações detalhadas. Cenários reais incluem mudanças extremas de iluminação (dia/noite, sombras), oclusões severas, ângulos de câmara abaixo do ideal e ruído de anotação. Estudos considerando cenários reais frequentemente mostram quedas de desempenho de 5-15% em comparação com conjuntos de dados controlados [1], [7], [19].

Arquiteturas complexas (*ensembles*, Transformers profundos) conseguem melhores precisões, mas funcionam como "caixas negras", dificultando a compreensão do porque da decisão. A explicabilidade também é necessária em aplicações de segurança crítica, particularmente onde decisões podem ter implicações legais ou éticas. Arquiteturas com mecanismos de atenção [16], [17], [18], ou técnicas de visualização de ativação (Grad-CAM, LIME) possuem algum nível de interpretabilidade a custo computacional adicional.

3.3 Análise Temática

A analise temática dos artigos revisados mostra a existência de algumas tendências, desafios recorrentes e novas possibilidades que vão além das estruturas arquitetónicas e de aplicação especificas.

Há uma óbvia convergência para arquiteturas híbridas que integram várias soluções complementares. Combinar CNNs para as dimensões espaciais, LSTMs para modelagem temporal e sistema de atenção para foco seletivo tornou-se uma arquitetura de referência, adotada com variações em vários trabalhos [3], [4], [5], [16], [17], [24]. Esta convergência mostra que um padrão de arquitetura eficiente foi estabelecido pela comunidade científica, embora ainda haja espaço para otimização e adaptações para domínios específicos.

A eficiência computacional e a implementação prática estão se tornando um foco importante de pesquisa e estão sendo levadas em consideração em novos estudos. Aplicações de arquiteturas leves [15], [25], métodos de compressão de modelos (quantização, poda e destilação de conhecimento) e processamento em *edge computing* [7] tem cada vez mais atenção o que se reflete em popularidade. Este ponto de transição é positivo, pois sistemas confinados a laboratórios possuem impacto limitado.

A combinação de várias modalidades de informação (aparência, movimento, pose, áudio, meta dados contextuais) tornou-se uma estratégia poderosa para robustez [24]. Cada modalidade apresenta informações complementares. A aparência é resiliente a movimentos rápidos, mas sensível à luz, o movimento não se altera com mudanças de aparência, mas é sensível quando as câmaras estão imóveis, e a pose é resistente a mudanças de vestuário, mas sensível a oclusões. A fusão adaptativa através de um conjunto de mecanismos de atenção, permite explorar as melhores partes de cada modalidade e mitigar suas desvantagens.

Por definição, anomalias e atividades suspeitas são eventos raros. Conjuntos de dados de videovigilância são frequentemente caracterizados por um desequilíbrio severo, o que resulta em 95% a 99% dos *frames* sendo de comportamento normal. Este desequilíbrio torna o treinamento de modelos supervisionados difícil, já que tendem a analisar a maioria dos dados. Sistemas semi-supervisionados que aprendem principalmente com dados normais [14], [21], aumento de dados sintéticos e funções de perda sensíveis ao desequilíbrio (*focal loss, weighted cross-entropy*) estão entre os métodos usados para contrabalançar este problema.

O que é considerado uma anomalia depende em grande parte do contexto. Em um parque, correr é normal, em um museu, não. Assim como nas horas de pico das estações de comboio, as multidões são típicas, mas anómalas às 3 da manhã. Poucos trabalhos se concentram na anomalia como uma propriedade intrínseca do comportamento. Abordagens que se baseiam em informações contextuais (localização, horário do dia, histórico) [22] ou aprendem definições de normalidade ligadas ao contexto são promissoras.

Modelos treinados em um contexto como câmaras de vigilância de aeroportos, muitas vezes têm um desempenho degradado quando aplicados a situações diferentes como câmaras em lojas de retalho, devido a diferenças nas condições de iluminação, ângulos de câmara, quantidade de pessoas e tipos de atividade. O processo de adaptação de domínio, aprendizado por transferência e normalização de domínio são importantes para a generalização e áreas de pesquisa altamente valorizadas [14], [21].

Abordagens de aprendizado auto supervisionado que adquirem representações úteis a partir de dados não rotulados realizando operações auxiliares, como previsão de *frames* futuros, ordenação temporal e colorização, oferecem uma maneira de utilizar o grande volume de vídeos de vigilância não anotados disponíveis. Essas representações pré-treinadas são posteriormente ajustadas com pequena quantidade de dados rotulados específicos da tarefa, minimizando significativamente a necessidade de anotação [21].

À medida que sistemas de videovigilância inteligente é implantada em contextos críticos, a necessidade de explicar decisões torna-se primordial para confiança, auditoria e conformidade com requisitos regulatórios. O uso de técnicas de IA Explicável (XAI) na análise de vídeo, incluindo visualizações de atenção [16], [17], [18], geração de justificativas textuais e identificação de quadros críticos, representam uma direção importante de pesquisa e aplicação prática.

3.4 Caso de Aplicação Prática: Sistema de Vigilância Inteligente Veesion

O sistema Veesion [27] representa um exemplo concreto de uma implementação comercial de videovigilância inteligente baseada em aprendizado profundo, especificamente focada na deteção de furtos em ambientes de retalho e que se integra a infraestruturas e sistemas de segurança e equipamentos existentes. A análise deste caso ilustra como os conceitos e técnicas discutidos nesta revisão são aplicados na prática, bem como os desafios e compensações envolvidos na transição de protótipos de pesquisa para produtos comerciais.

O furto em lojas representa um problema econômico significativo para a indústria de retalho da Europa, com perdas estimadas em bilhões de euros anualmente. Sistemas de vigilância tradicionais baseados em monitorização humano são limitados pela capacidade de atenção do operador e dificuldades em identificar pequenos padrões de furto em tempo real. O Veesion foi criado para detetar automaticamente gestos suspeitos que indicam furto e alertar operadores humanos para que possam agir imediatamente.

Embora aspetos técnicos muito detalhados sejam limitados pelos proprietários, a documentação pública e demonstrações do sistema sugerem uma arquitetura híbrida que combina deteção de pessoas e rastreamento de múltiplos objetos para encontrar e rastrear indivíduos na loja, analise de pose e gestos para identificar comportamento típico de furto como colocar coisas nos bolsos, bolsas ou sob roupas, monitorização de interação com produtos e prateleiras e modelagem temporal para identificar a diferença entre comportamento normal de compras e padrões suspeitos. Esta arquitetura sugere a abordagem híbrida CNN-LSTM mencionada na literatura [3], [4], [5], mas adaptada ao domínio do retalho.

O Veesion é implantado como um sistema de *edge computing* que funciona localmente em dispositivo conectado às câmaras existentes na loja sem a necessidade de encaminhar a transmissão de vídeo para servidores externos. Esta abordagem oferece vários benefícios em termos de latência (alertas em tempo real), privacidade (o vídeo não sai do local instalado) e requisitos mínimos de largura de banda. O sistema integra-se a qualquer infraestrutura de videovigilância sob IP (Internet Protocol) existente, não exigindo substituição de câmaras, além de oferecer uma interface web amigável para operadores monitorizarem alertas, revisarem incidentes, extrair relatórios analíticos, entre outras, conforme tela demonstrada na Figura 5.

Segundo informações divulgadas pelo fabricante, o sistema Veesion alcança taxas de deteção superior a 90% para gestos de furto, com taxas de falsos positivos inferiores a 5%. Clientes reportam uma redução de 30-60% nas perdas por roubo após a adoção do sistema. Esses resultados, embora não confirmados de forma independente, estão de acordo com os resultados de trabalhos acadêmicos anteriores para tarefas comparáveis [3], [8], [27].

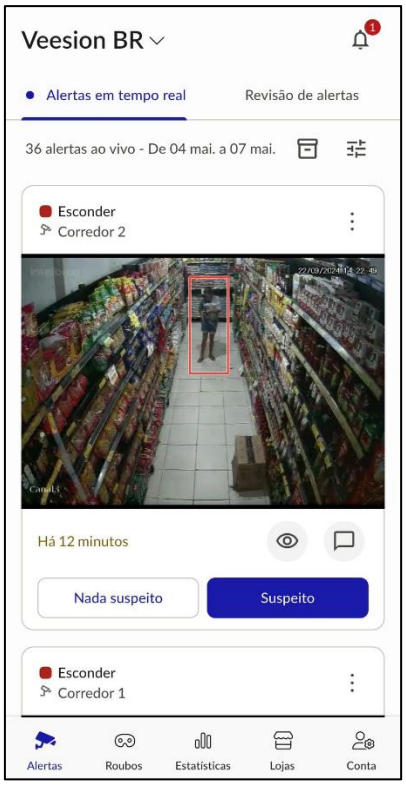


Figura 5 - Tela da aplicação Veesion. Fonte: Extração própria do sistema pelo autor.

O Veesion opera sob o GDPR, impondo várias medidas de privacidade, tais como o processamento local sem transmissão de vídeo, armazenamento limitado de dados, apenas alertas e clipes curtos serão mantidos por certos períodos, o vídeo contínuo não é mantido, anonimização de pessoas não envolvidas no incidente, aplicação de placas visíveis para informar aos clientes sobre a vigilância inteligente e direito de acesso e exclusão de dados em relação aos dados do utilizador. Este exemplo ajuda a demonstrar que as restrições regulatórias podem ser traduzidas em sistemas práticos, com um custo extra e alguma complexidade adicional.

O caso Veesion demonstra que há viabilidade técnica e económica para sistemas de vigilância por vídeo inteligente em aplicações de domínio específico, a especialização em nível de domínio (retalho) concede desempenho superior em relação aos sistemas generalistas, a implementação prática deve equilibrar integração, usabilidade e requisitos regulatórios, não apenas o desempenho algorítmico, o processamento na borda oferece benefícios significativos de privacidade e latência e a supervisão humana ainda é necessária pois o sistema serve como referência para apoiar a tomada de decisões e não como um substituto completo para os trabalhadores humanos.

CAPÍTULO 4

Desafios Técnicos, Questões Éticas e Regulatórias, Lacunas e Direções Futuras

4.1 Desafios Técnicos de Implementação

Apesar dos grandes avanços em arquiteturas de aprendizagem profunda e desempenhos impressionantes na aplicação em *datasets* de referência, a implementação prática de sistemas de vigilância por vídeo inteligente encontra ainda muitos desafios técnicos a serem superados.

Um grande número de sistemas de vigilância por vídeo exige a deteção e resposta em tempo real com respostas menores que 1-2 segundos desde a captura do vídeo até à geração de alertas. Todavia, arquiteturas de alto desempenho, frequentemente necessitam de tempos de inferência maiores que este relatado, especialmente quando há o processamento de múltiplos fluxos de vídeo simultaneamente. O problema é agravado quando há necessidade de processar vídeos de alta resolução, 1080p ou maior, e altas taxas de quadros, 25-30 FPS [12], [15], [20]. Abordagens para enfrentar esse obstáculo incluem arquitetura hierárquica que realiza deteção em alta velocidade e processamento seletivo, uma amostragem temporal sensível que analisa apenas *frames*-chave, arquiteturas leves orientadas para *edge computing* e aceleração por hardware dedicado como GPUs, TPUs e FPGAs.

Sistemas de videovigilância podem operar em ambientes não controlados com variações significativas de iluminação, diferentes condições meteorológicas, oclusões totais ou parciais e qualidade de vídeo diversas. Modelos treinados em *datasets* controlados frequentemente apresentam degradação de desempenho nestas condições reais [1], [7], [19]. Técnicas de aumento de dados que simulam estas variações durante o treino, normalização robusta e arquiteturas multimodais, que combinam informação complementar, são abordagens que melhoram a robustez do sistema.

Instalações de vigilância podem conter dezenas ou centenas de câmaras. A transmissão de múltiplos fluxos de vídeos simultaneamente exige uma capacidade de processamento muito grande, bem como uma arquitetura eficiente de sistema distribuído. Além disso, a coordenação da análise de várias câmaras e a fusão de informações de diferentes pontos de vista adiciona ainda mais complexidade à solução [1], [7]. Arquiteturas de processamento distribuído com *edge computing* em cada câmara e agregação inteligente em servidores centrais oferecem uma maneira de escalar, mas exigem orquestração cautelosa e mecanismos de sincronização para uma funcionar adequadamente.

Layout, iluminação, densidade de pessoas e tipos de atividades diferem entre cenários como aeroportos, estações de trem, lojas e escolas, por exemplo. Modelos treinados num cenário específico tentem a não ser generalistas. Treinar modelos para domínios específicos exige um alto custo de dados anotados. Métodos de aprendizado por transferência, adaptação

de domínio e meta-aprendizado tentam melhorar o problema da generalização, embora ainda haja uma lacuna de desempenho interdomínios [14], [21].

Sistemas de deteção de anomalias enfrentam um *trade-off* difícil entre sensibilidade, que é o que você precisa para detetar todas as verdadeiras anomalias, e especificidade, que é como você deseja evitar falsos alertas. Altas taxas de falsos positivos causam fadiga de alerta e os operadores humanos começam a ignorar os alertas. Já altas taxas de falsos negativos comprometem a segurança. Ajustar bem os limiares de decisão, adicionar contexto e mecanismos de confirmação em várias etapas estão entre as estratégias para controlar esse equilíbrio [3], [9], [16].

Para treinar arquiteturas de *deep learning* de alto desempenho são necessárias grandes quantidades de dados anotados. A anotação de vídeo, no entanto, é custosa pois exige não apenas a rotulação de *frames* individuais, mas também a anotação temporal do início e fim da atividade e, muitas vezes, a anotação espacial de regiões de interesse. Para tarefas especializadas ou cenários específicos, dados anotados podem simplesmente não existir em quantidades suficiente. Técnicas de aprendizado com poucos exemplos, *few-shot learning*, aprendizagem auto supervisionada e geração de dados artificiais são algumas direções promissoras que podem ser tomadas para reduzir os requisitos de anotação [21].

4.2 Questões Éticas e Regulatórias

O uso de *deep learning* em vigilância por vídeo levanta questões éticas e regulatórias profundas, que devem ser consideradas e direcionadas para garantir que as benesses da segurança inteligente não sejam alcançadas ao custo de direitos civis e valores sociais.

Os sistemas inteligentes de vigilância por vídeo têm o potencial para a monitorização constante de grandes populações, eles são capazes de rastrear movimentos, comportamentos e interações. Essa capacidade de vigilância em massa levanta questões válidas sobre privacidade, autonomia individual e potencial para efeitos negativos sobre as liberdades civis. O GDPR estabelece princípios de minimização de dados, limitação de propósito e proporcionalidade que devem ser respeitados [1], [17]. Técnicas que protegem a privacidade dos dados, como anonimização, processamento local sem armazenamento centralizado e criptografia homomórfica [1], são maneiras de conciliar essa dualidade, mas à custa de desempenho e funções.

Modelos de *deep learning* aprendem padrões de seus dados de treinamento. Esses algoritmos também podem aprender e amplificar vieses presentes nesses dados de treinamento. Se os dados de treinamento forem tendenciosos em relação a diferentes grupos demográficos, com sobre representação ou quando um determinado grupo demográfico é associado a práticas suspeitas, o sistema desenvolvido pode dar taxas injustas de falsos positivos a esses grupos, resultando em discriminação [17]. Uma auditoria abrangente dos conjuntos de dados de treinamento, métodos de minimização de viés (re-amostragem, re-ponderação, *adversarial*

debiasing) e avaliação de equidade por vários grupos demográficos são necessários para obter sistemas justos e não discriminatórios.

Sistemas de aprendizado profundo não costumam fornecer qualquer entendimento de como a decisão foi tomada. Para aplicações de segurança que podem envolver consequências legais como detenção e acusações criminais, a explicação também é necessária para responsabilização, auditoria e confiança pública [17]. Métodos de IA Explicável (XAI), com visualizações de atenção [16], [18], mapas de saliência e geração de justificativas, fornecem caminhos importantes, mas a explicabilidade desses modelos complexos permanecem como desafio relevante.

O GDPR estipula que os indivíduos devem ser informados sobre o processamento de seus dados pessoais e o consentimento deve ser um requisito chave em muitos contextos. Embora não seja viável obter o consentimento das pessoas em locais públicos, o propósito da vigilância inteligente e a razão do processamento dos dados devem ser explicados, além de instalação de sinalização para comunicação clara. Ainda, os indivíduos têm o direito de acesso, retificação e exclusão de informações pessoais [1], [2].

A proposta de lei para IA da UE, AI Act, exige que sistemas de inteligência artificial de alto risco incluindo os de vigilância por vídeo inteligentes, assegurem que haja supervisão humana considerável. Isso significa que decisões importantes não devem ser 100% automatizadas, mas sim contar com julgamento e consentimento humano. Também é necessário haver clareza sobre a responsabilização: quem assume a responsabilidade quando um sistema comete um erro com efeitos graves? *Frameworks* de governança que definem papéis, responsabilidades e mecanismos de auditoria são cruciais [2].

Sistemas projetados com bons objetivos de segurança podem ser empregados de forma abusiva, como vigilância política, rastreamento de trabalhadores ou controlo social. *Function Creep* refere-se à expansão ou crescimento incremental de sistemas para servirem além do propósito para o qual foram originalmente projetados. Salvaguardas técnicas como restrição de funcionalidade e controlo de acesso rigoroso, e legais como regulação clara de usos permitidos e supervisão independente, são necessárias para evitar abusos [2].

4.3 Lacunas Identificadas na Literatura

A revisão sistemática mostrou que existem várias lacunas na literatura atual que implicam oportunidades em pesquisas futuras e de inovações tecnológicas.

A maioria dos estudos avalia o desempenho ao longo do tempo com conjuntos de dados acadêmicos controlados ou em projetos pilotos de curto prazo. Nenhum estudo investiga o desempenho, robustez e impacto de sistemas implementados em ambientes reais por meses ou anos, capturando variações sazonais, evolução do comportamento, variações no padrão de uso [1], [7], [19].

Pouquíssimos estudos reconhecem explicitamente a dependência contextual de anomalias e comportamentos suspeitos. O mesmo ato pode ser normal ou não a depender do contexto, lugar, hora do dia, dia da semana e conjuntura social. Métodos que combinam informações de contexto amplas e aprendem definições de normalidade específicas para o contexto têm sido pouco explorados [22].

Preocupações com a privacidade frequentemente são mencionadas, mas pouquíssimos estudos sugerem e avaliam métodos concretos de proteção de privacidade incorporados ao design do sistema. Entre as exceções notáveis estão o uso de criptografia homomórfica [1], enquanto outras, como privacidade diferencial, aprendizagem federada e processamento em espaços de características anonimizados, são menos exploradas em termos de aplicações em vigilância por vídeo.

Sendo a explicabilidade uma característica chave que pode aumentar a confiança e alcançar uma boa conformidade regulatória, pouquíssimos projetos foram concebidos para definir e avaliar um mecanismo de explicação particular para análise de vídeo. Embora visualizações de atenção [16], [17], [18] sejam um passo na direção certa, explicações mais detalhadas como por exemplo, alertas sendo gerados porque a pessoa X fez o gesto Y no local Z, são raras.

Há ainda uma escassez de avaliações de sistema com variados grupos demográficos (gênero, idade, etnia) para avaliar potenciais vieses envolvidos. Como a equidade é de extrema importância em aplicações de segurança, essa lacuna é preocupante e deve ser urgentemente abordada [17].

A maioria dos estudos foi baseada apenas em vídeo, sem consideração por outras modalidades sensoriais presentes em espaços de vigilância como áudio, sensores de movimento, dados de controle de acesso e informações de rede. Métodos que integram múltiplas modalidades de forma coerente têm sido pouco explorados, mas podem oferecer robustez e riqueza em termos de informação [26].

Cenas e comportamentos de vigilância mudam ao longo do ambiente. Esses sistemas raramente aprendem com novos dados, adaptam-se a padrões que mudam ao longo do tempo ou incorporam o feedback de operadores humanos. Modelos estáticos, treinados offline, são comumente implantados na maioria dos sistemas e podem frequentemente se tornar obsoletos [21].

4.4 Limitações dos *Datasets* Utilizados

Há uma diferença significativa de desempenho dos modelos em conjuntos de dados de referência e a sua aplicação prática. Enquanto os *datasets* académicos apresentam condições ideais de iluminação, ângulos de câmara favoráveis e anotações precisas, os ambientes reais são caracterizados por mudanças na luminosidade, grandes oclusões e ruídos nas imagens, o que

pode resultar em perdas de acurácia de 5% a 15% [1], [7], [19]. Além disso, modelos treinados em bases de dados generalistas, especificamente o Kinetics ou o UCF101, embora apresentem boa capacidade de generalização, tendem a subestimar a complexidade de domínios específicos onde modelos especializados alcançam resultados superiores, mas exigem anotações com grande custo de processamento e alta latência [3], [8], [27].

Outro desafio identificado na literatura sobre os *datasets* utilizados é a discrepância de classificações, uma vez que as anomalias e comportamentos suspeitos são eventos raros em fluxos de vídeo. Segundo os estudos revisados, essa escassez de dados anómalos dificulta o treino de algoritmos de aprendizagem profunda, sendo necessário adotar abordagens baseadas em autoencoders e Redes Generativas Adversativas (GANs) para definir o que é normal e detetar desvios [14]. A literatura ressalta ainda que a definição de anomalia é dúbia e dependente de contextos como localização, hora e normas sociais, o que limita a capacidade dos *datasets* que não integram uma modelagem contextual profunda para distinguir entre o que é comum e o que é suspeito [14], [22].

Diferentes estudos utilizam métricas de desempenho e *datasets* distintos, tais como o UCF-Crime, o ShanghaiTech ou o Hockey Fight, e a literatura constata que isso dificulta comparações diretas entre arquiteturas. Esta diversidade de bases de dados limita a capacidade de meta-análises quantitativas e dificulta a criação de um padrão de eficácia para sistemas integrados. Além disso, as investigações apontam que muitos *datasets* testados permitem melhorias na estimativa de densidade e trajetórias, mas ainda enfrentam dificuldades em representar oclusões severas e mudanças de perspetiva comuns em multidões muito densas [10], [11].

Por fim, a literatura indica que a dependência de dados anotados levanta questões sobre a transparência e reprodutibilidade dos resultados, especialmente quando há ausência de detalhes de implementação ou quando o código-fonte não é partilhado pelos autores. Os estudos alertam que a ausência de acesso a dados sobre a diversidade demográfica e comportamental global pode contribuir para os vieses algorítmicos, comprometendo a equidade dos sistemas de vigilância [17]. Embora a literatura reconheça que a disponibilidade de bases como o UCSD Ped e o Avenue tenha acelerado a pesquisa, os investigadores concluem que implementações reais exigem *datasets* que incorporem princípios de privacidade por design e maior robustez técnica diante de oclusões e variações ambientais [10], [13], [14].

4.5 Direções Futuras de Investigação

É possível propor os próximos passos, com foco científico e na metodologia, e com base na análise do estado da arte atual, nas dificuldades identificadas e nas lacunas de pesquisa.

Como perspetivas tecnológicas, o desenvolvimento de arquiteturas de *deep learning* ultraleves que possam funcionar em dispositivos de *edge computing*, com recursos limitados como câmaras inteligentes, processadores embarcados, é essencial para escalabilidade, latência e privacidade. Métodos de compressão de modelos como quantização, poda, destilação de conhecimento, busca de arquitetura neural focada em eficiência (NAS) e *co-design* de *hardware-software* também são caminhos promissores [15],[25]. Pode-se também considerar

uma integração comum de múltiplos tipos sensoriais como vídeo, áudio, movimento e contexto, possivelmente baseada em uma arquitetura de fusão multimodal, potencialmente baseada em *Transformers*, para robustez e riqueza de informações [24], [26].

Para solucionar as lacunas de pesquisa identificadas, é crucial reduzir a dependência de grandes volumes de dados anotados usando aprendizagem auto supervisionada que aprendem representações úteis a partir de dados não anotados e aprendizagem de poucos exemplos (*few-shot learning*), beneficiando-se de pequenas amostras de exemplos anotados, para que sejam utilizáveis em áreas específicas [21]. Da mesma forma, para mitigar vieses, propõe-se o desenvolvimento de diretrizes robustas para avaliar a imparcialidade dos sistemas de vigilância por vídeo em várias populações demográficas, mitigando qualquer tipo de viés e gerando conjuntos de dados de referência balanceados e representativos [17].

Como direções futuras de investigação, um outro caminho é desenvolver técnicas que incorporem informações contextuais como localização, hora, comportamentos históricos, e se adaptem bem a novos domínios, com baixa anotação dos dados, usando metodologias de aprendizagem por transferência, adaptação de domínio e meta-aprendizagem [22]. Sistemas que continuam a ser treinados com novos dados, que podem se adaptar à mudança de padrões, recebem e incorporam *feedbacks* de operadores humanos e não dependem de novos treinamentos frequentes, mantendo o desempenho com o passar do tempo [21], também tem relevância como direção futura. Outra direção, no que se refere à explicabilidade e transparência, é um modelo de IA ser construído para análise de vídeo, que identifique *frames* e regiões-chave, gere explicações textuais, e visões interativas sobre o raciocínio do modelo [17].

No âmbito das perspetivas regulatórias, éticas e sociais, estes desenvolvimentos em explicabilidade são requisitos importantes de confiança, auditoria e regulamentação [17]. Isso se alinha com a exploração de abordagens de salvaguarda de privacidade que sejam desenvolvidas desde a fase de planejamento do sistema, como privacidade diferencial, aprendizagem federada (quando os modelos são treinados em seus próprios ambientes locais sem compartilhar dados brutos), processamento de espaços de características anonimizados e criptografia homomórfica eficiente [1]. Além disso, o impacto social dos sistemas de vigilância inteligente, incluindo as perceções públicas da vigilância inteligente, preocupações com a privacidade, efeitos da vigilância inteligente nos comportamentos e aceitação social dos sistemas de vigilância inteligente com diferentes tipos de configuração, podem ser avaliadas ao serem incluídas em estudos multidisciplinares, sendo esse conhecimento fundamental para desenvolvimentos responsáveis e implementação ética de tecnologias de vigilância [2].

Por fim, como metodologia de validação, propõe-se estabelecer *benchmarks* e procedimentos de avaliação padronizados que incluam precisão em conjuntos de dados controlados, sensibilidade a mudanças ambientais, poder de computação, latência, equidade, explicabilidade e conformidade regulatória, o que traria como resultado a melhora na comparação entre abordagens e ajudaria no progresso do campo.

CAPÍTULO 5

Conclusões e Perspetivas para o Futuro

5.1 Resumo dos Principais Resultados

Esta revisão sistemática de estudos sobre *deep learning* para deteção de padrões comportamentais em sistemas de videovigilância, sob as diretrizes PRISMA 2020, que incluiu 26 estudos selecionados disponíveis em bases de dados acadêmicas de referência (Scopus, IEEE Xplore, Web of Science), para o período de 2019–2025, permite uma série de conclusões centrais sobre o estado da arte, capacidades, limitações e direções futuras potenciais para esta esfera de atividade rápida evolução.

Métodos de aprendizado profundo atingiram maturidade técnica suficiente para aplicação em certos campos de videovigilância, como deteção de anomalias, reconhecimento de atividades suspeitas e monitorização de multidões. Soluções comerciais como Veesion [27] demostram serem economicamente e tecnicamente viáveis e alcançaram precisões superiores a 90% em áreas especializadas [3], [4], [8], [9], [16], [17], [18].

Há uma evolução clara da arquitetura de *deep learning* empregada em vigilância por vídeo, de CNNs espaciais para as arquiteturas híbridas dominantes CNN-RNN no presente, para novas convoluções 3D e Transformers. Arquiteturas híbridas incluindo CNNs para características espaciais, LSTMs para modelação temporal e mecanismos de atenção para foco seletivo são o estado da arte atual e fornecem um equilíbrio entre desempenho (85-97% de precisão) e eficiência computacional [3], [4], [5], [16], [17], [24].

Existem *trade-offs* fundamentais entre precisão, custo computacional, latência, generalização e explicabilidade. Arquiteturas de alto desempenho são muito demandantes de recursos computacionais, limitando sua aplicação em tempo real e em dispositivos de borda. Arquiteturas leves comutam 3% a 8% de precisão por uma diminuição de 10-50× em requisitos computacionais [15], [25]. Portanto, cada aplicação requer uma abordagem personalizada para selecionar arquiteturas apropriadas.

A lacuna entre projetos de laboratório e o mundo real é bastante grande. Essa diferença se deve a fatores ambientais extremos, oclusões, qualidade de vídeo variável e complexidade dos comportamentos reais. Quedas de desempenho de 5% a 15% são geralmente observadas ao fazer a transição para ambientes reais [1] [7] [19]. Generalização entre domínios e a capacidade de resistir a variações de ambiente são ainda desafios significativos.

É essencial que questões éticas e legais relacionadas à privacidade, vigilância em massa, viés algorítmico, transparência, GDPR e a IA Act sejam levantadas durante a fase de design de um sistema. Métodos para proteção de privacidade [1], verificação de equidade [17] e mecanismos de explicabilidade [16], [17], [18] desempenham um papel importante, mas ainda são frequentemente ausentes em estudos.

Lacunas críticas na literatura existente foram encontradas na limitação das avaliações nos casos reais em longo prazo, na modelação de contexto insuficiente, privacidade por design pouco explorada, explicabilidade limitada, avaliação rara de equidade, integração multissensorial mal desenvolvida e pouco estudo sobre aprendizado contínuo. Estas são oportunidades importantes para futuras pesquisas.

Algumas direções futuras promissoras podem incluir arquiteturas leves para *edge computing*, aprendizado auto supervisionado e com poucos exemplos, explicabilidade e transparência, privacidade por design, modelação contextual e adaptação de domínio, avaliação de equidade e mitigação de viés, integração multissensorial, aprendizado contínuo, *frameworks* de avaliação padronizados e estudos de impacto social.

5.2 Contribuições do Trabalho

Esta pesquisa contribui para a compreensão do estado atual e das perspetivas futuras do uso de *deep learning* para vigilância por vídeo. Foi conduzida uma revisão sistemática seguindo as diretrizes PRISMA 2020, com padrões claros, transparência, reprodutibilidade e com literatura relevante de bases de dados acadêmicas de referência. Os critérios claros de inclusão/exclusão, juntamente com o processo de seleção explícito, permitem que pesquisadores subsequentes repliquem ou ampliem a revisão.

Outra contribuição foi a organização combinada com sintetização estruturada das arquiteturas de aprendizado profundo mais comuns em vigilância por vídeo, incluindo CNNs, RNNs, LSTMs, GRUs, convoluções 3D, ConvLSTM, Transformers e arquiteturas híbridas, o estudo comparativo identifica vantagens, desvantagens e casos de aplicação relevantes.

Também houve o mapeamento completo das principais aplicações como deteção de anomalias, deteção de atividades suspeitas, monitorização de multidões e sistemas integrados, e síntese dos desempenhos relatados para mostrar o que é tecnicamente viável e maduro e o que ainda está em investigação.

Permitiu uma compreensão aprofundada e discussão dos principais *trade-offs* como precisão vs. eficiência, latência vs. contexto temporal, generalização vs. especialização, desempenho vs. Explicabilidade, e desafios técnicos como tempo real, robustez, escalabilidade e generalização, para guiar decisões sobre o design do sistema.

Foram abordadas as questões éticas como privacidade pessoal, vigilância em massa, viés, transparência, GDPR e Ia Act, que frequentemente são negligenciadas por revisões técnicas, mas que são muito relevantes para o crescimento responsável e a ética da implementação de tecnologias de vigilância.

Ocorreu a identificação sistemática de lacunas na literatura existente juntamente com a proposta de uma agenda de pesquisa futura organizada em dez direções prioritárias, para

orientar esforços de pesquisa e desenvolvimento. Ainda foi discutido o estudo de caso (Veesion) ilustrando o caminho dos protótipos de pesquisa para produtos comerciais, enfatizando aspetos de implementação, integração e conformidade regulatória, que são frequentemente excluídos de estudos acadêmicos.

5.3 Limitações

Apesar do rigor metodológico adotado, este estudo apresenta limitações que devem ser identificadas. A revisão cobre o período de 2019 a 2025 e publicações apenas em inglês. Estudos em outras línguas e o desenvolvimento até 2019 não foram incluídos. Embora o período coberto seja recente e a língua inglesa seja a mais comum em publicações de engenharia e ciência da computação, isso limitaria o número de relatórios podendo haver omissão de alguns estudos relevantes.

Embora Scopus, IEEE Xplore e Web of Science sejam bases de dados de referência que têm cobertura ampla, outras bases de dados não foram incluídas na pesquisa. Literatura cinzenta (relatórios técnicos, teses, *preprints*) também não foi incluída. Isso levou à exclusão de alguns estudos, mas foi dada preferência a artigos revisados por pares visando a qualidade do estudo.

Os estudos revisados usam métricas de desempenho e *datasets* diferentes, o que dificulta comparações diretas rigorosas. Houve esforço para sintetizar o desempenho de forma comparável, mas a heterogeneidade limita o potencial para meta-análise quantitativa.

Muitos estudos são limitados em fornecer detalhes práticos de implementação, hiperparâmetros ou código-fonte, especialmente os que se referem a sistemas comerciais ou aplicações proprietárias. Essa restrição limita a reprodutibilidade e a avaliação crítica de alguns métodos.

Questões éticas e regulatórias foram tratadas, mas a discussão nesta revisão é essencialmente técnica. Por mais importantes que sejam, as dimensões sociais, psicológicas e políticas da vigilância inteligente por vídeo são discutidas de forma limitada. Uma análise interdisciplinar ampliada seria benéfica.

5.4 Perspetivas Futuras

As perspetivas futuras para a aplicação do aprendizado profundo na vigilância por vídeo são promissoras e desafiadoras, refletindo o aumento das capacidades técnicas e as preocupações éticas e sociais, com foco em como o mercado deve se consolidar e projeções de novas aplicações práticas.

Tecnicamente, espera-se uma convergência contínua para arquiteturas híbridas e multimodais que combinem as forças de abordagens distintas. Transformers adaptados para vídeo serão cada vez mais prevalecentes à medida que as melhores técnicas de treinamento e arquiteturas otimizadas sejam desenvolvidas. *Edge computing* continuará se beneficiando da evolução de arquiteturas leves que permitem processamento local avançado e de baixo custo, inclusive energético. Uma maior maturidade das técnicas de aprendizado auto supervisionado ajudará significativamente a reduzir a necessidade de anotações de dados, permitindo uma democratização de sistemas especializados.

Como direções futuras da investigação, deverá haver mais casos de uso não explorados ou subexplorados, como saúde com a supervisão de pacientes e deteção de quedas em lares de idosos, educação com a análise de engajamento em aulas, o transporte com a monitorização de comportamento de motoristas, indústria com a segurança laboral, controlo de processos, entre outras. Da mesma forma, espera-se mais especialização com sistemas otimizados para contextos específicos que superem sistemas generalistas.

Na esfera das perspetivas regulatórias e éticas, projeta-se que o quadro legal seja mais rigoroso e transparente com a implementação completa do *AI Act* da UE e regulações semelhantes em outros países. Isso estabelecerá exigências mais rigorosas para transparência, explicabilidade, avaliação de equidade, bem como proteção de privacidade, possivelmente retardando a implementação de alguns sistemas, mas também aumentando a confiança pública. Do ponto de vista ético, o debate sobre o equilíbrio entre proteção e privacidade, os limites aceitáveis de vigilância e os meios de supervisão e responsabilização, continuará. O envolvimento público nesta conversa é essencial para garantir que o avanço tecnológico seja consistente com os valores sociais e respeite os direitos básicos.

Para mitigar as lacunas científicas, em termos de pesquisa, espera-se um aumento na interdisciplinaridade, com colaboração entre cientistas da computação, engenheiros, advogados, sociólogos e políticos. Apenas abordagens técnicas terão impacto reduzido, o desenvolvimento responsável da tecnologia de vigilância deve considerar quais são as implicações sociais, legais e éticas.

Por fim, o mercado de vigilância por vídeo inteligente deverá experienciar um crescimento significativo, impulsionado em grande parte por preocupações com a segurança, redução de custos de *hardware* e maturação de tecnologias relacionadas. A consolidação já é esperada para alguns fabricantes que estão estabelecendo posições dominantes por meio de ecossistemas interligados que combinam *hardware*, *software* e serviços.

5.5 Considerações Finais

O uso de *deep learning* em sistemas de vigilância por vídeo para deteção de padrões comportamentais é uma área ativa de pesquisa e tecnologia, que tem visto progresso considerável nos últimos anos e possui grande potencial para o futuro. As soluções tecnológicas atuais já permitem deteção eficiente de anomalias, reconhecimento de atividades suspeitas, bem

como monitorização de multidões, com precisões que em alguns ambientes são superiores a 90% e próximas à precisão humana.

No entanto, a transição de protótipos de laboratório para implantação em ambientes reais apresenta grandes desafios técnicos relacionados com o processamento em tempo real, robustez a variações no ambiente, escalabilidade e generalização entre domínios. Fundamentalmente, a implantação responsável e ética dessas tecnologias não pode ser assumida levianamente sem entender os aspetos relacionados à privacidade, vigilância em massa, vieses algorítmicos, transparência e conformidade com um ambiente regulatório cada vez mais crescente.

Para que isso aconteça, deve-se encontrar um equilíbrio delicado entre inovação tecnológica e proteção dos direitos fundamentais, entre segurança e privacidade, entre automação e supervisão humana de qualidade. Este meio-termo não pode ser alcançado apenas através de soluções tecnológicas, mas sim por meio de diálogo entre desenvolvedores de tecnologia, reguladores, legisladores e sociedade civil.

Esta revisão sistemática visou estabelecer uma base sólida de conhecimento sobre o atual estado da arte, identificar desafios e lacunas e oferecer recomendações para futuras pesquisas. Espera-se que este trabalho contribua para a criação de sistemas de vigilância por vídeo inteligentes que sejam eficazes na promoção da segurança, respeitosos com a privacidade e os direitos fundamentais, transparentes e auditáveis, justos e não discriminatórios, e também alinhados com valores sociais e éticos.

Embora a vigilância por vídeo inteligente baseada em *deep learning* possa ter um papel importante na promoção da segurança pública e privada, esse potencial só será plenamente aproveitado quando o desenvolvimento tecnológico for acompanhado de uma reflexão ética profunda, um enquadramento regulatório apropriado e um compromisso com o desenvolvimento e implementação responsáveis. Este é o desafio e a oportunidade apresentados para a comunidade científica, indústria, reguladores e sociedade.

Referências Bibliográficas

[1] M. Qaraqe et al., “PublicVision: A secure smart surveillance system for crowd behavior recognition,” *IEEE Access*, vol. 12, pp. 1-15, 2024, doi: 10.1109/ACCESS.2024. 3366693.

[2] D. Dansana et al., “AI-Powered Surveillance,” 2024, doi: 10.1201/9781003596776-38.

[3] M. Mangrolia et al., “Abnormal Activity Recognition in Private Places Using Deep Learning,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, no. 6, pp. 1-8, 2023, doi: 10.32628/cseit228688.

[4] S. Rananaware et al., “Deep Learning Approach for Suspicious Activity Detection from Surveillance Video,” International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 2, pp. 1-6, 2024, doi: 10.48175/ijarsct-22360.

[5] S. Sundaram et al., “Intelligent Video Surveillance for Human Behavior Analysis and Anomalous Activity Detection using Deep Learning,” in Proc. ICSCSA, 2025, doi: 10.1109/icscsa66339.2025.11171248.

[6] Y. Huang et al., “Video surveillance-based identification of abnormal behavior of people in key areas,” in Proc. MLKE, 2022, pp. 1-5, doi: 10.1109/mlke55170.2022.00049.

[7] V. Konda, “Enhancing Public Safety through Real-time Video Surveillance Analytics Using AI and Computer Vision,” International Journal of Security and Applied Technologies, vol. 10, no. 4, pp. 1-12, 2019, doi: 10.71097/ijsat.v10.i4.4681.

[8] A. Joshi, “Deep Learning Approach for Suspicious Activity Detection from Surveillance Video in Exam Hall,” International Journal For Science Technology And Engineering, vol. 13, no. 1, pp. 1-5, 2025, doi: 10.22214/ijraset.2025.71381.

[9] M. Radhika et al., “Video Surveillance and Deep Learning Enhancing Security through Suspicious Activity Detection,” in Proc. IACIS, 2024, doi: 10.1109/iacis61494.2024.10721938.

[10] K. Rezaee et al., “A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance,” Personal and Ubiquitous Computing, vol. 26, no. 5, pp. 1-20, 2021, doi: 10.1007/S00779-021-01586-5.

[11] C. Maio et al., “Deep-Learning-Based Action and Trajectory Analysis for Museum Security Videos,” Electronics, vol. 13, no. 7, art. 1194, 2024, doi: 10.3390/electronics13071194.

[12] R. Sathya et al., “Intelligent Video Surveillance System for Real Time Effective Human Action Recognition using Deep Learning Techniques,” in Proc. ICACRS, 2023, doi: 10.1109/icacrs58579.2023.10404670.

[13] M. Rezaei et al., “Real-time crowd behavior recognition in surveillance videos based on deep learning methods,” Journal of Real-time Image Processing, vol. 18, no. 4, pp. 1-15, 2021, doi: 10.1007/S11554-021-01116-9.

[14] S. K. Wastupranata et al., “Deep Learning for Abnormal Human Behavior Detection in Surveillance Videos—A Survey,” Electronics, vol. 13, no. 13, art. 2579, 2024, doi: 10.3390/electronics13132579.

[15] S. Amingad et al., “Enhancing Public Safety with AI-Powered Intelligent Surveillance: An Examination of Immediate Incident Detection and Rapid Response in Urban Settings,” in Proc. ICAICCIT, 2023, doi: 10.1109/icaiccit60255.2023.10465690.

[16] S. S. A. Zaidi et al., “Suspicious human activity recognition from surveillance videos using deep learning,” IEEE Access, vol. 12, pp. 1-15, 2024, doi: 10.1109/access.2024.3436653.

[17] A. M. Buttar et al., “Toward trustworthy human suspicious activity detection from surveillance videos using deep learning,” Soft Computing, vol. 27, no. 18, pp. 1-15, 2023, doi: 10.1007/s00500-023-07971-x.

[18] E. T. Selvi et al., “Suspicious actions detection system using enhanced CNN and surveillance video,” Electronics, vol. 11, no. 24, art. 4210, 2022, doi: 10.3390/electronics11244210.

[19] E. Verdú, “ConvGRU-CNN: Spatiotemporal Deep Learning for Real-World Anomaly Detection in Video Surveillance System,” International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, no. 5, pp. 1-10, 2023, doi: 10.9781/ijimai.2023.05.006.

[20] R. S. Thenge et al., “Action Recognition Using a Deep Neural Network for Video Surveillance,” Epj Web of Conferences, vol. 328, art. 01020, 2025, doi: 10.1051/epjconf/202532801020.

[21] A. Gupta et al., “Advances in Deep Learning for Video Anomaly Detection: A Comprehensive Review,” in Proc. ICETETSIP, 2025, doi: 10.1109/icetetsip64213.2025.11156929.

[22] V. Konda, “Pattern Recognition Techniques in Video Surveillance for Improved Security in Public Spaces,” International Journal of Security and Applied Technologies, vol. 10, no. 1, pp. 1-10, 2019, doi: 10.71097/ijsat.v10.i1.7106.

[23] J. M. Flores-Munguía et al., “Leveraging a Neuroevolutionary Approach for Classifying Violent Behavior in Video,” Computational Intelligence and Neuroscience, vol. 2022, art. 1279945, 2022, doi: 10.1155/2022/1279945.

[24] I. Hussain et al., “AI-driven behavior biometrics framework for robust human activity recognition in surveillance systems,” Engineering Applications of Artificial Intelligence, vol. 129, art. 107218, 2024, doi: 10.1016/j.engappai.2023.107218.

[25] R. Vijeikis et al., “Efficient Violence Detection in Surveillance,” Sensors, vol. 22, no. 6, art. 2216, 2022, doi: 10.3390/s22062216.

[26] K. Muhammad et al., “Fuzzy logic in surveillance big video data analysis: comprehensive review, challenges, and research directions,” ACM Computing Surveys, vol. 54, no. 3, art. 1-33, 2021, doi: 10.1145/3444693.

[27] Veesion, “Prévenez le vol. Agissez à temps.” 2025. [Online]. Disponível em: https://veesion.io/. Acedido em: 13 de abril de 2026.

[28] Fastech, “A History of CCTV From Military Origins to AI Surveillance.” 2026. [Online]. Disponível em: https://fastechltd.co.uk/. Acedido em: 24 de abril de 2026.

[29] Advanosys, “Security – Command Control Center.” 2026. [Online]. Disponível em: https://www.advanosys.com/. Acedido em: 24 de abril de 2026.