Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Coordenação de Pós-Graduação em Ciência da Computação

Dissertação de Mestrado

Reconhecimento de Atividades Humanas Violentas em Videovigilância Utilizando Redes Neurais Profundas e Delimitação de Área de Interesse

Jayne de Morais Silva

Campina Grande, Paraíba, Brasil

08/2022

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Reconhecimento de Atividades Humanas Violentas em Videovigilância Utilizando Redes Neurais Profundas e Delimitação de Área de Interesse

Jayne de Morais Silva

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Visão Computacional

Orientador: Prof. Dr. Eanes Torres Pereira

Campina Grande, Paraíba, Brasil

S586r Silva, Jayne de Morais.

Reconhecimento de atividades humanas violentas em videovigilância utilizando redes neurais profundas e delimitação de área de interesse / Jayne de Morais Silva. – Campina Grande, 2022.

126 f.: il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2022.

"Orientação: Prof. Dr. Eanes Torres Pereira".

Referências.

1. Inteligência Artificial. 2. Visão Computacional. 3 Reconhecimento de Atividades Humanas. 4. Violência. I. Pereira, Eanes Torres. II. Título.

CDU 004.8(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA MARIA ANTONIA DE SOUSA CRB 15/398

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO CIENCIAS DA COMPUTACAO
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

JAYNE DE MORAIS SILVA

RECONHECIMENTO DE ATIVIDADES HUMANAS VIOLENTAS EM VIDEOVIGILÂNCIA UTILIZANDO REDES NEURAIS PROFUNDAS E DELIMITAÇÃO DE ÁREA DE INTERESSE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 04/08/2022

Prof. Dr. EANES TORRES PEREIRA, UFCG, Orientador

Prof. Dr. HERMAN MARTINS GOMES, UFCG, Examinador Interno

Prof. Dr. PÉRICLES BARBOSA CUNHA DE MIRANDA, UFRPE, Examinador Externo

Documento assinado eletronicamente por **EANES TORRES PEREIRA**, **PROFESSOR 3 GRAU**, em 04/08/2022, às 11:01, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **HERMAN MARTINS GOMES**, **PROFESSOR 3 GRAU**, em 04/08/2022, às 11:19, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade,

informando o código verificador **2575011** e o código CRC **61886FBD**.

Referência: Processo nº 23096.048931/2022-96 SEI nº 2575011

Resumo

O crescimento da quantidade de câmeras de videovigilância implantadas para o monitoramento de ambientes nos últimos anos não é proporcional à capacidade humana de análise das cenas capturadas. As cenas capturadas podem conter evidências de ocorrências de crimes. No entanto, câmeras de videovigilância são pouco utilizadas para interromper ou prever atividades criminosas simultaneamente a suas ocorrências. Para tornar o combate ao crime mais eficiente, o reconhecimento de ações humanas poderia ser realizado automaticamente por meio de técnicas computacionais capazes de detectar e classificar os tipos de comportamentos humanos. Além disso, no cenário de reconhecimento de padrões em sistemas de videovigilância, outro grande desafio é definir um limiar entre eventos violentos e não-violentos em ambientes em constante mudança e de comportamentos com interpretações ambíguas, considerando o contexto em que são realizados. Por esse motivo, como a natureza das cenas capturadas a partir de câmeras de videovigilância é constituída em sua maior parte de comportamentos comuns ou não violentos, o monitoramento de cenas requer que a capacidade de análise e percepção de atos agressivos seja precisa e acurada. Neste trabalho, é apresentada uma proposta para a detecção de comportamentos humanos violentos através de técnicas de visão computacional, tendo como principal contribuição a delimitação da área de interesse do quadro por meio do filtro gaussiano, como também, a redução do espaço de características de entrada para o modelo, mantendo as características mais relevantes. Além disso, a proposta é capaz de reduzir em aproximadamente até 45% o uso de memória VRAM (*Video Random Access Memory*) durante a fase de treinamento. A abordagem proposta obteve acurácia de 86,5% na fase de teste com o conjunto de dados *RWF-2000* e superou a abordagem *baseline*, constituída por uma rede neural convolucional (CNN) treinada para a classificação de cenas humanas violentas, combinada com a técnica de corte da área de interesse dos quadros de vídeos. A abordagem também superou outras propostas do estado da arte no cenário de videovigilância. Análises estatísticas realizadas apontam a significância da melhoria dos resultados ao adotar-se o método proposto nesta pesquisa. A proposta também foi avaliada em conjuntos de dados de *benchmark* em cenários de brigas humanas.

Palavras-chave: Visão Computacional, Reconhecimento de Atividades Humanas, Violência.

Abstract

The growth in the number of video surveillance cameras deployed for monitoring environments in recent years is not proportional to the human capacity to analyze the captured scenes. Captured scenes may contain evidence of crime occurrences. However, video surveillance cameras are rarely used to stop or predict criminal activities simultaneously with their occurrences. To make crime fighting more efficient, the recognition of human actions could be performed automatically by means of computational techniques capable of detecting and classifying the types of human behavior. Moreover, in the scenario of pattern recognition in video surveillance systems, another major challenge is to define a threshold between violent and non-violent events in changing environments and behaviors with ambiguous interpretations, considering the context in which they are performed. For this reason, as the nature of the scenes captured from video surveillance cameras consists mostly of common or non-violent behaviors, scene monitoring requires precision and accuracy on the ability to analyze and perceive aggressive actions. In this work, a proposal for the detection of violent human behavior through computer vision techniques is presented, having as main contribution the delimitation of the area of interest of the frame through the Gaussian filter, as well as the reduction of the space of input features for the model, keeping the most relevant features. Furthermore, the proposal is able to reduce the use of VRAM (Video Random Access Memory) by approximately up to 45% during the training phase. The proposed approach obtained accuracy of 86.5% in the test phase with the RWF-2000 dataset and outperformed the baseline approach, consisting of a convolutional neural network (CNN) trained for the classification of violent human scenes, combined with the technique of cutting the area of interest from the video frames. The approach also outperformed other state-of-the-art proposals in the video surveillance scenario. A performed analysis, pointed statistical significance when adopting the method proposed in this research. The proposal was also evaluated on benchmark datasets in human fight scenarios.

Keywords: Computer Vision, Human Activity Recognition, Violence.

Agradecimentos

Agradeço a Deus, por ter me concedido saúde e força para superar dificuldades que surgiram durante o meu percurso.

Aos meus pais, minha mãe Vanuza Morais e meu pai Jair Silva, agradeço os ensinamentos, o incentivo, o apoio e, pela incansável dedicação e confiança, não medindo esforços para investir na minha educação.

Ao meu noivo Matheus Lacerda pelo carinho, por todo amor, ajuda e tranquilidade.

À minha prima Emília Lima e seu esposo Leandro Silva pelo acolhimento e pelas palavras amigas e motivadoras.

À minha família, em especial à minha irmã Emilly Jullyane pelo companheirismo e pelas alegrias.

Aos meus professores, em especial ao meu orientador, Eanes Pereira, que contribuiu com minha formação acadêmica, agradeço os ensinamentos, as orientações e a paciência.

Aos demais amigos e também aos integrantes do Laboratório de Percepção Computacional (LPC) da UFCG, que contribuíram direta ou indiretamente com este trabalho.

Ao time da coordenação do programa de pós-graduação, pelo incentivo, disponibilidade, eficiência e toda atenção dada.

À instituição e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo incentivo para o desenvolvimento e conclusão desse trabalho.

Conteúdo

Lista de Símbolos

RAM - *Random Access Memory*

HAR - *Human Action Recognition*

OMS - Organização Mundial da Saúde

WHO - *World Health Organization*

CNN - *Convolutional Neural Network*

HA - *Human Activity*

LIBRAS - Língua Brasileira de Sinais

GPS - *Global Positioning System*

LBP - *Local Binary Pattern*

DNN - *Deep Neural Network*

RNN - *Recurrent Neural Network*

AE - *Auto-Encoders*

GAN - *Generative Adversarial Networks*

LSTM - *Long Short-Term Memory*

EUROSUR - *European Border Surveillance System*

MRF - *Markov Random Field*

GMM - *Gaussian Mixture Model*

HMM - *Hidden Markov*

SciELO - *Scientific Electronic Library Online*

ACM - *Association for Computing Machinery Digital Library*

IEEE - *Institute of Electrical and Electronics Engineers*

ConvLSTM - *Convolutional Long Short-Term Memory*

IDT - *Improved Dense Trajectories*

IFV - *Improved Fisher Vectors*

HOG - *Histogram of Oriented Gradients*

TS - *Trajectory Shape*

HOF - *Histogram of Optical Flow*

MBH - *Motion Boundary Histogram*

SVM - *Support Vector Machine*

SIFT - *Scale-Invariant Feature Transform*

MoSIFT - *Motion SIFT*

KDE - *Kernel Density Estimation*

PDF - *Probability Density Function*

STIP - *Space-Time Interest Point*

TOFF - *Temporal Optical Flow Features*

DMN - *Deep Multi-Net*

VGG16 - *Visual Geometry Group*

MNAS - *Automated Mobile Neural Architecture Search*

KNN - *K-Nearest Neighbors*

ViF - *Violent Flow*

DI - *Dynamic Images*

MOG2 - *Mixture of Gaussians*

C3D - *Convolutional 3D Networks*

SGD - *Stochastic Gradient Descent*

MAC - *multiply-and-accumulate*

GPU - *Graphics Processing Unit*

TPU - *Tensor Processing Unit*

AUC - *Area Under Curve*

IC - Intervalo de Confiança

FN - *False Negative*

FP - *False Positive*

TP - *True Positive*

TN - *True Negative*

LDA - *Linear Discriminant Analysis*

BG - *Background Subtraction*

POT - *Pooled of Time series*

DMEI - *Diferential Motion Energy Image*

ReLU - *Rectified Linear Unit*

AUC - *Area Under Curve*

VRAM - *Video Random Access Memory*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Neste Capítulo introdutório, o contexto da problemática encontra-se detalhado na Seção 1.1. Na Seção 1.2, são apresentadas as motivações do presente trabalho, respectivamente. Na Seção 1.3, são descritos os objetivos geral e específicos. Por último, na Seção 1.5, é detalhado, de maneira geral, a estrutura desta dissertação.

1.1 Contextualização

Conforme Lin (2019) até o ano de 2019 já existiam cerca de 770 milhões de câmeras de videovigilância no mundo e, segundo o autor, estima-se que até o final do ano de 2021 a quantidade de câmeras de videovigilância seria superada em mais de 1 bilhão. Atualmente, existe uma discrepância entre a quantidade de dados de vídeo capturados continuamente pelas câmeras de videovigilância e a aptidão humana de analisar eficientemente essas informações visuais (LEJMI et al., 2019).

A natureza de cenas que constituem vídeos capturados de câmeras de videovigilância é constituída em sua maior parte de eventos cotidianos, considerados normais ou comuns. Desta forma, informações visuais de eventos raros ou considerados anômalos ocorrem em menor frequência, podendo passar despercebidos aos olhos humanos e não serem analisados eficientemente, quando comparados ao poder computacional executado por algoritmos capazes de classificar, detectar ou prever esse tipo de comportamento em quantidades massivas de cenas e em poucos instantes.

A Organização Mundial da Saúde (OMS)[1] define violência como "o uso intencional de força física ou poder, real ou em ameaça, contra si mesmo, outra pessoa, ou contra um grupo ou comunidade, que resulte ou tenha grande probabilidade de resultar em lesão, morte, dano psicológico, deficiência de desenvolvimento ou privação". A violência pode ser classificada em autodirigida, interpessoal ou coletiva e também pode variar conforme a maneira que os atos são acometidos na vítima, como por exemplo: violência física, psicológica, de gênero, doméstica, sexual, intrafamiliar, patrimonial, moral, entre outras. No entanto, embora exista uma definição teórica sobre o conceito de violência, a detecção automática de comportamento agressivo por computador apresenta alguns desafios devido à sua natureza subjetiva na definição do que deve ser considerado como violência (LEJMI et al., 2019). Por exemplo, a atividade de abraçar, a depender do contexto que é realizada, pode significar afeto ou o ato de imobilizar um sujeito durante uma briga. Segundo Lejmi et al. (2019), ao longo destes últimos anos, várias pesquisas sobre o Reconhecimento de Ações Humanas (HAR - *Human Action Recognition*) foram propostas, porém, a caracterização da violência tem sido comparativamente menos explicada.

Além da questão da ambiguidade de significado em uma atividade, outros desafios relacionados à qualidade das cenas capturadas, como a baixa qualidade fornecida pela maioria dos dispositivos de videovigilância, cenas coletadas de contextos que não representam a realidade, como também, conforme Beddiar et al. (2020), a oclusão de objetos ou partes do corpo durante a cena, variações de iluminação e de *background*, ruído, entre outros, também podem afetar a confiabilidade dos algoritmos.

O estudo sobre as atividades humanas em visão computacional é um campo que compreende a análise de comportamentos, relações e interações dos seres humanos/objetos obtidos a partir de imagens ou vídeos. No contexto de detecção de ações humanas por computador, a literatura em sua maior parte, concentra-se no reconhecimento de atividades humanas comuns ou cotidianas, como por exemplo: caminhar, correr, pular, jogar e etc. Essas atividades realizadas cotidianamente costumam ser capturadas com maior frequência do que eventos ou atividades anormais, que raramente ocorrem, tais como, acidentes, atividades suspeitas e crimes.

[1] *The VPA Approach*. Disponível em: <https://www.who.int/groups/violence-prevention-alliance/approach>. Acesso em: 15 de maio de 2022.

De acordo com Hussain, Sheng e Zhang (2019), diferentes técnicas têm sido propostas para o reconhecimento automático da atividade humana, tanto para abordagens baseadas em visão quanto para abordagens baseadas em sensores. No entanto, existem vários desafios relacionados, tais como: alto custo, complexidade computacional, questões relacionadas à privacidade, entre outros.

Em linhas gerais, o conceito de violência definido no escopo desta pesquisa é caracterizado por atos humanos constituídos por comportamentos físicos agressivos, como socos, chutes, entre outros, sendo categorizados, na maioria das vezes, como violência interpessoal. O direcionamento do propósito desta dissertação está voltado para a identificação e classificação de atos não agressivos e agressivos. Portanto, o contexto e desenvolvimento desta pesquisa encontra-se inserido no problema estudado pela área HAR. É importante ressaltar que o tema detecção de violência, de maneira geral, pode estar diretamente relacionado com estudos de outras linhas de pesquisa, como por exemplo, detecção de anomalias.

A presente pesquisa representa uma progressão no estudo exploratório de lacunas desafiadoras do contexto de detecção de comportamento humano violento. Além disso, são propostas contribuições que foram testadas e validadas e obtiveram resultados significativos e superiores ao método utilizado como *baseline* durante o desenvolvimento desta pesquisa. O código-fonte está disponível em no GitHub[2].

1.2 Motivação

Atualmente, a violência resulta em mais de 1,5 milhão de pessoas mortas no mundo a cada ano. Em uma publicação online da ONU de 2010[3], afirma-se que a violência está entre as principais causas de morte em todo o mundo para pessoas de 15 a 44 anos. Por outro lado, há uma grande parcela da população que sofre lesões não-fatais provocadas por violência, resultando em incapacidade temporária ou permanente, necessidade de cuidados intensivos, hospitalizações, reabilitação física e/ou mental de longo prazo, entre outras. Além disso, a exposição a qualquer tipo de trauma pode também acarretar um aumento no risco de doenças

[2]Repositório GitHub - ViolenceDetection. Disponível em <https://github.com/devjaynemorais/ViolenceDetection>. Acesso em 25 de junho de 2022.

[3]*Violence prevention: the evidence*. 2010. Disponível em: <https://www.who.int/publications/i/item/violence-prevention-the-evidence>. Acesso em: 16 de maio de 2022.

mentais e/ou crônicas, suicídio, uso de álcool e drogas, entre outros problemas. Por essas razões, a prevenção de violência, não se resume apenas em evitar a lesão física, mas abrange a contribuição em ganhos substanciais em saúde, sociais e econômicos (OMS, 2021).

Conforme o estudo realizado por Bischoff (2021), o aumento do número de câmeras pode não implicar causalidade direta no fator de redução no índice de criminalidade. Contudo, o monitoramento e análise das cenas capturadas a partir de câmeras de videovigilância, de maneira eficiente, pode contribuir fortemente na prevenção e identificação de crimes, acidentes ou atividades suspeitas.

Dessa forma, a importância do estudo e monitoramento de ações humanas em cenas capturadas por câmeras de videovigilância está relacionada com diversas tarefas, tais como: a prevenção e identificação de crimes e acidentes; o monitoramento de cenas que contenham elementos que indiquem presença humana em determinados locais, na maioria das vezes proibidos; a classificação de comportamentos ou atividades humanas suspeitas; a detecção de comportamentos anômalos. Contudo, devido à vasta quantidade de cenas capturadas continuamente, faz-se necessário o uso de recursos automatizados para análise e identificação das mesmas. Para isso, os modelos baseados em redes neurais profundas são comumente utilizados para lidar com problemáticas relacionadas à detecção de comportamentos alvo em imagens ou sequência de imagens. No entanto, o desenvolvimento e execução desses modelos é acompanhado por diversos desafios e limitações relacionadas a complexidade de generalização do problema, como também o uso massivo de recursos computacionais para o treinamento de modelos.

Uma estratégia para tentar diminuir o uso de recursos computacionais é reduzir a quantidade ou refinar dados submetidos a um modelo. No entanto, a redução da quantidade de amostras pode afetar diretamente a capacidade de generalização de modelos. Ao mesmo tempo, muitas características dos dados de entrada após carregadas e processadas, acabam sendo classificadas como irrelevantes durante o processo de aprendizagem de um modelo. Neste cenário, é muito oportuno uma proposta que vise à seleção e aprimoramento de características irrelevantes contidas nas amostras de entrada de vídeos de videovigilância.

Outro grande desafio, para a execução e treinamento de um modelo capaz de desempenhar essas funcionalidades, é o fato de serem exigidas altas demandas de recursos computacionais, como o uso de memória VRAM e processamento, que torna essa tarefa bastante desafiadora

de ser explorada. A carência de hardware adequado dificulta desde a reprodução de trabalhos do estado da arte ao desenvolvimento de propostas de soluções mais robustas. Não há dúvidas que o custo computacional e limitações de hardware são limitações recorrentes no campo de pesquisa e desenvolvimento e muitas investigações surgem com o intuito de amenizar a problemática.

Neste cenário, considerando as problemáticas e lacunas supramencionadas, esta dissertação apresenta um estudo sobre a área de HAR, acompanhado de uma revisão sobre as principais características e desafios da área de visão computacional aplicada à videovigilância. Posteriormente, foi desenvolvido um método de seletividade de região de interesse dos quadros de vídeos e em seguida, o método foi aplicado como etapa de pré-processamento em uma abordagem de detecção de violência existente no estado da arte. A abordagem compreende um modelo de classificação, uma Rede Neural Convolucional (CNN - *Convolutional Neural Network*), a qual foi retreinada nesta pesquisa utilizando estratégias de redução do uso de recursos computacionais. Por fim, são apresentadas análises estatísticas dos resultados que validam o sucesso da proposta em comparação com outras do estado da arte.

1.3 Objetivos

O principal objetivo deste trabalho é desenvolver um método de seleção de área de interesse em sequências de quadros de vídeos, que aprimore o reconhecimento de atividades humanas violentas em cenas obtidas por câmeras de videovigilância. No método proposto, ao invés de extrair características de toda a área dos quadros com igual relevância ou realizar um corte de uma área específica desses quadros, a área de baixo interesse dos quadros é delimitada através da aplicação do filtro gaussiano, sendo preservadas características originais da área de interesse. Para atingir o objetivo principal, foram definidos os seguintes objetivos específicos:

- Demonstrar a melhoria de resultados existentes no estado da arte, através da aplicação do método desenvolvido nesta pesquisa em um modelo *baseline* de detecção de violência em vídeo, utilizando uma abordagem de delimitação da área de interesse de quadros de vídeo por meio do filtro gaussiano;

- Impulsionar a pesquisa na área de reconhecimento de atividades humanas violentas em videovigilância através de redes neurais profundas, utilizando uma técnica de redução do uso de recursos computacionais durante o treinamento por meio de precisão mista;

1.4 Método

- Revisão bibliográfica do estado da arte da área de HAR em visão computacional, fundamentos de um sistema de videovigilância, características dos conjuntos de dados encontrados e sobre os principais desafios e lacunas;
- Implementação e validação das propostas de (1) estratégia de redução de características de entrada por meio do filtro gaussiano na área de baixo interesse e (2) uso de precisão mista nas camadas do modelo visando à redução de memória RAM, ambas implementadas e aplicadas a uma abordagem *baseline* previamente selecionada;
- Avaliação estatística dos resultados e avaliação dos resultados em conjuntos de dados de *benchmark* em detecção de brigas humanas.

1.5 Organização do Trabalho

Este trabalho está organizado da seguinte forma: No Capítulo 2, são apresentados a fundamentação teórica, a descrição de conceitos relacionados e desafios de pesquisa (Seção 2.1), o contexto em que a pesquisa encontra-se inserida (Seção 2.2) e a área de aplicação (Seção 2.3). No Capítulo 3, são apresentados uma visão sobre os trabalhos do estado da arte, técnicas e abordagens adotadas, assim como, um agrupamento e sumarização das mesmas. Posteriormente, no Capítulo 4, são descritos os materiais e métodos utilizados, com ênfase nos experimentos realizados e detalhes de implementação. No Capítulo 5, são relatados os resultados experimentais e exploração desses resultados, assim como uma exploração das falhas encontradas. Por último, este trabalho é concluído no Capítulo 6 com direcionamentos futuros para melhoria da abordagem e propostas de novos experimentos.

Capítulo 2

Fundamentação Teórica

Neste Capítulo, apresenta-se toda a fundamentação teórica associada a esta pesquisa. Na Subseção 2.1, é introduzida uma visão geral da área de pesquisa (*Human Activity Recognition* - HAR), assim como artefatos e conceitos gerais relacionados. Na Subseção 2.2, são explicados conceitos relacionados com as principais contribuições deste trabalho, evidenciando a linha tecnológica da pesquisa (Visão Computacional) onde a problemática está inserida e seus principais desafios e limitações. Por último, na Subseção 2.3, são apresentados os principais fundamentos e processos associados ao contexto da aplicação (Videovigilância).

2.1 Reconhecimento de Ações Humanas

Caracterizada pela complexidade e multiplicidade, a atividade humana diz respeito à conduta e aos comportamentos realizados por indivíduos. Desse modo, a análise de atividades humanas compreende o estudo das relações interativas entre seres humanos e/ou ambiente, que sejam plausíveis de serem observadas, dadas as circunstâncias em que ocorrem.

Conforme Beddiar et al. (2020), atividade humana é definida como:

> Mais especificamente, uma atividade humana (HA - *Human Activity*) se refere ao (s) movimento(s) de uma ou várias partes do corpo da pessoa. Isso pode ser atômico ou composto de muitas ações primitivas executadas em alguma ordem sequencial. Portanto, o reconhecimento da atividade humana deve permitir rotular a mesma atividade com o mesmo rótulo, mesmo quando realizada por pessoas diferentes em condições ou estilos diferentes. (BEDDIAR et al., 2020, p.1)

O reconhecimento de atividades humanas objetiva detectar determinados comportamentos

físicos do corpo, como, por exemplo, as ações de caminhar, pular, sentar, etc. Esses movimentos podem envolver uma parte específica do corpo como gestos realizados pelas mãos e podem conter ou não interação com outros indivíduos e/ou objetos.

2.1.1 Visão Geral da Área

O Reconhecimento de Atividade Humana (HAR - *Human Activity Recognition*) é uma área científica que tem como principal instrumento de estudo os padrões de atividades realizadas por agentes ou entre esses em uma determinada cena, com ou sem interação com outros agentes, sejam estes, humanos ou objetos. O HAR está envolvido no desenvolvimento de muitas aplicações, tais como, videovigilância, monitoramento doméstico, interação humano-computador (IHC), saúde.

Nas Figuras 2.1 e 2.2, conforme Beddiar et al. (2020), são apresentadas as pesquisas relacionadas com HAR entre os anos de 2010 e 2019. A Figura 2.1 evidencia o desbalanceamento entre a pequena quantidade de pesquisas dedicadas em discutir aspectos gerais do HAR e o grande número de pesquisas em taxonomias específicas e domínios de aplicação de HA. Na Figura 2.2, é apresentada a distribuição dos principais assuntos de HAR cobertos pelas pesquisas mencionadas, onde pode ser observado que apenas 5% das pesquisas possuem aprendizagem profunda em HAR como principal assunto.

Segundo Beddiar et al. (2020), a implementação de sistemas HAR é guiada por duas correntes principais de tecnologias de interação homem-computador: com base em contato e por métodos remotos. Os sistemas baseados em contato requerem interação física do usuário com a máquina, os dados podem ser emitidos a partir de fontes sensoriais como, por exemplo: telas multitoque, acelerômetros, sensores montados no corpo (eletrodos) ou vestíveis (roupas ou luvas especiais). Uma crítica forte a essa categoria de sistema é que na maioria das vezes esses sensores precisam ser acoplados ao corpo do usuário, fazendo com que o mesmo tenha uma experiência negativa por seu uso ser desconfortável ou intrusivo. Já os sistemas baseados em visão (métodos remotos) usam imagens ou sequências de vídeos gravados para reconhecer atividades.

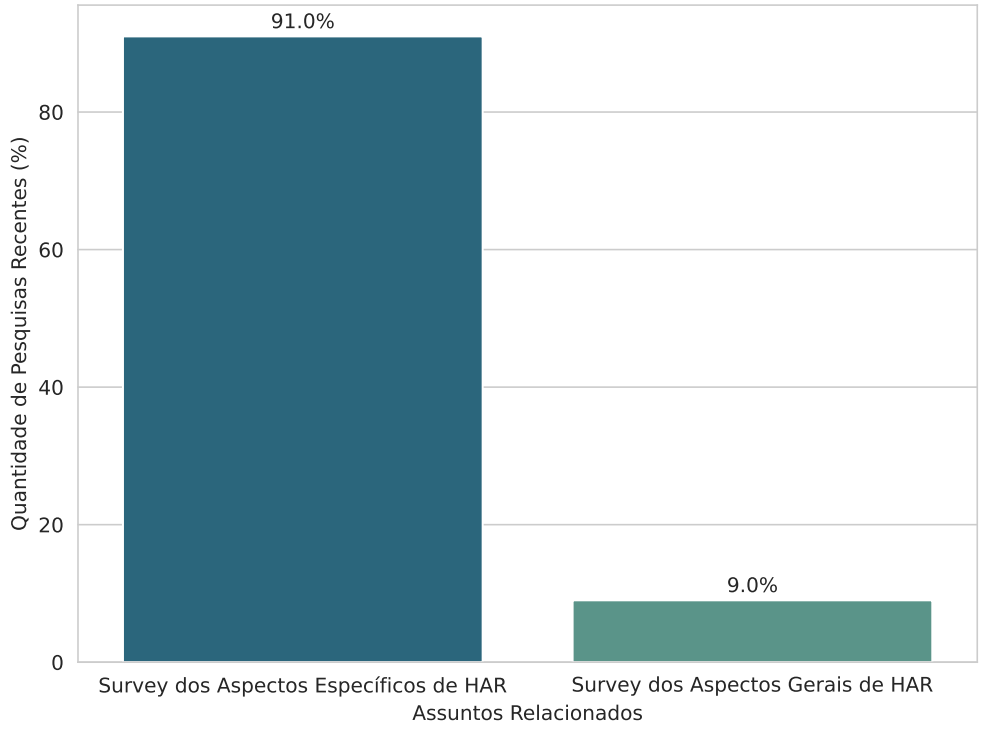

Figura 2.1: Pesquisas recentes em HAR. Fonte: Adaptado de (BEDDIAR et al., 2020, p.3).

2.1.2 Categorias de Atividade

Segundo Aggarwal e Ryoo (2011), as atividades humanas são classificadas nos seguintes níveis hierárquicos, conforme a escala de complexidade desde uma ação simples a eventos mais complexos: gestos, ações, interações e atividades em grupo. Enquanto Vrigkas e Kakadiaris (2015) apresentam outras categorias de atividade humana, como ação (e.g., ação atômica), comportamento e evento.

Dessa maneira, cada categoria de atividade possui um nível de complexidade diferente, que pode variar conforme, por exemplo, o número de indivíduos da cena ou quantidade de partes do corpo envolvidas na cena. Na Figura 2.3, é apresentada uma hierarquia entre essas categorias de atividades, da menos complexa a mais complexa, considerando os dois exemplos citados acima.

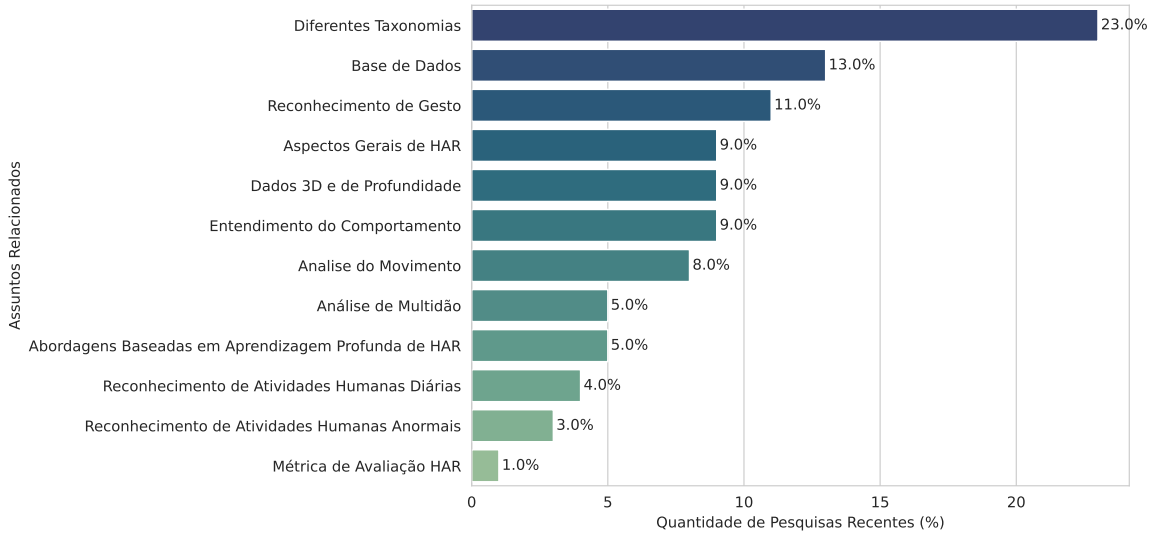

Figura 2.2: Pesquisas recentes em HAR. Adaptado de (BEDDIAR et al., 2020, p.3).

Figura 2.3: Categorias de atividades humanas que vão desde uma ação simples até um evento. Fonte: Adaptado de (BEDDIAR et al., 2020, p.17).

Atividades baseadas em gestos

Conforme a definição de Aggarwal e Ryoo (2011), os gestos possuem a capacidade de descrever um movimento significativo produzido pelas partes do corpo de um indivíduo, os quais concorrem com os elementos linguísticos e não obedecem a um sistema de restrições, como, por exemplo, levantar um braço. Beddiar et al. (2020) ainda complementam:

> Normalmente, um gesto é uma linguagem ou parte da comunicação não verbal que pode ser empregada para expressar ideias ou ordens significativas. Os gestos são um segundo tipo de atividade que pode ser consciente como "aplaudir", e inconsciente como "esconder o rosto com as mãos ao ficar tímido" ou "puxar a mão ao tocar em um material quente". Alguns gestos são universais, enquanto outros estão relacionados a contextos sociais e culturais bastante específicos. (BEDDIAR et al., 2020, p.18)

Atividades baseadas em ações ou comportamentos

A identificação e reconhecimento de comportamentos ou ações humanas pode ser muito útil na detecção de comportamentos anormais nos setores da saúde e segurança, seja em espaços públicos ou privados, para auxiliar na identificação de cenas agressivas, crimes, quedas, acidentes, etc. Outro exemplo de aplicação é a identificação de comportamentos de clientes em um determinado estabelecimento comercial relacionados aos interesses, preferências e marcas, com o objetivo de auxiliar no setor de varejo, marketing, controle de fluxo/tráfego em centros de atendimento, na tomada de decisão e direcionamentos estratégicos. Em casas inteligentes, o reconhecimento de atividades diárias pode auxiliar na detecção de intrusão ou no ajuste do ambiente conforme a atividade do usuário, ou através da presença ou ausência de comportamento humano.

De acordo com Aggarwal e Ryoo (2011), as atividades baseadas em ações ou comportamentos são realizadas de forma individual e normalmente compostas de múltiplos gestos organizados temporalmente, como, por exemplo, caminhar. Beddiar et al. (2020) também acrescentam que as atividades baseadas em ações ou comportamentos descrevem o conjunto de ações e reações físicas de indivíduos em situações específicas que são observáveis de fora e são relativas às suas emoções e estados psicológicos.

Atividades baseadas em interação

Algumas atividades também podem ser realizadas interagindo com outros indivíduos, como também interagindo com objetos ou usando objetos. A maneira como se interage com os demais indivíduos ou objetos pode resultar em atividades diferentes. A iteração pode ser realizada através de gestos ou alguma ação específica. Segundo a definição de Beddiar et al. (2020):

> Iterações são ações ou trocas recíprocas entre duas ou mais entidades, que modificam o comportamento de indivíduos ou objetos envolvidos na interação. São, em geral, atividades complexas de dois tipos: de humano para humano, como "beijar", ou de humano para objeto, como "cozinhar", que envolve vários utensílios de cozinha. (BEDDIAR et al., 2020, p.18)

Atividades em grupos

São atividades realizadas por um conjunto de indivíduos, como uma reunião, por exemplo. Pode envolver gestos, ações e interações. Devido à sua natureza, são mais propícias a serem complexas e possuem um nível de dificuldade maior para serem reconhecidas.

Outras atividades

Segundo Beddiar et al. (2020), ainda existem mais duas categorias de atividades humanas: as ações humanas elementares e os eventos. As ações elementares teriam o nível de complexidade menor em relação às outras atividades e consistem em atividades simples, as quais formam a base para a construção das outras categorias de atividades (exemplo, levantar a mão esquerda). Enquanto os eventos são atividades com o maior nível de complexidade e ocorrem em ambientes específicos que representam ações sociais entre indivíduos (festas, por exemplo).

Existem também as atividades humanas associadas ao contexto comunicativo, como as utilizadas durante a linguagem corporal/facial ou em Língua de Sinais. Essa categoria de atividade humana não verbal, abrange principalmente gestos, posturas, expressões faciais, movimento do corpo, direção do olhar e a proximidade entre os agentes da cena. Abordagens de visão computacional e inteligência artificial podem ser utilizadas durante a decodificação dessas expressões, como na pesquisa de Silva (2020), por exemplo.

No contexto de linguagem corporal, principalmente em contextos investigativos, o comportamento humano pode fornecer indícios que facilitam os objetivos do processo investigativo ou interrogatórios, tais como micro expressões faciais ou expressões corporais que podem estar associadas às emoções de culpa, desprezo, alegria, tristeza, etc.

As Línguas de Sinais são línguas espaço-visuais que possuem sua própria estrutura gramatical. A comunicação ocorre através da realização de sinais pelas mãos do indivíduo e geralmente, são acompanhadas de expressões faciais ou corporais, que podem representar afetividade ou valor gramatical (SILVA, 2020). Conforme Nascimento (2009), a Língua Brasileira de Sinais (LIBRAS), por exemplo, é formada por cinco parâmetros fonológicos: a configuração da mão (forma realizada pela mão e seus respectivos dedos), o ponto de articulação (local onde o sinal é realizado em relação ao corpo), a orientação da mão (direção

da palma da mão), o movimento (tipo, direcionalidade, forma e frequência) e as expressões não manuais (corporais e faciais).

Considerações

Cenas utilizadas no HAR podem conter diversos tipos de ruídos que dificultam a etapa de identificação do tipo de atividade, tais como, grande número de pessoas presentes na cena, variação do *background*, variação de iluminação, movimentação da câmera. Existem algumas partes desnecessárias durante o processamento da imagem ou vídeo, e devem ser removidas, pois levam a um maior tempo de execução ou memória necessária (JOUDAKI; SUNAR; KOLIVAND, 2015). Informações desnecessárias podem confundir classificadores de atividade humana. Um exemplo de conteúdo desnecessário é a quantidade de características coletadas do *background* da cena, que pode conter informações que não têm relação direta ou indiretamente com a atividade humana presente. A subtração de *background* é uma das principais técnicas para análise automática de vídeo, especialmente no domínio da vigilância por vídeo (BRUTZER; HöFERLIN; HEIDEMANN, 2011), (JOUDAKI; SUNAR; KOLIVAND, 2015).

2.1.3 Aquisição de Dados

O reconhecimento/detecção de atividade humana é realizado com base em informações recebidas por diferentes categorias de tecnologias. As tecnologias mais comuns usadas para a aquisição de dados são câmeras de videovigilância, câmeras de profundidade, Wi-Fi[1], sensores (acelerômetros, magnetômetro, sensor de movimento, sensor de profundidade) e RFID[2].

Além disso, Vrigkas e Kakadiaris (2015) categorizam as abordagens HAR com relação ao canal de origem que cada uma dessas abordagens emprega para o reconhecimento de atividades humanas em duas categorias: unimodal (usam dados de uma única modalidade)

[1]Wi-Fi. Tecnologia de rede sem fio que permite que computadores (laptops e desktops), dispositivos móveis (smartphones e dispositivos vestíveis) e outros equipamentos (impressoras e câmeras de vídeo) se conectem à Internet. O Wi-Fi permite que esses e muitos outros dispositivos troquem informações entre si, criando uma rede (CISCO, 2021).

[2]RFID. (*Radio-Frequency IDentification*). Um sistema de RFID é composto, basicamente, de uma antena, um transceptor, que faz a leitura do sinal e transfere a informação para um dispositivo leitor, e também um transponder ou etiqueta de RF (rádio frequência), que deverá conter o circuito e a informação a ser transmitida (TECMUNDO, 2009).

e multimodal (usam dados de diferentes fontes). Ainda segundo os autores, as abordagens unimodais são apropriadas para reconhecer atividades humanas com base em características de movimento, enquanto as abordagens multimodais podem ser categorizadas com base na implantação do sensor: i) vestível (por exemplo, acelerômetro, GPS e biossensores), ii) marcado com objeto (sensores são anexados a objetos de uso diário) e iii) detecção densa (marcado com ambiente/sem dispositivo). Conforme Chen et al. (2012), as abordagens de captura de informações sobre as atividades humanas podem ser classificadas em: baseadas em visão (uso de recursos de detecção visual) e baseadas em sensores de captura de movimento (uso de tecnologias de rede de sensores para monitorar atividades).

Tipos de dados

O reconhecimento de atividades humanas associado à natureza do tipo de dado de entrada, pode ser realizado com base em imagens estáticas (a atividade humana pode ser distinguível de outras com base nas características) ou em vídeo (mais indicada para o reconhecimento da atividade humana e requer informações relacionadas à ocorrência de eventos/quadros anteriores e posteriores). Conforme Pareek e Thakkar (2021), o principal foco do reconhecimento de ações humanas em imagens estáticas é identificar a ação de uma pessoa a partir de uma única imagem, sem considerar informações temporais para a caracterização da ação.

As características granulares da natureza do dado de entrada também podem variar conforme o escopo para o qual foi coletado, como, por exemplo: formato 2D ou 3D, infravermelho, mapa de calor, informações de pose do esqueleto humano, histogramas e etc. Essas características podem ser utilizadas de maneira isolada ou integradas, embora o custo e complexidade computacional aumentem quando usados de maneira integrada. Na pesquisa de Singh e Vishwakarma (2019), é evidenciado que os conjuntos de dados 2D enfrentam mais dificuldades com variações intraclasse, variações de iluminação, fundo desordenado, oclusões parciais e movimentos de câmera quando comparados com conjuntos de dados de profundidade.

2.2 Reconhecimento de Atividades Humanas Baseado em Visão Computacional

O reconhecimento de ações humanas por computador compreende um conjunto de procedimentos e técnicas, as quais estão ligadas à capacidade de percepção computacional em aprender a identificar padrões de execução de atividades humanas, em sua maioria através de aspectos visuais coletados a partir de imagens ou sequências de imagens. Possui como principal objetivo detectar a atividade humana e classificá-la em atividades específicas e, envolve uma vasta quantidade de tópicos profundamente enraizados, como, por exemplo, a detecção de alvo em movimento, o rastreamento de pessoas, a detecção de pessoas pela aparência, entre outras (VEZZANI; BALTIERI; CUCCHIARA, 2013).

Os métodos HAR, normalmente, são compostos por componentes inter-relacionados, que vão desde a segmentação de quadro de vídeo para detecção da ação e sua respectiva representação até o processo de aprendizagem que reconhece essas ações (BEDDIAR et al., 2020).

2.2.1 Processo de Extração de Características

Na pesquisa realizada por Beddiar et al. (2020), é evidenciada uma classificação das abordagens HAR, conforme o processo de extração de características, em abordagens baseadas em representação *handcrafted* e abordagens baseadas em aprendizado de características.

Métodos *Handcrafted*

As características *handcrafted* são extraídas através de algoritmos personalizados, isto é, criados especificamente para o problema em questão. Após as características serem extraídas, normalmente são fornecidas como entrada para modelos de classificação. Métodos baseados em características *handcrafted*, em sua grande parte, extraem características recomendadas por meio de um especialista, como um cientista de dados, por exemplo, (processo conhecido como *feature engineering*), que pode ser um procedimento caro e difícil, além de exigir conhecimento prévio para extrair as características discriminantes[3].

[3]Máquinas que aprendem (MAp): Como representar uma tarefa de aprendizagem de máquina: *handcrafted features* vs *feature learning*. Disponível em: <https://maquinasqueaprendem.com/2020/03/24/como-representar-

De acordo com Bux (2017) e Beddiar et al. (2020), os tipos de métodos baseados em características *handcrafted* para vídeos envolvem três etapas principais, as quais estão descritas abaixo e podem ser visualizadas na Figura 2.4:

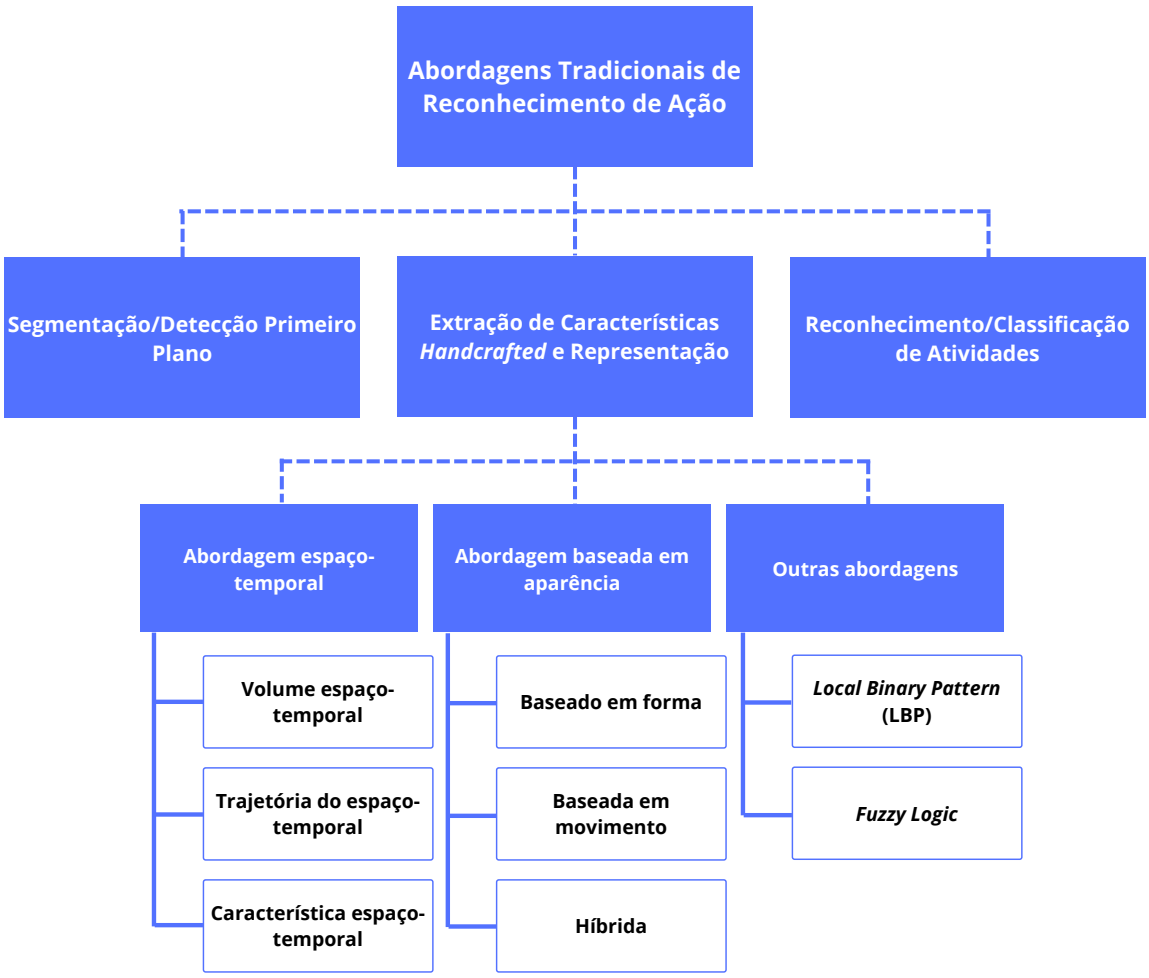

Figura 2.4: Abordagens tradicionais de reconhecimento para representação de ações. Fonte: Adaptado de (BUX, 2017, p.23).

1. A etapa inicial desse processo trata-se da segmentação da ação da sequência de quadros;

2. A segunda etapa é constituída pela extração de características desses quadros que são descritas e representadas por processos projetados por especialistas;

3. Por último, é realizada a classificação da ação, representada pelas características extraídas.

uma-tarefa-de-aprendizagem-de-maquina-handcrafted-features-vs-feature-learning/>. Acesso em: 22 de novembro de 2021.

Para a extração de características, os métodos baseados em características *handcrafted*, podem ser categorizados em representações espaço-temporais ou de aparência. De acordo com as definições de Bux (2017), as representações podem ser obtidas por meio de modelos categorizados e descritos na Tabela 2.1. Ainda segundo o autor, as abordagens baseadas em características espaço-temporais são mais adequadas para conjuntos de dados simples, enquanto para conjuntos de dados complexos é sugerida uma combinação de diferentes tipos de características.

Tabela 2.1: Classificação dos métodos *handcrafted*

Representação baseada em:	**Modelo baseado em:**	**Descrição**
Espaço-Temporal	Volume	Indicados para o reconhecimento de gestos ou ações simples, com uma ou poucas pessoas. Seu funcionamento dá-se através de uma medida de similaridade entre dois volumes para reconhecer a ação.
	Trajetória	Indicados para o reconhecimento de ações complexas com movimentos invariáveis, mas são desafiadoras para lidar com a localização de coordenadas 3D.
	Características	Extraem características espaço-temporais das duas abordagens acima para reconhecimento de ação humana.
Aparência	Forma	Capturam características geométricas da imagem humana, como, por exemplo, o contorno ou silhueta.
	Movimento	Utilizam fluxo óptico ou volume histórico do movimento para representar a ação.
	Híbrido	Combinam modelos baseados em forma e movimento para representar a ação.

Fonte: Adaptado da pesquisa de Bux (2017).

Ainda existem outras abordagens baseadas em descritores visuais de textura, como o *Local Binary Pattern* (LBP)[4] e abordagens acompanhados de um classificador genérico, como *Fuzzy Logic-Based*[5].

[4] *Local Binary Pattern*. Descritor de padrão de textura local de uma imagem. Disponível em <https://www.sciencedirect.com/topics/engineering/local-binary-pattern>. Acesso em 24 de junho de 2022.

[5] *Fuzzy Logic-Based*. Abordagem da computação baseada em "graus de verdade"ao invés de lógica booleana usual (verdadeiro ou falso). Disponível em: <https://www.techtarget.com/searchenterpriseai/definition/fuzzy-logic>. Acesso em: 24 de junho de 2022.

Métodos baseados em *Feature Learning*

Diferente dos métodos *handcrafted*, os métodos baseados em *feature learning* conseguem aprender as características relevantes e necessárias e é realizada através de um algoritmo[6]. De acordo com Nanni, Ghidoni e Brahnam (2017):

> É possível considerar as camadas profundas de uma rede neural convolucional (CNN - *Convolutional Neural Network*) como um extrator de características, assim como o SIFT, a grande diferença é que as características extraídas por uma CNN são aprendidas usando os dados em contraste com as características *handcrafted* que são projetadas de antemão por especialistas humanos para extrair um determinado conjunto de características. Como as características extraídas pelos níveis inferiores de uma CNN dependem fortemente do conjunto de treinamento, cuidados especiais devem ser tomados na seleção do conjunto de dados. (NANNI; GHIDONI; BRAHNAM, 2017, p.5)

Conforme o autor, os métodos baseados em *feature learning*, podem ser categorizados em abordagens baseadas em *Deep Learning* (modelos generativos/não supervisionados e discriminativos/supervisionados) e abordagens não baseadas em *Deep Learning* (como programação genética e *dictionary learning*). Essa divisão é apresentada abaixo pela Figura 2.5:

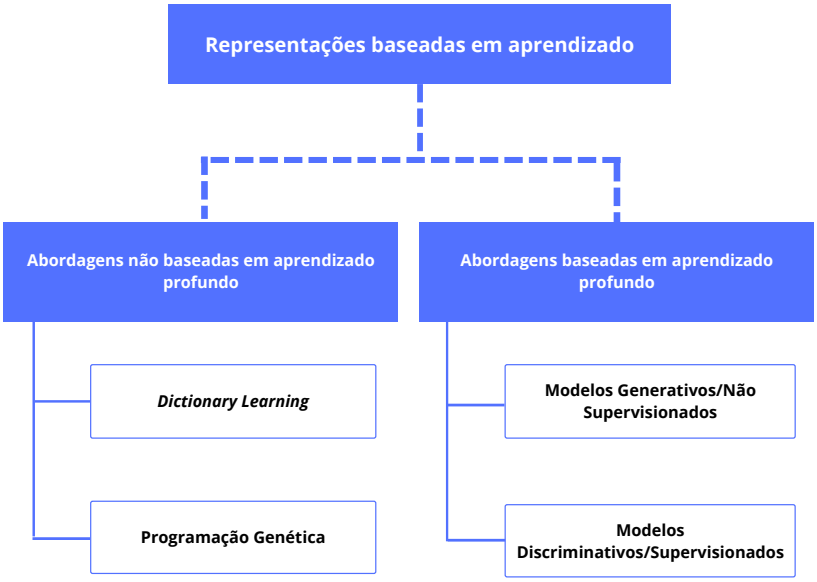

Figura 2.5: Abordagens de representação de ação baseadas em aprendizado. Fonte: Adaptado de (BUX, 2017, p.39).

[6]Máquinas que aprendem (MAp): Como representar uma tarefa de aprendizagem de máquina: *handcrafted features* vs *feature learning*. Disponível em: <https://maquinasqueaprendem.com/2020/03/24/como-representar-uma-tarefa-de-aprendizagem-de-maquina-handcrafted-features-vs-feature-learning/>. Acesso em: 22 de novembro de 2021.

Nas abordagens baseadas em *Deep Learning*, o principal objetivo dos modelos generativos é a geração de instâncias de saída com base na distribuição dos dados do conjunto de treinamento, considerando as características pertinentes a cada classe. Algumas das abordagens mais utilizadas são *Auto-Encoders* e *Generative Adversarial Networks* (GAN). Enquanto o objetivo dos modelos discriminativos é a classificação de dados de entrada com base na probabilidade das instâncias de pertencerem a cada categoria, com base no mapeamento de características por rótulo. As abordagens mais utilizadas são Redes Neurais Profundas (DNN - *Deep Neural Network*), Redes Neurais Convolucionais e Redes Neurais Recorrentes (RNN - *Recurrent Neural Network*). Também existem abordagens híbridas, que possibilitam o uso de um modelo discriminador e um modelo generativo.

Dentre as abordagens não baseadas em *Deep Learning*, a programação genética tem como princípio a busca em um espaço de soluções possíveis (sem nenhum conhecimento prévio) possibilitando a classificação de instâncias fundamentadas na descoberta de relações funcionais de características entre dados (BEDDIAR et al., 2020). Nos métodos com base em *dictionary learning*, o principal objetivo é buscar uma representação esparsa dos dados combinando elementos de maneira linear.

De acordo com Beddiar et al. (2020), os métodos baseados em aprendizado profundo:

> (...) permitem o reconhecimento de atividades de alto nível com estruturas complexas. No entanto, a alta complexidade computacional e os enormes requisitos de dados para a fase de treinamento ainda estão entre os problemas desafiadores sem solução. (BEDDIAR et al., 2020, p.12)

Estágios do reconhecimento (HAR)

Baseada nos conceitos apresentados por Beddiar et al. (2020), os sistemas de reconhecimento de atividades humanas são, em geral, compostos por três etapas principais e suas respectivas técnicas associadas, conforme apresentado a seguir:

1. Detecção: nesta etapa, os esforços são concentrados em determinar as partes do corpo a serem rastreadas ou reconhecidas. Algumas técnicas de detecção são descritas na Tabela 2.2.

2. Rastreamento: fornece uma conexão entre a sequência de imagens. Exemplos de técnicas de rastreamento são: baseadas em contornos (requerem um contraste entre o objeto rastreado

Tabela 2.2: Técnicas baseadas em Detecção. Fonte dos Dados: (BEDDIAR et al., 2020)

Técnica - Detecção	**Descrição**
Cor da Pele	Pode ser utilizada para detectar as partes do corpo, no entanto, pode apresentar problemas quando a cor da pele e dos objetos ou *background* são muito próximas.
Forma	A partir dos contornos das partes do corpo, no entanto, os objetos de *background* também podem obstruir as formas detectadas.
Valores de pixel	A aparência pode ser expressa em termos de mudança de valores de pixel entre imagens de uma sequência.
Modelos 3D	Tentam construir correspondências entre as características do modelo com base em várias características das imagens. Apresentam algumas limitações em termos de atributos de posicionamento precisas.
Movimento	O movimento pode ser detectado usando a diferença de brilho de *pixels* de duas imagens sucessivas
Difusão anisotrópica *	Pode ser usado para detectar e descrever padrões de atividade.

*Difusão anisotrópica. Técnica usada em filtros de detecção de bordas baseada na filtragem de ruídos de regiões internas e inibindo esse processo nos pixels de borda (VIEL; JR.; ZEFERINO, 2017).

e o *background* ou em características com base na correlação (requer a permanência da mesma vizinhança do objeto na sequência de imagens).

3. Classificação: consiste em interpretar a semântica da localização, da postura e da atividade. Exemplos de modelos de classificação são: Máquina de Vetores de Suporte (do inglês, *Support Vector Machine* - SVM), Classificador Bayesiano Naïve, Algoritmo de K-Vizinhos Mais Próximos, Redes Neurais, entre outros.

Nível de supervisionamento (HAR)

- Supervisionada: nesse caso os dados de treinamento são anotados com as respostas ou classes a serem previstas.

- Não Supervisionada: em alguns casos é extremamente custoso conseguir dados anotados, portanto, o ideal é que o algoritmo descubra sozinho as características discriminativas dos dados e agrupe-os baseado no perfil dessas características.

- Semi supervisionada: combinam as duas categorias de aprendizado acima, isto é, se beneficiam do poder discriminativo das abordagens supervisionadas e da capacidade das abordagens não supervisionados em localizar ações automaticamente (BEDDIAR et al., 2020).

Redes Neurais para Visão Computacional: conceitos básicos

As Redes Neurais são modelos computacionais inspirados na estrutura neural interconectada do cérebro humano para realizar aprendizado de máquina.

Existem diferentes arquiteturas de Redes Neurais, desde as mais utilizadas às outras arquiteturas especializadas para determinadas tarefas. As mais utilizadas são a *Deep Neural Network* (DNN), *Convolutional Neural Network* (CNN), *Long Short-Term Memory* (LSTM), *Auto-Encoder* (AE), *Generative Adversarial Network* (GAN), entre outras. Dentre essas, as CNNs se destacam pela capacidade de lidar com problemas relacionados com imagens, decorrente dos filtros contidos nas camadas de convolução, responsáveis por acentuar padrões locais em imagens, obtidos a partir da geração de diferentes mapas de atributos.

A convolução é a operação base que ocorre nas camadas convolucionais aplicada sob a matriz de *pixels* de uma imagem. Dada uma imagem de entrada, o valor de cada *pixel*, é recalculado com base em uma matriz de pesos (chamada filtro ou *kernel*) e a vizinhança de *pixels* locais desse determinado *pixel*, de forma que cada elemento do *kernel* e dessa matriz de vizinhança sejam multiplicados, para que posteriormente o resultado obtido seja somado e agrupado em um único *pixel* novamente. Esse processo é realizado a cada passo onde o *kernel* "desliza" abstratamente. Alguns parâmetros podem ser configurados para com o *kernel* selecionado, tais como: tamanho da matriz do *kernel*, *stride*[7] e *padding*[8].

Um exemplo da operação de convolução em um determinado *pixel*, pode ser visualizado na Figura 2.6 e uma visão geral de como ocorre o "deslizamento"do passo do *kernel* durante a convolução de uma imagem pode ser visualizada na Figura 2.7. Em uma CNN, existem diferentes tipos de camadas e cada um tem uma responsabilidade específica. Abaixo estão descritas alguns componentes da uma CNN:

- Camadas de Convolução: utilizadas para reduzir o tamanho espacial das características de entrada, onde os dados passam por vários filtros que "deslizam"[9] sobre a imagem, com o objetivo de identificar atributos locais relevantes nas imagens a medida que se altera o tamanho do passo do *kernel*.

[7] *Stride*. Representa o tamanho do passo de uma abstração de "deslizamento" do *kernel*, isto é, o tamanho do deslocamento do *kernel* com base na quantidade de *pixels*.

[8] *Padding*. Essa técnica de preenchimento do quadro de pixels da imagem final pode ser utilizada para manter o tamanho original da imagem, por exemplo.

[9] Janela deslizante (do inglês, *Sliding Window*). Técnica utilizada durante a operação de convolução entre uma "janela"(parte) de imagem de origem e um determinado *kernel* (HUANG et al., 2011).

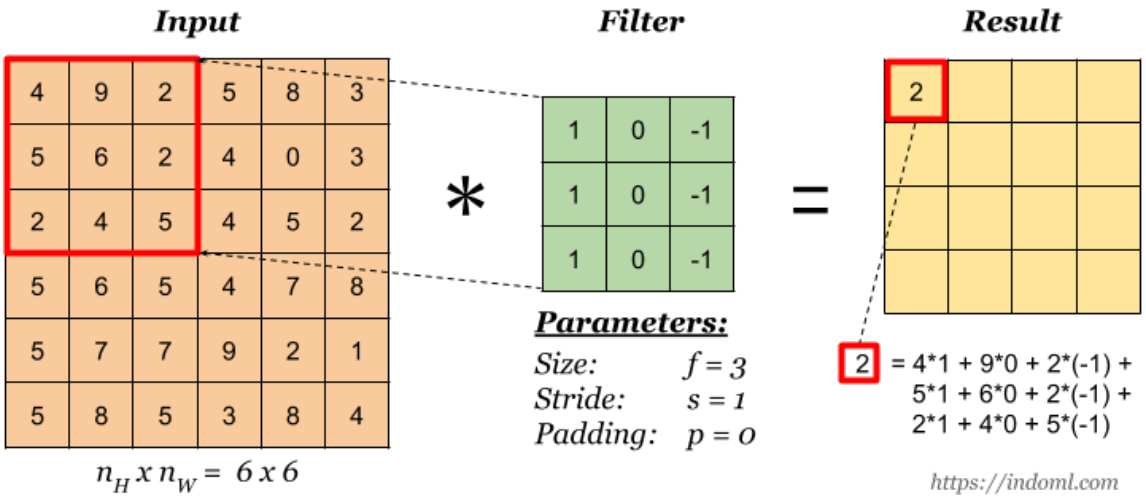

Figura 2.6: Operação de convolução. Fonte: (CHOULWAR, 2019)

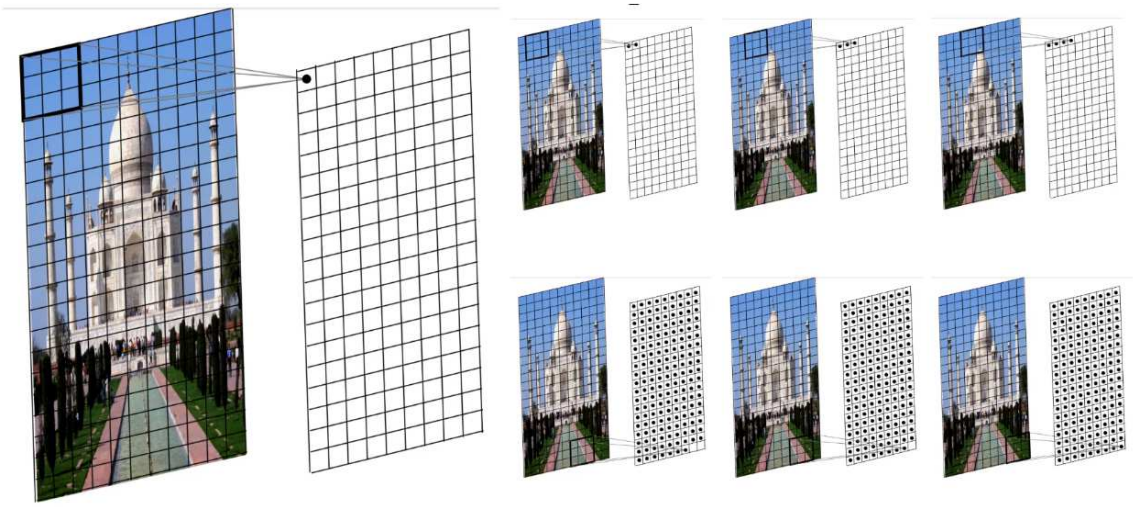

Figura 2.7: Representação de uma convolução em uma imagem. Fonte: (KHURANA, 2020)

- Função de Ativação: responsável por calcular um valor de saída para cada neurônio e a partir desse valor decidir ativá-lo ou não, esse valor é calculado através de uma transformação não linear nos dados, fazendo com que as redes neurais possam aprender além de relações lineares entre os dados e possam executar tarefas mais complexas.

- Função de Perda: também conhecida como função de custo, é usada para calcular a diferença entre o valor predito pelo modelo e o valor real esperado para o dado de entrada.

- Camada de Abstração (agrupamento): também utilizada para reduzir o dimensionamento das características, existem vários tipos de agrupamento, tais como *MaxPooling*, *AvgPooling*, entre outros;

- Camada Totalmente Conectada: normalmente, recebe como entrada as características extraídas pelas camadas de convolução e produz a decisão final de classificação com base no agrupamento dessas características.

- Otimizador: usado durante o treinamento do modelo para minimizar a função de perda e obter parâmetros de rede ótimos dentro de um tempo aceitável (LI et al., 2021).

Durante o treinamento de uma CNN, os hiper-parâmetros são aprendidos e ajustados nas camadas convolucionais e camadas totalmente conectadas. Conforme Li et al. (2021), os hiper-parâmetros a seguir podem interferir diretamente no desempenho do modelo: taxa de aprendizagem[10], número de épocas[11] de treinamento, tamanho do mini-lote[12], número de camadas e *kernels*[13] de convolução, tamanho do *kernel* de convolução.

2.2.2 Desafios e Limitações

De acordo com Beddiar et al. (2020), em sistemas de reconhecimento baseados na visão em geral, alguns métodos de detecção de atividades humanas com base na forma ou na aparência, como a segmentação colorimétrica, podem confundir as partes do corpo humano com os objetos da cena, ou não podem operar corretamente durante a variação na aparência ou no vestuário de pessoas.

A maioria das câmeras de videovigilância possui uma qualidade muito baixa em comparação às câmeras de outros dispositivos, como, por exemplo, câmeras de *smartphones*. Ainda conforme Beddiar et al. (2020), a variação da iluminação afeta a qualidade dos quadros de vídeos e pode dificultar o funcionamento de sistemas de reconhecimento, assim como, o processo de extração de características importantes para diferenciar as atividades humanas, como em cenas com ações humanas sutis.

Outros fatores que afetam a qualidade da imagem e, consequentemente, as informações contidas nela, são: a variação da iluminação, a variação de *background* (seja pela complexidade ou por ser móvel), variação de escala (distância do ator até o dispositivo), ruído e

[10]Taxa de aprendizagem. É um parâmetro que pode ser constante ou variável e determina o quanto os pesos da rede serão ajustados em relação à descida do gradiente. Se a taxa de aprendizagem for muito pequena, a velocidade para a convergência será lenta, se a taxa de aprendizagem for muito grande, os parâmetros oscilarão entre um ponto anterior e posterior da convergência.

[11]Época. Número de vezes que todo o conjunto de dados é submetido ao modelo.

[12]Mini-lote. Divisão dos dados de treinamento em uma parcela menor.

[13]*Kernel.* Uma matriz de pesos usada durante a convolução de uma imagem.

oclusão (seja de partes do corpo ou objetos, muitas vezes ocorridas também pelas próprias partes do corpo se obstruindo, ou de objetos ocluindo partes do corpo).

Além dos desafios relacionados com as características visuais dos elementos do vídeo, outro fator que torna complexo o processo de identificação de comportamentos violentos, criminosos ou agressivos, é a identificação do limiar onde um determinado comportamento humano deixa de ser natural ou normal e se torna criminoso. Devido à variedade de formas existentes de realizar uma mesma atividade ou gesto, como também, o fato de um mesmo comportamento realizado sob circunstâncias ou contextos diferentes, podem ser ocasionadas mudanças completamente diferentes do real significado das ações. Conforme Beddiar et al. (2020):

> A independência de gestos e a detecção de gestos a partir de fluxos de dados contínuos também constituem outro tipo de desafio, uma vez que ainda é difícil localizar temporalmente o gesto em vídeos longos e contínuos. Na verdade, os sistemas HAR ainda não são capazes de detectar e reconhecer vários gestos em diferentes condições de fundo e não são tolerantes com a escalabilidade e o crescimento dos gestos. (BEDDIAR et al., 2020, p.30)

Portanto, o reconhecimento de atividades humanas em partes específicas da cena ou de cenas específicas do vídeo, a previsão de atividades, o reconhecimento de mais de uma atividade realizada em simultâneo (principalmente em cenas com muitas pessoas), e a discriminação entre ações intencionais e involuntárias ainda são áreas muito desafiadoras (BEDDIAR et al., 2020).

2.3 Videovigilância

2.3.1 Análise de Atividades Violentas

Dada a importância do estudo da análise de comportamentos humanos, a segurança, geralmente, visa a proteção da liberdade social e recorre a artifícios para combater quaisquer manifestações que tentem limitá-la. Dentre esses artifícios, a vigilância, segundo a Comissão Europeia durante os trabalhos para estabelecer o Sistema Europeu de Vigilância de Fronteiras (*European Border Surveillance System* - EUROSUR 2011)[14], regulamentado

[14] *EUROSUR: Providing authorities with tools needed to reinforce management of external borders and fight cross-border crime*. Disponível em: <https://ec.europa.eu/commission/presscorner/detail/en/MEMO_11_896>. Acesso em: 10 de janeiro de 2020.

por EUROSUR (2011)[15], possui o objetivo de coletar informações sobre características e comportamentos em uma determinada área. Ainda baseado na definição fornecida por EUROSUR (2011), no contexto da aplicação de ferramentas de monitoramento nas operações de vigilância de fronteiras, o processo de análise de pessoas em segurança normalmente compreende algumas tarefas ligadas aos seguintes conceitos:

- Detecção: está relacionada à capacidade de distinção entre a presença de um objeto em uma imagem e o seu próprio *background* (ex.: a delimitação da região geométrica correspondente a um determinado objeto em uma imagem);
- Classificação: é a capacidade de determinar o tipo ou classe do objeto, conforme suas características (ex.: a rotulagem do objeto detectado na classe 'pessoa');
- Identificação: é a capacidade de atribuir informações discriminativas ao objeto, sem conhecimento prévio (ex.: atribuir um código unificado, características gênero e/ou idade ao objeto detectado);
- Reconhecimento: é a capacidade de determinar que o objeto detectado é exclusivo e específico, dado conhecimento prévio (ex.: a constatação de ocorrências do objeto detectado em uma base de dados, dada uma instância desse objeto para consulta);
- Verificação: objetiva confirmar a presença de um objeto específico detectado, dado um conhecimento prévio (ex.: a confirmação ou validação da identidade do objeto detectado).

Consequentemente, devido à vasta quantidade de câmeras de vigilância, a tarefa de supervisão humana sob os monitores de sequências de imagens capturadas, tem se tornado cada vez mais complexa e desafiadora, visto que circunstâncias anormais não ocorrem com tanta frequência, quando comparadas às atividades normais. Portanto, cresce a demanda por sistemas inteligentes capazes de detectar automaticamente eventos anômalos.

[15]Documento publicado pela Comissão Europeia que fornece uma descrição conceitual de alto nível de possíveis serviços para as aplicações comuns de ferramentas de vigilância a nível da União Europeia em apoio à vigilância de fronteiras. O Documento é intitulado como: *Application of surveillance tools to border surveillance 'concept of operations'*. Disponível em: <https://ec.europa.eu/research/participants/portal/doc/call/fp7/fp7-space-2012-1/31341-2011_concept_of_operations_for_the_common_application_of_surveillance_tools_in_the_context_of_eurosur_en.pdf>. Acesso em: 10 de janeiro de 2020.

Conforme Hussain, Sheng e Zhang (2019), a maneira mais básica e tradicional de reconhecimento de atividades é instalar câmeras de vigilância nas instalações e monitorar as atividades humanas. O monitoramento pode ser realizado por humanos (pessoa que assiste os vídeos e imagens das câmeras) ou por processo automático.

A análise de atividades humanas em videovigilância inclui aspectos promissores através do uso da tecnologia aliada a equipamentos de segurança para auxiliar o monitoramento humano. Dessa forma, a videovigilância é um componente essencial para quaisquer sistemas de defesa e prevenção de ocorrências que ameacem a segurança humana e possui como principal propósito, a identificação de possíveis alvos de interesse a partir de vídeos capturados de câmeras de segurança, tanto para alvos já conhecidos, quanto pela sua detecção a partir da análise de comportamentos violentos.

2.3.2 Fundamentos de um Sistema de Videovigilância

Nesta seção, são discutidas as principais funcionalidades pertencentes a um sistema de videovigilância, as quais compreendem tarefas interligadas aos conceitos supracitados e conforme a finalidade específica para o qual foi construído.

Processos contidos em um sistema de videovigilância

Segundo Mabrouk e Zagrouba (2018), na temática de reconhecimento de padrões em sistemas de videovigilância inteligentes há duas principais etapas, a representação e a modelagem do comportamento. Essas etapas são apresentadas na Figura 2.8.

A partir das definições do referido autor, o nível correspondente à representação do comportamento permite, geralmente, detectar e descrever o objeto em movimento na cena, em termos de processamento de baixo nível. Em sua primeira etapa, a extração de características, onde a região da cena é detectada com base em características relevantes de baixo nível, sejam características locais ou globais. As características locais podem representar com precisão o movimento local em um vídeo, no entanto, podem não produzir informações ao ter um índice alto de movimento. As características globais fornecem informações holísticas sobre uma determinada cena, porém podem fornecer informações irrelevantes em casos de *background* que apresentam ruídos ou desordenação.

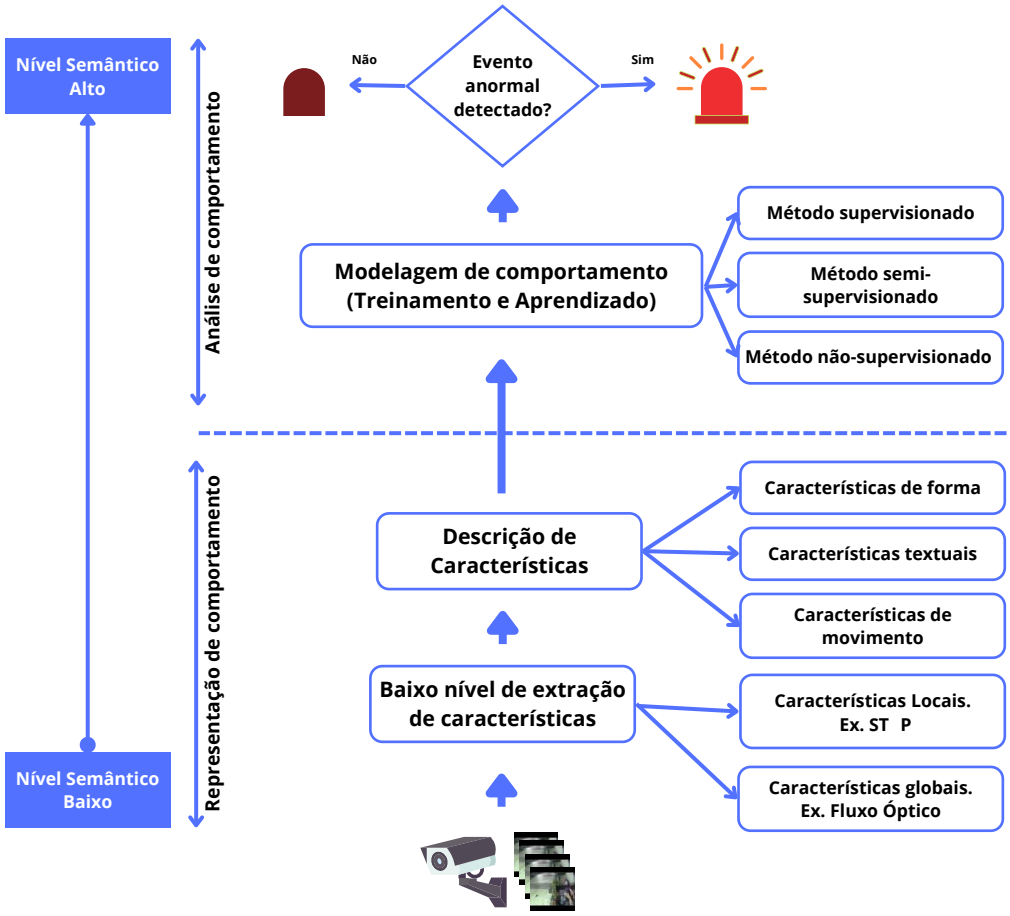

Figura 2.8: Funcionamento de um sistema inteligente de videovigilância. Fonte: Adaptado de (MABROUK; ZAGROUBA, 2018, p.4)

Ainda na etapa de representação de comportamento, cabe ao processo de descrição de características fornecer uma representação acerca da região de interesse da cena, com base nas respectivas características extraídas no processo anterior. Para essa representação do objeto alvo da cena, diferentes características podem ser extraídas, como o fluxo óptico, pontos de interesse, volume espaço-temporal, forma, textura, rastreamento de objeto e extração da trajetória. Em um nível mais alto de semântica, a etapa de modelagem de comportamento está voltada à capacidade de interpretação do tipo de ação realizada pelo objeto, assim como identificar se esse comportamento é normal ou não. Segundo Mabrouk e Zagrouba (2018):

> Escolher características que sejam robustas a transformações de cena (rotação, oclusão, fundos desordenados, etc.) e menos sensíveis às mudanças na aparência do objeto é essencial para capturar informações relevantes e discriminativas sobre o comportamento do objeto em movimento (MABROUK; ZAGROUBA, 2018, p.35).

O critério de escolha do método de classificação de comportamento depende dos tipos de amostras necessárias ao processo de aprendizagem, dentre eles estão: 1) os métodos supervisionados, os quais permitem modelar comportamentos a partir de amostras previamente rotuladas na fase de treinamento; 2) os métodos semi-supervisionados, sejam baseados em regras ou em modelos, isto é, determinar regras que definem normalidade a partir de padrões de amostras de comportamentos considerados normais ou através da construção de modelos que representam comportamentos comuns (modelos *Markov Random Field* (MRF), *Gaussian Mixture Model* (GMM) e *Hidden Markov* (HMM)), respectivamente. Por fim, 3) há os métodos não supervisionados, que possuem como principal característica a tarefa de aprender padrões de comportamento normais e anormais a partir de propriedades estatísticas extraídas de amostras não rotuladas.

De maneira geral, a análise de atividades humanas por computador relacionada à classificação de ação e comportamento, também ligadas ao reconhecimento de postura (estática), ação (curto prazo) ou comportamento global (longo prazo), exige uma gama de peculiaridades relacionadas tanto ao ambiente em que se é realizada a cena, quanto ao conteúdo presente nessa. Esse conjunto de peculiaridades, quando combinado com tarefas/procedimentos que requerem robustez ao local da cena (*indoor* ou *outdoor*), ao posicionamento da(s) câmera(s), à aparência do(s) objeto(s) detectado(s), às propriedades de resolução e atributos de cor das imagens obtidas, entre outros, dão a essa atividade um nível maior de complexidade.

Densidade da Cena e Interação

Fatores relevantes a serem considerados durante a modelagem e análise de comportamento são os aspectos relacionados à densidade da cena e à interação do objeto em movimento. A densidade da cena diz respeito às características visuais da cena relacionadas ao número de pessoas presentes, sendo essas classificadas como cenas lotadas ou sem aglomeração (com um único ator ou um pequeno número de atores).

Em cenas não aglomeradas e sem interação são considerados três importantes comportamentos anormais, (i) a detecção de queda humana, (ii) o ato de permanecer por longos períodos vagando em espaços públicos sem objetivo e (iii) a presença de seres humanos em locais não permitidos, respectivamente apresentadas na Figura 2.9. Em cenas não aglomeradas com interação, existe a facilidade de se analisar comportamentos agressivos, como brigas,

chutes, socos, etc. entre pessoas, o que é uma tarefa bastante relevante no contexto de detecção de comportamentos agressivos em videovigilância, exemplos desse tipo de cena podem ser visualizados na Figura 2.10.

Figura 2.9: Eventos anormais realizados por uma única pessoa. Fonte: (MABROUK; ZAGROUBA, 2018, p.22)

Figura 2.10: Exemplos de violência em cenas não lotadas. Fonte: (MABROUK; ZAGROUBA, 2018, p.23)

Em cenas com interação realizada em grupo e cenas aglomeradas, existe a dificuldade de se rastrear e analisar comportamentos humanos individualmente, devido à oclusão gerada pelo grande número de elementos presentes em uma única cena e apenas uma pequena região corresponde à representatividade do objeto alvo. Portanto, nesses casos, é mais adequado analisar o comportamento da multidão de forma holística. Exemplos de cenas de eventos anormais podem ser visualizados na Figura 2.11.

Figura 2.11: Eventos anormais em cenas lotadas. Fonte: (MABROUK; ZAGROUBA, 2018, p.24)

As **categorias de interação de objetos em movimento** podem ser classificadas em: 1) não possuir interação; 2) interação realizada entre dois indivíduos; 3) e interação realizada em grupo.

Reidentificação

A tarefa de reidentificação abordada por Vezzani, Baltieri e Cucchiara (2013) é uma ferramenta útil para a análise de pessoas em sistemas de videovigilância, devido à capacidade de se atribuir o mesmo identificador a diferentes instâncias do mesmo objeto, de forma independente da câmera ou do ponto de vista. Segundo o autor supracitado, a maioria das metodologias de reidentificação são ligadas às abordagens de rastreamento de pessoas e de reconhecimento biométrico, que têm como principal finalidade manter uma representação precisa do estado do objeto e buscar a identidade exata desse, respectivamente. As principais diferenças entre essas abordagens são dadas a partir de limitações/variações de características de continuidade espaço-temporal, dado um conjunto de quadros consecutivos de uma sequência de vídeo em uma alta taxa de quadros por segundo (fps) (ver em Tabela 2.3). Estas características são: a posição, o ponto de vista, a aparência e o perfil biométrico.

Tabela 2.3: Restrições de continuidade impostas pelas abordagens.

	Rastreamento de Pessoas	**Reidentificação de Pessoas**	**Reconhecimento Biométrico**
Continuidade de...			
Posição	x		
Ponto de Vista	x		
Aparência	x	x	
Perfil Biométrico	x	x	x

Fonte: Adaptado de Vezzani, Baltieri e Cucchiara (2013)

Ainda na pesquisa realizada por Vezzani, Baltieri e Cucchiara (2013), é proposta uma taxonomia sobre os aspectos computacionais e o *design* de aplicações baseados em abordagens ligadas à tarefa de reidentificação. Suas principais características são:

1. Configuração da(s) câmera(s), isto é, se a coleta de imagens é realizada através de uma mesma câmera ou de câmeras diferentes, ou ainda se estão tanto sobrepostas, quanto calibradas);

2. Cardinalidade do conjunto de amostras, compreendendo um único alvo ou múltiplos;

3. Assinatura, ou seja, o conjunto de características associadas à forma, textura, cor, posição, etc.

4. Adoção de um modelo corporal, seja esse um mapeamento local ou global ao nível espacial 2D ou 3D;

5. Exploração de técnicas de *machine learning* ao nível de imagem, assinatura ou *matching*[16];

6. Cenário do aplicativo para recuperação de imagem ou para rastreamento a curto/longo prazo.

A escolha de determinados fatores de aspectos e designs descritos acima pode influenciar nos demais, como, por exemplo, a escolha da configuração da câmera afeta diretamente a escolha da assinatura. Desse modo, a capacidade de detectar e reconhecer atividades humanas utilizando visão computacional depende intrinsecamente de que esses fatores sejam correlacionados de maneira sistemática e coerente.

2.4 Considerações

A maioria das câmeras de videovigilância possuem uma qualidade muito baixa em comparação às câmeras de outros dispositivos, como, por exemplo, câmeras de *smartphones*. A baixa resolução de vídeos pode dificultar o processo de extração de características importantes para diferenciar os tipos de atividades humanas, como cenas com ações humanas sutis.

Outros fatores que afetam a qualidade da imagem e consequentemente, as informações contidas nela, são: a variação da iluminação, a variação de *background* (seja pela complexidade ou por ser móvel), variação de escala (distância do ator até o dispositivo), ruído e oclusão (seja de partes do corpo ou objetos, muitas vezes ocorridas também pelas próprias partes do corpo se obstruindo, ou de objetos ocluindo partes do corpo). Outros desafios ainda são mencionados por Beddiar et al. (2020), como:

> Hoje em dia, a restrição de memória, alto número de atualização de parâmetros, coleta e fusão de grandes dados variantes multimodais para o processo de treinamento, bem como a implantação de diferentes arquiteturas de métodos baseados em aprendizagem profunda em *smartphones* ou dispositivos vestíveis ainda são

[16] *Matching*. Técnica usada para realizar a correspondência de dados, como entre duas imagens, por exemplo.

> questões não resolvidas em sistemas HAR *deep learning* (BEDDIAR et al., 2020, p.31).

Além dos desafios relacionados com as características visuais dos elementos do vídeo, outro fator que torna complexo o processo de identificação de comportamentos violentos, criminosos ou agressivos, é a identificação do limiar em que um determinado comportamento humano deixa de ser natural ou normal e se torna violento. Devido à variedade de formas existentes de realizar uma mesma atividade ou gesto, como também, o fato de que um mesmo comportamento pode ser realizado sob circunstâncias ou contextos diferentes, um mesmo comportamento pode ter significados completamente diferentes. Portanto, a discriminação entre ações intencionais e involuntárias ainda é uma área muito desafiadora de abordar (BEDDIAR et al., 2020).

Capítulo 3

Pesquisas Relacionadas

Neste Capítulo, são apresentados os métodos de detecção de comportamentos violentos e a síntese das principais abordagens utilizadas em pesquisas do estado da arte. Nas Seções 3.1 e 3.2, são apresentados os métodos de busca dos trabalhos e evidenciadas as principais caraterísticas das pesquisas relacionadas, respectivamente. Na Subseção 3.2.1, são sintetizados os principais tipos de dados utilizados durante o processo de extração de características em vídeos. Na Subseção 3.2.2, são sintetizados os conjuntos de dados utilizados.

3.1 Levantamento do Estado da Arte

A organização das atividades para a construção da base teórica para a elaboração desta pesquisa, é composta por três etapas inter-relacionadas. Ambas foram realizadas através da revisão de publicações de artigos extraídos a partir dos repositórios *Scientific Electronic Library* Online (SciELO)[1], Google Scholar[2], *Association for Computing Machinery Digital Library* (ACM)[3], *Institute of Electrical and Electronics Engineers* (IEEE)[4], Springer[5] e arXiv®[6].

Considerando a temática deste trabalho, os termos de busca utilizados em sua maior parte foram combinações entre os termos: **"violence detection"**, "surveillance cameras",

[1]SciELO. <https://www.scielo.org/>. Acesso em: 02 de julho de 2022.
[2]Google Scholar. <https://scholar.google.com.br/>. Acesso em: 02 de julho de 2022.
[3]ACM. <https://dl.acm.org/>. Acesso em: 02 de julho de 2022.
[4]IEEE. <https://www.ieee.org/>. Acesso em: 02 de julho de 2022.
[5]Springer. <https://www.springer.com/> Acesso em: 02 de julho de 2022.
[6]arXiv®. <https://arxiv.org/>. Acesso em: 02 de julho de 2022.

"videos", "surveillance videos", "violence recognition", "anomaly detection", "violence pattern", "activity recognition" e "violent activity detection". No entanto, também foram coletados trabalhos além do motor de busca, como no caso de alguns trabalhos contidos no referencial bibliográfico de outros trabalhos selecionados.

Para a apresentação dos resultados da pesquisa bibliográfica foi adotada a abordagem qualitativa, realizada através de percepções e análises dos trabalhos relacionados, com o objetivo de identificar e entender os problemas e dificuldades enfrentadas na área e delimitar a temática do presente trabalho. Os critérios de seleção dos respectivos trabalhos ocorreram em função das palavras-chave presentes nos títulos ou *abstract*, ano de publicação (entre 2014 e 2021), leitura do *abstract* e a leitura integral do trabalho.

3.2 Detecção e Reconhecimento de Atividades Violentas

Conforme Pareek e Thakkar (2021), sistemas de videovigilância visual dependem da detecção de eventos anômalos e a videovigilância pode ser usada pelas organizações para prevenir crimes ou para inspecionar a cena do crime. Existem diferentes métodos para a detecção de eventos anômalos, desde métodos tradicionais para a extração de características até métodos baseados em *deep learning*.

A principal diferença entre esses métodos é a abordagem de extração-aprendizado de características. Os métodos tradicionais contam com uma etapa conhecida como *hand-crafted*, onde as características são extraídas separadamente por meio de algoritmos de *machine learning* e na maioria das vezes, as características podem ser extraídas e estruturados antes da fase de aprendizagem. Já em métodos baseados em *deep learning*, as características são extraídas e selecionadas durante a etapa de aprendizagem do modelo neural, onde geralmente, são submetidos dados não estruturadas (quadros de vídeos, por exemplo) como entrada para o modelo. Existem também, abordagens híbridas, onde existem etapas de extração de um ou vários tipos de características (RGB, fluxo óptico, pontos-chave do corpo humano, entre outros) através de algoritmos de aprendizado de máquina e posteriormente, dados como entrada em modelos de aprendizado profundo.

Métodos Tradicionais

Na pesquisa desenvolvida por Deniz et al. (2014), é proposto um método que tem como principal característica e contribuição, a utilização de padrões de aceleração extrema a partir de trajetórias de pontos rastreados para a detecção de ações específicas. Uma aceleração extrema ocorre quando acontece algum movimento desfocado entre quadros consecutivos, o que consequentemente implica em uma mudança no conteúdo da imagem. Para esse fim, foi utilizada a função de *Hanning*[7] e a *Fast Fourier Transform*[8] com o objetivo de detectar uma elipse na representação da imagem do espectro de potência do segundo quadro (*transformada de Radon*[9]), usada para indicar a ocorrência de um movimento repentino. Por fim, para extrair as características de aceleração, desaceleração e potência são calculados histogramas.

Na pesquisa desenvolvida por Bilinski e Bremond (2016), para a detecção de violência em vídeos é proposta uma extensão dos *Improved Fisher Vectors*[10] (IFV). Para isso, são extraídas características espaço-temporais locais dos vídeos através das *Improved Dense Trajectories* (IDT)[11]. Segundo os autores, após os pontos de interesse serem rastreados e extraídos para um campo de fluxo óptico, os volumes de vídeo espaço-temporais são extraídos em torno das trajetórias detectadas e representados por: *Histogram of Oriented Gradients* (HOG), que contém informações de aparência; *Trajectory Shape* (TS); *Histogram of Optical Flow* (HOF) e *Motion Boundary Histogram* (MBH), descritores de informação de movimento. Após a fase de representação do vídeo, o reconhecimento de violência é realizado pelo classificador linear *Support Vector Machines* (SVMs). Por fim, a identificação dos intervalos de sequência de vídeo em que ocorrem violência foi baseada na abordagem *Sliding Window* e avaliada em diferentes locais e escalas.

Na pesquisa desenvolvida por Xu et al. (2014), é empregado o algoritmo *Motion* SIFT

[7] Função de *Hanning*. Função utilizada para atenuar distorções em processamento de sinal digital.

[8] *Fast Fourier Transform* (FFT). Algoritmo para se calcular a Transformada Discreta de Fourier. "Converte um sinal em componentes espectrais individuais e assim fornece informações de frequência sobre o sinal". Disponível em: <https://www.nti-audio.com/pt/suporte/saber-como/transformacao-rapida-de-fourier-fft>. Acesso em: 04 de janeiro de 2020.

[9] Transformada de Radon. Utilizada na análise de projeções de objetos sobre linhas retas, fornece embasamento matemático para, por exemplo, tomografia computadorizada.

[10] IFV. Estratégia de codificação de vídeo constituída por descritores formados pelo agrupamento de características locais em uma representação global.

[11] *Improved Dense Trajectories* (IDT). Baseado na pesquisa de Wang e Schmid (2013) e Bilinski e Bremond (2016), trata-se de uma estratégia de representação de vídeo, em que o rastreamento dos pontos de interesse são extraídos utilizando um campo de fluxo óptico denso a partir de uma amostragem densa.

(MoSIFT)[12] para o processo de extração de características de baixo nível. Em seguida, é utilizado o método de seleção de características baseado em *Kernel Density Estimation* (KDE) para inferir a função de densidade de probabilidade adjacente (PDF) objetivando eliminar o ruído selecionando as características mais representativas. Na etapa de processamento das características extraídas é adotado o método de codificação esparsa[13] emparelhado com o procedimento de *max pooling* para gerar uma representação de vídeo de alto nível a partir de características locais. Para a geração dos resultados experimentais, foi adotado o *Support Vector Machine* (SVM) como classificador e conjuntos de dados utilizados que compreendem cenas com densidade lotada e não lotada.

Já na pesquisa desenvolvida por Ding et al. (2014), é proposto um modelo 3D *ConvNets* para detecção de violência em vídeo, aplicando convolução além das camadas 2D, nas camadas temporais para se obter informações espaciais. Durante a etapa de aprendizagem sem conhecimento prévio, a rede opera diretamente nos *pixels* da entrada, utilizando os descritores *Space-Time Interest Point* (STIP) e MoSIFT. Os experimentos foram realizados em conjunto de dados formado por vídeos com cenas de luta e de não ocorrência de luta de jogos de Hockey.

Considerações Sobre os Métodos Tradicionais

Para a extração de características dos vídeos proposta por Nievas et al. (2011), utilizaram descritores de características espaço-temporais STIP e MoSIFT. Enquanto, a proposta desenvolvida por Xu et al. (2014) além do MoSIFT, foi utilizado o método *Kernel Density Estimation* (KDE) para a seleção das características e para inferir a Função de Densidade de Probabilidade (do inglês, *Probability Density Function* - PDF) adjacente e posteriormente, o processamento das características através de codificação esparsa. No entanto, o método MoSIFT para a extração de características requer um longo tempo de processamento, conforme mencionado na pesquisa de Chu e Tanaka (2011) e Deb, Arman e Firoze (2018).

Para o processo de descrição de características, na pesquisa desenvolvida por Nievas et al. (2011) é utilizada a abordagem tradicional baseada em *Bag-of-Words*[14]. A princípio, essa

[12]MoSIFT. Descritor responsável por extrair as características da imagem via SIFT padrão e posteriormente, gerar um histograma analógico de fluxos ópticos, capturando assim padrões locais distintos de forma e movimento de uma atividade.

[13]Codificação Esparsa. Método usado para encontrar automaticamente a representação dos dados sem perder nenhuma informação representativa desses. Disponível em: <https://blog.metaflow.fr/sparse-coding-a-simple-exploration-152a3c900a7c>. Acesso em: 02 de julho de 2022.

[14]*Bag-of-Words* (BoW). É um modelo de linguagem estatística usado para analisar texto e documentos com

abordagem não considera a relação semântica entre as palavras. No entanto, um comportamento pode possuir ou não violência dependendo da semântica em que se é analisada (ex.: movimentos de uma luta esportiva e uma briga entre os adversários dessa luta, acontecem no mesmo cenário, mas as características semânticas podem diferir, isto é, a diferença entre um comportamento decorrente de violência intencional e um comportamento agressivo esportivo). No entanto, Bilinski e Bremond (2016), demonstram contornar em partes essa dificuldade e superar a característica padrão de *Bag-of-Words*, ao utilizar a estratégia de codificação de vídeo através dos Vetores Aprimorados de Fisher (do inglês, *Improved Fisher Vector* - IFV) obtidos através da associação de características locais em uma representação global (holística).

Durante a fase de classificação de comportamentos, os autores Nievas et al. (2011), Deniz et al. (2014), Bilinski e Bremond (2016) e Xu et al. (2014) utilizaram o classificador discriminativo *Support Vector Machine* (SVM). Além disso, Deniz et al. (2014) apresentaram experimentos realizados com o classificador Adaboost. No entanto, existem alguns desafios encontrados na utilização de SVM, como, por exemplo, limitações de velocidade e tamanho (no treinamento e no teste) e, a representação discreta de dados. Mais especificamente, essas limitações estão relacionadas à necessidade de um tempo computacional considerável para prever uma tarefa quando o número de classes aumenta, como também à demora para o processo de treinamento quando aplicado à classificação de dados em várias classes, conforme relatado por Shah et al. (2017) e Devi, Kumar e Shankar (2019).

Portanto, os trabalhos supramencionados possuem como objetivo a detecção de comportamentos violentos em vídeos. Os mesmos apresentam resultados satisfatórios quando comparados aos demais trabalhos da literatura. A maior parte das limitações desses são correspondentes ao custo computacional dos métodos utilizados, assim como à complexidade existente de delimitação da fronteira de comportamento violento e não violento, o que ainda é um aspecto bastante desafiador na temática abordada.

base na contagem de palavras. O modelo não leva em conta a ordem das palavras em um documento. O BoW pode ser implementado como um dicionário Python com cada chave definida para uma palavra e cada valor definido para o número de vezes que a palavra aparece em um texto. Disponível em: <https://www.codecademy.com/learn/dscp-natural-language-processing/modules/dscp-bag-of-words/cheatsheet>. Acesso em: 11 de agosto de 2022.

Métodos baseados em Deep Learning e Híbridos

Características baseadas nos canais RGB e Fluxo Óptico

Redes Neurais Convolucionais (*Convolutional Neural Network* - CNN ou ConvNet) são uma categoria de arquitetura de Aprendizado Profundo (*Deep Learning*). Normalmente, é aplicada no contexto de processamento e análise de imagens digitais, visando a extração e mapeamento de características e sumarização desse mapeamento para realizar uma determinada tarefa. Na pesquisa desenvolvida por Ding et al. (2014), é apresentado um modelo 3D CNN (*ConvNets*), que possui a operação de convolução também aplicada nas camadas temporais para se obter as informações de características espaciais. Uma CNN possui a capacidade de fornecer um descritor de alta qualidade do conteúdo visual de uma imagem (GAILLARD; EGYED-ZSIGMOND; GRANITZER., 2018, p.1).

Já na pesquisa desenvolvida por Cheng, Cai e Li (2021) é posposto o banco de dados RWF-2000 (*Real-World Fighting*), com cenas de mundo real capturadas a partir de câmeras de videovigilância, o qual encontra-se disponível para download, mediante solicitação aos autores. No *pipeline* de coleta de dados, os vídeos são coletados do YouTube com base em palavras-chave previamente definidas e relacionadas à violência, posteriormente, são cortados em 5 segundos a 30 fps e submetidos a processamentos de remoção de clips irrelevantes. Cada amostra de vídeo é anotada como violento ou não violento.

Fluxo óptico é um método de detecção e estimativa do movimento de intensidades da imagem, a partir de uma sequência de quadros de vídeo, podendo ser usado para segmentar regiões no plano da imagem, que podem estar associadas a objetos em movimento. A partir da transição de quadros consecutivos é possível calcular também a direção e amplitude do movimento. Ainda na pesquisa desenvolvida por Cheng, Cai e Li (2021), é apresentado um método de detecção de violência em vídeos de vigilância, onde um mecanismo de *pooling* exclusivo é empregado por meio do fluxo óptico, pois na falta deste, as informações de movimento poderiam ser inúteis por conta do mecanismo de *coarse pooling*. O mecanismo proposto é uma estratégia de *pooling* autoaprendida, que, com base nesse ramo de fluxo óptico, determina o que o modelo deve preservar ou eliminar. Nesse trabalho, é utilizada uma estratégia de seleção de quadros através da estratégia de amostragem uniformizada, isto é, uma redução da quantidade de quadros de um vídeo de forma sistemática. A variação experimental desse trabalho também é constituída por mudanças arquiteturais de rede neural

que podem utilizar diferentes tipos de características (RGB, Fluxo óptico). Os resultados foram avaliados nas bases RWF2000, *Hockey Fight* (NIEVAS et al., 2011), *Movies* (NIEVAS et al., 2011) e Violent Flows (HASSNER; ITCHER; KLIPER-GROSS, 2012).

O foco do trabalho de Aktı, Tataroğlu e Ekenel (2019) é a detecção de atividades que envolvam luta, isto é, um evento que envolve duas ou mais pessoas, as quais lutam em um grau que deve ser interferido. A abordagem proposta integra o modelo Xception (CHOLLET, 2017), redes Bi-LSTM e camada de atenção. Para a extração de características são utilizadas duas abordagens, fluxo óptico e representações baseadas em CNN. Na camada de classificação, a abordagem é desenvolvida usando LSTM bidirecional (Bi-LSTM) com uma camada de autoatenção para melhorar o desempenho. A etapa de pré-processamento desse trabalho contempla uma amostragem uniformizada dos quadros de vídeos, com o objetivo de serem selecionadas uma quantidade fixa de quadros. Além da combinação de modelos, foram realizadas variações experimentais entre a quantidade de quadros selecionados durante a amostragem uniformizada, sendo 5 e 10 quadros. No experimento, foram utilizadas as bases de dados *Hockey Fight* (NIEVAS et al., 2011), *Movies* (NIEVAS et al., 2011) e *Surveillance Camera Fight* (AKTı; TATAROğLU; EKENEL, 2019).

Na pesquisa de Rendón-Segador et al. (2021) é apresentada uma arquitetura denominada ViolenceNet, para classificar uma ação como violenta ou não violenta. A arquitetura é composta por um codificador espaço-temporal de rede DenseNet-121, uma camada de autoatenção multi-head [39], uma camada LSTM 2D de convolução bidirecional (BiConvLSTM2D) - célula recorrente com dois estados, que permite obter informações de sequências para frente e para trás no tempo simultaneamente e 4 camadas *fully connected*. Os experimentos, para representar implicitamente dimensão temporal, foram realizados com duas categorias de entradas, o primeiro lote com fluxo óptico e o segundo lote com subtração de quadros adjacentes (fluxo pseudo-óptico). Os resultados foram avaliados em 4 bases de dados que mais aparecem nos estudos de detecção de comportamentos violentos: *Hockey Fights* (NIEVAS et al., 2011), *Movie Fights* (NIEVAS et al., 2011), *Violent Flows (HASSNER; ITCHER; KLIPER-GROSS, 2012) and Real Life Violence Situations* (RLVSs) (SOLIMAN et al., 2019). Ainda conforme o autor:

> A principal diferença entre os dois métodos é que o fluxo pseudo-óptico também representa aqueles *pixels* que não se moveram entre dois quadros consecutivos. Se o mesmo pixel em ambos os quadros não mudasse de valor, ao subtrair os dois

> quadros esse pixel ficava preto, independentemente de ter se movido. Para o método de fluxo óptico, os *pixels* que ficariam pretos são aqueles que não se moveram entre dois quadros consecutivos (RENDÓN-SEGADOR et al., 2021, p.8).

Na pesquisa desenvolvida por Ullah et al. (2021), dado um vídeo de entrada, o mesmo é pré-processado e segmentado para extrair características de alvos importantes (como, por exemplo, veículos e pessoas). Para este procedimento é utilizado um modelo Mask-RCNN, cujo objetivo é descartar quadros inoperantes da cena a partir da seletividade/filtragem de quadros que contenham pelo menos um objeto, fato que consequentemente reduz a carga computacional.

Foi realizado um processo de aumento de dados baseado nas transformações de filtro Gaussiano invertido horizontalmente, inversão de cores baseada na intensidade dos *pixels*, rotação aleatória de grau e inversão vertical. A etapa de extração de características é composta de duas fases: na primeira por características de fluxo óptico temporal (do inglês, *Temporal Optical Flow Features* - TOFF) e na segunda utiliza-se o modelo Darknet-19 (REDMON, 2013). Posteriormente, ambos os mapas de características extraídos são concatenados e dados como entrada em um modelo de rede LSTM para a detecção de violência. Os resultados foram avaliados nas bases de dados *Hockey Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e *Fight Surveillance Camera* (AKTı; TATAROğLU; EKENEL, 2019), este último consiste em cenas diurnas e noturnas, com filmagens internas e externas tornando-o mais desafiador.

Na pesquisa desenvolvida por Mumtaz, Sargano e Habib (2020) apresenta uma arquitetura Deep Multi-Net (DMN), baseada em dois modelos pré-treinados de fluxos de execução paralelos de classificação de imagens, AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012) e GoogLeNet (SZEGEDY et al., 2015). Ambos os modelos são treinados no conjunto de dados ImageNet ILSVRC[15] e, por transferência de aprendizado e validação cruzada, cada rede possui características pré-aprendidas distintas e, são integradas para formar um sistema de aprendizado. A fusão desses modelos pretende detectar violência em vídeo. As bases de dados utilizadas foram o *Hockey Fights* (NIEVAS et al., 2011) e o *Movie Fights* (NIEVAS et al., 2011). Os resultados dessa abordagem superaram os modelos AlexNet e GoogleNet, tanto em acurácia, quanto em tempo de aprendizado de características.

[15]ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Disponível em: <https://www.image-net.org/challenges/LSVRC/>. Acesso em: 03 de julho de 2022.

Já na pesquisa realizada por Ehsan e Mohtavipour (2020), é apresentada a arquitetura Vi-Net, baseada na Rede Convolucional Profunda utilizada para a tarefa de detecção de violência em sequências de vídeos, a partir de vetores de fluxo óptico extraídos dos padrões de movimentos. Nesse trabalho o autor ainda faz referência ao descritor ViF (vetores de fluxo óptico), que extraído através de métodos *hand-crafted*, não forneceu resultados consideráveis. Portanto, o Vi-Net é uma combinação entre o ViF e a arquitetura CNN, onde o movimento é calculado a partir de subtração de vetores de fluxo óptico para dois quadros consecutivos e dado como entrada na rede CNN. Os resultados foram avaliados nas bases *Hockey Fights* (NIEVAS et al., 2011), *Movie Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012).

Já na pesquisa realizada por Mugunga et al. (2021), é explorada uma técnica de *Deep Learning* que se baseia no modelo *Visual Geometry Group* (VGG16) (SIMONYAN; ZISSERMAN, 2014) pré-treinado na base de dados ImageNet (DENG et al., 2009) e *Convolutional Long Short-Term Memory* (ConvLSTM) para a detecção de violência em conjuntos de dados de videovigilância. Dado um vídeo como entrada, é utilizada a VGG-16 pré-treinada para extrair características e as sequências de características são processadas por camadas ConvLSTM. Na sequência ocorre uma concatenação das sequências de ConvLSTM e são passadas por camadas *fully connected* para classificação. Os resultados foram validados nas bases UCF-crime (SULTANI; CHEN; SHAH, 2018), RWF-2000 (CHENG; CAI; LI, 2021) e quatro *benchmarking datasets* de brigas: *Hockey Fights* (NIEVAS et al., 2011), *Movie Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e BEHAVE (BLUNSDEN; FISHER, 2010).

Na proposta de Honarjoo, Abdari e Mansouri (2021), foram utilizadas duas redes pré-treinadas, ResNet-50 (HE et al., 2016) e VGG16 (SIMONYAN; ZISSERMAN, 2014), como extratores de características. Posteriormente, foi utilizado o *Pooled of Time series* (PoT) (RYOO; ROTHROCK; MATTHIES, 2015) para agrupar as características temporais por vídeos e submetidas em uma rede neural totalmente conectada para a classificação. Para cada rede pré-treinada foi realizado um experimento diferente.

A proposta de Jahlan e Elrefaei (2021) é um método para detectar violência física envolvendo duas ou mais pessoas. Inicialmente, foi aplicada uma técnica de seleção de quadros (dividindo todos os quadros de vídeo em grupos e, em seguida, escolhendo um

quadro aleatoriamente de cada grupo). As características são extraídas a partir da diferença de dois quadros adjacentes e dadas como entrada em uma rede *Automated Mobile Neural Architecture Search* (MNAS) (TAN et al., 2019) pré-treinada e ConvLSTM (SHI et al., 2015) (para extrair e agregar características espaço-temporais discriminantes no nível de quadro que permitem a análise de movimento local no vídeo). Após a extração dos mapas de características, foi utilizada a fusão de duas camadas *pooling* (*max pooling* e *average pooling*) visando a captura de características mais discriminantes.

Posteriormente, as magnitudes das características são colocadas em um mesmo intervalo e é aplicada uma técnica de redução de dimensionalidade, através da *Linear Discriminant Analysis* (LDA). Por fim, modelos de aprendizado de máquina (*Random Forest*, *Support Vector Machine* e *K nearest neighbor*) são utilizados para classificar as características como violência ou não violência. O desempenho do método proposto é treinado e avaliado nos três conjuntos de dados de referência *Hockey Fights* (NIEVAS et al., 2011), *Movie Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012), como também combinados.

O principal foco da pesquisa de Mohtavipour, Saeidi e Arabsorkhi (2021) é o impacto das características de entrada em uma rede profunda CNN *multi-stream* de detecção de violência. A arquitetura da rede, inclui tanto a parte *handcrafted* para a extração de características, quanto de aprendizado profundo para a classificação dos dados.

Já na pesquisa desenvolvida por Mohtavipour, Saeidi e Arabsorkhi (2021), é proposta uma estrutura de detecção de violência com base em características de aparência, velocidade de movimento e representações de imagem, derivados de métodos *handcrafted*. Ou seja, um *framework* com os fluxos espacial (ao nível de escala de cinza), temporal (diferença de magnitude de fluxo óptico) e espaço-temporal baseado na imagem de diferença de energia de movimento (do inglês, *Diferential Motion Energy Image* - DMEI). Os resultados foram validados nas bases *Hockey Fight* (NIEVAS et al., 2011) e *Movies Fights* (NIEVAS et al., 2011) (bases de dados sem aglomeração) e *Violent Flows* (ViF) (HASSNER; ITCHER; KLIPER-GROSS, 2012) (base de dados com aglomeração).

No estudo de Caetano et al. (2017), é proposta a inserção de mais um fluxo temporal no modelo *Very Deep Two-Stream* (WANG et al., 2015), que já possui o fluxo RGB e de fluxo óptico. O novo fluxo é baseado em imagens calculadas a partir da magnitude e orientação

do fluxo óptico, a *Magnitude-Orientation Stream* (MOS), assumindo que informações que podem ser descritas pela relação espacial contida na vizinhança local. A arquitetura de rede empregada foi a rede convolucional profunda VGG-16 (SIMONYAN; ZISSERMAN, 2014) e os resultados foram obtidos a partir das bases UCF101 (SOOMRO; ZAMIR; SHAH, 2012) e HMDB51 (KUEHNE et al., 2011).

Características baseadas em pontos chaves do corpo humano

Um modelo de rede neural profunda para identificar atividades individuais violentas é proposto por Naik e Gopalakrishna (2021), para isso foi utilizada uma CNN baseada na região da máscara (Mask-RCNN modificada), detecção de pontos-chave e memória de longo prazo (LSTM). Ao localizar o humano na cena, com base na geração da caixa delimitadora desta, são também gerados 17 pontos-chave de estimativa de pose do ser humano, com a máscara sob a forma do corpo e dados como entrada na LSTM. Os resultados foram avaliados nos *datasets* KTH (SCHULDT; LAPTEV; CAPUTO, 2004), Weizmann (BLANK et al., 2005) e a uma base gerada pelos próprios autores.

A estimativa de pose do corpo humano (*Human Pose Estimation*, é baseada em pontos-chave, pode ser realizada com base nas observações de partes do corpo humano e suas dependências. Existem diferentes abordagens (BAZAREVSKY; GRISHCHENKO, 2020), (CAO et al., 2019) e (TAYLOR et al., 2012) para lidar com o problema de estimativa de *landmarks* do corpo humano, como as metodologias (VEZZANI; BALTIERI; CUCCHIARA, 2013) baseadas em modelos espaciais 2D ou 3D (BAZAREVSKY; GRISHCHENKO, 2020), modelos baseados em esqueleto (CAO et al., 2019) ou contorno (HE et al., 2017), possibilitando o mapeamento, rastreamento e previsão dessas coordenadas que descrevem o corpo humano e possivelmente suas atividades.

Características baseadas a partir da geração de imagens dinâmicas

São utilizadas para representar segmentos de vídeos, de forma que as informações da evolução temporal e ordenação temporal sejam mantidas, mesmo sob a condição de variação das ações das cenas. Os vídeos são resumidos a uma ou poucas imagens RGB, as quais são construídas com o método *Rank Pooling* (método de agrupamento temporal onde as características ao nível de quadro evoluem ao longo do tempo em um vídeo) e portanto, são geradas imagens que possuem a aparência do movimento.

Na pesquisa realizada por Roman e Chávez (2020), um método de detecção e localização

da violência em sequências de vídeo baseado em 4 etapas principais: inicialmente, dado um vídeo, o mesmo é resumido em uma imagem através da estratégia de *rank pooling*, para representar características de movimento através de uma imagem dinâmica (BILEN et al., 2017); posteriormente, essas imagens são dadas como entrada em um modelo de classificação CNN; caso o vídeo seja classificado como violento, é gerada a máscara de saliência a partir de um modelo fracamente supervisionado proposto por Dabkowski e Gal (2017); e por fim, a etapa de refinamento, que processa as máscaras anteriores visando obter apenas as regiões de interesse (as regiões que ocorrem comportamentos violentos).

Nesse experimento também foram avaliados resultados em diferentes tamanhos de sequências de quadros para a formação das imagens dinâmicas no AlexNet e também variações no número de imagens dinâmicas no ResNet-50. Foi avaliada nos seguintes datasets: *Hockey Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e UCFCrime2Local (LANDI; SNOEK; CUCCHIARA, 2019).

Na pesquisa realizada por Jain e Vishwakarma (2020), a extração de características é realizada por meio do *fine-tuning* do modelo *Inception-Resnet-V2* pré-treinado no conjunto de dados ImageNet, que recebe como entrada o vídeo RGB e é transformado em características de movimento de Imagens Dinâmicas RGB (DI - *Dynamic Image*) com foco nos padrões de movimento do objeto, mantendo a cinética de longo prazo. A proposta foi avaliada nos datasets *Hockey Fights* (NIEVAS et al., 2011), *Movie Fights* (NIEVAS et al., 2011) e *Real-Life Violence* (SOLIMAN et al., 2019), embora esse último seja bastante novo para o domínio e não seja usado extensivamente ainda.

Características baseadas em regiões de interesse

A proposta apresentada por Nasaruddin et al. (2020) é um método de detecção de anomalias em comportamentos humanos, utilizando uma abordagem de detecção de área de atenção da cena através de subtração de *background*, ao contrário dos métodos *full-frame*. Neste método, dado um vídeo de entrada, as regiões de interesse são localizadas através de um algoritmo que incorpora a subtração de *background* (utilizada para recuperar o primeiro plano com as regiões de atenção candidatas, retendo as bordas das áreas observadas) com o filtro bilateral (alivia o ruído dos quadros). Após essa etapa de pré-processamento, o modelo proposto usa apenas a região de interesse obtida a cada quadro.

Os autores ainda compararam o método de subtração de *background* com as abordagens

de Mistura de Gaussianas (do inglês, *Mixture of Gaussians* - MOG2) (ZIVKOVIC, 2004) e K-Vizinhos Mais Próximos (KNN) (ZIVKOVIC; HEIJDEN, 2006) e puderam perceber que os dois últimos eram sensíveis a confundir ruído como candidato a movimento, por exemplo. Foi utilizado o extrator de características o modelo pré-treinado *Convolutional 3D Networks* (C3D) (TRAN et al., 2015), os quais são dados como entrada em uma rede neural *fully connected* que gera saídas e pontuação de anomalia. Esse método foi treinado e testado na base UCF-Crime (SULTANI; CHEN; SHAH, 2018).

Outras abordagens

Na pesquisa realizada por Martínez-Mascorro, Ortiz-Bayliss e Terashima-Marín (2020), é apresentada uma análise comparativa de abordagens de detecção de comportamento humano violento, combinando abordagens de treinamento e classificação, isto é, mais especificamente visa compreender se amostras de comportamento violento devem ser treinadas em grupo ou segregadas por natureza da atividade suspeita. Para esse experimento foram coletadas amostras da base UCF-Crime (SULTANI; CHEN; SHAH, 2018) correspondentes a 4 tipos distintos de crimes: furto em lojas, roubo, incêndio e abuso. As abordagens de treinamento e classificação exploradas foram: treinamento binário com classificação binária, treinamento multiclasse com classificação binária e treinamento multiclasse com classificação multiclasse. Nesse experimento, também foram realizadas variações como o número de filtros das camadas convolucionais e uma série de execuções distintas para validar os resultados.

A principal contribuição apresentada por Martínez-Mascorro, Ortiz-Bayliss e Terashima-Marín (2020), diz respeito aos resultados obtidos quando as amostras são agrupadas independente do tipo de crime durante o treinamento, que evidenciam uma melhora considerável de acurácia durante a classificação. Sendo que o treinamento binário junto à classificação binária alcançou resultados de até 24,5% maiores quando comparados aos melhores modelos de outras abordagens.

Na pesquisa realizada por Wu et al. (2020), é proposto um método para modelar as relações entre os trechos de vídeos, a HL-Net, que é formada por três fluxos paralelos (*branches*): (1) o holístico (captura dependências de longo alcance usando a similaridade anterior); (2) o localizado (captura a relação posicional local usando a proximidade anterior); e (3) de pontuação (captura dinamicamente a proximidade de pontuação prevista).

A entrada do modelo proposto por Wu et al. (2020) é multimodal (audiovisual). Para

a extração de características, foram utilizadas duas redes, C3D (TRAN et al., 2015) e I3D (CARREIRA; ZISSERMAN, 2017). Foram extraídas características da camada fc6 do C3D pré-treinado na base Sports-1M (KARPATHY et al., 2014) e extraídas características através do parâmetro *global_pool* do I3D pré-treinado na base de dados Kinetics-400[16]. Para as características de áudio foi utilizada a rede VGGish (HERSHEY et al., 2017) pré-treinada em um conjunto de dados extraídos do Youtube. Nesse trabalho também é proposta a base *XD-Violence* (WU et al., 2020), com 4.754 vídeos contando com 6 categorias de violência com sinais de áudio e rotulados fracamente.

3.2.1 Extração de Características

Sintetizando as principais características dos trabalhos relacionados e seus respectivos resultados relacionados à métrica de acurácia[17], os quais podem ser evidenciados nos Apêndices A e B, pode-se notar que existem abordagens baseadas em métodos *full frame*[18] (DING et al., 2014) (CHENG; CAI; LI, 2021) (EHSAN; MOHTAVIPOUR, 2020) e métodos focados em áreas alvo dos vídeos (NASARUDDIN et al., 2020) (ROMAN; CHáVEZ, 2020). Nos métodos baseados em *full frames*, são normalmente consideradas todas as características extraídas de maneira holística dos quadros, enquanto que nos métodos baseados em área de interesse são consideradas características extraídas de maneira específica (ULLAH et al., 2021), (ROMAN; CHáVEZ, 2020) (NASARUDDIN et al., 2020) isto é, de áreas dos quadros específicas e, para isso algumas técnicas tais como, máscara de saliência (ROMAN; CHáVEZ, 2020) ou remoção de *background* (NASARUDDIN et al., 2020) são aplicadas. A abordagem focada em área alvo, pode ser considerada menos custosa em termos de aprendizagem do modelo. No entanto, pode ser que a sua fase de pré-processamento tenha um custo computacional maior que a abordagem *full frame*.

Outro ponto importante a ser considerado é a informação temporal, para isso são utilizados modelos baseados em arquitetura LSTM (ULLAH et al., 2021) (NAIK; GOPALA-

[16]Kinetics-400. Disponível em: <https://www.deepmind.com/open-source/kinetics>. Acesso em: 03 de julho de 2022

[17]Como se trata de um problema de classificação, normalmente essa é a métrica reportada pelos trabalhos do estado da arte. Um fator relevante, que deve ser considerado, é que essa métrica pode enviesar os resultados do modelo quando os dados avaliados não estão balanceados.

[18]*Full Frame*. Que opera sob todas as características visuais presentes no quadro, ou seja, os dados de entrada não possuem filtragem ou seleção de áreas de interesse.

KRISHNA, 2021), LSTM bidirecional (AKTı; TATAROğLU; EKENEL, 2019) (RENDÓN-SEGADOR et al., 2021) (MUGUNGA et al., 2021) ou ConvLSTM (JAHLAN; ELREFAEI, 2021). A informação temporal também pode ser extraída por meio do fluxo óptico (CHENG; CAI; LI, 2021) (RENDÓN-SEGADOR et al., 2021) (EHSAN; MOHTAVIPOUR, 2020) ou através de imagens dinâmicas (JAIN; VISHWAKARMA, 2020) (ROMAN; CHáVEZ, 2020).

Fluxo Óptico

Informações de fluxo óptico podem fornecer estimativas de direções e intensidades de movimentos agressivos ou violentos, sejam produzidos por partes específicas do corpo humano e/ou envolvendo outras pessoas ou objetos utilizados durante o comportamento violento, como utilizado por Ding et al. (2014), Cheng, Cai e Li (2021), Rendón-Segador et al. (2021), Ehsan e Mohtavipour (2020) e Caetano et al. (2017).

Região de Interesse

Regiões de interesse ou de atenção são áreas em quadros definidas a partir da intensidade do fluxo de movimento. Existem determinadas regiões dos quadros de um vídeo, que permanecem constantes durante a execução do mesmo e na maioria das vezes não agregam informações importantes para a identificação de comportamentos humanos agressivos, que são caracterizados pela intensidade de movimentos realizados pelo corpo. Desta forma, abordagens (NASARUDDIN et al., 2020) baseadas na identificação de áreas de interesse em uma cena e a subtração do *background* (segundo plano da cena) dessas áreas constantes têm o potencial de reduzir o tamanho e a quantidade de informações triviais submetidas aos modelos e, consequentemente, a redução de tempo e recursos de processamento de dados.

Segmentação

Segmentação é o processo de decomposição de uma *imagem* em segmentos, regiões, objetos, contornos, linhas, entre outros, isto é, em conjuntos de pixels de maneira isolada e seguindo um determinado critério, visando a simplificação da representação da imagem e suas análises. Normalmente, a segmentação é utilizada para encontrar ou extrair objetos e formas em imagens, como na pesquisa realizada por Naik e Gopalakrishna (2021), Roman e

Chávez (2020) e Ullah et al. (2021), que utilizaram abordagens baseadas em Mask R-CNN (HE et al., 2017) para a extração de características segmentadas. As principais estratégias de segmentação de imagens são baseadas em descontinuidade (descontinuações bruscas de intensidade) ou por similaridade (critério de similaridade de *pixels*).

3.2.2 Conjuntos de Dados

Nesta etapa, foi realizado um aprofundamento direcionado para a identificação, análise e sumarização das principais características das bases de dados utilizadas nesses trabalhos do estado da arte e suas respectivas lacunas observadas. A sumarização das bases de dados encontra-se no Apêndice B.

As bases de dados analisadas constituem o domínio de *Human Action Recognition* (HAR). Na pesquisa de Blank et al. (2005) e Schuldt, Laptev e Caputo (2004), as atividades encontradas são mais genéricas (ações de pular, andar correr, caminhar, etc.) e não envolvem necessariamente violência humana ou captura a partir de câmeras de videovigilância. Já na pesquisa de Kuehne et al. (2011) e Soomro, Zamir e Shah (2012), embora estejam contidas no agrupamento mencionado, existem classes específicas do conjunto de dados relacionadas a violência física, tais como as ações de chutar e socar, respectivamente.

Os cenários das amostras de vídeos em que ocorrem violência incluem desde filmes (NIEVAS et al., 2011), esportes (NIEVAS et al., 2011), pessoas atuando em cenários controlados (BLUNSDEN; FISHER, 2010) (SCHULDT; LAPTEV; CAPUTO, 2004) (BLANK et al., 2005), ruas (HASSNER; ITCHER; KLIPER-GROSS, 2012) (SOLIMAN et al., 2019), escolas (SOLIMAN et al., 2019) (HASSNER; ITCHER; KLIPER-GROSS, 2012), multidão (HASSNER; ITCHER; KLIPER-GROSS, 2012) e variados (WU et al., 2020), (KUEHNE et al., 2011), (SOOMRO; ZAMIR; SHAH, 2012), (LANDI; SNOEK; CUCCHIARA, 2019), (SULTANI; CHEN; SHAH, 2018) e (CHENG; CAI; LI, 2021). Os conjuntos de dados *UCF-Crime* (SULTANI; CHEN; SHAH, 2018), *UCFCrime2Local* (LANDI; SNOEK; CUCCHIARA, 2019), *RWF-2000* (CHENG; CAI; LI, 2021) e *XD-Violence* (WU et al., 2020) contém vídeos capturados de câmeras de videovigilância, além de conter diferentes tipos de violência humana, tais como briga, abuso, assalto, entre outros.

A maioria dos conjuntos de dados *Hockey Fight* (NIEVAS et al., 2011), *Movies Fights* (NIEVAS et al., 2011), *Violent Flow* (HASSNER; ITCHER; KLIPER-GROSS, 2012), *RWF-*

2000 (CHENG; CAI; LI, 2021), *Surveillance Fight* (AKTı; TATAROğLU; EKENEL, 2019), *UCF-Crime* (SULTANI; CHEN; SHAH, 2018) e *Real Life Violence Situations* (RLVSs) (SOLIMAN et al., 2019) possuem seus rótulos associados a duas categorias, as quais normalmente representam presença ou ausência de violência. Ainda existem as bases *XD-Violence* (WU et al., 2020), *UCF-Crime* (SULTANI; CHEN; SHAH, 2018) e *UCFCrime2Local* (LANDI; SNOEK; CUCCHIARA, 2019) que fornecem o agrupamento do tipo de violência baseado em rotulagem fraca, isto é, um rótulo por vídeo, visto que os vídeos podem ter duração longa e conter cenas de violência e não violência em uma mesma amostra de vídeo, por exemplo.

As formas de anotações de vídeos mais comuns são caracterizadas ao nível de vídeo (SOOMRO; ZAMIR; SHAH, 2012), (CHENG; CAI; LI, 2021) e (AKTı; TATAROğLU; EKENEL, 2019) (SOLIMAN et al., 2019) (*trimmed*, geralmente são vídeos com a duração curta) ou ao nível de quadro (SULTANI; CHEN; SHAH, 2018) (*untrimmed*), ou seja, por possuírem uma duração mais longa, normalmente a hora de início e término de atividades violentas possuem anotações ao nível de quadro (CHENG; CAI; LI, 2021). As bases de dados *UCFcrime2local* (LANDI; SNOEK; CUCCHIARA, 2019), *UCF-Crime* (SULTANI; CHEN; SHAH, 2018) e *XD-Violence* (WU et al., 2020) possuem uma rotulagem fraca, isto é, são constituídas por uma variação de vídeos com diferentes durações e por uma anotação ao nível de intervalo de quadros (*untrimmed*). Esse tipo de base é mais comum para tarefas associadas a aprendizado semi-supervisionado ou não supervisionado, como a localização de violência no eixo temporal do vídeo ou detecção de violência como anomalia. Utilizar amostras de dados com anotações *untrimmed* para o cenário de aprendizado supervisionado, pode ocasionar problemas para a aprendizagem dos modelos, em casos em que o tipo de atividade contida nas cenas variar ao longo do vídeo.

No contexto de detecção de violência, um dos principais problemas das bases de dados é o uso de atores para realizar cenas violentas fictícias/artísticas (seja em cenário controlado ou de diferentes ambientes) (SCHULDT; LAPTEV; CAPUTO, 2004) e (BLANK et al., 2005), baixa qualidade de resolução, baixa quantidade de amostras, como na pesquisa de Nievas et al. (2011). Na pesquisa realizada por Sultani, Chen e Shah (2018) e Landi, Snoek e Cucchiara (2019) há exemplos de conjuntos de dados que apresentam cenas de violência realística. Outros problemas relacionados ao reconhecimento de ações humanas, conforme Khurana e Kushwaha (2018), estão ligados à baixa variação de intraclasse e a alta variação

interclasse.

Na baixa variação interclasse, comportamentos violentos e não violentos possuem muitas semelhanças em determinadas situações, como, por exemplo, o padrão de comportamento entre uma cena de imobilização de um sujeito e uma cena de um abraço entre sujeitos. Ao mesmo tempo, na alta variação intraclasse, amostras de cenas de uma determinada classe (violência) podem diferir muito entre os padrões de comportamentos realizados, como em cenas violentas de assalto, roubo, abuso, briga, entre outras.

Segundo Nweke et al. (2018), embora o processo de coleta e rotulação manual de dados seja algo custoso e existam poucos conjuntos de dados de referência, muitos pesquisadores coletam seus próprios conjuntos de dados. No entanto, esses conjuntos de dados podem não ser generalizados o suficiente para novos cenários ou para diferentes tipos de comportamento violentos. Para a construção da base de dados RWF-2000 proposta por Cheng, Cai e Li (2021), por exemplo, os critérios de seleção para os vídeos contidos na base não se limitam a tipos de violência específicos. Dessa forma, a base contém cenas variadas de briga, roubo, tiroteio, sangue, assalto, entre outros. As amostras dessa base foram coletadas a partir do YouTube[19] e posteriormente, cortadas de maneira que cada amostra fosse constituída por 5 segundos com 30 FPS.

[19] Youtube. Disponível em: <https://youtube.com/>. Acesso em: 15 de junho de 2022.

Capítulo 4

Metodologia

Esta pesquisa é uma extensão da abordagem proposta por Cheng, Cai e Li (2021). Nesta dissertação são propostas melhorias relacionadas ao método de delimitação de área de interesse dos quadros de vídeo, como também a adição de técnicas que reduzem o uso de recursos computacionais durante a etapa de treinamento. Para atingir o objetivo desta pesquisa e esclarecer as metas iniciais, a metodologia é constituída de duas fases:

- na Seção 4.1, é apresentada a exploração da abordagem proposta por Cheng, Cai e Li (2021), a qual foi utilizada como *baseline* nesta pesquisa,
- na Seção 4.2 são apresentadas as principais mudanças e diferenciais entre o método adotado por Cheng, Cai e Li (2021) e o método proposto nesta pesquisa.;
- a Seção 4.3, apresenta o conjunto de métodos e experimentos necessários para a realização desta pesquisa, assim como a definição das métricas de avaliação utilizadas e os detalhes de implementação.

4.1 Baseline

Em (CHENG; CAI; LI, 2021), é apresentada uma abordagem baseada em CNN para a detecção de violência em vídeos no domínio de videovigilância. Essa abordagem possui o repositório de códigos e a base de dados disponíveis publicamente (base de dados disponível mediante contato aos autores), que facilita o processo de reprodutibilidade da pesquisa. Essa pesquisa também é relativamente recente no estado da arte e obteve resultados competitivos.

Por esses motivos, essa abordagem será utilizada como o método *baseline* para o desenvolvimento da proposta desta dissertação. O fluxo de execução da abordagem *baseline* é composto por algumas etapas de fluxo de execução sintetizadas na Figura 4.1 sendo detalhadas na sequência. Esse fluxo de execução também pode ser visualizado como pseudo-código comentado através do Algoritmo 1.

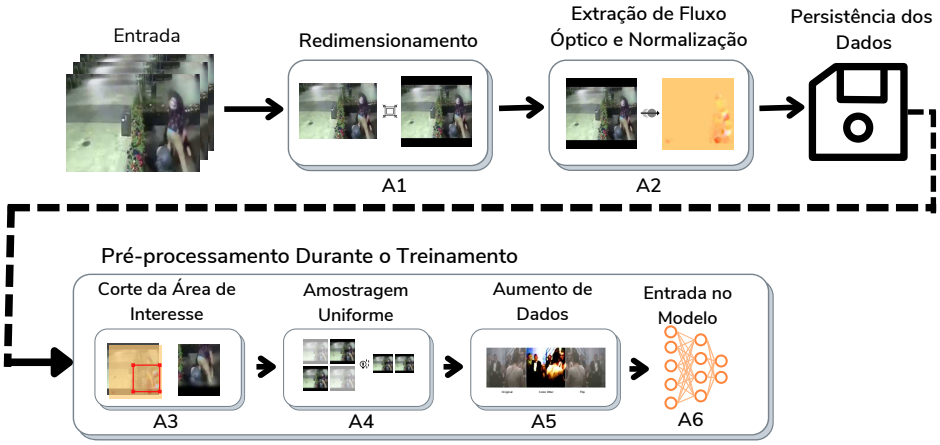

Figura 4.1: Representação do fluxo de execução proposto em (CHENG; CAI; LI, 2021). Fonte: A Autora (2022) baseado em (CHENG; CAI; LI, 2021).

Inicialmente, as amostras de vídeos da base de dados RWF-2000 (CHENG; CAI; LI, 2021) são carregadas e redimensionadas para 224×224 *pixels* (A1). Posteriormente, nesta abordagem, para seleção da área de interesse dos quadros de vídeos, foi implementada uma estratégia com o objetivo de reduzir a quantidade de características de entrada na rede. Nessa estratégia, a região de interesse é obtida através do método Gunner Farneback (FARNEBäCK, 2003) usado para calcular uma estimativa de movimento e obter o fluxo óptico denso entre quadros vizinhos (A2). Na Figura 4.2, é possível visualizar um exemplo de representação do método Gunner Farneback aplicado a um quadro.

Posteriormente, é realizada uma etapa de normalização (A3) para eliminar tanto o movimento da câmera, quanto informações de fluxo óptico destoantes como, por exemplo, uma área pode ser detectada com uma magnitude de movimento alta com base nas informações de fluxo óptico, mas possuir pouca relevância. Esse fato descrito pode ser ocasionado por eventos com velocidade e movimentos extremos, os quais não estão necessariamente ligados à área de interesse principal (como em cenas com atividades humanas violentas em uma determinada área do quadro e objetos não relacionados diretamente a essa atividade também

Algoritmo 1 Delimitação da AI Cheng, Cai e Li (2021)

Require: Matriz **M** de estrutura $(f \times w \times h \times of)$ e Matriz **S** de estrutura $(f \times w \times h \times rgb)$.
Ensure: $f = 64$, $w = 224$, $h = 224$ ▷ f é a quantidade de quadros
▷ w é largura do quadro de vídeo
▷ h é a altura do quadro de vídeo
Ensure: $of = 2$, $rgb = 3$ ▷ of é a quantidade de canais do Fluxo Ótico
▷ rgb é a quantidade de canais RGB

function BBOXEXRACTION(**M**, **S**)
  $bboxWidth \leftarrow 112$
  $bboxHeight \leftarrow 112$
  $bboxLimitX \leftarrow 56$ ▷ limite de localização do eixo x da AI[1]
  $bboxLimitY \leftarrow (w - bboxLimitX) - 1$ ▷ limite de localização do eixo y de AI
  T: $(224 \times 224 \times 2)$
  $VideoSeq \leftarrow [0, 0, 0]$ ▷ inicialização de uma matriz de 3 dimensões (w, h, rgb)
  $T \leftarrow \sum_{i=0}^{|M|} M_i$ ▷ soma da magnitude do fluxo ótico do quadro individual
  $threshold \leftarrow \overline{M}$ ▷ definição de um limiar com base na média de valores
  for $magnitude \in T$ **do**
    if $magnitude < threshold$ **then**
      $T_{[magnitude]} \leftarrow 0$ ▷ filtragem de valores com base no limiar
    end if
  end for
  $T_{eixo\_x} \leftarrow \sum_{i=0}^{|T|} T_{[1]}$ ▷ *flattening*[2] da largura (1×224)
  $T_{eixo\_y} \leftarrow \sum_{i=0}^{|T|} T_{[0]}$ ▷ *flattening* da altura (224×1)
  $X \leftarrow PDF(T_{eixo\_x})$[3] ▷ captura as probabilidades dos valores de altura da AI
  $Y \leftarrow PDF(T_{eixo\_y})$ ▷ captura as probabilidades dos valores de largura da AI
  $C \leftarrow RandomChooseCandidates(n \leftarrow 10, probability \leftarrow X)$
  $V \leftarrow RandomChooseCandidates(n \leftarrow 10, probability \leftarrow Y)$
  $xl \leftarrow max(bboxLimitX, min(\overline{C}, bboxLimitY))$ ▷ extremidades de largura
  $yl \leftarrow max(bboxLimitX, min(\overline{V}, bboxLimitY))$ ▷ extremidades de altura
  $x \leftarrow ((xl - bboxWidth/2), (yl - bboxHeight/2))$ ▷ p. inicial[4] da largura da AI
  $y \leftarrow ((xl + bboxWidth/2), (yl + bboxHeight/2))$ ▷ p. inicial da altura da AI
  for $f \in S$ **do** ▷ para cada quadro da sequência
    $VideoSeq_{[f]} \leftarrow corteAI(f, x, y)$ ▷ retorna o quadro correspondente a AI
  end for
  return $VideoSeq$
end function

[1] AI: Área de Interesse.

[2] *flattening*: Reduz valores compostos de múltiplas dimensões para apenas uma dimensão.

[3] PDF: Função de densidade de probabilidade.

[4] p. inicial: ponto inicial.

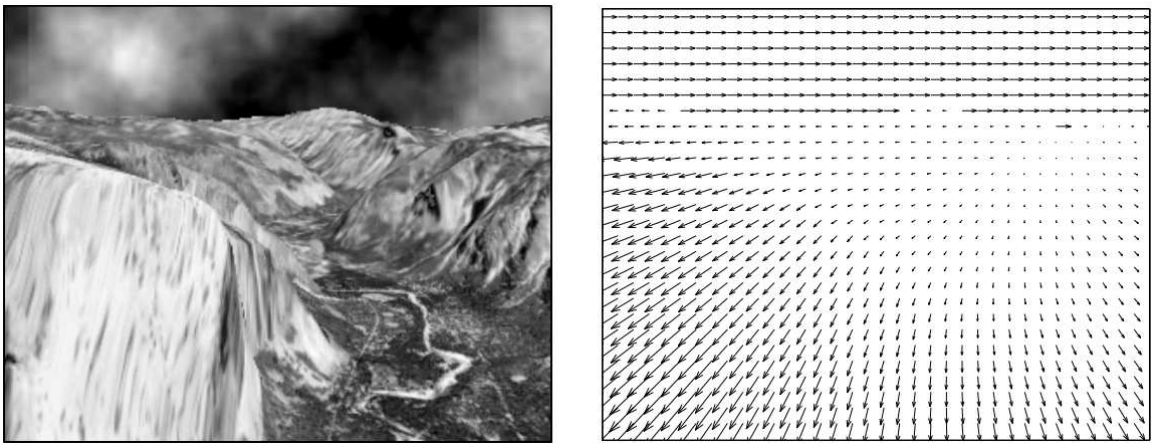

Figura 4.2: Imagem original à esquerda e o campo de velocidade estimado correspondente à direita. Fonte: (FARNEBäCK, 2003).

presentes na cena, como um carro trafegando em alta velocidade no *background* do quadro).

Por último, a dimensão de dados de toda a base resultante é composta por 5 canais: 2 canais correspondentes aos dados de fluxo óptico (componentes horizontal e vertical) e 3 canais correspondentes aos dados RGB dos quadros. Todos estes canais são armazenados para submetê-los à fase de treinamento.

4.1.1 Etapa de Treinamento Proposta por Cheng, Cai e Li (2021)

Como conjuntos de dados compostos por sequências de quadros, além de representarem movimento contínuo, também possuem informações adicionais e possivelmente redundantes devido à alta correlação com quadros vizinhos, então, muitos pesquisadores se dedicam a fundir adequadamente as informações espaciais e temporais (CHENG; CAI; LI, 2021).

No fluxo implementado pelo referido autor, antes dos dados armazenados da etapa anterior serem fornecidos como entrada no fluxo de treinamento, os mesmos são submetidos a uma etapa de uniformização das amostras (A4) com o objetivo de fundir e diminuir a quantidade de dados de entrada. Nessa etapa, para cada vídeo de entrada, é gerado um subvídeo de tamanho fixo a partir de uma seleção de quadros amostrados esparsamente em um intervalo uniforme.

Os vídeos da base RWF-2000 possuem 30 FPS. O tamanho fixo (tamanho alvo das amostras que serão submetidas ao modelo) é previamente definido para conter 64 quadros. Para um melhor entendimento desta técnica, na Figura 4.3 é exemplificada a seleção dos quadros.

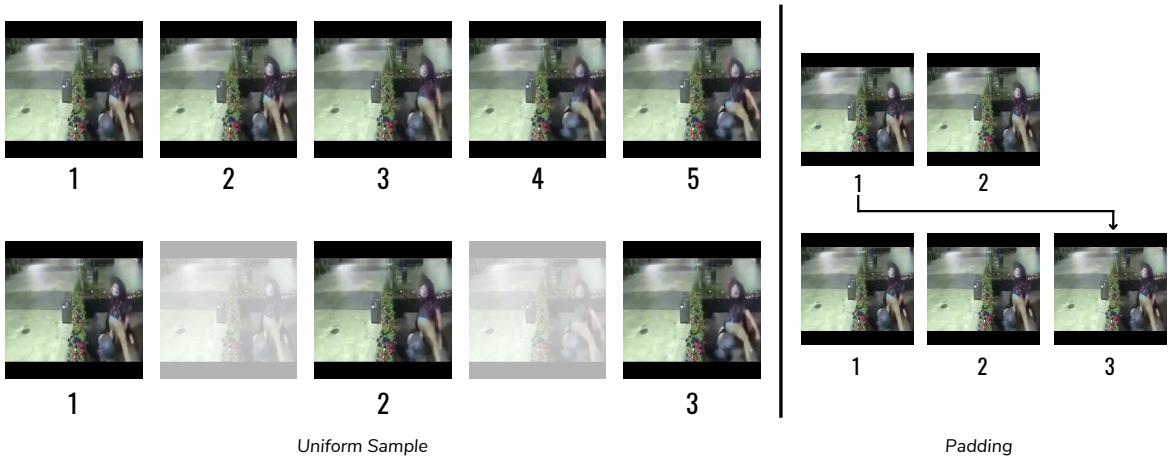

Figura 4.3: Representação da técnica de uniformização de amostras (à esquerda) e a técnica de *padding* aplicada. Fonte: A autora (2022).

Por exemplo: dado um vídeo de entrada de 200 quadros, é realizado um cálculo para identificar o tamanho do intervalo da sequência de quadros da qual deverá ser selecionado 1 quadro por vez, desta forma $200/64 = 3{,}125$ quadros. Esse resultado significa que a cada 3 quadros da sequência de 200 quadros, será selecionado 1 quadro e descartados os restantes, até completar 64 quadros no total. Por isso os quadros são "comprimidos" de maneira uniforme e não aleatória.

É importante também mencionar que há uma técnica de preenchimento (*padding*) associada para realizar o preenchimento dos quadros. Caso após a divisão, a amostra não possua quadros suficientes para formar o seguimento de 64 quadros, essa técnica introduz os quadros iniciais do vídeo no final da sequência.

Na fase de delimitação da área de interesse, para a remoção de informações destoantes, é definido e aplicado um limiar com base na média geral das intensidades dos *pixels* de todo o quadro como critério de filtragem e são eliminadas todas as informações que não atendem a esse limiar (linha 9 do Algoritmo 1). Também são selecionados aleatoriamente 10 candidatos de quadros de vídeos para fornecer informações para a geração do mapa geral de intensidade dos *pixels*.

Dadas as informações de fluxo óptico e sabendo que as informações de quadros de vídeos consecutivos são altamente correlacionadas, para cada quadro também é calculado um mapa de intensidade de *pixels* para indicar as intensidades de movimento (linha 8 do Algoritmo 1). Posteriormente, um mapa de intensidades geral é obtido com base na soma de valores

por eixo (linhas 15 e 16 do Algoritmo 1). Na sequência, esses valores são normalizados a partir da soma das probabilidades de todos os mapas de intensidade dos quadros nas direções horizontal e vertical (linhas 17 e 18 do Algoritmo 1, respectivamente).

Por fim, como a região de interesse da cena para reconhecer comportamentos violentos está geralmente concentrada em uma área menor ou específica, através desse mapa de calor geral é possível identificar essa região com base na intensidade de movimento e então, um videoclipe de tamanho fixo é gerado com base no corte dessa área de interesse. O comprimento alvo dos videoclipes é 64 quadros e o tamanho das regiões recortadas é 224×224 *pixels*.

Além disso, antes das amostras serem submetidas ao modelo, a cada época são introduzidas amostras de quadros gerados pela técnica de *data augmentation online*[1], sendo elas o (1) *flip* e a (2) alteração de brilho, saturação, contraste e matiz, ambas utilizadas na tentativa de reduzir o ajuste excessivo do modelo aos dados simulando variações do mesmo cenário de maneira aleatória. Exemplos de efeitos aplicados aos quadros podem ser vistos na Figura 4.4.

Figura 4.4: Amostras de quadros originais (à direita), com os tipos de aumento *collor jittter* (centro) e *flip* (à direita). Fonte: A autora (2022).

[1] *Data Augmentation Online*. Técnica de aumentação de dados que transformam dados em tempo real durante o treinamento (SHORTEN; KHOSHGOFTAAR, 2019).

4.2 Abordagem Proposta

A abordagem proposta nesta pesquisa trata-se de um método de pré-processamento para a seleção de área de interesse de quadros de vídeo, baseado em filtro gaussiano. O objetivo desse método é refinar o mapa de características submetido ao modelo de detecção de violência humana, aplicando o filtro gaussiano na área de baixo interesse, ao invés de eliminar toda a área de baixo interesse através do corte da área, por exemplo. Uma abstração dessa etapa pode ser visualizada na Figura 4.5, em que a etapa de corte (A3) é substituída por uma etapa de seleção da área de interesse e aplicação do filtro gaussiano (B3), a seleção da área de interesse é composta de uma área 112×112 *pixels* (50% da área total do quadro de 224×244 *pixels*). A hipótese é que com a técnica de delimitação da área de interesse através de filtro gaussiano, seja possível manter informações importantes do contexto do *background* da cena, ao mesmo tempo em que características de *background* recebem um grau de relevância menor ao suavizar a área de baixo interesse.

Essa mudança também pode ser visualizada através da substituição da linha 26 do Algoritmo 1[2] pela aplicação do filtro gaussiano na área de baixo interesse proposta nesta pesquisa e descrita pelo Algoritmo 2. O Algoritmo 2 é responsável pela dinâmica entre a delimitação e permanência das características originais da área de interesse, como também, pela aplicação do filtro gaussiano nas demais áreas.

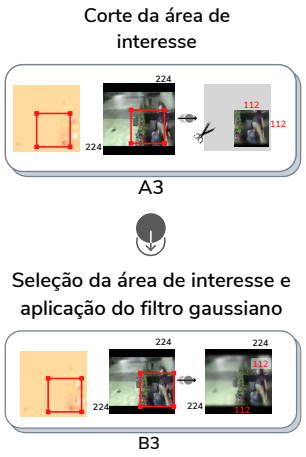

Figura 4.5: Representação do processo do corte da área de interesse proposto por Cheng, Cai e Li (2021) (A3) e o processo de delimitação de área de interesse proposta nesta pesquisa (B3). Fonte: A autora (2022).

[2]O Algoritmo 1 foi inicialmente proposto por Cheng, Cai e Li (2021).

Algoritmo 2 Aplicação do filtro gaussiano

Require: Matriz M de estrutura $(f \times w \times h \times of)$ e Matriz S de estrutura $(f \times w \times h \times rgb)$.
Ensure: $f = 64$, $w = 224$, $h = 224$ ▷ f é a quantidade de quadros
▷ w é largura do quadro de vídeo
▷ h é a altura do quadro de vídeo
Ensure: $of = 2$, $rgb = 3$ ▷ of é a quantidade de canais do Fluxo Ótico
▷ rgb é a quantidade de canais RGB
Ensure: x ▷ ponto inicial da largura da AI
Ensure: y ▷ ponto inicial da alturra da AI

```
1: function BBOXEXRACTION(M, S)
Require: (...)                                   ▷ Incluir códigos até a linha 24 do Algoritmo 1
2:     for f ∈ S do                              ▷ para cada quadro da sequência
3:         img_blur ← ApplyKernelGaussian(S[f])  ▷ aplicação do filtro gaussiano
4:         mask ← S[f]                           ▷ cópia do quadro
5:         color ← (255, 255, 255)
6:         mask ← drawRetangle(mask, x, y, color)  ▷ retângulo delimitador da AI
7:         for p ∈ S[f] do                       ▷ para cada pixel do quadro
8:             if mask == color then             ▷ se ambos possuírem o mesmo valor
9:                 VideoSeq[f][p] ← S[f][p]      ▷ retorne pixel original
10:            else
11:                VideoSeq[f][p] ← img_blur[f][p]  ▷ retorna pixel com efeito gaussiano
12:            end if
13:        end for
14:    end for
15:    return VideoSeq
16: end function
```

[1] AI: Área de Interesse.
[4] p. inicial: ponto inicial.

Outras mudanças propostas por esta pesquisa são relacionadas à melhoria de desempenho do modelo, mais especificamente à redução do uso de recursos computacionais e consequentemente, a redução do tempo de treinamento do modelo. Sendo elas:

- O uso de precisão mista na arquitetura no modelo para diminuir o uso de memória VRAM (detalhamento na Seção 4.3.3);
- A redução do tamanho do *batch* de 16 para 8;
- O desacoplamento das etapas realizadas juntas ao treinamento por Cheng, Cai e Li (2021), possibilitando que apenas o aumento online dos dados e o treinamento do modelo sejam realizados durante essa etapa. A representação dessa mudança pode ser visualizada na Figura 4.6, onde as informações das etapas de corte da área de interesse (A3) e a amostragem

uniforme (A4) são persistidas localmente, antes do início do treinamento.

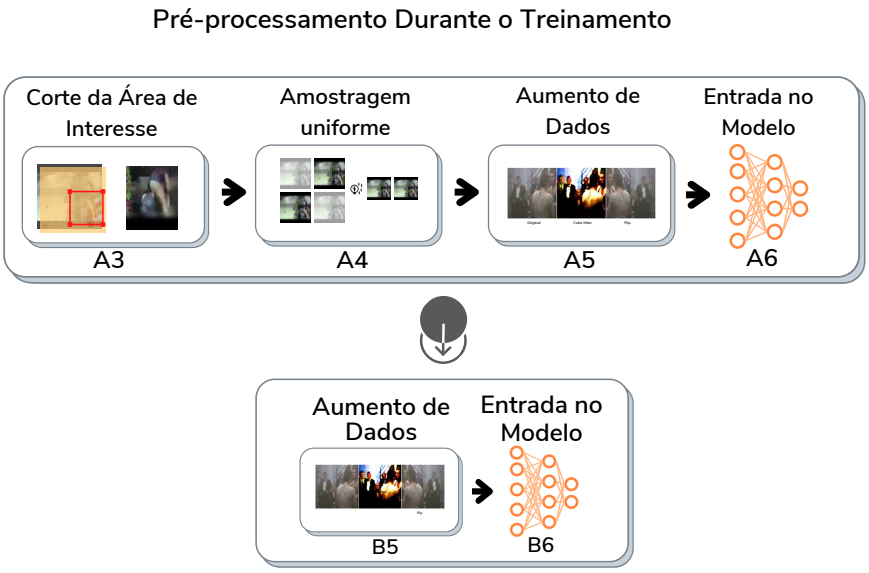

Figura 4.6: Representação da mudança aplicada durante a etapa de treinamento do modelo. Fluxo superior proposto por Cheng, Cai e Li (2021) e fluxo inferior proposto nesta pesquisa. Fonte: A Autora (2022).

A necessidade dessas mudanças ocorreu principalmente devido a limitações relacionadas às restrições de uso contínuo de recursos computacionais disponibilizados. Portanto, a implementação foi otimizada de forma que apenas os processos essenciais ao treinamento, fossem de fato executados durante a fase de treinamento, como o aumento de dados (B5), que possuem valores de parâmetros variáveis por época. Uma representação geral de todo o fluxo de execução após a aplicação das mudanças anteriormente detalhas, pode ser visualizado na Figura 4.7.

4.3 Experimentos

Na Subseção 4.3.1, são detalhados a justificativa do conjunto de dados utilizado, a arquitetura de modelo *baseline* proposta por Cheng, Cai e Li (2021) e detalhes de implementação. Na Subseção 4.3.2, é detalhado o primeiro experimento da proposta de redução do mapa de características ao aplicar a técnica de suavização através da aplicação do filtro gaussiano. Em sequência, na Subseção 4.3.3, é apresentado o segundo experimento que consiste na aplicação

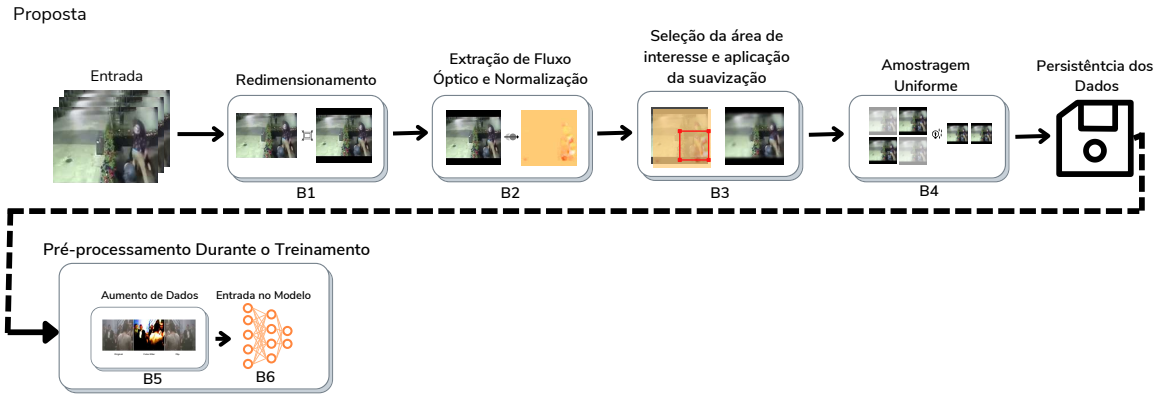

Figura 4.7: Diagrama de fluxo de execução da proposta deste trabalho. Fonte: A autora (2022).

de uma técnica de redução do uso de recurso de memória VRAM durante o treinamento, que ajudou a superar desafios relacionados aos limites de recursos computacionais disponíveis.

4.3.1 Metodologia Experimental

Conjunto de Dados

Devido à complexidade, abrangência e variedade dos atos violentos, para a definição do escopo desta pesquisa e visando tornar a proposta o mais fidedigna possível à realidade, serão delimitados para o escopo da mesma, atos de violência física realizados em ambientes distintos e capturados de câmeras de videovigilância, que contenham cenas legítimas/reais de comportamentos violentos, agressivos e/ou suspeitos de crime. A base de dados selecionada para os experimentos desenvolvidos neste trabalho é a *RWF-2000* (CHENG; CAI; LI, 2021) pelo motivo mencionado acima e pelos critérios de: disponibilidade para *download* do conjunto de dados via solicitação direta aos autores, volume considerável de amostras de vídeos em comparação às demais bases, o fato do conteúdo dos vídeos ser fiel ao contexto de videovigilância e ao pelo menos um dos tipos de comportamentos violentos (briga humana) e, por fim, ser a mesma base de dados utilizada nos experimentos realizados por Cheng, Cai e Li (2021) (facilitando a comparação dos resultados reproduzidos).

Arquitetura

Embora em (CHENG; CAI; LI, 2021), o principal modelo apresentado seja o *FlowGated*, para o desenvolvimento do presente trabalho, foi adotada a arquitetura considerando apenas a versão com o canal RGB (utilizando informações de fluxo óptico apenas para estimar e definir a área de interesse), pois essa possui menos parâmetros que a versão *FlowGated*, quanto maior a quantidade de parâmetros em um modelo, maior o uso de recursos computacionais. A arquitetura é formada por 4 blocos de CNNs com *MaxPooling* em cascata, 1 *MaxPooling*, 3 blocos de CNNs com *MaxPooling* em cascata, a *Fully Connected Layer* com *Dropout* de 20%, gerando a saída para a camada *SoftMax* e possui a função de ativação ReLU na última camada.

Detalhes de Implementação Experimental

Para o particionamento das amostras de vídeos foram consideradas 1600 amostras de vídeo para treinamento (80%) e 400 para teste (20%). Tanto no conjunto de treinamento, quanto no conjunto de teste, a quantidade de amostras foi balanceada por classe (violência e não violência), mesmo após a divisão. O treinamento do modelo foi submetido a 30 épocas. A taxa de aprendizagem foi inicializada com o valor de 0,01 e aplicada de maneira dinâmica, isto é, a cada época, a taxa de aprendizagem é reduzida em 1/10 do valor original, como a taxa de aprendizagem varia, foi usado como critério de parada as 30 épocas fixas. Foi utilizado o otimizador SGD com *momentum* (0,9) e decaimento da taxa de aprendizado (1e-6).

Os detalhes de implementação acima descritos, são os mesmos definidos por Cheng, Cai e Li (2021), salvo a redução do tamanho do *batch* de 16 para 8 para tornar possível a reprodução do experimento, dada a limitação de recursos computacionais no ambiente experimental disponível. Outro fator importante a ser mencionado é que tanto para a reprodução da pesquisa (CHENG; CAI; LI, 2021), quanto para a execução dos experimentos a serem detalhados nas subseções a seguir, o fluxo de execução já compreenderá as etapas de pré-processamento desacopladas da etapa de treinamento, conforme listada na Seção 4.2.

Para os experimentos, a arquitetura de rede foi treinada sem utilizar o recurso de pesos pré-treinados. Os recursos de hardware utilizados para este experimento são provenientes da plataforma *Google Colaboratory Pro*[3], com uma instância de máquina possuindo uma

[3] *Google Colaboratory Pro*. Disponível em: <https://colab.research.google.com/signup>. Acesso em: 15 de

placa de vídeo Tesla P100, 16GB de memória de vídeo, processador Intel(R) Xeon(R) CPU @ 2.30GHz de 2 cores e 26GB de memória RAM.

4.3.2 Aplicação do Filtro Gaussiano para a Delimitação da Área de Interesse

A maioria dos resultados de última geração utiliza entradas multicanal (por exemplo, imagens RGB, fluxos ópticos, mapas de aceleração) (CHENG; CAI; LI, 2021). No entanto, a capacidade das arquiteturas CNNs 3D existentes é extremamente limitada com alto custo computacional e demanda de memória, dificultando o treinamento de uma CNN 3D muito profunda (QIU; YAO; MEI, 2017). Na pesquisa realizada por Qiu, Yao e Mei (2017), por exemplo, é proposta uma simulação de convoluções 3D com convoluções espaciais 2D mais conexões temporais 1D.

Na tentativa de reduzir o impacto de características localizadas em área de baixo interesse das cenas, na pesquisa desenvolvida por Nasaruddin et al. (2020), é apresentada uma técnica de detecção da área de atenção visual a partir da remoção de *background* de vídeos de videovigilância, de forma que sejam submetidas para o modelo CNN 3D apenas as regiões de interesse obtidas a cada quadro. Em outras palavras, a região de baixo interesse ficará desfocada e não produzirá nenhuma característica visual importante durante o processo de extração, treinamento e inferência. Nessa abordagem, é utilizado um algoritmo que incorpora a subtração de *background* com filtro bilateral (filtro para a suavização do ruído) baseada em texturização, visando a extração das regiões de interesse.

Baseado na ideia anterior, para obter esse efeito de redução de detalhes de baixo interesse nos quadros submetidos ao modelo, nesta pesquisa é utilizado o desfoque/suavização gaussiana, como resultado de uma convolução de uma imagem com uma função gaussiana. Dessa maneira, serão utilizadas operações espaciais para filtrar determinadas frequências de componentes da matriz da imagem, de forma que ainda sejam mantidas informações relevantes, como bordas e contornos. Um exemplo de filtro gaussiano aplicado em toda imagem pode ser visualizado na Figura 4.8

A aplicação do filtro gaussiano é a convolução de uma imagem com a função gaussiana,

junho de 2022.

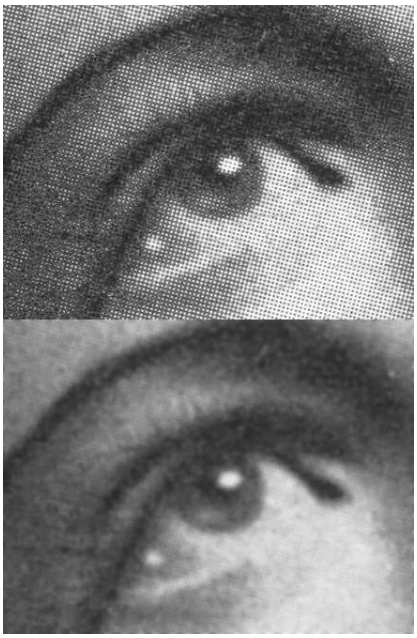

Figura 4.8: Exemplo do filtro gaussiano aplicado a uma imagem. Imagem original (acima) e imagem com o filtro gaussiano (abaixo). Fonte: Wikipedia (2022).

representada pela Equação 4.1 em sua variação bidimensional, para calcular a transformação a ser aplicada em cada pixel da imagem. Dessa forma, é obtida uma distribuição de dados de uma matriz gaussiana, que é utilizada como máscara durante a convolução aplicada à imagem original.

$$G(x, y) = \frac{1}{2\pi\sigma^2}\epsilon^{-\frac{x^2+y^2}{2\sigma^2}}, \tag{4.1}$$

onde:

- x: conjunto com n valores, tal que $-\infty < x < \infty$, que representa a distância da origem no eixo horizontal;
- G: distribuição gaussiana dos valores de X, onde X representa o conjunto de pares ordenados (x, y);
- y: distância da origem no eixo vertical;
- σ: largura do *kernel* gaussiano, isto é, o desvio padrão[4] dos valores de X, tal que $\sigma > 0$;

[4]O desvio padrão influencia como os *pixels* vizinhos do *pixel* de interesse afetam o resultado dos cálculos, portanto, quanto maior o valor, maior será o efeito de suavização.

Baseado na abordagem de Nasaruddin et al. (2020) de utilizar uma região de atenção em cada quadro, como entrada para o processo de aprendizagem do modelo e aplicar o efeito de desfoque no restante da área de baixo interesse, nesta pesquisa, também será utilizada a delimitação de uma área de interesse em cada quadro.

Para isso, como ponto de partida, a proposta do modelo denominado por "*Only RGB*" em (CHENG; CAI; LI, 2021) será reproduzida e retreinada no mesmo conjunto de dados, respeitando a divisão de treinamento e teste. Para o desfoque da área de baixo interesse, será utilizada a biblioteca OpenCV (OPENCV, 2022) através da função *GaussianBlur* utilizando *kernel* gaussiano com tamanho empiricamente definido para (21, 21) e desvio padrão nas direções de X e Y calculados também a partir do tamanho do *kernel*[5]. Na Figura 4.9, é possível visualizar o filtro gaussiano aplicado em uma área de baixo interesse.

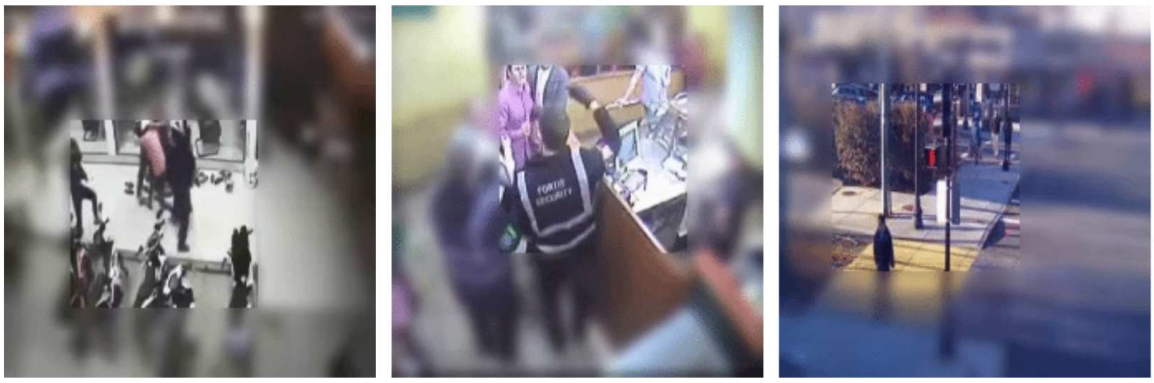

Figura 4.9: Exemplos de área de interesse de imagens evidenciadas pelo filtro gaussiano. Fonte: A autora (2022).

4.3.3 O Uso de Precisão Mista para a Redução do Uso de Memória VRAM

O *IEEE Standard for Floating-Point Arithmetic* (IEEE 754) (ZURAS et al., 2008) é um padrão técnico que define formatos aritméticos (ponto flutuante, binários, finitos, infinitos e etc.) e de intercâmbio (codificações), regras de arredondamento, operações e tratamento de exceções (divisão por zero, *overflow* e etc.). Dentre os tipos de formatos aritméticos, estão tipos de dados, tais como: *binary16*, *binary32*, *binary64*, *binary128*, *binary256*, *decimal32*,

[5] *Gaussian Blurring*. Disponível em: <https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html> Acesso em: 12 de junho de 2022.

decimal64 e *decimal128*.

Neste protocolo, é definido que o tipo de dado *float16*, conhecido como *half precision* (deve ser formado por até 11 bits de casas decimais e 5 bits para o expoente, gerando 2^{10} possibilidades de dígitos binários). O tipo de dado *float32*, também chamado por *single precision* pode possuir até no máximo 24 bits para casas decimais e 8 para o expoente), resultando em 2^{23} possibilidades de combinações de dígitos binários.

Atualmente, a maioria dos modelos usa o tipo de dado *float32*, que consome 32 bits de memória (GOOGLE, 2022). Aceleradores modernos podem executar operações mais rapidamente nos tipos de dados de 16 bits, pois possuem unidades de hardware especiais (*Tensor Cores*) em Unidades Gráficas de Processamento (*Graphics Processing Unit* - GPUs) NVIDIA para acelerar as multiplicações e convoluções da matriz *float16*. Portanto, a execução de cálculos de 16 bits e a leitura em memória podem ser realizadas mais rapidamente (GOOGLE, 2022).

As operações de acumulação múltipla (*multiply-and-accumulate* - MAC) são operações que realizam multiplicação de valores e adição do produto resultante em um acumulador. Os neurônios de uma DNN, são formados por operações MAC fundamentais e capazes de computar e atualizar os pesos sinápticos, que representam o grau de importância de uma determinada entrada em relação aquele neurônio. No entanto, o custo computacional requerido pelas unidades MAC é elevado (LANGROUDI et al., 2019).

Neurônios são unidades de processamento simples, que recebem um conjunto de valores de entrada (x), que representam as características de amostras do conjunto de treinamento para computar valores de predição ($\hat{y}$). Cada unidade possui seu próprio conjunto de parâmetros formados pelo vetor de pesos sinápticos (W) e um viés (b), que sofrem alterações durante o processo de aprendizado. A atualização dos pesos e viés dos neurônios é realizada com base no erro conforme o resultado gerado. Esse procedimento encontra-se ilustrado na Figura 4.10 e é descrito matematicamente na Equação 4.2.

$$y = \sigma(\sum_{i} w_i x_i + b) \tag{4.2}$$

Após esse processo linear, é realizada uma transformação não linear pela função de ativação (σ), possibilitando que a rede neural possa aprender mais do que relações lineares entre as variáveis dependentes e independentes, isto é, aprender e executar tarefas mais

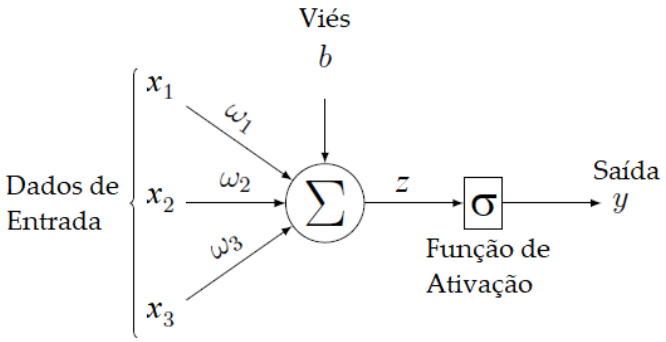

Figura 4.10: Representação de um neurônio artificial. Fonte: (GIBARU, 2019).

complexas. Dessa forma, as funções de ativação (σ) decidem se a informação recebida pelo neurônio é relevante ou se deve ser ignorada, mais genericamente, se um neurônio deve ser ativado ou não.

Portanto, uma vez que todas as amostras do conjunto de treinamento foram submetidas ao processo anteriormente descrito, é possibilitado que os parâmetros de peso e viés finais sejam devidamente ajustados através de um algoritmo de otimização, chamado descida do gradiente. A descida de gradiente é uma técnica de otimização que modifica os pesos iterativamente para cada dado de treinamento, conforme o erro observado (GIBARU, 2019). Na pesquisa desenvolvida por Nandakumar et al. (2020), é evidenciado que é possível acelerar o treinamento de DNNs reduzindo a precisão dos valores de pesos usados das operações MAC, desde que sejam mantidas informações de gradiente em precisão alta.

Conforme Langroudi et al. (2019), na tentativa de solucionar a lacuna descrita, foram desenvolvidas algumas iniciativas voltadas para comprimir o tamanho dos modelos e reduzir os requisitos de computação, tais como: poda de redes neurais profundas (YAZDANI et al., 2018), destilação e aritmética de baixa precisão (NANDAKUMAR et al., 2020).

Nesse contexto, os métodos de precisão mista combinam o uso de diferentes formatos numéricos em uma carga de trabalho computacional (NVIDIA, 2022). Mais especificamente, a precisão mista representa a quantidade de bits que podem ser alocados para casas decimais de um *float*. A precisão mista (GOOGLE, 2022) utiliza uma mistura entre os tipos de dados *float16* e *float32*, de maneira que dois aspectos importantes sejam mantidos: os cálculos das camadas devem ser realizados em *float16* e as variáveis em *float32*, com o objetivo de manter a estabilidade numérica[6].

[6]Estabilidade Numérica. O termo “estabilidade numérica” refere-se a como a qualidade de um modelo é

Para manter a mesma qualidade, alguns cálculos e variáveis precisam ainda permanecer em *float32* devido a problemas relacionados a *underflow* e *overflow*. Como os tipos de dados *float16* e *float32* possuem faixas de valores divergentes, de maneira que a faixa de valores do tipo *float16* é inferior a do tipo *float32*, é possível que a multiplicação entre valores do tipo *float16* resulte em alguma instância de resultado superior ao limite da faixa permitida para o tipo *float16*. Desta maneira, o resultado obtido ultrapassará o limite máximo, causando o que é chamado de *overflow*.

Também é possível ocorrer o cenário inverso, isto é, a obtenção de valores inferiores ao intervalo delimitado para o tipo de dado em questão. Um exemplo desse caso durante o treinamento de um modelo é quando a função de perda tende a 0 (quando a representação da faixa de valores para o tipo *float16* possuir um intervalo inferior com início em 0). Esse cenário é chamado de *underflow* e para evitá-lo, no caso do exemplo, é importante realizar o dimensionamento da escala dos valores de perda do modelo.

Existem vários benefícios em usar formatos numéricos com menor precisão do que o ponto flutuante de 32 bits (NVIDIA, 2022). A utilização de técnica de precisão mista possibilita uma menor alocação e uso de recursos de memória, permitindo um treinamento mais rápido em muitos modelos e de maneira igualitária em relação às métricas de avaliação dos mesmos. O Keras fornece uma API que permite a combinação dos tipos *float16* com *float32*. O uso dessa API pode melhorar o desempenho em mais de 3 vezes em GPUs modernas e 60% em TPUs, sem causar impacto nas métricas de avaliação do modelo (GOOGLE, 2022).

Para utilizar a precisão mista do Keras[7], será necessário criar uma política especificando os tipos de dados que devem ser utilizados, seja global através do *tf.keras.mixed_precision.Policy* ou local diretamente no construtor da função *Softmax*[8] do modelo. A representação da arquitetura utilizando precisão mista pode ser visualizada na Figura 4.11, na área em verde-claro é aplicada a precisão mista *float16* e na área em verde-escuro *float32*.

Devido a limitações de recursos computacionais, não foi possível realizar o experimento

afetada pelo uso de um tipo de dado de menor precisão em vez de um tipo de dados de maior precisão. "Uma operação é 'numericamente instável' em *float16* se executá-la em um desses tipos de dados, fizer com que o modelo tenha uma precisão de avaliação pior ou outras métricas em comparação com a execução da operação em *float32*". (GOOGLE, 2022)

[7] *Keras - Mixed Precision*. Disponível em: <https://keras.io/api/mixed_precision/>. Acesso em 14 de junho de 2022.

[8] *Softmax*. Função de ativação bastante utilizada em problemas de classificação, responsável por transformar as saídas das camadas para cada classe para valores entre 0 e 1.

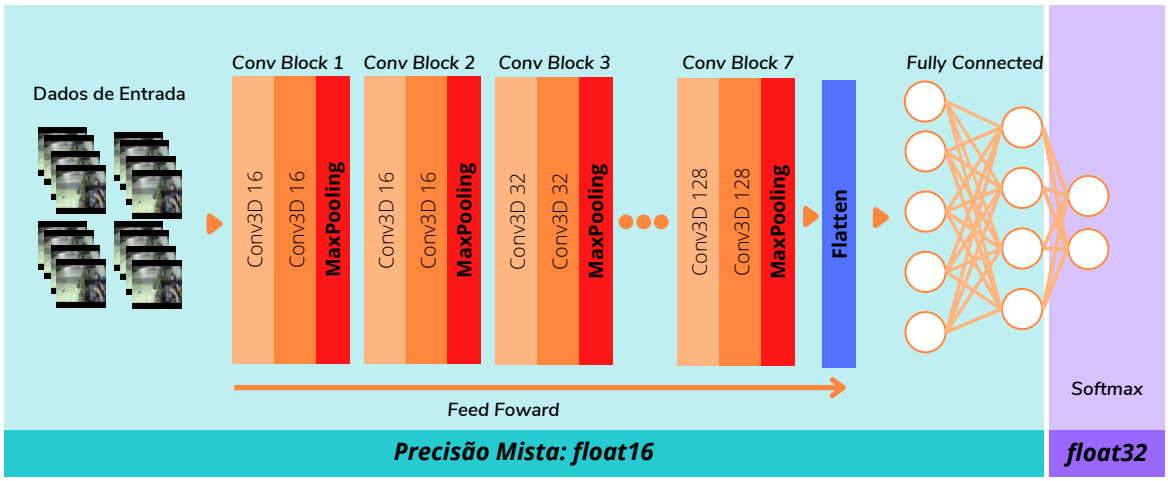

Figura 4.11: Representação da aplicação da precisão mista na arquitetura do modelo proposto por Cheng, Cai e Li (2021). Fonte: A autora (2022).

considerando o modelo *baseline* sem precisão mista com a etapa de desfoque gaussiano nos quadros de vídeo. Os resultados obtidos pela metodologia proposta podem ser visualizados na Seção 5.

4.3.4 Definição das Métricas de Avaliação

Nesta Seção, são definidas as métricas de avaliação para os experimentos. Na presente proposta, o conceito de classe positiva está relacionado a classe correspondente às amostras de vídeos que contém eventos violentos, enquanto a classe negativa corresponde a eventos normais.

As métricas frequentemente utilizadas para avaliar o desempenho do reconhecimento de atividades humanas são F1, acurácia média, precisão e revocação. Antes de sumarizar e descrever essas métricas, é importante definir os seguintes conceitos com aplicação no contexto de detecção de violência.

- Verdadeiro Positivo (*True Positive - TP*): Ações violentas reais e preditas corretamente.
- Falso Negativo (*False Negative - FN*): Ações violentas reais preditas incorretamente como não violentas.
- Verdadeiro Negativo (*True Negative - TN*): Ações não violentas reais e preditas corretamente.

- Falso Positivo (*False Positive - FP*): Ações não violentas reais preditas incorretamente como violentas.

Desta forma, no cenário de reconhecimento de atividades violentas, um alto impacto de Falsos Positivos pode ocasionar muitos casos de alarmes falsos (detecção de violência em cenas que não existe violência), assim como, um alto impacto de Falsos Negativos pode indicar um modelo que não consegue identificar as atividades violentas presentes em vídeo. Portanto, para avaliar o desempenho do método proposto, além da matriz de confusão contendo os indicadores supramencionados, são utilizadas também as seguintes métricas:

- Acurácia: quantidade de previsões corretas em relação ao número total de amostras. Seu uso é indicado quando a quantidade de amostras de classes é balanceada.

  Acurácia $= \frac{TP+TN}{TP+TN+FP+FN}$

- Precisão: probabilidade dos casos em que uma atividade prevista como positiva ser de fato, a sua real atividade. A precisão é uma boa medida para ser utilizada quando os custos dos Falso Positivos são impactantes.

  Precisão $= \frac{TP}{TP+FP}$

- Revocação: calcula quantos dos Positivos Reais o modelo consegue classificar como Positivo (Verdadeiro Positivo). A revocação é uma boa medida para ser utilizada quando os custos dos Falsos Negativos são impactantes.

  Revocação $= \frac{TP}{TP+FN}$

- F1-Score: é calculado a partir da precisão e da revocação, considerando a média harmônica entre essas duas métricas. Atinge seu melhor valor em 1 e pior em 0.

  F1 $= \frac{2\times Precision\times Recall}{Precision+Recall} = \frac{2\times TP}{2\times TP+FP+FN}$

- Área Sob a Curva (*Area Under Curve* - AUC): usada para avaliar o desempenho do modelo na distinção entre classes positivas e negativas. Quanto mais próximo o valor de AUC de 1, melhor o desempenho.

- Perda: quantifica o erro produzido pelo modelo. O valor de perda será calculado pela função de custo *categorical_crossentropy*, a qual é normalmente utilizada em problemas

que envolvem modelos de classificação multiclasse, isto é, com um ou mais rótulos de saída, por meio de um esquema de codificação categórico[9]. Essa função de perda é empregada no modelo *baseline* proposto na pesquisa desenvolida por Cheng, Cai e Li (2021).

Função de Perda

A função de perda pode ser utilizada para avaliar o desempenho de um modelo tanto no conjunto de dados de treinamento, quanto no de validação. A perda de treinamento é uma medida que quantifica como o modelo se ajusta aos dados de treinamento, já o valor de perda de validação, quantifica o desempenho do modelo em um conjunto de dados composto por amostras não utilizadas no processo de aprendizagem, ou seja, em novos dados. A análise do valor de perda também é utilizada para diagnosticar possíveis ajustes a serem implementados, tais como *overfitting*[10] e *underfitting*[11]. Um bom modelo[12], normalmente, apresenta valores de perda de treinamento e validação que diminuem e se estabilizam em um ponto específico.

[9]Codificação Categórica. Esquema de codificação que converte um vetor de inteiros para uma matriz binária por categoria. Disponível em: <https://www.tensorflow.org/api_docs/python/tf/keras/utils/to_categorical>. Acesso em: 14 de junho de 2022.

[10]*Overfitting*. Quando o valor de perda de validação é maior que a perda de treinamento, indicando um modelo sobreajustado que não consegue generalizar para novos dados.

[11]*Underfitting*. Onde ambos os valores de perda são altos, indicando um modelo subajustado que não consegue modelar com precisão os dados e treinamento.

[12]Training and Validation Loss in Deep Learning. Dsiponível em: <https://www.baeldung.com/cs/training-validation-loss-deep-learning>. Acesso em: 14 de junho de 2022.

Capítulo 5

Resultados e Discussão

Nesta seção, são apresentados os resultados do modelo de detecção de violência, como também são apresentadas discussões e análises estatísticas comparativas para avaliar o grau de significância das contribuições apresentadas em ambos os experimentos e a avaliação da abordagem proposta em bases desafiadoras do estado da arte. Por fim, são discutidos os desafios e lacunas ainda em aberto, assim como os benefícios apresentados pela proposta apresentada. Como os dados de treinamento e validação são balanceados, utilizaremos as métricas de acurácia e função de perda para avaliar o desempenho dos modelos.

As Figuras 5.1, 5.2 e 5.3 representam os valores de acurácia obtidos em treinamento e validação para os experimentos (1) *baseline* (CHENG; CAI; LI, 2021), (2) *MP*: *baseline* com precisão mista e (3) *MP+B*: *baseline* com precisão mista + área de interesse delimitada pelo filtro gaussiano, respectivamente. Já as Figuras 5.4, 5.5 e 5.6 representam os valores de perda obtidos em treinamento e validação para os experimentos (1) *baseline* (CHENG; CAI; LI, 2021), (2) *MP*: *baseline* com precisão mista e (3) *MP+B*: *baseline* com precisão mista + área de interesse delimitada pelo filtro gaussiano, respectivamente.

Para uma melhor visualização do desempenho do treinamento e validação do modelo, nas linhas das Figuras 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6 foi aplicado um efeito de suavização[1]. As linhas representam as tendências de valores obtidos em treinamento e validação do modelo. É possível visualizar as curvas sem aplicação do efeito de suavização, representadas por uma tonalidade da cor mais fraca. Nos experimentos (1) *baseline*, (2) *MP* e (3) *MP+B*, as fases de

[1] As linhas em cores mais claras dos gráficos representam a suavização da tendência original dos dados, submetidos a um efeito de *smoothing* de valor 0,35.

treinamento e validação tiveram comportamentos semelhantes para os três experimentos. Os principais comportamentos identificados foram:

- A curva de validação atinge um pico, respectivamente nas épocas 16, 13 e 20;
- O valor de acurácia do treinamento nas respectivas épocas acima é semelhante ou pouco inferior ao de validação.
- O desempenho da fase de validação inicia um decaimento após essas respectivas épocas, caracterizando um possível cenário de *overfitting*, enquanto na fase de treinamento continua em crescimento;
- Nas figuras correspondentes ao desempenho do valor de perda (Figuras 5.4, 5.5 e 5.6), é possível perceber que ao atingir as respectivas épocas supramencionadas, o valor de perda em validação tende a crescer, enquanto em treinamento continua a diminuir e tender a 0. Fato que fortalece a hipótese de *overfitting* a partir dessas respectivas épocas.

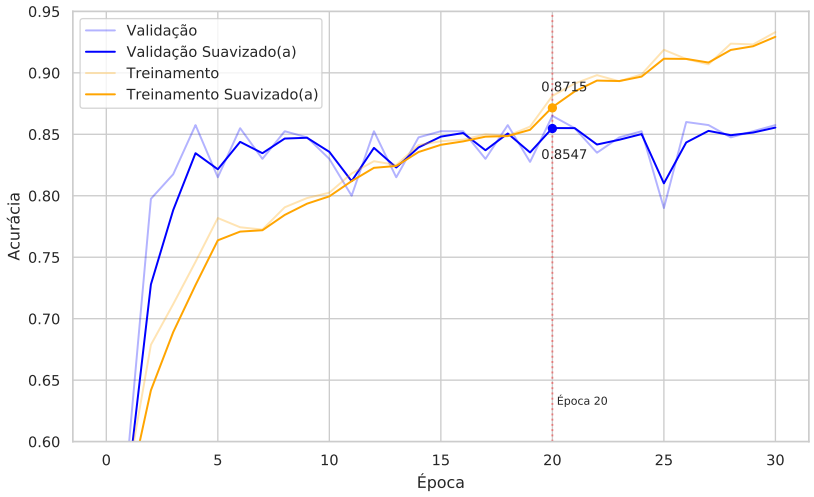

Figura 5.1: Experimento: *MP+B* - Acurácia por época (validação). Fonte: A autora (2022)

Neste caso, consideramos como a melhor versão do modelo aquela cuja época manteve sua perda em validação baixa e estável e a maior pontuação de acurácia em validação. Na Tabela 5.1, são sumarizados e apresentados, os resultados de cada um dos três experimentos considerando a versão com 30 épocas fixas e também a época que atingiu o maior pico de acurácia no cenário de validação. Nessa mesma tabela, são evidenciados além das épocas, valores das métricas de acurácia e valor de perda, a quantidade de memória utilizada durante

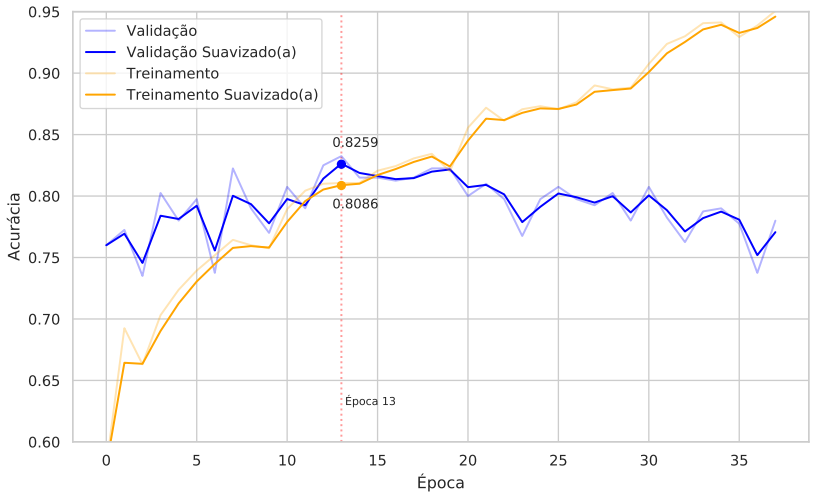

Figura 5.2: Experimento: *MP* - Acurácia por época (validação). Fonte: A autora (2022)

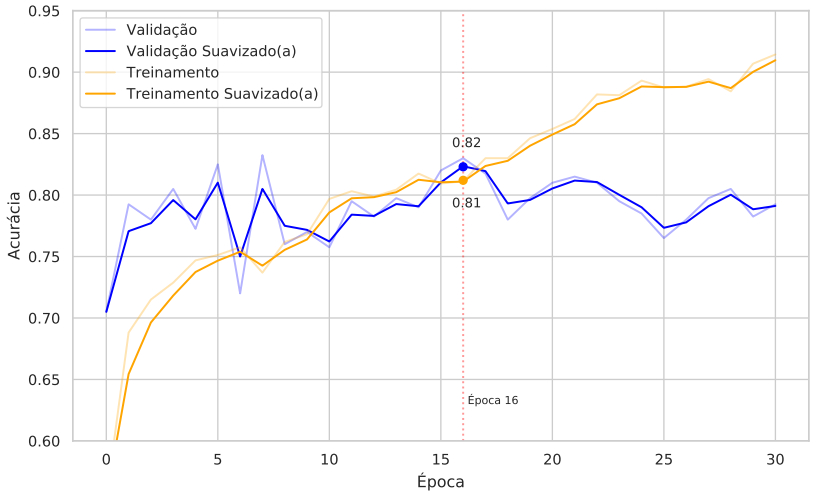

Figura 5.3: Experimento: *baseline* - Acurácia por época (validação). Fonte: A autora (2022)

a execução do treinamento e validação, representado pela unidade de MebiByte (MiB)[2], o tempo de treinamento e validação e o total de parâmetros.

Inicialmente, é possível observar uma divergência entre os valores apresentados por Cheng, Cai e Li (2021) (*paper*) e a reprodução realizada nesta pesquisa (*baseline* - 30 épocas). As possíveis hipóteses de causa dessas divergências podem ser:

1. A aleatoriedade dos parâmetros dos tipos de aumento de dados (etapa A5 e B5 das Figuras 4.1 e 4.7, respectivamente) utilizados durante o treinamento;

2. A aleatoriedade utilizada durante o corte dinâmico para selecionar 10 posições candidatas

[2] 1 Mebibyte (MiB) corresponde a 2^{20} bytes ou 1048576 bytes ou 1,04858 Megabyte.

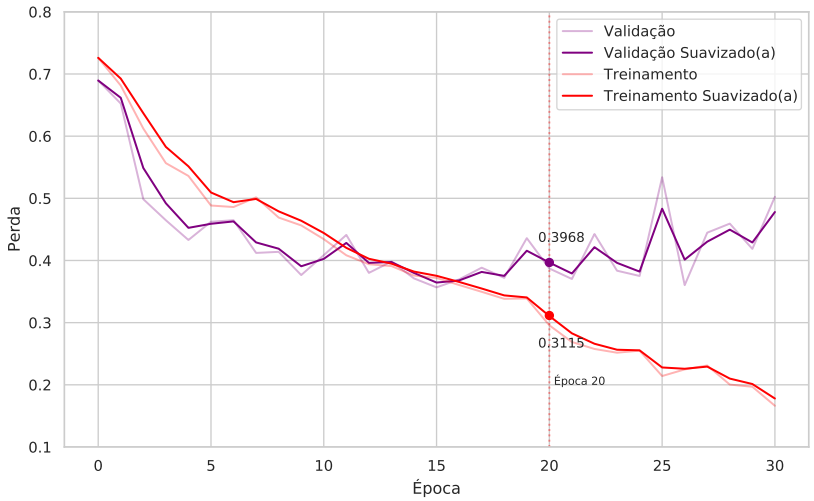

Figura 5.4: Experimento: *MP+B* - Valor de perda por época (validação). Fonte: A autora (2022)

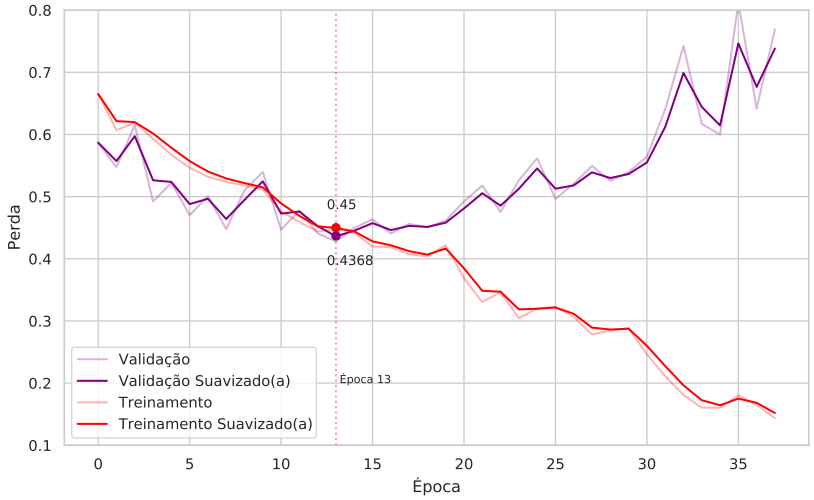

Figura 5.5: Experimento: *MP* - Valor de perda por época - Precisão Mista - (validação). Fonte: A autora (2022)

para a área do quadro de vídeo (linhas 16 e 17 do Algoritmo 1);

3. A reestruturação da sequência de experimentos, antes realizada durante o treinamento. No entanto, por questões de limitações de recursos computacionais, foi necessário segmentar essas etapas, realizando-as em uma etapa anterior ao início do treinamento.

No experimento considerando o modelo *baseline* (*baseline*) em comparação com os demais que utilizam precisão mista aplicada (*MP* e *MP+B*), é possível notar a redução, quase que em dobro, do uso de memória utilizada durante o treinamento e validação. Outro ponto

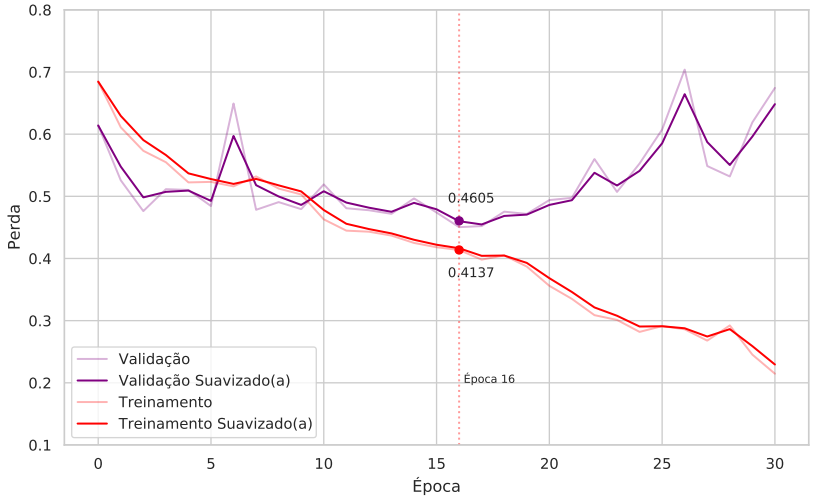

Figura 5.6: Experimento: *baseline* - Valor de perda por época (validação). Fonte: A autora (2022)

Experimento	**Acurácia *Train* (%)**	**Acurácia Validação (%)**	**Época**	**Perda *Train***	**Perda Validação**	**Memória (MiB)**	**Tempo**	**Total de Parâmetros**
paper	89,5	84,5	-	-	-	-	-	
baseline	91,44	79,25	30	0,2146	0,6744	15,981 MiB	06:15:42	**248.402**
	81,12	83	16	0,4137	0,4504		03:07:06	
MP	90,75	80,75	30	0,2462	0,5642	**8,787 MiB**	05:25:31	**248.402**
	81,06	83,25	**13**	0,4484	0,4279		**02:15:28**	
MP+B	**93,31**	85,75	30	**0,1663**	0,5025	**8,787 MiB**	05:47:11	**248.402**
	88,13	**86,50**	20	0,2957	**0,3871**		03:44:27	

paper: versão *baseline* (resultados reportados no *paper* original); ***baseline***: versão apresentada na Seção 4.1 (apenas reprodução do *paper*); ***MP***: versão *baseline* com precisão mista apresentada na Seção 4.3.3; ***MP+B***: versão com precisão mista + efeito gaussiano apresentada na Seção 4.3.2; ***Train***: Treinamento; **Memória**: Memória utilizada.

Tabela 5.1: Resultados experimentais

relevante e de comportamento semelhante a ser notado é o impacto no tempo utilizado para esse processo, considerando a versão dos modelos treinados em 30 épocas fixas. Nesse mesmo cenário, também nota-se que o experimento utilizando o desfoque gaussiano necessitou de um tempo superior ao que não utiliza o desfoque, mas ainda inferior ao experimento *baseline* sem a precisão mista.

Na Figura 5.7, ao analisar a matriz de confusão da etapa de validação, é possível evidenciar que tanto os Verdadeiros Positivos, quanto os Verdadeiros Negativos estão em quantidade superior quando comparados com os Falsos Negativos e Falsos Positivos. Isso significa que, o modelo é capaz de distinguir entre comportamentos violentos e não-violentos, pontuando mais em cenas de não-violência (90%) do que cenas-violentas (75,5%). No entanto, de modo

geral, a taxa de erro (tipo I e tipo II)[3] é baixa (Falso Positivo: 10% e Falso Negativo: 24,5%) se comparada com a taxa de acertos.

Para entender a capacidade de classificação do modelo por classe, uma análise mais aprofundada dos resultados, considerando as métricas de precisão, revocação e f1-score foi realizada. As métricas obtidas pelo experimento (*MP+B*) encontram-se na Tabela 5.2, o que novamente pode-se confirmar que de todas as cenas de violência no cenário de validação, o modelo conseguiu classificar corretamente em 75,5% dos casos. Por outro lado, considerando todas as classificações realizadas pelo modelo como cenas violentas, o modelo esteve correto em 83,30% das vezes. Dessa forma, o modelo possui uma precisão maior para a detecção de violência, do que para a detecção de cenas em que não há violência.

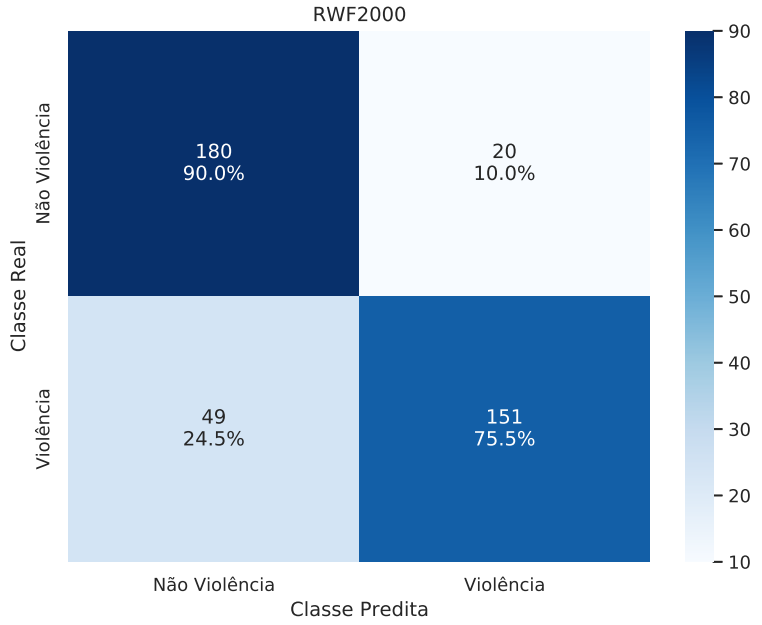

Figura 5.7: Matriz de confusão (validação). Fonte: A autora (2022).

5.1 Análise de Resultados

Até este ponto, obtivemos indícios de que as abordagens experimentais das versões de modelo com precisão mista (*MP+B*) são superiores ao método *baseline* em termos de valores

[3]Erro Tipo I: conhecido também como "falso alarme", em que o modelo prevê a amostra de entrada como classe positiva, mas a classe real é negativa (Falso Positivo); Erro Tipo II: caso em que a amostra de entrada é referente à classe positiva, mas o modelo classificou como classe negativa (Falso Negativo).

	Precisão	**Revocação**	**F1-score**
Não Violência	78,60%	90,00%	83,92%
Violência	88,30%	75,50%	81,40%
Macro	83,45%	82,75%	82,66%

Tabela 5.2: Resultados das métricas obtidas na fase de validação do modelo.

de acurácia e de perda. Para comprovarmos, será adotado um método estatístico para avaliar se as diferenças entre as abordagens é de fato, significativa.

Para isso, utilizaremos os resultados por época do experimento *baseline*, isto é, a reprodução do experimento "*RGB Only*" contido na pesquisa desenvolvida por Cheng, Cai e Li (2021), como o grupo de controle e referência. Inicialmente, os valores de acurácia por época obtidos serão usados como as amostras para comparação entre os experimentos. Também foi realizada uma segunda análise estatística, considerando os valores obtidos pela função de perda também por época entre os experimentos, para comparação entre o grupo de controle e os grupos de teste.

Considera-se a diferença das médias de valores de acurácia e de perda de cada grupo, como recurso para medir o tamanho do efeito. Desta forma, para as comparações experimentais dos grupos de teste *MP* e *MP+B*, são ambas realizadas com o mesmo grupo de controle, o *baseline*. A análise é apresentada por intermédio do gráfico de estimativa de *Cumming* [4], tanto para a análise dos valores de acurácia, quanto para os valores de perda.

Conforme Alves (2013), o *bootstrap* é uma técnica estatística, computacional intensiva de reamostragem, introduzida por Efron e Tibshirani (1994) em 1979, com finalidade de obter informações de características da distribuição de alguma variável aleatória. Segundo Claridge-Chang e Ho (2017), devido ao Teorema do Limite Central, as médias de amostras aleatórias independentes (como a reamostragem da diferença de médias) se aproximarão de uma distribuição normal, mesmo que a população subjacente não seja normalmente distribuída. Desta forma, nesta pesquisa será utilizado o T-Test como teste de hipótese estatística.

Dadas 5000 amostras reamostradas via *bootstrap*, os valores de *p-value* relatados são as probabilidades de observar os tamanhos do efeito, se a hipótese nula de diferença zero for verdadeira. Para cada *p-value* de permutação, foram realizadas 5000 remanejamentos

[4] *Shared control Cumming plot*. Gráfico que apresenta as diferenças médias entre um único grupo de controle e cada um dos grupos de intervenção. Disponível em: <https://www.estimationstats.com/#/user-guide/shared-control> Acesso em: 14 junho 2022.

dos rótulos de controle e teste. Os tamanhos de efeito e Intervalos de Confiança (ICs) são relatados como: tamanho do efeito [limite inferior da largura do IC; limite superior].

Para a geração das figuras e estimativas, foi utilizada a plataforma *Estimation Stats*[5] (HO et al., 2019) com a configuração de *shared control*. Para ambas as figuras, Figura 5.8 e Figura 5.9, a distribuição dos dados de entrada está localizada nos eixos superiores, enquanto as diferenças entre médias estão apresentadas como distribuições de amostragem *bootstrap* e localizadas nos eixos inferiores. Cada diferença média é representada como um ponto e cada intervalo de confiança de 95% é indicado pelas extremidades das barras de erro verticais.

5.1.1 Acurácia em Validação

A representação da análise estatística para os valores de acurácia atingidos por época e por experimento, durante a fase de validação do modelo, constam na Figura 5.8. Os tamanhos de efeito e Intervalos de Confiança (ICs) estão descritos a seguir:

- A diferença média não pareada entre *baseline* e a *MP* é 0,00675 com IC de 95% [-0,00625; 0,0197].
- A diferença média não pareada entre *baseline* e o *MP+B* é de 0,0307 com IC de 95% [-0,0124; 0,0516].

Ao analisar as estatísticas das amostras apresentadas anteriormente e a Figura 5.8, por meio de inferência estatística via estimativa de intervalo de confiança, com o intervalo de 95% de confiança, é estimado que exista uma diferença positiva com o valor de 0,0307 para a população, entre o desempenho da abordagem *MP+B* e a abordagem *baseline*. No entanto, embora tenha sido obtida uma diferença positiva entre as abordagens, ao avaliar o intervalo de confiança [-0,0124; 0,0516], percebe-se que é possível que no intervalo de confiança estejam contidos casos em que há diferença negativa entre as médias (valores do IC abaixo de 0), como também casos em que não há diferença entre as médias, pois o 0 encontra-se inserido neste intervalo.

Considerando o cenário acima e a falta de evidências estatísticas suficientes para identificar o grau de significância dessa diferença positiva, será explorado também o método de inferência

[5] *Estimation Stats*. Disponível em: <https://www.estimationstats.com/>. Acesso em 14 junho 2022.

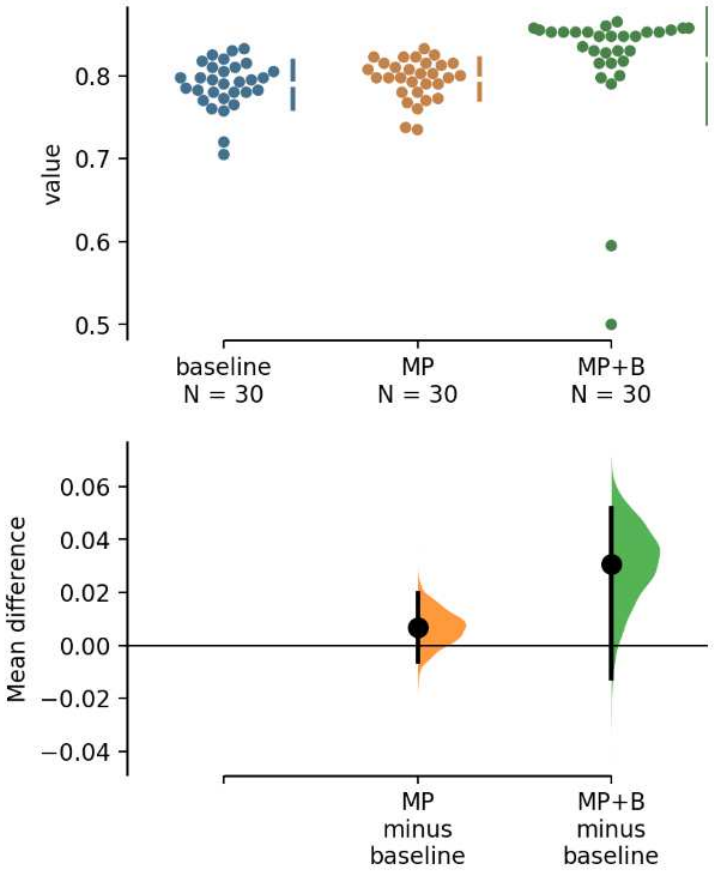

Figura 5.8: Diferença entre as médias do grupo de controle e os experimentos. Fonte: A autora (2022) com base na pesquisa desenvolvida por Ho et al. (2019) (2019). Legenda: "*Mean difference*": diferença das Médias; "*value*": valor; "*minus*": menos; "*baseline*": versão apresentada na Seção 4.1; "*MP*": versão *baseline* com precisão mista apresentada na Seção 4.3.3; "*MP+B*": versão com precisão mista + efeito gaussiano apresentada na Seção 4.3.2. Fonte: Imagem gerada através do *estimationstats.com* (HO et al., 2019).

estatística via teste de hipóteses usando T-test (normalmente utilizado para comparar médias amostrais). As hipóteses encontram-se descritas na Tabela 5.3 e os *p-values* obtidos são:

- Para os métodos *baseline* e o *MP*, o *p-value* do T-test de permutação bilateral é 0,332;
- Para os métodos *baseline* e o *MP+B*, o *p-value* do T-test de permutação bilateral é 0,04.

Tipo de Hipótese	**Descrição**
Hipótese Nula	Não há diferença entre os dois grupos.
Hipótese Alternativa	As médias de diferenças são diferentes.

Tabela 5.3: Definição das hipóteses para as amostras de acurácia

Conclusão: o *p-value* no valor de 0,04 indica que a hipótese nula deve ser rejeitada. Isso significa que, considerando 95% de confiança, em um total de apenas 4% das simulações geradas, o modelo nulo gera um efeito igual ou superior ao efeito da amostra do modelo real. Contudo, para 96% (a maioria) dos casos ocorre o cenário inverso, isto é, o modelo nulo gera um efeito menor que o modelo real. Portanto, pode ser concluído que existe uma diferença positiva e significativa entre as abordagens *MP+B* e *baseline*, tornando a abordagem *MP+B* plausível de possuir o melhor desempenho entre os métodos apresentados. Nesse mesmo sentido, outra conclusão que pode ser considerada para as diferenças de médias entre os métodos *baseline* e a *MP* é que com *p-value* de 0,332 não há evidências estatísticas suficientes para que a hipótese nula seja rejeitada.

5.1.2 Valor de Perda em Validação

A representação da análise estatística para os valores de perda atingidos por época e por experimento, durante a fase de validação do modelo, constam na Figura 5.9. Os tamanhos de efeito e Intervalos de Confiança (ICs) estão descritos a seguir:

- A diferença média não pareada entre *baseline* e a *MP* é -0,022 com 95%CI [-0,0517; 0,00342];
- A diferença média não pareada entre *baseline* e o *MP+B* é de -0,0882 com 95%CI [-0,121; -0,0503].

O padrão de valores avaliados relacionados à métrica de perda é levemente diferente da métrica de acurácia, enquanto buscamos valores altos definidos no intervalo [0-100] para a acurácia, buscamos valores baixos de perda definidos no intervalo [0-1]. Desta forma, a diferença das médias para a métrica de acurácia esperada é uma diferença positiva, já para a métrica de perda é esperada uma diferença negativa, ou seja, que os valores de perda encontrados nas propostas *MP* e *MP+B* sejam menores que a da proposta *baseline*.

Ao analisar as estatísticas das amostras apresentadas anteriormente e a Figura 5.9, através de inferência estatística via estimativa de intervalo de confiança, com o intervalo de 95% de confiança, é estimado que exista uma diferença negativa com o valor de -0.0503 para a população, entre o desempenho da abordagem *MP+B* e a abordagem *baseline*. Nesse caso,

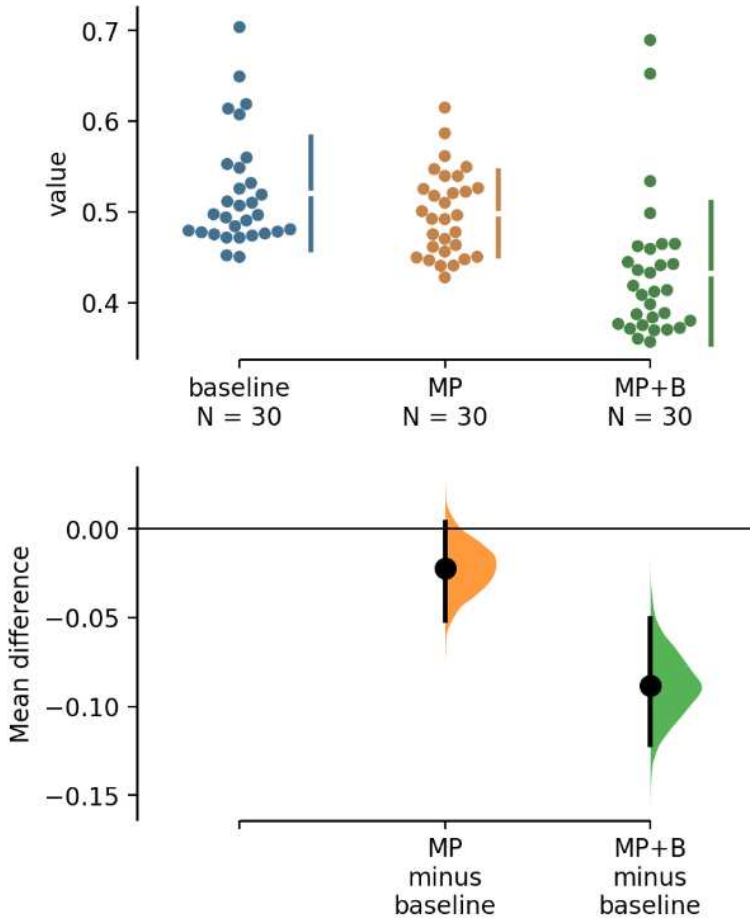

Figura 5.9: Diferença entre as médias do grupo de controle e os experimentos. Fonte: A autora (2022) com base na pesquisa desenvolvida por Ho et al. (2019) (2019). Legenda: "*Mean difference*": diferença das Médias; "*value*": valor; "*minus*": menos; "*baseline*": versão apresentada na Seção 4.1; "*MP*": versão *baseline* com precisão mista apresentada na Seção 4.3.3; "*MP+B*": versão com precisão mista + efeito gaussiano apresentada na Seção 4.3.2. Fonte: Imagem gerada através do *estimationstats.com* (HO et al., 2019).

como na faixa de valores no intervalo de confiança não está contido o valor nulo (0) e nem valores positivos, têm-se evidências estatísticas que indicam que essa diferença negativa seja de fato, significativa. É plausível que o método *MP+B* novamente possua o melhor desempenho entre os demais métodos apresentados.

Contudo, não é possível obter evidência similar para as diferenças das médias entre os métodos *baseline* e a *MP* e, por isso, será necessário recorrer ao método de inferência estatística via teste de hipóteses via T-test para identificar o nível de significância dessa diferença. As hipóteses encontram-se descritas na Tabela 5.4 e os *p-values* obtidos são:

- Para os métodos *baseline* e o *MP*, o *p-value* do T-test de permutação bilateral é 0,124;
- Para os métodos *baseline* e o *MP+B*, o *p-value* do T-test de permutação bilateral é 0.

Tipo de Hipótese	**Descrição**
Hipótese Nula	Não há diferença entre os dois grupos.
Hipótese Alternativa	As médias de diferenças são diferentes.

Tabela 5.4: Definição das hipóteses para as amostras de valor de perda

Conclusão: o *p-value* no valor de 0 indica que a hipótese nula deve ser rejeitada. Isso significa que, considerando 95% de confiança, nenhuma das simulações geradas (0%), o modelo nulo gera um efeito igual ou superior ao efeito da amostra do modelo real e para todos os casos (100%) ocorre o cenário inverso, isto é, o modelo nulo gera um efeito menor que o modelo real. Portanto, pode ser comprovado, mais uma vez, que o método *MP+B* é plausível de possuir o melhor desempenho entre os métodos avaliados. Nesse mesmo sentido, outra conclusão que pode ser considerada para as diferenças de médias entre os métodos *baseline* e a *MP* é que com *p-value* de 0,124 a hipótese nula deve ser aceita para este caso. Portanto, não há evidências estatísticas suficientes para que a hipótese nula seja rejeitada.

Conclusão Final: após as análises e explorações estatísticas, conclui-se que o método proposto na Seção 4.3.2 (*MP+B*), que utiliza a abordagem de desfoque por filtro gaussiano na região de baixo interesse com a técnica de precisão mista possui desempenho superior ao próprio método *baseline* proposto na pesquia desenvolvia por Cheng, Cai e Li (2021) descrito na Seção 4.1.

5.1.3 Outros Cenários

Para a comparação de resultados entre as abordagens propostas nesta pesquisa e as abordagens encontradas no estado da arte, foram considerados resultados de abordagens em que, na metodologia de treinamento, estivesse contido ao menos um conjunto de dados com cenas de violência capturadas de câmeras de videovigilância. A comparação de resultados obtidos entre este trabalho e outras abordagens do estado da arte pode ser observada na Tabela 5.5.

É importante salientar que, além dos dados apresentados na Tabela 5.5, na pesquisa realizada por Mugunga et al. (2021) foi alcançado um resultado de acurácia de 92,4%

(utilizando *VGG-16* pré-treinada e blocos *ConvLSTM*) no conjunto de dados *RWF-2000* (CHENG; CAI; LI, 2021), mesmo com a abordagem sendo treinada em um conjunto de dados de cenas de filme, o *Movies Fights* (NIEVAS et al., 2011). Outro caso semelhante, também ocorreu na pesquisa desenvolvida por Aktı, Tataroğlu e Ekenel (2019), onde foi utilizado o conjunto de dados *Hockey* (NIEVAS et al., 2011) para treinamento e variações experimentais (*Bi-LSTM*, *Xception*, *attention*) alcançando resultados entre 68% e 72% na base e dados *Fight Surveillance Camera* (AKTı; TATAROğLU; EKENEL, 2019).

Pesquisa	***RWF-2000***	***Hockey***	***Movies***	***Violent Flows***	***SCF***	**Base de Treino**
Ding et al. (2014)		91				*RWF-2000*
Cheng, Cai e Li (2021) (RGB)	84,5					*RWF-2000*
Cheng, Cai e Li (2021) (OPT)	75,5					*RWF-2000*
Cheng, Cai e Li (2021) (P3D)	87,25	98	100	88,87		*RWF-2000*
Cheng, Cai e Li (2021) (C3D)	85,75					*RWF-2000*
Ullah et al. (2021)		98		98,2	74	*Hockey*, *Violent Flows* e *SCF*
Baseline*	79,25	65,4	72,64	56,67	68,18	*RWF-2000*
MP*	83,25	73,8	78,61	62,08	69,23	*RWF-2000*
MP+B*	86,5	80,2	78,61	64,17	73	*RWF-2000*

***** Experimentos desenvolvidos neste trabalho.
SCF. *Surveillance Camera Fight*; ***baseline***: versão apresentada na Seção 4.1 (reprodução de Cheng, Cai e Li (2021)); ***MP***: versão *baseline* com precisão mista apresentada na Seção 4.3.3; ***MP+B***: versão com precisão mista + efeito gaussiano apresentada na Seção 4.3.2; **AU**: amostragem uniformizada;

Tabela 5.5: Avaliação dos resultados de acurácia em cenários relacionados

Considerando o aspecto de anotação de dados ao nível de vídeo *trimmed* adotado no escopo do trabalho *baseline*, a avaliação de resultados do método proposto nesta pesquisa e a comparação de resultados com abordagens do estado da arte foi realizada nas bases *RWF-2000* (CHENG; CAI; LI, 2021), *Hockey* (NIEVAS et al., 2011), *Movies Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e *Fight Surveillance Camera* (AKTı; TATAROğLU; EKENEL, 2019). Além disso, essas bases são constituídas, em sua maior parte, de cenas de violência humana.

Devido à escassez de bases de dados de cenas de comportamentos humanos em videovigilância ao nível de anotação *trimmed*, algumas bases de dados com cenas de comportamentos violentos em outros cenários também costumam ser avaliadas pelo estado da arte, tais como, *Hockey* (NIEVAS et al., 2011), *Movies Fights* (NIEVAS et al., 2011) e *Violent Flows* (HAS-

SNER; ITCHER; KLIPER-GROSS, 2012). É esperado que o modelo desenvolvido nesta pesquisa, perca desempenho ao ser avaliado em cenários desafiadores e possivelmente desconhecidos para o mesmo, como em cenas de jogos de hóquei (*Hockey*), multidões (*Violent Flows*) e cenas de filmes (*Movies Fights*). No entanto, embora os resultados obtidos pelos métodos propostos nesta pesquisa sejam inferiores aos obtidos por abordagens do estado da arte pra os conjuntos de dados, *Movies Fights*, *Hockey* e *Violente Flows*, um ponto relevante, é a evidência de que a contribuição do uso da técnica de precisão mista combinada com o método de delimitação de área de interesse por meio do filtro gaussiano supera os resultados da abordagem *baseline*, mesmo em cenários desafiadores e possivelmente desconhecidos.

Em especial, as bases de dados *Fight Surveillance Camera* (AKTı; TATAROğLU; EKENEL, 2019) e *RWF-2000* (CHENG; CAI; LI, 2021) são constituídas de cenas de violência humana capturadas de vídeos de câmeras de videovigilância e, por isso, representam melhor o cenário da problemática, para qual o escopo deste trabalho foi construído. Dessa forma, essas bases de dados são mais adequadas para serem utilizadas na avaliação da finalidade do método proposto neste trabalho. A exploração de possíveis falhas encontra-se na seção posterior.

5.2 Análise de Falhas

Com o intuito de identificar possíveis problemas e vieses no classificador proposto no presente trabalho, foi realizada uma avaliação das predições erradas, sendo essas as predições FP (Falsas Positivas) e as predições FN (Falsas Negativas). Predições FP são caracterizadas pela atribuição errônea do modelo, que determina que o vídeo em questão pertence à classe positiva (violência), no entanto, a classe real do vídeo é a classe negativa (não violência). Já as predições FN são caracterizadas pela atribuição errônea do modelo, que determina que o vídeo em questão pertence à classe negativa (não violência), no entanto, a classe real do vídeo é a classe positiva (violência). Para a análise de falhas a seguir, foi coletada uma amostra com um total de 150 amostras preditas incorretamente pelo modelo. As amostras foram coletadas aleatoriamente de cada uma das seguintes bases (30 amostras por base): *RWF-2000* (CHENG; CAI; LI, 2021) (conjunto de teste), *Hockey* (NIEVAS et al., 2011), *Movies Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e

Fight Surveillance Camera (AKTı; TATAROğLU; EKENEL, 2019).

Inicialmente, foram avaliadas as taxas de FN e FP por base de dados e consolidadas na Figura 5.10. Para as bases *RWF-2000* (conjunto de teste) (CHENG; CAI; LI, 2021), *Hockey* (NIEVAS et al., 2011), *Movies Fights* (NIEVAS et al., 2011), *Violent Flows* (HASSNER; ITCHER; KLIPER-GROSS, 2012) e *Fight Surveillance Camera* (AKTı; TATAROğLU; EKENEL, 2019) foram obtidos casos de FN's quase o triplo do valor quando comparadas aos casos de FP's, isto significa que, considerando a amostra analisada, nessas bases o modelo classificou erroneamente mais vezes casos de erros do tipo crítico (cenas de violência despercebidas), do que tipos de erro do tipo "alarme falso" (cenas de não violência apontadas como cenas violentas). No entanto, o cenário se torna diferente ao observar a base *Violent Flows*, nesse caso, a hipótese é de que em cenas de multidão, devido à movimentação constante de indivíduos, a classificação esteja sendo impactada pela "textura" holística de movimento na área de interesse.

Para uma análise mais aprofundada a fim de se obter hipóteses de possíveis causas de falhas encontradas, as amostras coletadas foram submetidas à uma etapa de categorização de possíveis tipos de causas de erros a partir da observação da cena. Para isso, definimos sete possíveis tipos de causas de falhas que podem afetar o desempenho do modelo, sendo elas: Movimento Abrupto (Figura 5.11), Multidão (Figura 5.12), Movimento ou Interferência na Câmera (Figura 5.13), Baixa Qualidade (Figura 5.14), Área do Filtro Gaussiano Mal Localizada (Figura 5.15), Rótulo Original Errado (contém anotação divergente do que é apresentado nas cenas do vídeo, Figura 5.16), Desconhecido (Figura 5.17).

É possível notar, ao visualizar a Figura 5.18 que, de maneira geral, a maioria dos erros críticos (tipo II) encontrados estão relacionados a possível causas advindas do filtro gaussiano não delimitar a área de interesse corretamente, como também situações de possíveis causas não identificadas (desconhecidas). Para erros de falso alarme (tipo I) a maior incidência de problemas é provocada em cenas de multidão e/ou em cenas que contém movimentação ou interferência da câmera.

Ao observar a Figura 5.19, é também possível notar que o fator de baixa qualidade parece afetar menos o desempenho do modelo do que as demais possíveis causas. Desta forma, é possível concluir que é importante refinar a técnica de localização da área de interesse para reduzir o uso de recurso computacional e aumentar o desempenho do modelo, como também

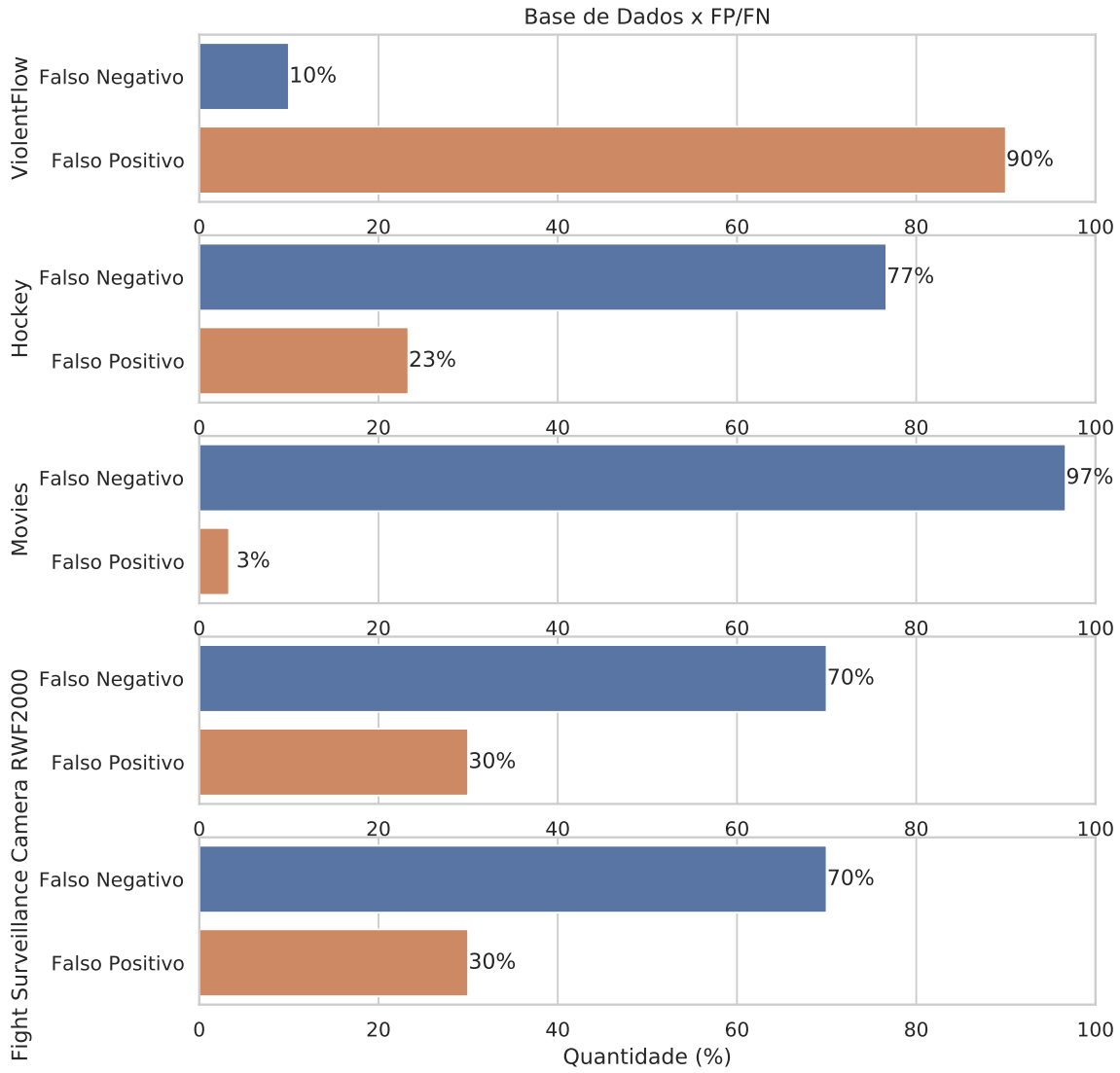

Figura 5.10: FP's e FN's por base de dados. Fonte: A autora (2022).

explorar técnicas de estabilização da cena de vídeo em casos de movimentação abrupta e movimentação da câmera. Também é importante investigar mais sobre o quesito de multidão, agrupando as falhas ao nível de quantidade de pessoas contidas na cena, tipos de fluxos de movimentos e a distância da câmera até à multidão.

Figura 5.11: Movimento abrupto

Figura 5.12: Multidão

Figura 5.13: Movimento ou interferência na câmera

Figura 5.14: Baixa qualidade

Figura 5.15: Área do filtro gaussiano mal localizada

Figura 5.16: Rótulo original errado

Figura 5.17: Desconhecido

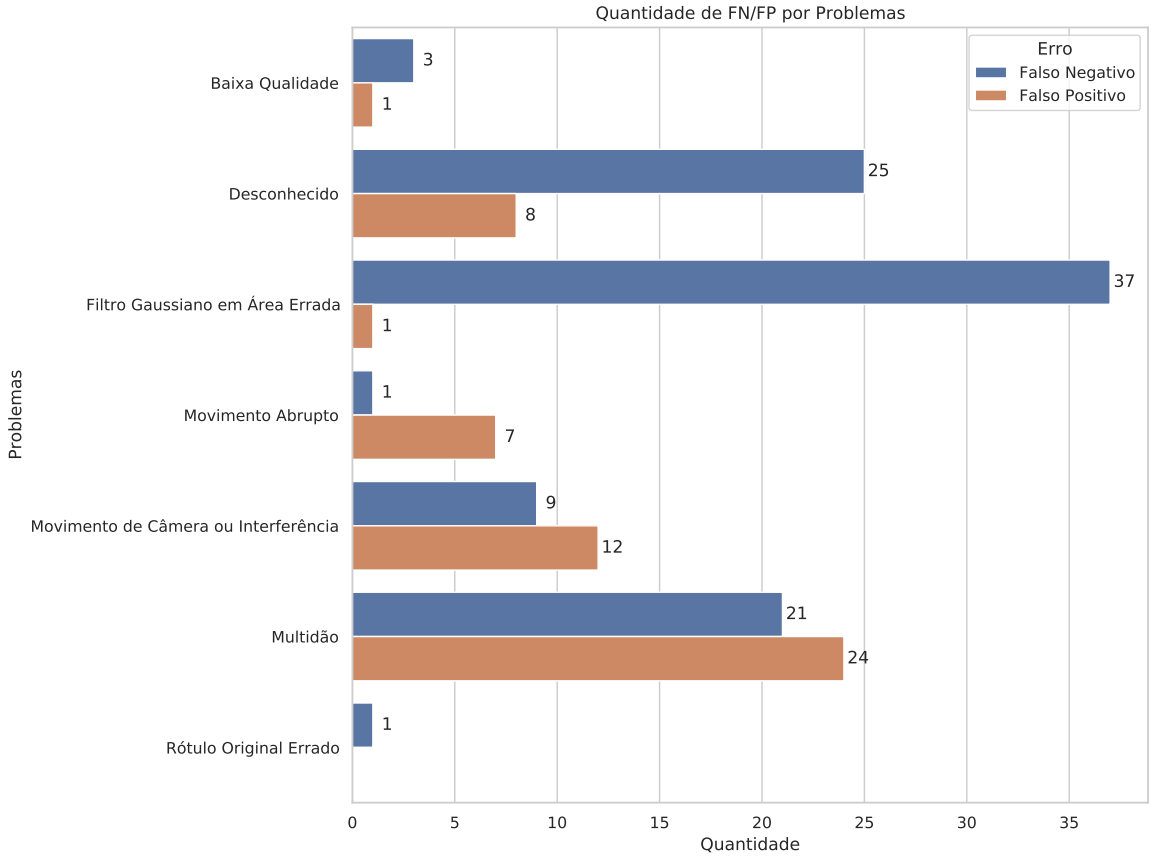

Figura 5.18: Quantidade de FP's e FN's por possível causa de falha. Fonte: A autora (2022).

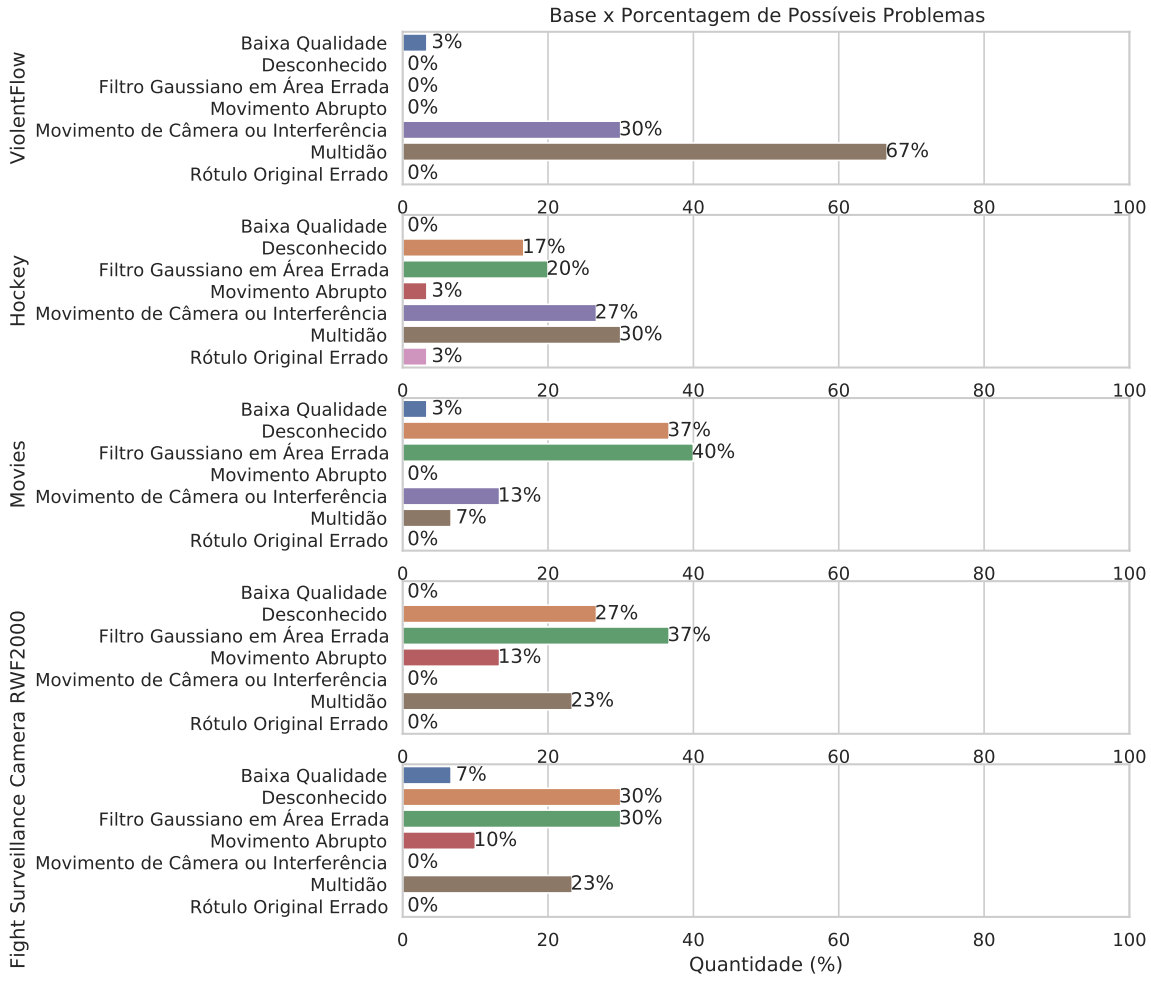

Figura 5.19: Possíveis causas de falhas por base. Fonte: A autora (2022).

Capítulo 6

Conclusão

Neste Capítulo, são apresentadas as conclusões dessa dissertação. Na Seção 6.1, são apresentadas as considerações finais relacionadas às contribuições alcançadas ao longo do desenvolvimento da pesquisa. Na Seção 6.3, são apresentadas pontos de possíveis ameaças à validade dos resultados obtidos. Na Seção 6.4, são apresentados pontos para a continuidade da pesquisa realizada.

6.1 Considerações Finais

A detecção de violência humana em vídeos de videovigilância é uma tarefa desafiadora, pois envolve desde questões subjetivas relacionadas ao limiar entre comportamento violento e não-violento a questões que interferem em características visuais da cena. Além dessas situações que dificultam o aprendizado de modelos, há também questões relacionadas ao conteúdo da cena que é submetida ao modelo, isto é, nem tudo que está presente na cena está relacionado com a área de interesse onde ocorre o comportamento violento.

Na tentativa de mitigar os desafios e problemáticas descritas no Capítulo 1, foi desenvolvido um método que possui duas principais contribuições, conforme descritas nas Seções 4.3.2 e 4.3.3. O método de delimitação da área de interesse de quadros de vídeo desenvolvido neste trabalho, possibilita alcançar resultados mais promissores em cenários reais de cenas capturadas de câmeras de videovigilância. Neste trabalho, também foi evidenciado que o custo computacional e as limitações de hardware são desafios recorrentes no campo de pesquisa e desenvolvimento. Contudo, foi demonstrado que o uso da técnica de precisão

mista durante o treinamento de uma rede neural profunda é capaz de reduzir o consumo de recursos computacionais, como a redução de aproximadamente 45% do uso de VRAM.

Os resultados obtidos, avaliados e apresentados no Capítulo 5, superaram a proposta *baseline* de detecção de violência, que compreende a etapa de pré-processamento por meio de corte da área de interesse. Os resultados foram validados em bases *benchmarks* e comparados com outras abordagens do estado da arte. Portanto, é esperado que a abordagem aqui proposta possa contribuir na identificação de cenas violentas em videovigilância com maior desempenho e alavancar estudos sobre otimização de recursos computacionais para o treinamento de modelos de redes neurais.

6.2 Contribuições

As contribuições desta pesquisa são:

- Desenvolvimento, aplicação e avaliação de uma abordagem de pré-processamento para a seleção de área de interesse em uma sequência de quadros de vídeo, baseada na aplicação do filtro gaussiano nas áreas de baixo interesse e preservação das características originais da área de interesse. Esse método possibilitou uma melhoria estatisticamente significativa na métrica de acurácia e no valor de perda de uma abordagem *baseline* do estado da arte, composta por uma CNN treinada para a classificação de violência em videovigilância;
- Aplicação de método de manipulação de precisão de casas decimais, durante a fase de treinamento do modelo, visando à redução do uso de recursos de memória e tempo de processamento, sem impactos negativos nas métricas de avaliação do modelo;
- Identificação da importância e contribuição com a pesquisa na identificação de soluções que visem à redução do uso de recursos computacionais disponíveis para treinamento de modelos;
- Contribuição com a pesquisa e desenvolvimento de métodos para a detecção de comportamentos violentos em vídeos capturados em câmeras de videovigilância.

6.3 Ameaças à Validade

Os resultados obtidos a partir da reprodução da abordagem *baseline* (CHENG; CAI; LI, 2021) (RGB), descrita na Seção 4.1, foram usados como "grupo de controle" para auxiliar a avaliação dos impactos das contribuições. Portanto, os resultados desta pesquisa não podem ser generalizados para cenas de violências obtidas por meio de outros dispositivos que não sejam câmeras de videovigilância, como câmeras de celulares, por exemplo.

Outro ponto relevante a ser considerado é que a base de dados utilizada para treinamento do modelo *baseline* (CHENG; CAI; LI, 2021) é constituída de comportamentos violentos, mas com foco maior em cenas agressivas ou de brigas entre indivíduos. Portanto, é possível que o modelo perca desempenho ao predizer cenas com tipos de violência que possuam um padrão diferente de agressão, como assalto sem interação física, violência através de um arremesso de objeto distante do alvo, atropelamento com veículos, entre outros.

É relevante salientar também que, devido ao tamanho de modelo reproduzido e à quantidade de amostra de dados de vídeos, o ambiente utilizado para a realização dos experimentos não supriu a necessidade de recursos computacionais necessários para a realização de testes empregando ou integrando redes neurais mais robustas, como a adição de camadas LSTM diretamente no modelo, mesmo utilizando a técnica de precisão mista (Subseção 4.3.3). Além disso, também não foi possível obter êxito ao incrementar os canais de fluxo óptico (2 canais) aos canais RGB (3 canais), totalizando 5 canais (versão "*Flow Gated Network*" (CHENG; CAI; LI, 2021)), pois a alocação de memória de vídeo (VRAM) necessária para iniciar o treinamento ultrapassou a quantidade disponibilizada no ambiente da plataforma utilizada.

6.4 Propostas para Pesquisas Futuras

A partir dos desafios e resultados elencados nesta dissertação, a seguir são apresentadas algumas propostas de pesquisas a serem realizadas, a saber:

- Avaliar a proposta desenvolvida na Seção 4.3.2 na versão *Flow Gated* (CHENG; CAI; LI, 2021), levando em conta o *branch* arquitetural especialista em informações temporais de fluxo óptico;
- Testar a alternativa de extrair características temporais por meio de um *branch* arquitetural

especialista, formado por camadas *Long short-term memory* (LSTM) ou a extração de características através de um modelo LSTM pré-treinado; Assim como um estudo mais aprofundado para identificar a melhor localização para a inserção e integração dessas camadas na arquitetura atual e/ou o uso de técnicas de redução do mapa de características temporais obtido pelas camadas LSTM.

- Testar a adição de características baseadas na máscara de saliência humana (Mask R-CNN (HE et al., 2017)) ou detecção de objetos (REDMON; FARHADI, 2018) para aperfeiçoar a garantia de que na área de interesse exista pelo menos um humano, possibilitando a redução dos erros de classificação ao submeter uma área de interesse que não há comportamentos humanos;

- Delimitar o tamanho da área de interesse dos quadros de vídeo de maneira dinâmica, com base no posicionamento de áreas em que ocorrem movimentos humanos;

- Avaliar os efeitos da variação do valor do tamanho do *kernel* gaussiano na proposta descrita na Seção 4.3.2, assim como, a exploração de outros filtros de remoção de ruído (KUMAR; SODHI, 2020);

- Analisar uma amostra maior de cenas de falhas e realizar análises estatísticas para consolidar as conclusões.

- Ampliar o escopo da abordagem proposta utilizando o conjunto de dados proposto na pesquisa desenvolvida por Sultani, Chen e Shah (2018) e a investigação de um classificador de tipos de violência.

- Descentralizar o treinamento do modelo através da técnica *Federated Learning*[1], permitindo que o algoritmo seja treinado em vários dispositivos descentralizados e sem a necessidade de partilhar amostras de dados locais;

- Avaliar o desempenho e precisão do modelo ao submetê-lo em cenários de teste com conjuntos de dados de videovigilância desbalanceados;

[1] *Federated Learning* Disponível em: <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>. Acesso em: 13 de agosto de 2022.

- Ampliar a otimização do modelo CNN através da técnica de poda para reduzir a quantidade de parâmetros e o uso de recurso de memória VRAM.

Referências Bibliográficas

AGGARWAL, J. K.; RYOO, M. S. Human activity analysis: A review. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 43, n. 3, p. 16:1–16:43, abr. 2011. Citado 3 vezes nas páginas 9, 10 e 11.

AKTı Şeymanur; TATAROğLU, G. A.; EKENEL, H. K. Vision-based fight detection from surveillance cameras. In: IEEE. *2019 Ninth International Conference on Image Processing Theory, Tools e Applications (IPTA)*. [S.l.], 2019. p. 1–6. Citado 12 vezes nas páginas 39, 40, 47, 49, 83, 84, 85, 108, 114, 118, 119 e 121.

ALVES, E. J. *Métodos de bootstrap e aplicações em problemas biológicos*. Dissertação (Mestrado Profissional em Matemática Universitária), 2013. Citado na página 77.

BAZAREVSKY, V.; GRISHCHENKO, I. On-device, real-time body pose tracking with mediapipe blazepose. *Google AI Blog*, 2020. Citado na página 43.

BEDDIAR, D. R. et al. Vision-based human activity recognition: a survey. *Multimedia Tools e Applications*, Springer, v. 79, n. 41, p. 30509–30555, 2020. Citado 17 vezes nas páginas xii, xv, 2, 7, 8, 9, 10, 11, 12, 15, 16, 19, 20, 23, 24, 31 e 32.

BILEN, H. et al. Action recognition with dynamic image networks. *IEEE transactions on pattern analysis e machine intelligence*, IEEE, v. 40, n. 12, p. 2799–2813, 2017. Citado na página 44.

BILINSKI, P.; BREMOND, F. Human violence recognition e detection in surveillance videos. In: IEEE. *2016 13th IEEE International Conference on Advanced Video e Signal Based Surveillance (AVSS)*. [S.l.], 2016. p. 30–36. Citado 2 vezes nas páginas 35 e 37.

BISCHOFF, P. *Surveillance camera statistics: which cities have the most CCTV cameras?* 2021. Https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities/. Citado na página 4.

BLANK, M. et al. Actions as space-time shapes. In: *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*. [S.l.: s.n.], 2005. p. 1395–1402. Citado 4 vezes nas páginas 43, 48, 49 e 126.

BLUNSDEN, S.; FISHER, R. B. The behave video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, British Machine Vision Association, v. 4, n. 1-12, p. 4, 2010. Citado 3 vezes nas páginas 41, 48 e 122.

BRUTZER, S.; HöFERLIN, B.; HEIDEMANN. Evaluation of background subtraction techniques for video surveillance. In: IEEE. *CVPR 2011*. [S.l.], 2011. p. 1937–1944. Citado na página 13.

BUX, A. *Vision-based human action recognition using machine learning techniques*. Tese (Doutorado em Filosofia), 2017. Citado 4 vezes nas páginas xii, 16, 17 e 18.

CAETANO, C. et al. Activity recognition based on a magnitude-orientation stream network. In: IEEE. *2017 30th SIBGRAPI Conference on Graphics, Patterns e Images (SIBGRAPI)*. [S.l.], 2017. p. 47–54. Citado 4 vezes nas páginas 42, 47, 109 e 115.

CAO, Z. et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis e Machine Intelligence*, 2019. Citado na página 43.

CARREIRA, J.; ZISSERMAN, A. Quo vadis, action recognition? a new model e the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision e Pattern Recognition*. [S.l.: s.n.], 2017. p. 6299–6308. Citado na página 46.

CHEN, L. et al. Sensor-based activity recognition. *IEEE Transactions on Systems, Man, e Cybernetics, Part C (Applications e Reviews)*, v. 42, n. 6, p. 790–808, 2012. Citado na página 14.

CHENG, M.; CAI, K.; LI, M. Rwf-2000: An open large scale video database for violence detection. In: IEEE. *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.], 2021. p. 4183–4190. Citado 39 vezes nas páginas viii, xii, xiii, 38, 41, 46, 47, 48, 49, 50, 51, 52, 53, 54, 57, 58, 59, 60, 61, 62, 64, 68, 70, 71, 73, 77, 82, 83, 84, 85, 93, 105, 106, 113, 114, 118, 119, 120 e 121.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision e pattern recognition*. [S.l.: s.n.], 2017. p. 1251–1258. Citado na página 39.

CHOULWAR, A. *The Art of Convolutional Neural Network*. 2019. <https://medium.com/@achoulwar901/the-art-of-convolutional-neural-network-abda56dba55c>. Acesso em: 24 de junho de 2022. Citado 2 vezes nas páginas xii e 22.

CHU, S.; TANAKA, J. Hand gesture for taking self portrait. In: . [S.l.: s.n.], 2011. v. 6762, p. 238–247. Citado na página 36.

CISCO. *O que é Wi-Fi?* 2021. [Online; Acesso em: 22 de novembro de 2021]. Disponível em: <https://www.cisco.com/c/pt_br/products/wireless/what-is-wifi.html>. Citado na página 13.

CLARIDGE-CHANG, A.; HO, J. *WHAT IS ESTIMATION?* 2017. <https://www.estimationstats.com/#/background>. [Online; Acesso em: 24 de agosto de 2022]. Citado na página 77.

DABKOWSKI, P.; GAL, Y. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, v. 30, 2017. Citado na página 44.

DEB, T.; ARMAN, A.; FIROZE, A. Machine cognition of violence in videos using novel outlier-resistant vlad. In: *2018 17th IEEE International Conference on Machine Learning e Applications (ICMLA)*. [S.l.: s.n.], 2018. p. 989–994. Citado na página 36.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision e pattern recognition*. [S.l.], 2009. p. 248–255. Citado na página 41.

DENIZ, O. et al. Fast violence detection in video. In: IEEE. *2014 International Conference on Computer Vision Theory e Applications (VISAPP)*. [S.l.], 2014. v. 2, p. 478–485. Citado 2 vezes nas páginas 35 e 37.

DEVI, B.; KUMAR, S.; SHANKAR, V. G. Anadata: A novel approach for data analytics using random forest tree e svm. In: *Computing, Communication e Signal Processing*. [S.l.]: Springer, 2019. p. 511–521. Citado na página 37.

DING, C. et al. Violence detection in video by using 3d convolutional neural networks. In: SPRINGER. *International Symposium on Visual Computing*. [S.l.], 2014. p. 551–558. Citado 8 vezes nas páginas 36, 38, 46, 47, 83, 105, 113 e 118.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994. Citado na página 77.

EHSAN, T. Z.; MOHTAVIPOUR, S. M. Vi-net: A deep violent flow network for violence detection in video sequences. In: IEEE. *2020 11th International Conference on Information e Knowledge Technology (IKT)*. [S.l.], 2020. p. 88–92. Citado 8 vezes nas páginas 41, 46, 47, 110, 115, 118, 119 e 120.

EUROSUR. *Application of surveillance tools to border surveillance 'concept of operations' (European Commission, Enterprise e Industry 8 January 2012)*. 2011. [Online; Acesso em: 25 de dezembro de 2019]. Citado na página 25.

FARNEBäCK, G. Two-frame motion estimation based on polynomial expansion. In: SPRINGER. *Scandinavian conference on Image analysis*. [S.l.], 2003. p. 363–370. Citado 3 vezes nas páginas xii, 52 e 54.

GAILLARD, M.; EGYED-ZSIGMOND, E.; GRANITZER., M. Cnn features for reverse image search. *Document numerique*, Lavoisier, v. 21, n. 1, p. 63–90, 2018. Citado na página 38.

GIBARU, O. *Neural Network*. 2019. <https://oliviergibaru.org/courses/ML_NeuralNetwork>. Acesso em: 13 de junho de 2022. Citado 2 vezes nas páginas xiii e 66.

GOOGLE. *Precisão Mista*. 2022. <https://www.tensorflow.org/guide/mixed_precision>. Acesso em: 11 de junho de 2022. Citado 3 vezes nas páginas 65, 66 e 67.

HASSNER, T.; ITCHER, Y.; KLIPER-GROSS, O. Violent flows: Real-time detection of violent crowd behavior. In: IEEE. *2012 IEEE Computer Society Conference on Computer Vision e Pattern Recognition Workshops*. [S.l.], 2012. p. 1–6. Citado 10 vezes nas páginas 39, 40, 41, 42, 44, 48, 83, 84, 85 e 120.

HE, K. et al. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 2961–2969. Citado 3 vezes nas páginas 43, 48 e 94.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision e pattern recognition*. [S.l.: s.n.], 2016. p. 770–778. Citado na página 41.

HERSHEY, S. et al. Cnn architectures for large-scale audio classification. In: IEEE. *2017 ieee international conference on acoustics, speech e signal processing (icassp)*. [S.l.], 2017. p. 131–135. Citado na página 46.

HO, J. et al. Moving beyond p values: data analysis with estimation graphics. *Nature methods*, Nature Publishing Group, v. 16, n. 7, p. 565–566, 2019. Citado 4 vezes nas páginas xiv, 78, 79 e 81.

HONARJOO, N.; ABDARI, A.; MANSOURI, A. Violence detection using pre-trained models. In: IEEE. *2021 5th International Conference on Pattern Recognition e Image Analysis (IPRIA)*. [S.l.], 2021. p. 1–4. Citado 7 vezes nas páginas 41, 112, 116, 118, 119, 120 e 123.

HUANG, F.-C. et al. High-performance sift hardware accelerator for real-time image feature extraction. *IEEE Transactions on Circuits e Systems for Video Technology*, IEEE, v. 22, n. 3, p. 340–351, 2011. Citado na página 21.

HUSSAIN, Z.; SHENG, M.; ZHANG, W. E. Different approaches for human activity recognition: A survey. *CoRR*, abs/1906.05074, 2019. Citado 2 vezes nas páginas 3 e 26.

JAHLAN, H. M. B.; ELREFAEI, L. A. Mobile neural architecture search network e convolutional long short-term memory-based deep features toward detecting violence from video. *Arabian Journal for Science e Engineering*, Springer, p. 1–15, 2021. Citado 7 vezes nas páginas 41, 47, 107, 114, 118, 119 e 120.

JAIN, A.; VISHWAKARMA, D. K. Deep neuralnet for violence detection using motion features from dynamic images. In: IEEE. *2020 Third International Conference on Smart Systems e Inventive Technology (ICSSIT)*. [S.l.], 2020. p. 826–831. Citado 5 vezes nas páginas 44, 47, 109, 115 e 118.

JOUDAKI, S.; SUNAR, M. S. B.; KOLIVAND, H. Background subtraction methods in video streams: a review. In: IEEE. *2015 4th International Conference on Interactive Digital Media (ICIDM)*. [S.l.], 2015. p. 1–6. Citado na página 13.

KARPATHY, A. et al. Large-scale video classification with convolutional neural networks. In: *CVPR*. [S.l.: s.n.], 2014. Citado na página 46.

KHURANA, P. *The convolution operation*. 2020. Https://prvnk10.medium.com/the-convolution-operation-48d72a382f5a. Acesso em: 24 de junho de 2022. Citado 2 vezes nas páginas xii e 22.

KHURANA, R.; KUSHWAHA, A. K. S. A deep survey on human activity recognition in video surveillance. In: IEEE. *2018 International Conference on Research in Intelligent e Computing in Engineering (RICE)*. [S.l.], 2018. p. 1–5. Citado na página 49.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012. Citado na página 40.

KUEHNE, H. et al. Hmdb: a large video database for human motion recognition. In: IEEE. *2011 International conference on computer vision*. [S.l.], 2011. p. 2556–2563. Citado 3 vezes nas páginas 43, 48 e 125.

KUMAR, A.; SODHI, S. S. Comparative analysis of gaussian filter, median filter e denoise autoenocoder. In: *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)*. [S.l.: s.n.], 2020. p. 45–51. Citado na página 94.

LANDI, F.; SNOEK, C. G.; CUCCHIARA, R. Anomaly locality in video surveillance. *arXiv preprint arXiv:1901.10364*, 2019. Citado 4 vezes nas páginas 44, 48, 49 e 123.

LANGROUDI, H. F. et al. Cheetah: Mixed low-precision hardware & software co-design framework for dnns on the edge. *arXiv preprint arXiv:1908.02386*, 2019. Citado 2 vezes nas páginas 65 e 66.

LEJMI, W. et al. Challenges e methods of violence detection in surveillance video: A survey. In: . [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 1 e 2.

LI, Z. et al. A survey of convolutional neural networks: analysis, applications, e prospects. *IEEE transactions on neural networks e learning systems*, IEEE, 2021. Citado na página 23.

LIN, L. *A World With a Billion Cameras Watching You Is Just Around the Corner*. 2019. <https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402>. Acesso em: 04 de julho de 2022. Citado na página 1.

MABROUK, A. B.; ZAGROUBA, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, v. 91, p. 480 – 491, 2018. Citado 4 vezes nas páginas xii, 26, 27 e 29.

MARTíNEZ-MASCORRO, G. A.; ORTIZ-BAYLISS, J. C.; TERASHIMA-MARíN, H. Detecting suspicious behavior: How to deal with visual similarity through neural networks. *arXiv preprint arXiv:2007.15235*, 2020. Citado na página 45.

MOHTAVIPOUR, S. M.; SAEIDI, M.; ARABSORKHI, A. A multi-stream cnn for deep violence detection in video sequences using handcrafted features. *The Visual Computer*, Springer, p. 1–16, 2021. Citado 6 vezes nas páginas 42, 109, 114, 118, 119 e 120.

MUGUNGA, I. et al. A frame-based feature model for violence detection from surveillance cameras using convlstm network. In: IEEE. *2021 6th International Conference on Image, Vision e Computing (ICIVC)*. [S.l.], 2021. p. 55–60. Citado 10 vezes nas páginas 41, 47, 82, 110, 115, 118, 119, 120, 121 e 122.

MUMTAZ, A.; SARGANO, A. B.; HABIB, Z. Fast learning through deep multi-net cnn model for violence recognition in video surveillance. *The Computer Journal*, 2020. Citado 5 vezes nas páginas 40, 110, 115, 118 e 119.

NAIK, A. J.; GOPALAKRISHNA, M. T. Deep-violence: individual person violent activity detection in video. *Multimedia Tools e Applications*, Springer, v. 80, n. 12, p. 18365–18380, 2021. Citado 5 vezes nas páginas 43, 47, 106, 114 e 126.

NANDAKUMAR, S. et al. Mixed-precision deep learning based on computational memory. *Frontiers in Neuroscience*, v. 14, 2020. Citado na página 66.

NANNI, L.; GHIDONI, S.; BRAHNAM, S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, Elsevier, v. 71, p. 158–172, 2017. Citado na página 18.

NASARUDDIN, N. et al. Deep anomaly detection through visual attention in surveillance videos. *Journal of Big Data*, SpringerOpen, v. 7, n. 1, p. 1–17, 2020. Citado 8 vezes nas páginas 44, 46, 47, 62, 64, 109, 115 e 122.

NASCIMENTO, S. P. de F. D. *Representações lexicais da língua de sinais brasileira: uma proposta lexicográfica*. Tese (Doutorado) — Universidade de Brasília, 2009. Citado na página 12.

NIEVAS, E. B. et al. Violence detection in video using computer vision techniques. In: *Computer Analysis of Images e Patterns*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 332–339. Citado 14 vezes nas páginas 36, 37, 39, 40, 41, 42, 44, 48, 49, 83, 84, 85, 118 e 119.

NVIDIA. *Deep Learning Performance Documentation*: Train with mixed precision. 2022. <https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html>. Acesso em: 13 de junho de 2022. Citado 2 vezes nas páginas 66 e 67.

NWEKE, H. F. et al. Data fusion e multiple classifier systems for human activity detection e health monitoring: Review e open research directions. *Information Fusion*, v. 46, 06 2018. Citado na página 50.

OMS. *Injuries e violence*. 2021. <https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence>. Citado na página 4.

OPENCV. *OpenCV*: Open source computer vision. 2022. <https://docs.opencv.org/4.x/>. Acesso em: 12 de junho 2022. Citado na página 64.

PAREEK, P.; THAKKAR, A. A survey on video-based human action recognition: recent updates, datasets, challenges, e applications. *Artificial Intelligence Review*, Springer, v. 54, n. 3, p. 2259–2322, 2021. Citado 2 vezes nas páginas 14 e 34.

QIU, Z.; YAO, T.; MEI, T. Learning spatio-temporal representation with pseudo-3d residual networks. In: *proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 5533–5541. Citado na página 62.

REDMON, J. *Darknet: Open source neural networks in c*. 2013. Citado na página 40.

REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. *arXiv*, 2018. Citado na página 94.

RENDÓN-SEGADOR, F. J. et al. Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence. *Electronics*, Multidisciplinary Digital Publishing Institute, v. 10, n. 13, p. 1601, 2021. Citado 9 vezes nas páginas 39, 40, 47, 108, 114, 118, 119, 120 e 123.

ROMAN, D. G. C.; CHáVEZ, G. C. Violence detection and localization in surveillance video. In: IEEE. *2020 33rd SIBGRAPI Conference on Graphics, Patterns e Images (SIBGRAPI)*. [S.l.], 2020. p. 248–255. Citado 9 vezes nas páginas 43, 46, 47, 48, 108, 114, 118, 120 e 123.

RYOO, M. S.; ROTHROCK, B.; MATTHIES, L. Pooled motion features for first-person videos. In: *Proceedings of the IEEE Conference on Computer Vision e Pattern Recognition*. [S.l.: s.n.], 2015. p. 896–904. Citado na página 41.

SCHULDT, C.; LAPTEV, I.; CAPUTO, B. Recognizing human actions: a local svm approach. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* [S.l.], 2004. v. 3, p. 32–36. Citado 4 vezes nas páginas 43, 48, 49 e 126.

SHAH, J. H. et al. Facial expressions classification e false label reduction using lda e threefold svm. *Pattern Recognition Letters*, 2017. Citado na página 37.

SHI, X. et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, v. 28, 2015. Citado na página 42.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, SpringerOpen, v. 6, n. 1, p. 1–48, 2019. Citado na página 56.

SILVA, E. P. da. *Facial expression recognition in Brazilian sign language using facial action coding system: Reconhecimento de expressões faciais na língua de sinais brasileira por meio do sistema de códigos de ação facial*. Tese (Doutorado em Engenharia Elétrica), 2020. Citado na página 12.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Citado 2 vezes nas páginas 41 e 43.

SINGH, T.; VISHWAKARMA, D. K. Human activity recognition in video benchmarks: A survey. In: *Advances in Signal Processing e Communication*. [S.l.]: Springer, 2019. p. 247–259. Citado na página 14.

SOLIMAN, M. M. et al. Violence recognition from videos using deep learning techniques. In: IEEE. *2019 Ninth International Conference on Intelligent Computing e Information Systems (ICICIS)*. [S.l.], 2019. p. 80–85. Citado 5 vezes nas páginas 39, 44, 48, 49 e 123.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. Citado 4 vezes nas páginas 43, 48, 49 e 124.

SULTANI, W.; CHEN, C.; SHAH, M. Real-world anomaly detection in surveillance videos. In: *Proceedings of the IEEE conference on computer vision e pattern recognition*. [S.l.: s.n.], 2018. p. 6479–6488. Citado 8 vezes nas páginas 41, 45, 48, 49, 94, 112, 116 e 122.

SZEGEDY, C. et al. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision e pattern recognition*. [S.l.: s.n.], 2015. p. 1–9. Citado na página 40.

TAN, M. et al. Mnasnet: Platform-aware neural architecture search for mobile. In: *Proceedings of the IEEE/CVF Conference on Computer Vision e Pattern Recognition*. [S.l.: s.n.], 2019. p. 2820–2828. Citado na página 42.

TAYLOR, J. et al. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: *2012 IEEE Conference on Computer Vision e Pattern Recognition*. [S.l.: s.n.], 2012. p. 103–110. Citado na página 43.

TECMUNDO. *Como funciona a RFID?* 2009. [Online; Acesso em: 22 de novembro de 2021]. Disponível em: <https://www.tecmundo.com.br/tendencias/2601-como-funciona-a-rfid-.htm>. Citado na página 13.

TRAN, D. et al. Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 4489–4497. Citado 2 vezes nas páginas 45 e 46.

ULLAH, F. U. M. et al. An intelligent system for complex violence pattern analysis e detection. *International Journal of Intelligent Systems*, Wiley Online Library, 2021. Citado 8 vezes nas páginas 40, 46, 48, 83, 109, 115, 118 e 120.

VEZZANI, R.; BALTIERI, D.; CUCCHIARA, R. People reidentification in surveillance e forensics: A survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 46, n. 2, p. 29:1–29:37, dez. 2013. Citado 3 vezes nas páginas 15, 30 e 43.

VIEL, F.; JR., F. W.; ZEFERINO, C. A. Sistema integrado para o processamento do filtro de difusão anisotrópica em fpga. *Revista de Sistemas e Computação-RSC*, v. 7, n. 2, 2017. Citado na página 20.

VRIGKAS, C. N. M.; KAKADIARIS, I. A. A review of human activity recognition methods. *Frontiers in Robotics e Artificial Intelligence*, v. 2, 11 2015. Citado 2 vezes nas páginas 9 e 13.

WANG, H.; SCHMID, C. Action recognition with improved trajectories. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2013. p. 3551–3558. Citado na página 35.

WANG, L. et al. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. Citado na página 42.

WIKIPEDIA. *Gaussian blur*. 2022. <https://en.wikipedia.org/wiki/Gaussian_blur>. Acesso em: 12 de junho de 2022. Citado 2 vezes nas páginas xiii e 63.

WU, P. et al. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2020. p. 322–339. Citado 8 vezes nas páginas 45, 46, 48, 49, 111, 115, 122 e 125.

XU, L. et al. Violent video detection based on mosift feature e sparse coding. In: IEEE. *2014 IEEE International Conference on Acoustics, Speech e Signal Processing (ICASSP)*. [S.l.], 2014. p. 3538–3542. Citado 3 vezes nas páginas 35, 36 e 37.

YAZDANI, R. et al. The dark side of dnn pruning. In: *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. [S.l.: s.n.], 2018. p. 790–801. Citado na página 66.

ZIVKOVIC, Z. Improved adaptive gaussian mixture model for background subtraction. In: IEEE. *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* [S.l.], 2004. v. 2, p. 28–31. Citado na página 45.

ZIVKOVIC, Z.; HEIJDEN, F. V. D. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, Elsevier, v. 27, n. 7, p. 773–780, 2006. Citado na página 45.

ZURAS, D. et al. Ieee standard for floating-point arithmetic. *IEEE Std*, v. 754, n. 2008, p. 1–70, 2008. Citado na página 64.

Apêndice A

Síntese dos Trabalhos Relacionados

Neste apêndice é apresentada uma síntese com as principais características das pesquisas relacionadas à temática desta pesquisa.

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Ding et al. (2014)	RGB		Espaciais e de movimento	CNN 3D	
Cheng, Cai e Li (2021)	RGB + Fluxo Óptico	Branch único formado por uma CNN	RGB		248.402 parâmetros

Continuação na página seguinte. . .

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Cheng, Cai e Li (2021)	RGB + Fluxo Óptico	Branch único formado por uma CNN	Fluxo Óptico		248.258 parâmetros
Cheng, Cai e Li (2021)	RGB + Fluxo Óptico	Blocos que simulam convoluções 3D reduzindo a quantidade de parâmetros	- *Fusion Pseudo-3D* (P3D) - Canais RGB + Fluxo Óptico - Ramificação do canal de fluxo óptico para ajudar a construir mecanismos de *pooling*	*Flow Gated Network* (CNN)	272.690 parâmetros
Cheng, Cai e Li (2021)	RGB + Fluxo Óptico	Convoluções 3D tradicionais	*Fusion C3D* Canais RGB + Fluxo Óptico	*Flow Gated Network* (CNN)	507.155 parâmetros
Naik e Gopalakrishna (2021)	Segmentação		17 pontos-chave do corpo humano forma do corpo humano	Mask-RCNN modificada + LSTM	Máscara do corpo humano

Continuação na página seguinte...

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Jahlan e Elrefaei (2021)	RGB + Fluxo Óptico	Classificador *Random Forest*	CNN + Magnitude do movimento em características espaciais	*Automated Mobile Neural Architecture Search* (MNAS) pré-treinada + ConvLSTM	Redução de dimensionalidade e mapa de características com base na diferença entre quadros
Jahlan e Elrefaei (2021)	RGB + Fluxo Óptico	Classificador *Support Vector Machine*	CNN + Magnitude do movimento em características espaciais	*Automated Mobile Neural Architecture Search* (MNAS) pré-treinada + ConvLSTM	Redução de dimensionalidade e mapa de características com base na diferença entre quadros
Jahlan e Elrefaei (2021)	RGB + Fluxo Óptico	Classificador *K Nearest Neighbor*	CNN + Magnitude do movimento em características espaciais	*Automated Mobile Neural Architecture Search* (MNAS) pré-treinada + ConvLSTM	Redução de dimensionalidade e mapa de características com base na diferença entre quadros

Continuação na página seguinte...

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Aktı, Tataroğlu e Ekenel (2019)	RGB		Fight-CNN	Xception + redes Bi-LSTM + camada de atenção	
Rendón-Segador et al. (2021)	RGB + Fluxo Óptico	RGB e Fluxo Óptico e pseudo Fluxo Óptico	RGB para Fluxo Óptico	DenseNet-121 + camada de autoatenção + BiConvLSTM2D + Classificador *Fully Connected*	4,5 M Parâmetros
Roman e Chávez (2020)	Imagens Dinâmicas		Imagem dinâmica e máscara de saliência	Classificação: CNN Localização: Yolo V3 + Mask R-CNN	

Continuação na página seguinte...

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Mohtavipour, Saeidi e Arabsorkhi (2021)	RGB + Fluxo Óptico		Hand-Crafted: *Spation-temporal* (DMEI), temporal, *spation*	CNN	motion energy Discussão de resultados interessantes
Caetano et al. (2017)	RGB + Fluxo Óptico		Magnitude-Orientation Stream (MOS) + (Espacial/temporal/espaial e temporal)	CNN + Fluxo Óptico	
Ullah et al. (2021)	Segmentação		Darknet *Residual Optical Flow* CNN M-LSTM	Mask-RCNN	Darknet CNN *Residual Optical Flow Human Boxes*
Nasaruddin et al. (2020)	Região de Atenção		C3D Subtração de background Extração da região de atenção visual bilateral	*Fully conected neural network*	
Jain e Vishwakarma (2020)	Imagens Dinâmicas		Imagens Dinâmicas Inception Resnet v2 pré-treinada	*Fine tuning*	

Continuação na página seguinte...

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Mumtaz, Sargano e Habib (2020)	RGB		*Transfer learning* com o modelo DCNN (para classificação de imagem)	Deep Multi-Net (DMN) CNN Model composto pelas redes pré-treinadas AlexNet GoogleNet	
Ehsan e Mohtavipour (2020)	RGB + Fluxo Óptico		Fluxo Óptico	CNN	
	Espacial e de movimento				
Mugunga et al. (2021)	RGB		VGG-16 pré-treinada	3 blocos ConvLSTM + *Fully conected*	

Continuação na página seguinte...

Trabalho	Característica dominante	Característica Predominante	Extração de Características	Modelo	Diferenciais Utilizados
Wu et al. (2020)	RGB	4 ramos	Imagem: C3D (pré-treinada no Sports-1M) + I3D (pré-treinada no Kinesis-400) Áudio: VGGish (pré-treinada com dados do YouTube)	3 ramos (holístico, localizado e de pontuação dinâmica)	
Wu et al. (2020)	RGB	Online	Imagem: C3D (pré-treinada bo Sports-1M) + I3D (pré-treinada no Kinesis-400) Áudio: VGGish (pré-treinada com dados do YouTube)	4 ramos (holístico, localizado e de pontuação dinâmica)	
Wu et al. (2020)	RGB	C3D	Imagem: C3D (pré-treinada bo Sports-1M) + I3D (pré-treinada no Kinesis-400) Áudio: VGGish (pré-treinada com dados do YouTube)	5 ramos (holístico, localizado e de pontuação dinâmica)	

Continuação na página seguinte. . .

Trabalho	**Característica dominante**	**Característica Predominante**	**Extração de Características**	**Modelo**	**Diferenciais Utilizados**
Honarjoo, Abdari e Mansouri (2021)	RGB		ResNet50 pré-treinada	*Fully conected* + NN	
Honarjoo, Abdari e Mansouri (2021)	RGB		VGG16 pré-treinada POT (agrupamento de características temporais)	*Fully conected* + NN	
Sultani, Chen e Shah (2018)	RGB		C3D pré-treinada	*Fully conected*	

Apêndice B

Síntese dos Resultados dos Trabalhos Relacionados (Acurácia)

Neste apêndice é apresentada uma sumarização dos resultados da métrica de acurácia, obtidos em conjuntos de dados utilizados em pesquisas relacionadas à temática desta pesquisa.

Trabalho	**Hockey**	**Movie**	**Violent Flows**	**XD-Violence**	**BEHAVE**	**RLVS**	**RWF-2000**	**Weizmann**	**KTH**	**UCF-Crime**	**UCFCrime2Local**
Ding et al. (2014)	91										
Cheng, Cai e Li (2021)							84,5				
Cheng, Cai e Li (2021)							75,5				

Continuação na página seguinte...

Trabalho	Hockey	Movie	Violent Flows	XD-Violence	BEHAVE	RLVS	RWF-2000	Weizmann	KTH	UCF-Crime	UCFCrime2Local
Cheng, Cai e Li (2021)	98	100	88,87				87,25				
Cheng, Cai e Li (2021)							85,75				
Naik e Gopala-krishna (2021)								73,1	93,4		
Jahlan e Elrefaei (2021)	99,3	100	95								
Jahlan e Elrefaei (2021)	99	100	96								
Jahlan e Elrefaei (2021)	99	100	96								
Aktı, Tataroğlu e Eke-nel (2019)	97,5	100									
Rendón-Segador et al. (2021)	99,20 ± 0,6%	100 ± 0%	96,90 ± 0,5%			95,60 ± 0,6%					
Roman e Chávez (2020)	96,40 ± 0,3%		92,0 ± 0,14%								79,1 ± 0,191%
Mohtavipour, Saeidi e Arabsorkhi (2021)	100	100	99,35								

Continuação na página seguinte...

Trabalho	Hockey	Movie	Violent Flows	XD-Violence	BEHAVE	RLVS	RWF-2000	Weizmann	KTH	UCF-Crime	UCFCrime2Local
Caetano et al. (2017)											
Ullah et al. (2021)	99		98,4								
Nasaruddin et al. (2020)										95,74	
Jain e Vishwakarma (2020)	93,33										
Mumtaz, Sargano e Habib (2020)	98,32	99,58									
Ehsan e Mohtavipour (2020)	98	99	94								
Mugunga et al. (2021)	99,1	100	98,4		99,3		92,4			99,1	
Wu et al. (2020)				78,64						82,44	
Wu et al. (2020)				73,67							
Wu et al. (2020)				67,19							

Continuação na página seguinte...

Trabalho	Hockey	Movie	Violent Flows	XD-Violence	BEHAVE	RLVS	RWF-2000	Weizmann	KTH	UCF-Crime	UCFCrime2Local
Honarjoo, Abdari e Mansouri (2021)	96	100	94			97					
Honarjoo, Abdari e Mansouri (2021)	95,5	100	96			96					
Sultani, Chen e Shah (2018)										75,41	

Apêndice C

Síntese das Bases de Dados

Neste apêndice é apresentada uma descrição das características dos conjuntos de dados, utilizados em pesquisas relacionadas à temática desta pesquisa.

Autores	**Base**	**Tipos de Violência/Ação**	**Rótulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Nievas et al. (2011)	*Hockey Fight*	Brigas, Lutas	"fight"; "no fight"	Ding et al. (2014), Cheng, Cai e Li (2021), Jahlan e Elrefaei (2021), Aktı, Tataroğlu e Ekenel (2019), Rendón-Segador et al. (2021), Roman e Chávez (2020), Mohtavipour, Saeidi e Arabsorkhi (2021), Ullah et al. (2021), Jain e Vishwakarma (2020), Mumtaz, Sargano e Habib (2020), Ehsan e Mohtavipour (2020), Mugunga et al. (2021), Honarjoo, Abdari e Mansouri (2021)	*IceHockey*	Detecção de comportamentos humanos violentos	1000 Clipes (500/500)

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Nievas et al. (2011)	*Movies Fights* ou Películas	Brigas, Lutas	"*fight*"; "*non fight*"	Cheng, Cai e Li (2021), Jahlan e Elrefaei (2021), Aktı, Tataroğlu e Ekenel (2019), Rendón-Segador et al. (2021), Mohtavipour, Saeidi e Arabsorkhi (2021), Mumtaz, Sargano e Habib (2020), Ehsan e Mohtavipour (2020), Mugunga et al. (2021), Honarjoo, Abdari e Mansouri (2021)	Moveis e esportes	Detecção de comportamentos humanos violentos	201 Clipes de filmes de ação

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Hassner, Itcher e Kliper-Gross (2012)	*Violent Flow* ou *Violent Flows*	Brigas, Lutas	"*violence*"; "*no violence*"	Cheng, Cai e Li (2021), Jahlan e Elrefaei (2021), Rendón-Segador et al. (2021), Roman e Chávez (2020), Mohtavipour, Saeidi e Arabsorkhi (2021), Ullah et al. (2021), Ehsan e Mohtavipour (2020), Mugunga et al. (2021), Honarjoo, Abdari e Mansouri (2021)	*Streets*, *school and sports*	Comportamento violento em multidão	- 246 videos do YouTube (123/123) (aproximadamente 1-7 segundos) - violência vs não violência

Continuação na página seguinte…

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Cheng, Cai e Li (2021)	RWF2000	Brigas, Lutas	"fight"; "non fight"	Cheng, Cai e Li (2021), Mugunga et al. (2021)	Variados	Detecção de comportamentos humanos violentos	- Coletados o Youtube (clipes entre 5s a 30 fps) - Violência e não violência - 2000 clipes (300.000 quadros)
Aktı, Tataroğlu e Ekenel (2019)	*Fight Surveillance Camera*	Brigas, Lutas	"fight"; "non fight"			Detecção de comportamentos humanos violentos em câmeras de videovigilância	- 300 vídeos (150/150); - 2 segundos de duração; - Apenas cenas relacionadas a lutas; - Coletados do Youtube.

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Sultani, Chen e Shah (2018)	UCF-Crime	14 diferentes ações	"anomalia"; "normal"	Nasaruddin et al. (2020), Mugunga et al. (2021), Wu et al. (2020), Sultani, Chen e Shah (2018)	CCTV Câmera	Detecção de comportamentos humanos violentos em câmeras de videovigilância, como anomalia	- 1900 vídeos (128 horas); - Anomalias realísticas.
Blunsden e Fisher (2010)	BEHAVE	Brigas, Lutas	10 cenários: "In-Group"(IG); "Approach"(A); "Walk-Together"(WT); "Split"(S); "Ignore"(I); "Following"(FO); "Chase"(C); "Fight"(FI), "Run-Together"(RT); "Meet"(M)	Mugunga et al. (2021)	Pessoas atuando em vários tipos de interações	Identificação de comportamento de múltiplas pessoas	- 25 fps; - 640x480px; - 4 clipes; - Envolvem vídeos de luta e outros comportamentos considerados normais.

Continuação na página seguinte...

Autores	Base	Tipos de Violência/Ação	Rotulos	Trabalhos que usam essa base	Ambiente	Domínio de aplicação	Descrição da Base
Soliman et al. (2019)	*Real Life Violence Situations* (RLVS)	Brigas de rua reais	"violence"; "no violence"	Rendón-Segador et al. (2021), Honarjoo, Abdari e Mansouri (2021)	Variado, como por exemplo ruas, presídios e escolas	Reconhecimento de violência em vídeos	- 2000 vídeos (1000/1000); - Os clipes de violência envolvem lutas em diversos ambientes, como ruas, presídios e escolas; - Entre 3-7 segundos; - 480p - 720p.
Landi, Snoek e Cucchiara (2019)	UCFCrime2Local	6 ações	Anomaly e Normal utilizadas no trabalho oficial, mas estão disponíveis as labels "arrest", "assault", "burglary", "robbery", "stealing"e "vandalism".	Roman e Chávez (2020)	CCTV Câmera	Detecção de comportamentos humanos violentos em câmeras de videovigilância, como anomalia	100 anormal / 200 normais.

Continuação na página seguinte...

Autores	Base	Tipos de Violência/Ação	Rotulos	Trabalhos que usam essa base	Ambiente	Domínio de aplicação	Descrição da Base
Soomro, Zamir e Shah (2012)	UCF101	Socos	101 rótulos associados a Interação Humano-Objeto, Somente Movimento Corporal, Interação Humano-Humana, Tocando instrumentos musicais e Esportes		presença de grandes variações no movimento da câmera, aparência e pose do objeto, escala do objeto, ponto de vista, fundo desordenado, condições de iluminação, etc	Reconhecimento de ações humanas	- 13320 vídeos de ações realistas; - 25 FPS e 320 × 240 de resolução; - Os vídeos possuem em média 7,21 segundos; - 1600 minutos de duração total.

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Kuehne et al. (2011)	HMDB51	Exemplo: Socos e chutes	51 ações associadas a *General facial actions*, *Facial actions with object manipulation*, *General body movements*, *Body movements with object interaction* e *Body movements for human interaction*.		Variância entre diferentes partes do corpo visíveis na cena, movimento de câmera, ponto de vista da câmera, número de pessoas envolvidas na cena e qualidade de vídeo	Reconhecimento de ações humanas	- 6849 vídeos; - Mínimo de 101 ações por ações.
Wu et al. (2020)	XD-Violence	6 tipos de violência	"abuse"; "car accident"; "explosion"; "fighting"; "riot and shooting".	Wu et al. (2020)	movies, sports, games, hand-held cameras, CCTV car cameras	Detecção de violência em vídeos	- 4754 vídeos; - 217 horas; - Rótulo fraco (áudio e imagem).

Continuação na página seguinte...

Autores	**Base**	**Tipos de Violência/Ação**	**Rotulos**	**Trabalhos que usam essa base**	**Ambiente**	**Domínio de aplicação**	**Descrição da Base**
Schuldt, Laptev e Caputo (2004)	KTH Human Action Dataset	6 ações	"walking"; "jogging; "running";"boxinghandwaving";"handclapping"	Naik e Gopalakrishna (2021)	Participação de 25 sujeitos em quatro cenários diferentes: ao ar livre, ao ar livre com variação de escala, ao ar livre com roupas diferentes e dentro de casa.	*Human action recognition in real outdoor conditions*	2391 sequências. Todas as sequências foram tiradas em background homogêneo e com uma câmera estática. Possuem uma duração de 4 segundos em média.
Blank et al. (2005)	Weizmann Human Action Dataset	10 ações	"walk";"run"; "jump"; "gallop"; "sideways"; "bend"; "one-hand wave"; "two-hands wave"; "jump in place"; "jumping Jjack"; "skip"	Naik e Gopalakrishna (2021)	Pessoas andando em vários cenários difíceis na frente de diferentes fundos não uniformes	Ações humanas	90 sequências de vídeo de 9 pessoas diferentes realizando 10 ações simples.