UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

ALLAN COSTA NASCIMENTO DOS SANTOS

Análise de modelos de visão computacional para detecção de quedas baseado em vídeo RGB e infravermelho

NITERÓI

2023

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

ALLAN COSTA NASCIMENTO DOS SANTOS

Análise de modelos de visão computacional para detecção de quedas baseado em vídeo RGB e infravermelho

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações.

Orientadora:
Natalia Castro Fernandes

Co-orientador:
Flávio Luiz Seixas

NITERÓI

2023

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S237a Santos, Allan Costa Nascimento dos
Análise de Modelos de Visão Computacional para Detecção de Quedas Baseado em Vídeo RGB e Infravermelho com Alta Sensibilidade / Allan Costa Nascimento dos Santos. - 2023.
100 f.: il.

Orientador: Natalia Castro Fernandes.
Coorientador: Flavio Luiz Seixas.
Dissertação (mestrado)-Universidade Federal Fluminense, Escola de Engenharia, Niterói, 2023.

1. Detecção de queda. 2. Visão computacional. 3. Reconhecimento de atividade humana. 4. Rede neural convolucional. 5. Produção intelectual. I. Fernandes, Natalia Castro, orientadora. II. Seixas, Flavio Luiz, coorientador. III. Universidade Federal Fluminense. Escola de Engenharia.IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

ALLAN COSTA NASCIMENTO DOS SANTOS

Análise de Modelos de Visão Computacional para Detecção de Quedas Baseado em Vídeo RGB e Infravermelho com Alta Sensibilidade

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações.

Aprovada em 29 de maio de 2023.

BANCA EXAMINADORA

Profª. Natalia Castro Fernandes, D.Sc. – Orientador, UFF

Prof. Flavio Luiz Seixas, D.Sc. – Orientador, UFF

Profª. Dianne Scherly Varela de Medeiros, D.Sc. – UFF

Prof. Alexandre Sztajnberg, D.Sc. – UERJ

Niterói

2023

À Jesus Cristo, minha família, amigos, meus irmãos em Cristo e professores.

Agradecimentos

Agradecimentos, primeiramente aos orientadores Natalia Castro Fernandes e Flavio Luiz Seixas pela paciência, amizade e cooperação no desenvolvimento do trabalho. Agradeço ao grupo de trabalho dos laboratórios MídiaCom e GTECCOM – UFF, ao PPGEET - UFF, aos participantes do grupo do projeto Monitora UFF, aos professores Tadeu Ferreira, Diogo Menezes, Débora Christina, Vinícius Nunes, Carlos Alberto Malcher, Marcos Tadeu, Dianne Scherly e Cledson Oliveira. Aos meus colegas, entes queridos e amigos Alessandro Aparecido Milan, Tiago Bornia de Castro, Fábio Henrique Moreira dos Anjos, Jorge Barros, Luana Villafuerte, Thiago Ribeiro Aragão, Lorran Davi Azarany, Adriano Busson, Jonatas Bento, Lucas Rosa Laureano Madeira, Andreia dos Santos da Silva, Ana Paula Azevedo de Oliveira, Patrícia Levino, Luiz Rocha, Joacir de Oliveira e Marister Monteiro, agradecimentos não apenas às contribuições nesse trabalho, mas também na minha vida.

Agradecimentos à casa Bem Estar - Lar Israelita para Cuidados de Idosos, aos funcionários, aos residentes, à Advá Griner e Adriano Ferreira de Menezes pelo apoio à pesquisa. Agradeço a nossa faculdade, bem como a sua agremiação e a infraestrutura proporcionada. Agradeço aos meus pais Antônio Carlos dos Santos e Marina Costa do Nascimento, ao meu irmão Arthur Costa Nascimento dos Santos, aos meus familiares, a congregação Cristã, amigos que oraram por mim e entes queridos que deram todo amor, apoio, carinho e inspiração para persistir e sobrepor as dificuldades e, finalmente, ao nosso Senhor e Salvador Jesus Cristo, razão da nossa fé e esperança. A todos que participaram, direta ou indiretamente do desenvolvimento deste trabalho de pesquisa, enriquecendo o meu processo de aprendizado.

Resumo

As quedas são um grave problema de saúde pública e as pessoas com mais de 65 anos estão entre as mais vulneráveis a lesões graves decorrentes de quedas. Há ainda o fato de que as quedas podem afetar negativamente a mentalidade do idoso, resultando em baixa autoestima, pois ele se torna dependente de uma pessoa que o monitora constantemente, além das constantes idas ao hospital. Uma abordagem natural e prática para pessoas idosas com vulnerabilidade em locomoção e que precisam de assistência imediata mediante uma queda. Portanto, este trabalho propõe e avalia modelos de visão computacional para melhorar o monitoramento e a segurança de indivíduos com risco de queda, como idosos ou pessoas com mobilidade reduzida. O modelo compreende uma rede neural generativa, blocos convolucionais espaço-temporais, cálculo do fluxo óptico, uma técnica para rastrear a região de interesse e uma rede neural *feed-forward* para calcular a pontuação de anomalia. Também é relevante analisar o modelo para trabalhar com gravações infravermelho, pois quedas também podem ocorrer em ambientes com pouca luz. A análise consistiu na aplicação de diversos filtros e técnicas de processamento de imagem em diferentes combinações, buscando encontrar um modelo que satisfizesse uma alta sensibilidade e um alto F1 Score. O modelo de rede neural final utilizando câmera RGB atinge 99,21% de sensibilidade e F1 Score de 0.98, enquanto o modelo utilizando câmera infravermelha atinge 100% de sensibilidade e 0.98 de F1 Score, superando outras propostas da literatura. A técnica de pontuação de anomalias mostrou-se uma técnica de aprendizado que se adapta bem e é capaz de identificar a queda mesmo com a exposição de novos cenários de vídeo, sendo ideal para o uso do sistema em situações reais.

Palavras-chave: Detecção de queda, visão computacional, fluxo óptico, rede neural convolucional, rede adversária generativa, assistência médica, reconhecimento de atividade humana.

Abstract

Falls are a serious public health problem and people over 65 are among the most vulnerable to serious injuries from falls. There is also the fact that falls can negatively affect the elderly person's mentality, resulting in low self-esteem, as they become dependent on a person who constantly monitors them, in addition to constant trips to the hospital. A natural and practical approach for elderly people with vulnerable mobility and who need immediate assistance after a fall. Therefore, this work proposes and evaluates computer vision models to improve the monitoring and safety of individuals at risk of falling, such as the elderly or people with reduced mobility. The model comprises a generative neural network, spatiotemporal convolutional blocks, optical flow calculation, a technique to track the region of interest, and a *feed-forward* neural network to calculate the anomaly score. It is also important to analyze the model for working with infrared recordings, as falls can also occur in low-light environments. The analysis consisted of applying several filters and image processing techniques in different combinations, seeking to find a model that satisfied high sensitivity and a high F1 Score. The final neural network model using an RGB camera reaches 99.21% sensitivity and an F1 Score of 0.98, while the model using an infrared camera reaches 100% sensitivity and a 0.98 F1 Score, surpassing other proposals in the literature. The anomaly scoring technique proved to be a learning technique that adapts well and is capable of identifying the drop even with new video scenarios, being ideal for using the system in real situations.

Keywords: Fall Detection, Computer Vision, Optical Flow, Convolutional Neural Network, Generative Adversarial Network, Healthcare, Human Activity Recognition.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ADL	*Activities of Daily Life*	1
IR	*Infrared*	2
RGB	*Red Green Blue*	2
CNN	*Convolutional Neural Network*	2
ROI	*Region of Interest*	2
IoT	*Internet of Things*	11
WBAN	*Wireless Body Area Networks*	11
GCN	*Graph Convolutional Networks*	28
BS	*Background Subtraction*	29
TMF	*Temporal Median Filtering*	3
CVSC	*Convolutional Video Stream Combination*	34
CNT	*Count*	3
WSN	*Wireless Sensor Network*	10
GAN	*Generative Adversarial Network*	22
ML	*Machine Learning*	5
CV	*Computer Vision*	22
AI	*Artificial Intelligence*	22
OFC	*Optical Flow Computation*	2
EHR	*Electronic Health Record*	8
OCR	*Optical Character Recognition*	15
MOG	*Gaussian Mixture*	3
IoU	*Intersection over Union*	46
COCO	*Common Objects in Context*	45
R-FCN	*Region-based Fully Convolutional Networks*	45

Sumário

Capítulo 1

Introdução

Grandes melhorias na expectativa de vida têm sido a tendência predominante para os países desenvolvidos e de alta renda ao longo dos séculos 20 e 21 [10]. Em nove anos, houve 141.308 admissões de lares de idosos para hospitais *non-Veterans Administration* por lesões relacionadas a quedas, com 38,8% fraturas de quadril, 35,7% outras fraturas e 11,1% lesões intracranianas, com um custo médio de US$31.507 por admissão [11]. O aumento global da expectativa de vida humana criou a necessidade de tecnologia de saúde e monitoramento remoto adequado para idosos [3]. Um dos maiores problemas de saúde dos idosos são as quedas que ocorrem em casa durante as *Activities of Daily Life*s (ADLs) [3]. No Brasil, o número de idosos nestá aumentando e, até 2025, o Brasil será o 6º no mundo em quantidade de idosos [12].

Esses dados indicam uma necessidade em aumentar a qualidade de vida e independência dos idosos. Contudo, mesmo em lares de idosos, com assistência frequente, estima-se que a incidência de quedas seja 13,1% [11]. Devido à potencial gravidade da queda, é necessário que o evento seja detectado o quanto antes a fim de evitar riscos ainda piores à saúde do idoso.

Esta realidade levou ao desenho de sistemas de detecção de quedas que empregam câmeras de vigilância. Por outro lado, os sistemas baseados em vídeo têm limitações na detecção de quedas devido a mudanças no fundo da imagem, objetos de fundo, iluminação e movimento da câmera [9]. Algumas abordagens não funcionam em ambientes pouco iluminados, por exemplo, quando o idoso está no quarto à noite e precisa se levantar para ir ao banheiro ou tomar remédios. Outro problema relevante é que muitos desses algoritmos não são adequados para detectar quedas usando câmeras infravermelhas, não podendo funcionar em ambientes escuros [6]. Embora o efeito da iluminação na precisão e resolução da cor possa ser mitigado, é difícil mitigar seu efeito na faixa dinâmica de

imagens de câmeras *Infrared*s (IRs) [13]. A porção da capacidade disponível para a carga gerada pelos fótons da banda visível é reduzida quando os fótons atingem o plano da imagem. Ao contrário do sinal, o ruído não pode ser subtraído. De fato, qualquer operação que corrija o sinal do pixel usando os sinais dos pixels vizinhos apenas adiciona a contribuição de sua potência de ruído ao pixel. No entanto, a degradação pode ser estimada a partir do conjunto de eficiência quântica das câmeras *Red Green Blue* (RGB) e IRs [13].

Portanto, este trabalho propõe e avalia técnicas de pré-processamento para melhorar o modelo de redes neurais para detecção de quedas proposta em [9], visando como caso de uso a detecção de quedas de idosos [14]. Diferente de outras propostas da literatura, o modelo proposto detecta quedas em ambientes com alta iluminação com câmeras RGBs ou mesmo sem iluminação com o IRs. Assim, o sistema utiliza estrategicamente as câmeras domésticas em áreas onde os idosos com mobilidade reduzida apresentam maior risco de queda. Este sistema ajuda a aumentar a auto-estima do usuário. Os sistemas de detecção de quedas são importantes para reduzir as consequências de lesões graves, permitindo que a pessoa possa ser socorrida mais rapidamente pelos profissionais de saúde. Cabe ressaltar que o sistema pode ser utilizado tanto por idosos como por cadeirantes e pessoas com deficiência, ou qualquer pessoa com mobilidade reduzida e consequentemente mais elevado risco de queda.

O trabalho utiliza a *Convolutional Neural Network* (CNN) de [9] como base para calcular a probabilidade de evento de queda. Nesse modelo, que foi treinado com imagens de câmeras térmicas, a queda representa, na imagem, uma mudança espaço-temporal na posição do indivíduo. Quanto maior o erro de reconstrução, maior a chance da presença de uma anomalia. Como há uma mudança repentina, essa mudança gerará um erro de reconstrução significativo naquele momento. No modelo de [9], aplica-se o *Optical Flow Computation* (OFC) [15] para calcular a variação de pixels e, portanto, o movimento. Foi realizado o rastreamento de pessoas para extrair a região de interesse (*Region of Interest* (ROI)) utilizando a técnica *Region-based Fully Convolutional Network* (R-FCN).

O modelo proposto consiste em modificar o pré-processamento das imagens para estender o modelo de [9] para uso com câmeras RGB e infravermelho. Para tanto, foram feitas análises com diversas combinações de filtros e técnicas de processamento de imagem, visando uma alta sensibilidade, mas sem incorrer em alta taxa de falsos positivos. Nesse sentido, observou-se que para o cálculo do ROI, o fluxo óptico com o limite Otsu [16] apresentou maior eficiência para separar o fundo escuro da imagem, tendo impacto signi-

ficativo no resultado final da detecção da queda.

Além disso, a aplicação de um filtro de Kalman corrigiu as mudanças de pixels causadas por variações de iluminação e, assim, conseguindo rastrear a pessoa através do fluxo de imagens tanto RGB quanto infravermelho. Cabe destaque que, nesse cenário, o projeto do conjunto de filtros é crítico para se obter um sistema capaz de operar tanto com câmeras RGB e quanto IR [13]. O pré-processamento proposto possibilita a utilização de imagens RGB e IR.

As combinações de filtros testadas nesse trabalho foram: o filtro de fechamento [17], o filtro de abertura [17], o filtro de dilatação [18], o filtro de desfoque [19], o filtro de limiarização [20]. Também foram utilizadas as técnicas de subtração de fundo: *Temporal Median Filtering* (TMF) [21], *Count* (CNT) [2] e *Gaussian Mixture* (MOG)2 [22]. Os resultados mostram que o impacto dos filtros sobre a detecção de queda depende da combinação de filtros usados.

O processamento da imagem resultante das combinações dos filtros produziram impactos significativos na sensibilidade da detecção de queda. A combinação do filtro de abertura com a técnica de subtração de fundo CNT, modelo chamado de CVSC CNT2 ('O') resultou no modelo com a maior sensibilidade (1.0) e acurácia de 0.875. Assim, os resultado mostram que o impacto dos filtros na acurácia e sensibilidade aumenta ainda mais com a adição das técnicas de subtração de fundo. Isso ocorre porque as técnicas de subtração de fundo diminuem os erros causados por ruídos e objetos de fundo. Com relação ao tempo de processamento total médio, o modelo que obteve menor tempo de processamento foi o CVSC1 ('C', 'O', 'T'), com média de 26.21477s para detectar a queda. Esse modelo utiliza a combinação dos filtros de fechamento, abertura e limiarização, mas obteve uma acurácia de apenas 0.5 e sensibilidade de 0.0. O modelo CVSC1 ('C',) que utilizou apenas o filtro de fechamento tem tempo de processamento de 91.76557s. O modelo CVSC1 ('O',) que utilizou apenas o filtro de abertura tem tempo de processamento de 98.966918s. O modelo CVSC1 ('T',) que utilizou apenas o filtro de limiarização tem tempo de processamento de 39.49132s. Assim, os resultados mostram a combinação dos filtros também impactam sobre o tempo de processamento e que o filtro de limiarização diminui consideravelmente o tempo de processamento total. Essa redução do tempo ocorre porque a imagem resultante do filtro de limiarização é uma imagem binária com valores de pixel de 0 ou 255.

O restante desta dissertação está estruturado da seguinte forma. O Capítulo 2 trata do aprendizado de máquina na saúde e no processamento de imagens. O Capítulo 3 descreve

os trabalhos relacionados à detecção de queda e a revisão bibliográfica. O Capítulo 4 descreve o modelo de filtragem e pré-processamento proposta. O Capítulo 5 descreve as diferentes combinações analisadas e os resultados obtidos. Por fim, são apresentadas as conclusões no Capítulo 6 e os trabalhos futuros no Capítulo 7.

Capítulo 2

Aprendizado de máquina na saúde e no processamento de imagens

Esse capítulo trata do aprendizado de máquina na saúde, sistemas de detecção de quedas através de video monitoramento e o processamento de imagem necessário para essas aplicações. Esse campo de pesquisa utiliza algoritmos e técnicas de inteligência artificial para análise e interpretação de dados relacionados com aplicações médicas e no cuidado ao idoso. Serão apresentados diversos casos de como o aprendizado de máquina pode ser aplicado, fornecendo um sistema de segurança e assistência médica continuada, como o monitoramento remoto de pacientes e os sistemas de detecção de queda. Esse capítulo descreve os modelos de aprendizado de máquina na área de visão computacional. Esse capítulo também descreve as técnicas de detecção e rastreamento de pessoas em vídeo. Para detectar uma queda é necessário rastrear a movimentação do idoso no ambiente monitorado pela câmera. O capítulo conclui a apresentando as vantagens e desvantagens dos modelos de detecção de queda através de visão computacional.

2.1 Aprendizado de máquina na área da saúde

O setor de saúde possui diferentes aplicações de aprendizado de máquina. Contudo, diferentemente de outros cenários de aplicação, como em entretenimento ou em redes sociais, na saúde, os algoritmos precisam ser mais robustos. É necessário haver verificações para a implantação do *Machine Learning* (ML) com considerações sobre questões legais e responsabilidade [23, 24]. Muitas questões na área de saúde não se alinham perfeitamente com os avanços atuais no aprendizado de máquina. Existem várias questões que precisam ser abordadas para garantir a conformidade com as regulamentações e a proteção dos

pacientes, como por exemplo a privacidade e a proteção dos dados. O uso de algoritmos de ML na saúde requer o manuseio adequado de informações sensíveis do paciente. É necessário garantir a privacidade e a segurança desses dados. A *General Data Protection Regulation* (GDPR)[1] na União Europeia e a *Health Insurance Portability and Accountability Act* (HIPAA)[2] nos Estados Unidos são regulamentações relevantes para considerar. Os modelos de ML podem ser influenciados por viés, resultando em disparidades e discriminação em determinados grupos de pacientes. É essencial garantir que os algoritmos sejam justos e equitativos [25, 26, 27]. Quando ocorrem erros ou danos causados pelo uso de ML na saúde, questões de responsabilidade legal surgem. Determinar quem é responsável por erros diagnósticos ou tratamentos inadequados pode ser complexo quando envolve sistemas de ML. O artigo [28] analisa os aspectos legais da responsabilidade no contexto da IA na medicina. Muitos modelos de ML, como redes neurais profundas, são algoritmos que suas decisões são difíceis de explicar. No entanto, na área da saúde, é importante que os médicos e os pacientes possam entender as decisões tomadas pelos algoritmos. O artigo [29] discute as questões relacionadas à interpretabilidade dos modelos de ML.

Por outro lado, vários problemas de saúde estão sendo resolvidos com a aprendizagem semi-supervisionada ou não supervisionada [30]. Frequentemente, essas técnicas são aplicadas para identificar relações causais. Alguns ilustram como a aprendizagem semi-supervisionada e não supervisionada têm sido utilizadas com sucesso na área da saúde para identificar relações causais, descobrir padrões e ajudar no desenvolvimento de tratamentos mais eficazes. Os autores de [31] utilizaram técnicas de aprendizagem não supervisionada para identificar genes relacionados ao risco de desenvolver determinadas doenças. Eles demonstraram que a abordagem não supervisionada pode ajudar na identificação de marcadores genéticos relevantes para condições específicas. Na área de oncologia, a aprendizagem não supervisionada tem sido amplamente aplicada para identificar subtipos de câncer com base em dados genômicos. Por exemplo, no artigo [32], os autores utilizaram técnicas de *clustering* não supervisionado para identificar subtipos de câncer de mama, o que permitiu uma melhor compreensão da heterogeneidade da doença. A aprendizagem semi-supervisionada tem sido aplicada para identificar relações causais em doenças complexas, como a esclerose múltipla. No artigo [33], os autores utilizaram técnicas de aprendizagem semi-supervisionada para inferir relações causais entre diferentes variáveis biológicas relacionadas à esclerose múltipla. No artigo [34], os autores aplicaram técnicas de aprendizagem não supervisionada para descobrir associações entre características

[1] *https://gdpr-info.eu/*
[2] *https://www.hhs.gov/hipaa/index.html*

clínicas em grandes conjuntos de dados de saúde. Eles identificaram padrões ocultos e agruparam os pacientes em subgrupos com base em características semelhantes, permitindo uma melhor compreensão das relações entre diferentes variáveis clínicas. No artigo [35], os autores utilizaram métodos de aprendizagem semi-supervisionada para extrair informações valiosas de registros eletrônicos de saúde. Eles aplicaram técnicas de mineração de dados não supervisionadas para identificar fatores de risco e relações causais em doenças cardiovasculares, ajudando a aprimorar a pesquisa e a prestação de cuidados médicos nessa área.

Além disso, a implementação de sistemas baseados em dados de saúde tem desafios adicionais de caráter prático, como a integração novas tecnologias nos fluxos de trabalho das unidades de saúde. No artigo [36], os autores discutem os desafios práticos da implementação de algoritmos de inteligência artificial em sistemas de saúde. Eles abordam questões como a integração dos algoritmos em fluxos de trabalho clínicos, a necessidade de uma colaboração efetiva entre médicos e cientistas da computação, e a importância de validar e interpretar os resultados gerados pela IA. No artigo [37], os autores exploram os desafios da implementação de registros eletrônicos de saúde. Eles destacam a importância da consideração dos fluxos de trabalho clínicos existentes ao integrar esses registros, a necessidade de treinamento e educação adequados para os profissionais de saúde e a importância da interoperabilidade entre sistemas de saúde para facilitar a troca de informações.

Outra questão de relevância na aplicação de aprendizado de máquina para a saúde é que os dados do paciente são particularmente sensíveis, então a desidentificação é necessária e a negociação acordos de compartilhamento de dados consomem tempo. Na área da saúde, é comum encontrar desafios relacionados a dados ausentes e discrepâncias nas distribuições de dados de treinamento em comparação com os dados de teste. Isso ocorre devido à complexidade do sistema de saúde e às peculiaridades dos registros médicos e dados clínicos. Os dados de saúde podem ser incompletos devido a vários motivos, como erros de registro, informações não coletadas ou pacientes perdidos para acompanhamento. A presença de dados ausentes pode afetar negativamente o treinamento e desempenho dos modelos de aprendizado de máquina. Métodos para lidar com dados ausentes incluem exclusão de casos com dados ausentes, imputação de valores ausentes com técnicas estatísticas ou uso de algoritmos específicos projetados para lidar com dados faltantes [38]. Em muitos cenários de saúde, os dados de treinamento podem ser coletados em diferentes contextos, instituições ou populações, o que pode resultar em discrepâncias nas distribuições de dados em relação aos dados de teste. Por exemplo, um modelo trei-

nado em dados de uma região geográfica específica pode não generalizar bem para outra região devido a diferenças demográficas, culturais ou de práticas médicas. Essa discrepância de distribuição pode levar a uma queda no desempenho e à falta de generalização dos modelos [38].Sistemas de *Electronic Health Record* (EHR) podem ser difíceis de integrar, pois cada sistema comercial é ligeiramente diferente, gerando diferenças semânticas e sintáticas que podem inviabilizar a utilização conjunta de dados de diferentes fontes [39].

Na última década, os registros de saúde passaram de principalmente em papel para principalmente eletrônicos [1]. A evolução para o uso de EHR está sendo rápida. Nos EUA, o uso de EHR aumentou cerca de 9 vezes desde 2008, passando de 9,4% dos hospitais utilizando EHR para quase 84% em 2015, conforme mostrado na Figura 2.1 [1].

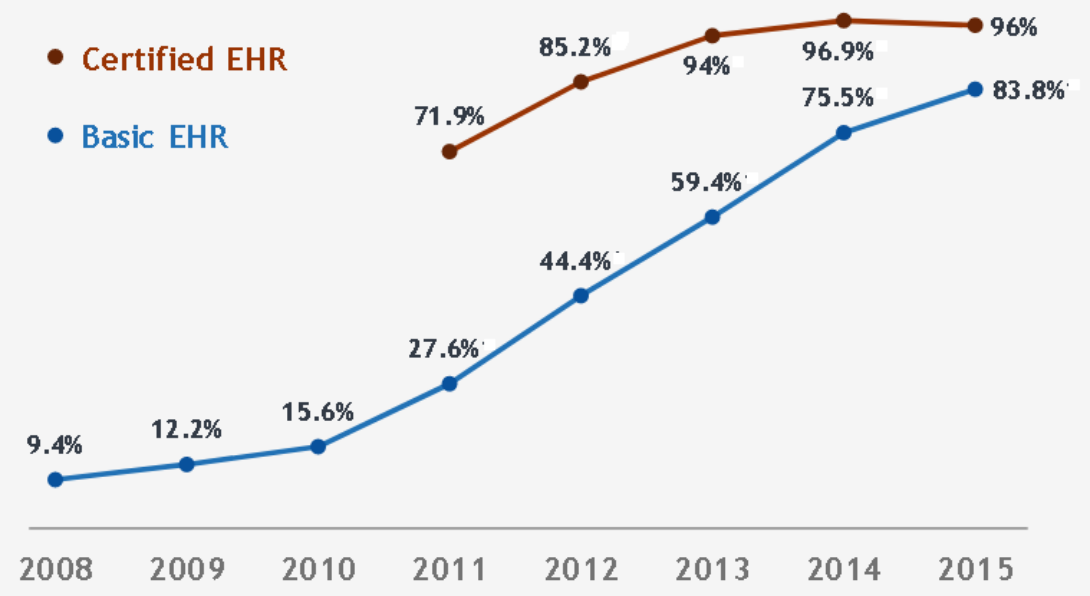


Figura 2.1: A adoção de EHR por hospitais de 2008-2015 nos Estados Unidos. Fonte: [1].

O aprendizado de máquina pode tornar os fluxos de trabalho de saúde digital, baseada em EHR, mais eficientes. As técnicas de processamento de imagens para radiografias de tórax e eletrocardiogramas podem ajudar a reduzir o número de consultas especializadas [39]. No lado administrativo, o aprendizado de máquina pode automatizar a documentação e os processos de faturamento [40]. No artigo [40], os autores aplicaram aprendizado profundo para procedimentos de auditoria, ilustrando como os recursos de aprendizado profundo para compreensão de texto, reconhecimento de fala, reconhecimento visual e análise de dados estruturados se encaixam no ambiente de auditoria. O aprendizado de máquina pode ser aplicado no lado administrativo da saúde para automatizar processos de documentação e faturamento, melhorando a eficiência e reduzindo erros. No artigo [41], os autores desenvolveram um sistema de aprendizado de máquina para au-

tomatizar o processo de codificação de faturas médicas. O modelo foi treinado em dados históricos de faturas e conseguiu identificar automaticamente os códigos corretos para cada procedimento, agilizando o processo de faturamento e reduzindo erros humanos.

Dispositivos vestíveis permitem monitoramento contínuo, potencialmente permitindo um melhor gerenciamento de doenças crônicas [42, 43, 44]. A miniaturização de dispositivos que podem ser usados como vestíveis ou acoplados ao corpo humano possibilitou o surgimento de novas redes de comunicação. Essas redes são compostas por pequenos dispositivos com baixo consumo de energia e capacidade de comunicação sem fio. A principal função desse tipo de rede é monitorar e atuar sobre o corpo humano e o meio ambiente. A composição dessa rede feita por nós sensores e atuadores que interagem com um nó coordenador. O nó coordenador é responsável por coletar e enviar as informações necessárias aos nós sensores/atuadores, por sua vez, o nó coordenador oferece a comunicação com outros tipos de redes [45]. Com mais dados, é possível entender melhor a progressão de doenças crônicas. O sistema de apoio à decisão clínica *Clinical Decision Support System* (CDSS) é aplicado ao diagnóstico de Demência, Doença de Alzheimer e Transtorno Cognitivo Leve [46]. O modelo de decisão foi construído com base em diretrizes clínicas aplicadas ao diagnóstico das doenças, utilizando uma representação multinível, visando assegurar a legibilidade do especialista do domínio e interoperabilidade semântica com outros sistemas de informação.

Métodos de aprendizado de máquina podem ajudar a identificar subtipos de doenças [47], procurar estruturas moleculares ideais para locais de ligação [48] e facilitar novos projetos de ensaios clínicos [49]. No artigo [47], os autores aplicaram técnicas de aprendizado de máquina para identificar subtipos de câncer de mama com base em dados de expressão genética. Eles utilizaram o algoritmo de aprendizado não supervisionado para agrupar os tumores de acordo com padrões de expressão gênica semelhantes, revelando subtipos moleculares distintos de câncer de mama com implicações clínicas. No artigo [48], os autores propuseram uma abordagem baseada em aprendizado profundo para prever a afinidade de ligação de compostos a alvos proteicos. O modelo desenvolvido foi treinado em uma variedade de dados estruturais de proteínas e compostos químicos, permitindo a identificação de potenciais inibidores de proteínas com alta precisão. No artigo [49], os autores utilizaram técnicas de aprendizado de máquina para otimizar o protocolo de um ensaio clínico relacionado ao câncer de próstata. Eles desenvolveram um modelo preditivo que incorporava dados genômicos e clínicos para prever a progressão da doença em diferentes estágios e identificar alvos moleculares potenciais. Essas informações foram utilizadas para aprimorar o protocolo do ensaio clínico, selecionando os pacientes

mais relevantes e aumentando as chances de sucesso do estudo.

Além disso, com o aprendizado de máquina, é possível realizar esses ensaios clínicos em escala. Isso facilita a previsão da evolução de uma determinada enfermidade. No artigo [50], os autores desenvolveram um modelo de aprendizado de máquina que integra dados clínicos e patológicas para prever a evolução do câncer de mama. O modelo foi treinado em uma grande quantidade de dados de pacientes e foi capaz de identificar fatores de risco e prever a probabilidade de metástase do câncer de mama com alta precisão.

As redes inteligentes de sensores fornecem uma tecnologia de monitoramento da saúde de forma ininterrupta [51]. Com o advento das redes móveis 5G de baixa potência, será possível fornecer meios de comunicação de alta disponibilidade e alta taxa de transferência necessários para redes de sensores utilizadas no monitoramento de pacientes e transferência de dados médicos [52, 53].

Atualmente, um grande volume de dados é gerado devido ao crescente uso de tecnologias de sensores em vários domínios de aplicativos em tempo real na assistência médica. Além disso, devido a algumas características sociais como o envelhecimento da sociedade, observa-se um aumento substancial na quantidade de dados que precisa ser processada nesses sistemas [54]. As *Wireless Sensor Network*s (WSNs) são um conjunto de sensores especializados distribuídos espacialmente que simultaneamente monitoram, registram e comunicam dados representando medições de variáveis ambientais ou de um determinado sistema. Uma possível utilização desse tipo de rede de sensores é o monitoramento da saúde de um indivíduo ou de um conjunto de indivíduos [51].

Uma das possibilidades do uso dos dispositivos móveis para o sensoriamento de pacientes é o sistema de detecção de quedas. Ao analisar as vibrações, variações de altura e de velocidade do dispositivo corporal, o sistema pode determinar o estado atual dos idosos e, em caso de risco, enviar um alerta para a família ou uma mensagem para profissionais e médicos cadastrados [55]. Para tal, é possível utilizar o sensor acelerômetro em um *smartphone* [51]. Os dados em tempo real são recuperados, processados e analisados por um sistema de detecção de queda *on-line* em execução no próprio *smartphone*. No artigo [56], os autores desenvolveram um sistema de detecção e avaliação de quedas em idosos com base em sensores corporais. O sistema usou sensores de aceleração e giroscópio incorporados em um colete vestível. Algoritmos de aprendizado de máquina foram aplicados para analisar os dados dos sensores e identificar eventos de queda. O sistema também avaliou a gravidade da queda com base em parâmetros como a aceleração de impacto. O estudo demonstrou a eficácia do sistema na detecção de quedas e na avaliação

das circunstâncias relacionadas.

Através desse método, é possível desenvolver um sistema de monitoramento para idosos através de meios técnicos [51]. Essa aplicação é importante porque o envelhecimento da população tornou-se uma característica marcante da sociedade nos últimos anos [57]. No entanto, os sensores comerciais de detecção de queda costumam ser caros [58]. Além disso, no caso de idosos e pessoas com deficiências cognitivas, é muito complicado garantir que a pessoa irá portar o dispositivo constantemente.

Nesse sentido, torna-se necessário um sistema de detecção de quedas mais econômico, adaptável e confiável para detectar quedas e enviar alarmes a uma autoridade apropriada. Os sistemas usam sensores de *smartphones* como acelerômetros, giroscópios e magnetômetros para detectar quando a queda aconteceu [59]. Outras abordagens usam sensores vestíveis, usando uma lógica semelhante ao uso de um telefone celular para detectar e notificar quedas [60]. Uma crítica comum aos sistemas baseados em *smartphones* e sensores vestíveis de detecção de queda em lares de idosos é que não podemos garantir que os idosos sempre carreguem os dispositivos necessários com eles [51].

Como a *Internet of Things* (IoT) pode existir de forma integrada com o ambiente, ela tem grande potencial para apoiar os objetivos de melhoria dos serviços de saúde [61]. E-Health é um conceito amplo que engloba todas as inovações tecnológicas que impactam a área da saúde. A E-Health desenvolve uma aplicação de Internet, utilizada em conjunto com outras tecnologias de informação, com o objetivo de melhorar as condições dos processos clínicos e de tratamento dos doentes e proporcionar melhores condições ao Sistema de Saúde. Ou seja, o E-Health melhora o fluxo de informações por meios eletrônicos para melhorar a prestação de serviços, e a coordenação dos sistemas de saúde, como aplicativos especializados, monitoramento do estado de idosos ou pacientes, serviços baseados em localização e muitos aplicativos de smartphones são exemplos reais de contribuições tecnológicas na área da saúde [62, 61]. *Dias et al* [63] propuseram um sistema de detecção de quedas IoT que pode ser introduzido em ambientes hospitalares, asilos de forma rápida e com baixo custo. A queda é detectada por meio de um acelerômetro integrado e o local é identificado por meio de transceptores *Zigbee*. O sistema também possui uma rede *Zigbee* utilizada para emissão de alarmes. À semelhança deste sistema *Zigbee*, a proposta deste artigo também oferece uma elevada taxa de sucesso na detecção de quedas e na identificação do local da queda em relação ao quarto onde ocorreu.

Outra modalidade de dispositivos IoT são os sensores que formam as *Wireless Body Area Networkss* (WBANs). Eles são capazes de se comunicar entre si e com outros senso-

res ou dispositivos [64]. WBANs permite a detecção e monitoramento de sinais vitais em corpos humanos. Em geral, são aplicados em contextos médicos onde as aplicações recebem os dados recolhidos, analisam-nos e enviam informação aos profissionais de saúde [64]. Esses aplicativos são capazes de monitorar e alertar condições críticas de saúde. Outro grupo de pesquisa discute o monitoramento do ambiente para detecção de quedas, não dependendo se o idoso possui algum dispositivo, como por meio do uso de câmeras de monitoramento ou sistemas de detecção de quedas baseados em WiFi [6] [3].

2.2 Técnicas de processamento de imagem

2.2.1 Pré-processamento da imagem

Uma representação para imagens é uma matriz na qual os elementos representam a intensidade ou a cor em uma posição correspondente na imagem [2]. No caso mais simples, um plano de elementos sensores é organizado em uma grade uniformemente espaçada, e cada elemento mede a quantidade de luz que incide sobre ele, como representado na Figura 2.2. Esta representação simples é usada na maioria das linguagens de programação e pode ser mapeado em *arrays* bidimensionais, o que possibilita uma forma bastante natural de trabalhar com imagens. Uma possível desvantagem dessa representação é que ela não depende do conteúdo da imagem. Em outras palavras, não faz diferença se a imagem contém apenas um par de linhas ou é uma cena complexa porque a quantidade de memória necessária é constante e depende apenas das dimensões da imagem. Regiões em uma imagem podem ser representadas usando uma máscara lógica na qual a área dentro da região recebe o valor *true* e a área sem o valor, *false*. Como esses valores podem ser representados por um único bit, essa matriz costuma ser chamada de *"bitmap"*.

Em uma imagem binária, os pixels podem assumir exatamente um de dois valores. Esses valores geralmente são considerados como representando o "primeiro plano" e o "plano de fundo" da imagem, embora esses conceitos muitas vezes não sejam aplicáveis a cenas naturais.

Neste trabalho, foca-se em regiões conectadas na imagem e como isolar e descrever tais estruturas. A tarefa é desenvolver um procedimento para encontrar o número e o tipo de objetos contidos em uma imagem. Considerando cada pixel isoladamente, não é possível determinar quantos objetos existem na imagem, onde estão localizados e quais pixels pertencem a quais objetos. Portanto, o passo é encontrar cada objeto agrupando todos os pixels que pertencem a ele. No caso mais simples, um objeto é um grupo de pixels

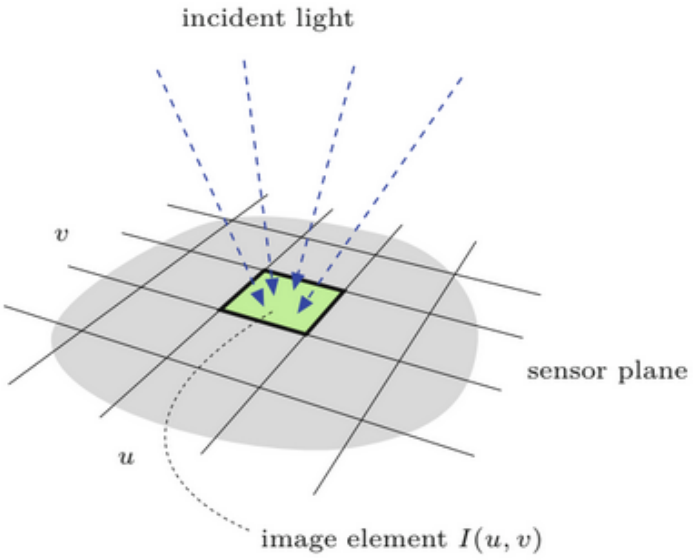


Figura 2.2: A geometria dos elementos sensores é diretamente responsável pela amostragem espacial da imagem contínua. Fonte: [2].

de primeiro plano que se tocam, ou seja, uma região binária conectada ou "componente".

Na busca por regiões binárias, as tarefas mais importantes são descobrir quais pixels pertencem a quais regiões, quantas regiões existem na imagem e onde essas regiões estão localizadas. Essas etapas geralmente ocorrem como parte de um processo chamado rotulagem de região ou coloração de região [2]. Durante esse processo, os pixels vizinhos são agrupados de maneira gradual para construir regiões nas quais todos os pixels dessa região recebem um número exclusivo (“rótulo”) para identificação. No método de marcação de região sequencial, a imagem é percorrida de cima para baixo, marcando as regiões à medida que são encontradas. Mas antes é preciso definir a vizinhança para determinar quando dois pixels estão "conectados"um ao outro, pois sob cada definição, pode-se acabar com resultados diferentes. Uma vez que as regiões em uma imagem tenham sido encontradas, o próximo passo é encontrar os contornos das regiões. Como tantas outras tarefas no processamento de imagens, à primeira vista basta seguir ao longo da borda da região.

Os pixels ao longo da borda de uma região binária podem ser identificados usando operações morfológicas simples e imagens de diferença (ou subtração) [2] [65]. Deve-se ressaltar, no entanto, que este processo marca apenas os pixels ao longo do contorno, o que é útil, por exemplo, para fins de reconhecimento/identificação. O algoritmo obtém uma sequência ordenada de coordenadas de pixel de borda para descrever o contorno de uma região. As regiões conectadas da imagem contêm o contorno externo, mas, devido aos

buracos, podem conter arbitrariamente muitos contornos internos. Dentro de tais buracos, regiões menores podem ser encontradas, que novamente terão seus próprios contornos externos, e por sua vez essas regiões podem conter outros buracos com regiões ainda menores, e assim por diante de maneira recursiva. Uma complicação adicional surge quando as regiões são conectadas por partes que diminuem até a largura de um único pixel. Nesses casos, o contorno pode percorrer o mesmo pixel mais de uma vez e em diferentes direções. Portanto, ao traçar um contorno a partir de um ponto inicial e retornar ao ponto inicial não é uma condição suficiente para encerrar o processo de traçado do contorno [2]. Outros fatores, como a direção atual ao longo da qual os pontos de contorno estão sendo percorridos, devem ser levados em consideração.

Uma maneira aparentemente simples de determinar um contorno é proceder em analogia ao processo de dois estágios [2] [65], ou seja, primeiro identificar as regiões conectadas na imagem e depois, para cada região, proceder ao seu redor, partindo de um pixel selecionado de sua borda. Da mesma forma, um contorno interno pode ser encontrado a partir de um pixel de borda de um buraco da região. O algoritmo combina localização de contorno e rotulagem de região em um único processo. Ele combina os conceitos de rotulação sequencial de regiões e traçado de contorno tradicional [2] em um único algoritmo capaz de executar ambas as tarefas durante a passagem pela imagem. Ele identifica e rotula regiões e traça seus contornos internos e externos. O algoritmo não requer estruturas de dados complicadas e é relativamente eficiente. Com a rotulagem sequencial da região, a imagem é percorrida do canto superior esquerdo para o canto inferior direito. Essa travessia garante que todos os pixels na imagem sejam eventualmente examinados e atribuídos a um rótulo apropriado.

Em uma determinada posição na imagem, podem ocorrer os seguintes casos:

- A transição de um pixel de fundo para um pixel de primeiro plano não marcado anteriormente significa que esse pixel está na borda externa de uma nova região. Um novo rótulo é então atribuído e o contorno externo associado é percorrido e marcado. Além disso, todos os pixels de fundo diretamente na borda da região são marcados com o rótulo especial -1.

- A transição de um pixel de primeiro plano para um pixel de fundo não marcado significa que esse pixel está em um contorno interno. A partir desse plano, o contorno interno é percorrido e seus pixels são marcados com rótulos da região circundante. Além disso, todos os pixels de fundo adjacentes recebem novamente o valor de rótulo especial -1.

- Quando um pixel de primeiro plano não está em um contorno, o pixel vizinho à esquerda já foi rotulado e esse rótulo é propagado para o pixel atual.

O algoritmo percorre a imagem linha por linha para que um novo contorno interno ou externo seja traçado. Os rótulos dos elementos da imagem ao longo do contorno, bem como os pixels vizinhos do primeiro plano são armazenados. Isso simplifica o processo de traçar os contornos da região externa, pois não é necessário nenhum tratamento especial nas bordas da imagem. Combinar a marcação de região e seguimento de contorno é particularmente adequado para o processamento de grandes imagens [2]. Os pixels isolados e seções finas são tratados corretamente pelo algoritmo ao seguir os contornos. Os contornos externos são como linhas passando pelos centros dos pixels do contorno, e os contornos internos linhas em brancas, por exemplo, para distinguir.

A comparação e classificação de regiões binárias é amplamente utilizada, por exemplo, em *Optical Character Recognition* (OCR). A característica de uma região é medida numérica ou qualitativa específica que é computável a partir dos valores e coordenadas dos pixels que compõem a região [2]. Por exemplo, uma das características mais simples é seu tamanho ou área. Esse é o número de pixels que compõem uma região. Para descrever uma região de forma compacta, diferentes características são combinadas em um vetor de características. Este vetor é então utilizado como uma espécie de "assinatura" da região que pode ser utilizada para classificação ou comparação com outras regiões. As melhores características são aquelas que são simples de calcular e não são facilmente influenciadas por mudanças irrelevantes, como por exemplo translação, rotação e dimensionamento [66] [2]. Depois de aplicar os filtros e o detector de características à imagem de entrada para gerar o mapa de característica que será utilizado pela CNN. A profundidade de um filtro em uma CNN deve corresponder à profundidade da imagem de entrada. O número de canais de cores no filtro deve permanecer igual ao da imagem de entrada. Diferentes filtros são criados para cada um dos três canais para uma imagem colorida. Os filtros para cada camada são inicializados aleatoriamente com base na distribuição Normal ou Gaussiana. As camadas iniciais de uma rede convolucional extraem características de alto nível da imagem, portanto, usam menos filtros. À medida que percorre as camadas mais profundas, o número de filtros aumenta pelo menos para duas vezes o tamanho do filtro da camada anterior. Os filtros das camadas mais profundas aprendem mais características, mas são computacionalmente muito custosos.

Ao clarear uma imagem adicionando um valor constante a todos os três canais RGB, não necessariamente pode-se dizer que isso atinge o efeito desejado de tornar a imagem

mais brilhante. Em alguns casos pode-se ver efeitos colaterais indesejáveis. Na verdade, adicionar o mesmo valor a cada canal de cor não apenas aumenta a intensidade aparente de cada pixel, mas também pode afetar o matiz e a saturação do pixel. As coordenadas de cromaticidade ou até mesmo proporções de cores mais simples podem primeiro ser calculadas e usadas após a manipulação da luminância para recalcular uma imagem RGB com o mesmo matiz e saturação [67]. Da mesma forma, o balanceamento de cores (por exemplo, para compensar a iluminação incandescente) pode ser realizado multiplicando cada canal com um fator de escala diferente ou pelo processo mais complexo de mapeamento do espaço de cores [2].

2.2.1.1 Projeto de filtro

A principal diferença entre filtros e operações de ponto é que os filtros geralmente usam mais de um pixel da imagem de origem para computar cada novo valor de pixel [2][67]. Muito usados na tarefa de suavizar uma imagem. As imagens parecem nítidas principalmente em locais onde a intensidade local aumenta ou diminui acentuadamente (ou seja, onde a diferença entre os pixels vizinhos é grande). Por outro lado, uma imagem borrada ou confusa pode ter a função de intensidade local suave. A primeira ideia para suavizar uma imagem poderia ser simplesmente substituir cada pixel pela média de seus pixels vizinhos. Essa média local simples já exibe todos os elementos importantes de um filtro típico. Em particular, é um chamado filtro linear, que é uma classe muito importante de filtros [2][67].

Primeiro, os filtros diferem das operações pontuais principalmente por usar não um único pixel de origem, mas um conjunto deles para computar cada pixel resultante. As coordenadas dos pixels de origem são fixas em relação à posição atual da imagem e geralmente formam uma região contínua. O tamanho da região do filtro é um parâmetro importante do filtro porque especifica quantos os pixels originais contribuem para cada valor de pixel resultante e, portanto, determinam a extensão espacial (suporte) do filtro [2][67]. Por exemplo, o filtro de suavização usa uma região de suporte 3×3 centrada na coordenada atual. Filtros semelhantes com suporte maior, como 5×5, 7×7 ou até mesmo 21×21 pixels, teriam efeitos de suavização mais fortes. A forma da região do filtro não é necessariamente quadrática ou mesmo retangular. Na verdade, uma região circular (em forma de disco) seria preferida para obter um efeito de desfoque isotrópico (isto é, um que seja o mesmo em todas as direções da imagem) [2][67]. Outra opção é atribuir pesos diferentes aos pixels na região de suporte, de modo a dar maior ênfase aos

pixels mais próximos do centro da região. Além disso, a região de suporte de um filtro não precisa ser contínua e pode nem conter o próprio pixel original (imagine um filtro em forma de anel, por exemplo). Teoricamente, a região do filtro pode até ser de tamanho infinito [2].

A Figura 2.3 ilustra a aplicação do filtro de caixa linear 3 × 3 em uma imagem em tons de cinza modificada com "ruído de sal e pimenta". No canto superior esquerdo está a imagem antes do filtro e no canto superior direito imagem filtrada. A baixo estão as figuras com detalhes ampliados para mostar que os pixels de ruído individuais são nivelados.

Figura 2.3: Exemplo da aplicação do filtro linear 3 x 3 em uma imagem em escala de cinza. Fonte: [2].

2.2.2 Mascaramento da Região de Interesse

ROI-Masking, ou *Region of Interest Masking*, é um conceito utilizado no processamento de imagem para destacar uma região de interesse específica em uma imagem e ignorar o restante [68]. É uma técnica comumente empregada quando se deseja realizar análises ou operações apenas em uma parte específica da imagem. A *ROI-Masking* envolve a criação de uma máscara que define a região de interesse, geralmente por meio de uma forma

geométrica, como um retângulo, círculo ou polígono. Em geral, a máscara é uma matriz binária em que os pixels dentro da região de interesse são marcados como "1"(branco) e os pixels fora da região de interesse são marcados como "0"(preto).

Ao aplicar a máscara à imagem original, todos os pixels correspondentes aos valores "1"na máscara são preservados, enquanto os pixels correspondentes aos valores "0"são eliminados ou atribuídos um valor de fundo. Isso cria uma nova imagem, chamada de imagem mascarada, contendo apenas a região de interesse definida pela máscara. A *ROI-Masking* é útil em várias aplicações, como detecção de objetos, segmentação de imagem, extração de características e medição de propriedades específicas de uma região. Ela permite que os algoritmos de processamento de imagem se concentrem apenas nas áreas relevantes, economizando tempo de processamento e evitando informações desnecessárias [68].

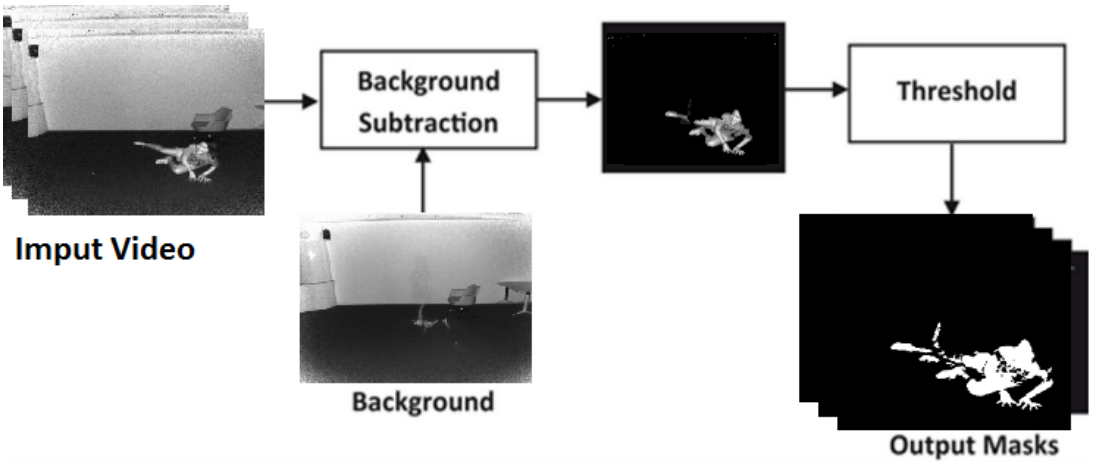


Figura 2.4: Procedimento de remoção de fundo e filtragem para se obter a imagem de máscara.

A Figura 2.4 mostra o procedimento geral de pré-processamento para se obter a imagem mascarada que representa a pessoa em um vídeo. Na Figura 2.4, o bloco 'Threshold' representa o filtro que será utilizado para se obter a imagem suavizada. Esse trabalho explora a variação de diferentes técnicas de subtração de fundo, filtragem e mascaramento, bem como em diferentes ordens de aplicação, para serem avaliados na sensibilidade do modelo de detecção de queda final.

2.2.3 Técnicas de remoção de fundo

As técnicas de remoção de fundo são usadas no processamento de imagem para isolar um objeto de interesse em uma imagem, removendo o fundo ou as áreas indesejadas ao redor dele [2]. Essas técnicas são amplamente aplicadas em várias áreas, como edição de fotos,

segmentação de objetos e detecção de objetos. Por exemplo, o Chroma Keying é uma técnica de chave de cor. Essa técnica envolve a seleção e remoção de uma cor específica do fundo da imagem [69]. É utilizada em produções de vídeo e fotografia, onde um fundo de cor sólida, geralmente verde ou azul, é usado como cenário. O objeto de interesse é capturado na frente desse fundo, e o algoritmo pode remover todas as áreas da imagem que correspondem à cor do fundo, substituindo-as por uma nova imagem ou fundo.

Outro exemplo são os algoritmos baseados em segmentação [70]. Esses algoritmos analisam as propriedades dos pixels da imagem para distinguir entre o objeto de interesse e o fundo. Existem várias técnicas de segmentação que podem ser usadas, como segmentação baseada em limiar, segmentação por crescimento de região, segmentação por detecção de borda, entre outras [70]. Esses algoritmos podem ser baseados em propriedades de cor, textura, forma ou outras características dos pixels para separar o objeto do fundo. Uma vez que a segmentação é realizada, os pixels correspondentes ao fundo podem ser substituídos por uma cor sólida, um fundo transparente ou uma nova imagem de fundo.

A eficácia das técnicas de remoção de fundo pode variar dependendo das características da imagem, como iluminação, complexidade do objeto e qualidade da captura. Além disso, em casos mais complexos, como cabelos ou objetos com bordas irregulares, a remoção de fundo pode ser um desafio e exigir técnicas mais avançadas, como aprendizado de máquina ou técnicas baseadas em modelos estatísticos [2].

2.2.4 Técnicas de detecção de pessoas

As técnicas de detecção de pessoas em imagens são usadas para identificar e localizar a presença de pessoas em uma imagem. Essas técnicas são amplamente aplicadas em várias áreas, como vigilância por vídeo, análise de multidões, reconhecimento facial e condução autônoma [71, 67]. Redes neurais convolucionais têm sido amplamente utilizadas na detecção de pessoas em imagens devido à sua capacidade de aprender representações complexas e hierárquicas das imagens [72]. Elas têm mostrado resultados bons e alcançaram altos níveis de desempenho em tarefas de detecção de pessoas [72].

Uma descrição básica de como as CNNs são aplicadas na detecção de pessoas envolve um conjunto de dados preparado, contendo imagens que contêm pessoas e imagens sem pessoas [71]. As imagens podem ser redimensionadas para um tamanho específico e convertidas para escala de cinza ou mantidas em cores, dependendo da abordagem adotada. Um modelo de CNN é projetado para a tarefa de detecção de pessoas. Esse modelo geralmente consiste em camadas convolucionais, camadas de pooling e camadas totalmente

conectadas [67]. As camadas convolucionais são responsáveis por aprender características relevantes nas imagens, enquanto as camadas totalmente conectadas são responsáveis por tomar decisões finais. A CNN é treinada usando o conjunto de dados rotulado. Durante o treinamento, os pesos da rede são ajustados iterativamente por meio de algoritmos de otimização, como o gradiente descendente, para minimizar a diferença entre as previsões da rede e as anotações do conjunto de dados [73].

Na etapa de detecção, ou seja, após a rede CNN estar treinada, a imagem de entrada é pré-processada para se adequar à entrada da rede, como redimensionamento e normalização [74]. A imagem pré-processada é passada pela rede neural convolucional, que gera uma saída que representa a probabilidade da presença de uma pessoa em diferentes regiões da imagem. A saída da rede é analisada e um threshold é aplicado para determinar as regiões que são consideradas como detecções de pessoas [67]. Além disso, técnicas de pós-processamento, como supressão de não-máximos, podem ser aplicadas para eliminar detecções redundantes ou sobrepostas [75, 67].

As CNNs podem aprender características discriminativas complexas, como formas corporais, partes do corpo e contextos, que são essenciais para a detecção precisa de pessoas em várias poses, tamanhos e ambientes [72]. A detecção de pessoas em tempo real requer um bom equilíbrio entre precisão e eficiência computacional, pois é necessário processar as imagens em tempo real, especialmente em aplicações como vigilância ou detecção de queda [73, 2].

2.2.5 O Fluxo Óptico

O fluxo óptico [76] [67] é o padrão de movimento aparente de objetos de imagem entre quadros consecutivos causado pelo movimento do objeto. É um campo vetorial 2D, onde cada vetor é um vetor de deslocamento que descreve o movimento dos pontos do primeiro quadro para o segundo, considerando que as intensidades de pixel de um objeto não mudam entre quadros consecutivos e que os pixels vizinhos têm movimento semelhante .

O movimento da imagem resulta da projeção do movimento de pontos ambientais que se movem em relação ao plano de imagem de um sensor. Tanto o sensor quanto o ponto filmado são livres para se mover de forma independente. O fluxo óptico (também chamado de velocidade da imagem) é uma aproximação calculada para este movimento de imagem supondo que as mudanças nas intensidades espaço-temporais na sequência são devidas ao movimento relativo do sensor e do ponto do ambiente. O calculo do fluxo óptico é um problema fundamental em Visão Computacional e tem muitas aplicações. O objetivo é

calcular um campo de movimento que alinha os pixels de uma imagem com os de outra. A estimativa de movimento é aplicada ao vídeo onde toda uma sequência de quadros está disponível para executar esta tarefa. Como o movimento do pixel é principalmente horizontal, as inclinações das trilhas de pixel individuais (texturizadas), que correspondem às suas velocidades horizontais, podem ser vistas claramente.

Considerando $I(x, y, t)$ o primeiro *frame*, na equação 2.1, x e y representam o *pixel* e t a dimensão de tempo. Se esse *pixel* se move uma distância (dx, dy) em um tempo dt no próximo *frame*. Como o objeto tem a mesma intensidade, o *pixel* é o mesmo:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{2.1}$$

Pela aproximação da série de *Taylor* do lado direito, removendo os termos comuns e divido por dt para obter a equação em 2.2:

$$f_x u + f_y v + f_t = 0 \tag{2.2}$$

Onde f_x, f_y, u e v são definidos como na equação em 2.3a 2.3b 2.3c 2.3d:

$$f_x = \frac{\partial f}{\partial x} \tag{2.3a}$$

$$f_y = \frac{\partial f}{\partial y} \tag{2.3b}$$

$$u = \frac{\partial x}{\partial t} \tag{2.3c}$$

$$v = \frac{\partial y}{\partial t} \tag{2.3d}$$

A equação 2.2 é chamada de equação de fluxo óptico, onde f_x e f_y são gradientes de imagem. O f_t é o gradiente ao longo do tempo. Não é possível resolver esta equação com duas variáveis (u, v) desconhecidas. Mas, o método *Lucas-Kanade*[76] [67][15] é um dos métodos para resolver esse problema. Considerando que todos os *pixels* vizinhos terão movimento semelhante, O método *Lucas-Kanade* considera uma matriz 3x3 ao redor do ponto. Todos os 9 pontos têm o mesmo movimento. Pode-se encontrar(fx, fy, ft) para esses 9 pontos. Então agora basta resolver 9 equações com duas variáveis desconhecidas. Uma melhor solução é obtida com o método de ajuste dos mínimos quadrados.

Então, a partir de alguns pontos para rastrear, calcula-se os vetores de fluxo óptico

desses pontos. Então, aplicando *Lucas-Kanade*, obtém o fluxo óptico. O *OpenCV* [16], uma biblioteca famosa para trabalhar com visão computacional utilizando a linguagem de programação *python*, fornece tudo isso em uma única função, *cv.calcOpticalFlowPyrLK()*. Movimento e estrutura 3D podem ser inferidos a partir de campos de velocidade 2D ou campos de deslocamento 2D ou diretamente de derivadas da intensidade. O fluxo óptico é usado para realizar detecção de movimento e segmentação de objetos. Também pode ser usado para calcular codificação compensada de movimento. O fluxo óptico calculado foi usado para uma imagem específica em uma sequência para calcular a próxima imagem da sequência.

2.2.6 Generative Adversarial Network

A Figura 2.5 mostra um esquema que ilustra didaticamente a relação da *Generative Adversarial Network* (GAN) dentro da área de *Artificial Intelligence* (AI), *Computer Vision* (CV) e ML, bem como suas dependências.

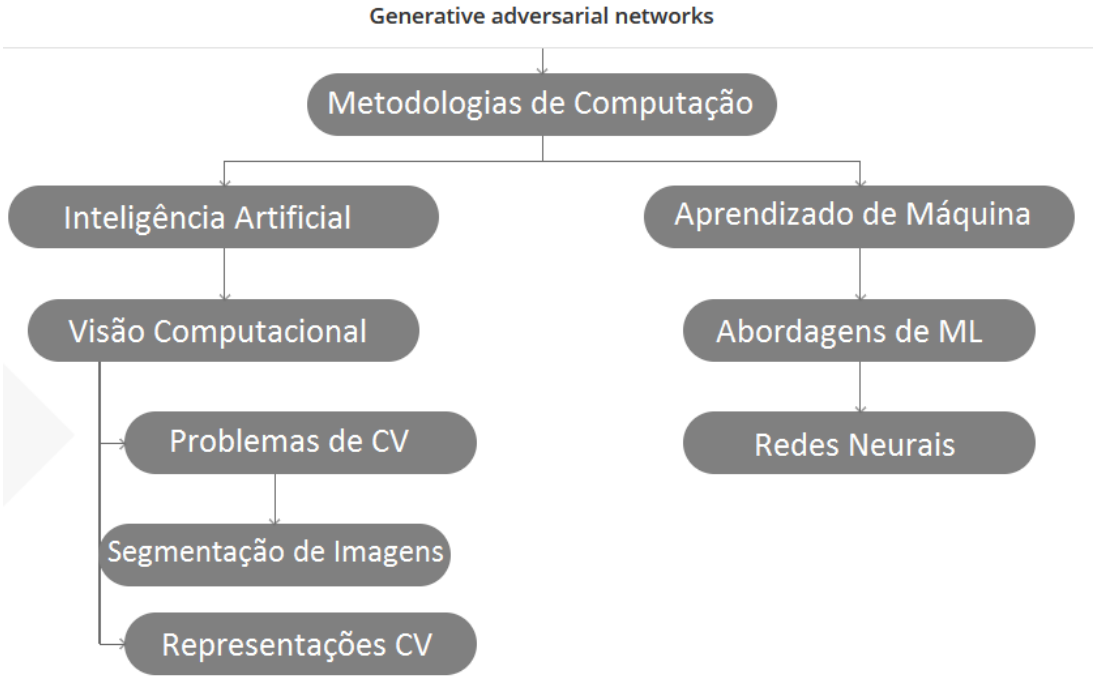


Figura 2.5: Representação da GAN dentro da área de Inteligência Artificial

As redes adversárias generativas são baseadas em um jogo, no sentido de teoria dos jogos, entre dois modelos de aprendizado de máquina, normalmente implementados usando redes neurais [65]. Uma rede chamada gerador define o $pmodel(x)$ implicitamente. O gerador não é necessariamente capaz de avaliar a função de densidade *pmodel*. Para algumas variantes de GANs, a avaliação da função de densidade é possível. Qualquer modelo

de densidade tratável para o qual a amostragem é tratável e diferenciável poderia ser treinado como gerador de GAN, mas isso não é necessário [65]. Em vez disso, o gerador deve ser capaz de extrair amostras da distribuição *pmodel*. O gerador é definido por uma distribuição a priori $p(z)$ sobre um vetor z que serve como entrada para a função do gerador $G(z;\theta(G))$ onde $\theta(G)$ é um conjunto de parâmetros que podem ser aprendidos definindo a estratégia do gerador no jogo. O vetor de entrada z pode ser pensado como uma fonte de aleatoriedade em um sistema determinístico, análogo à semente do gerador de números pseudo-aleatórios. A distribuição anterior $p(z)$ é normalmente uma distribuição não estruturada, como uma distribuição gaussiana de alta dimensão ou uma distribuição uniforme sobre um hipercubo [65]. As amostras z desta distribuição são então apenas ruído. A principal função do gerador é aprender a função $G(z)$ que transforma esse ruído não estruturado z em amostras realistas.

O outro jogador neste jogo é o discriminador. O discriminador examina amostras x e retorna alguma estimativa $D(x;\theta(D))$ se x é real (extraído da distribuição de treinamento) ou falso (extraído do *pmodel* executando o gerador). Na formulação original de GANs, essa estimativa consiste em uma probabilidade de que a entrada seja real em vez de falsa, assumindo que a distribuição real e a distribuição falsa são amostradas com a mesma frequência [65]. Outras formulações existem, mas de um modo geral, ao nível das descrições verbais e intuitivas, o discriminador tenta prever se a entrada foi real ou falsa.

Cada jogador incorre em um custo: $J(G)(\theta(G),\theta(D))$ para o gerador e $J(D)(\theta(G),\theta(D))$ para o discriminador [65]. Cada jogador tenta minimizar seu próprio custo. Grosso modo, o custo do discriminador o encoraja a classificar corretamente os dados como reais ou falsos, enquanto o custo do gerador o encoraja a gerar amostras que o discriminador classifica incorretamente como reais. Muitas formulações específicas diferentes desses custos são possíveis. Na versão original de GANs, $J(D)$ foi definido como o *log-likelihood* negativo que o discriminador atribui aos rótulos real e falso dado a entrada para o discriminador [65]. Em outras palavras, o discriminador é treinado como um classificador binário regular. O trabalho original sobre GANs oferecia duas versões do custo para o gerador. Uma versão, hoje chamada minimax GAN (M-GAN), definiu um custo $J(G) = -J(D)$. M-GAN define o custo do gerador invertendo o sinal do custo do discriminador [65]. Outra abordagem é o GAN não saturado (NS-GAN), para o qual o custo do gerador é definido invertendo os rótulos do discriminador. Em outras palavras, o gerador tenta minimizar a probabilidade logarítmica negativa (o *negative log-likelihood*) que o discriminador atribui aos rótulos errados. O último ajuda a evitar a saturação do gradiente durante o treinamento do modelo.

GANs são como falsificadores e policiais: os falsificadores fazem dinheiro falso enquanto a polícia tenta prender os falsificadores e continua permitindo a circulação de dinheiro legítimo. A competição entre falsificadores e a polícia leva a uma falsificação de dinheiro cada vez mais realista, até que eventualmente os falsificadores produzem falsificações perfeitas e a polícia não consegue distinguir entre dinheiro real e falso. Ilustrativamente, no modelo proposto, os *autoencoders* são os falsificadores, ou seja, os geradores.

Capítulo 3

Trabalhos relacionados

Existem várias abordagens para monitorar a atividade humana que podem ser empregadas para detectar quando ocorreu um evento de queda humana. Os sistemas de detecção de eventos baseados no monitoramento de atividades foram desenvolvidos usando diferentes técnicas de detecção [62]. Os sistemas de detecção de queda irão identificar a queda e alertar o contato de emergência (previamente determinado) de forma passiva. Isso ajuda muito, pois a severidade do acidente depende muito do tempo que a pessoa fica deitada após a queda, então se a pessoa cair e perder a consciência, alguém ficará sabendo imediatamente para tomar uma ação. Alguns sistemas usam sensores de smartphones como acelerômetros, giroscópios e magnetômetros para detectar quando a queda aconteceu [62]. A Figura 3.1 ilustra o esquema geral de um sistema de detecção de quedas baseado em sinais de radiofrequência. Cardenas et al [3] se concentram na etapa de detecção do sistema. Na etapa de detecção, um sinal de RF é transmitido por um transmissor através de um ambiente interno e captado por um receptor. Foi implementada uma plataforma operando na faixa de frequência de sistemas WiFi - IEEE 802.11. Esta plataforma permite experimentar diferentes orientações e posições das antenas. Os sinais transmitidos são recolhidos por um analisador de espectro que permite registar uma série de amostras da dispersão do sinal no domínio da frequência em todo o ambiente interior. Com esta informação é possível calcular os espectrogramas do sinal. Os experimentos são conduzidos em um ambiente controlado para evitar sinais de interferência gerados por reflexões de objetos em movimento que não sejam o alvo. O ambiente interno inclui escada de três degraus. Quando as quedas ocorrem o sistema registra e detecta as dispersões de sinal durante esse evento.

Entre essas tecnologias, destaca-se a iniciativa recente do uso de *Channel State Information* (CSI) de redes *Wi-Fi* para monitorar pacientes remotamente, quedas e podendo

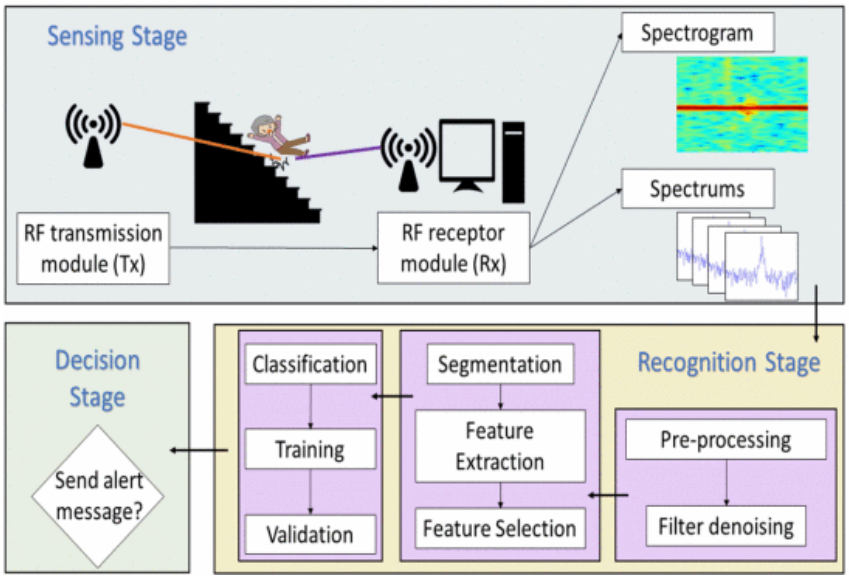


Figura 3.1: Esquema geral de um sistema de detecção de quedas baseado em sinais de radiofrequência. Fonte: [3].

fornecer meios para obter um poderoso pacote de informações médicas de forma não invasiva e com baixo custo [77, 78]. O sinal CSI representa a resposta em frequência do canal (*Channel Frequency Response* - CFR) para cada subportadora entre os pares de antenas de transmissão e recepção[79]. O CSI pode capturar as interferências que o corpo humano causa no sinal eletromagnético nos domínios do tempo e da frequência e em domínios espaciais. Essas informações podem ser usadas para diferentes aplicações, como a detecção da presença humana, detecção de movimentos, identificação humana, detecção de queda, reconhecimento de gestos, localização humana e monitoramento de sinais vitais e das condições de saúde [77, 79]. Para monitoramento da movimentação do paciente, as subportadoras do OFDM (*Orthogonal Frequency Division Multiplexing*) são usadas como vários sensores para detectar a mudança física de uma pessoa. Uma análise de forma de onda CSI permite detectar atividades mínimas do corpo humano, como a respiração, os batimentos cardíacos, dentre outros.

Liu et al [4] usam restrição de fluxo óptico e imagens reconstruídas para prever quadros futuros. Eles usam um modelo baseado em CNN para estimativa de fluxo, o que pode facilitar a retropropagação e realizar treinamento com imagens RGBs. Seguindo essa ideia, nesse trabalho é utilizada uma rede espaço-temporal para reconstrução de fluxo, que recebe como entrada uma janela de quadro do fluxo óptico. Esta rede calcula quadros de fluxo óptico denso para dois quadros consecutivos. O fluxo é combinado na direção e magnitude dos eixos horizontal e vertical para formar uma imagem tridimensional.

A detecção de anomalias em vídeos refere-se à identificação de eventos que não estão

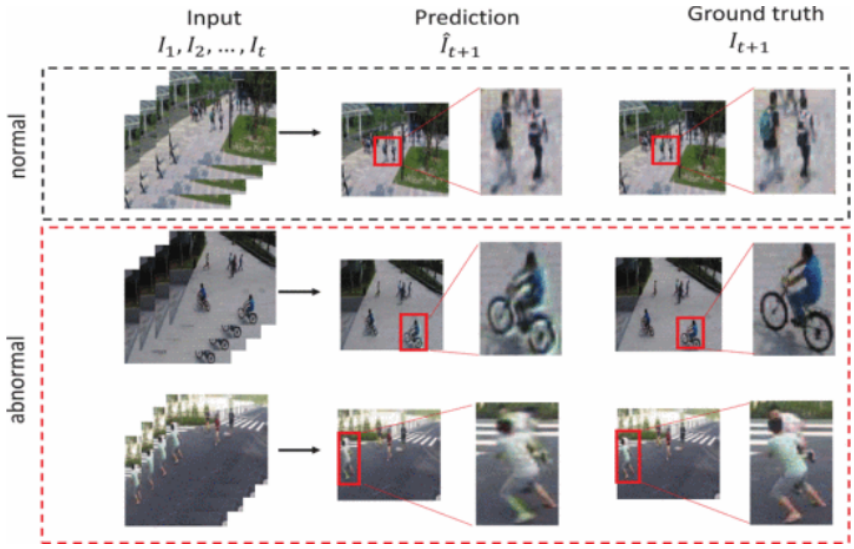


Figura 3.2: Alguns quadros preditos e sua verdade básica (ground truth) em eventos normais e anormais. Fonte: [4].

de acordo com o comportamento esperado. No entanto, quase todos os métodos existentes abordam o problema minimizando corrigir os erros de reconstrução dos dados de treinamento, o que pode não garantir um erro de reconstrução maior para um evento anormal. A Figura 3.2 ilustra essa idéia: nela a região é zona de caminhada, quando os pedestres estão andando na área, os quadros podem ser bem previstos. Enquanto para alguns eventos anormais (uma bicicleta ou dois homens brigando), as previsões são borradas e com distorção de cores. *Liu et al* [4] propõem abordar a anomalia como problema de detecção dentro de uma estrutura de previsão de vídeo. Este trabalho aproveita a diferença entre um quadro futuro e previsto e sua verdade fundamental para detectar um evento anormal para prever um quadro futuro com maior qualidade para eventos normais. Esse trabalho apresenta melhor visualização em cores. Além das restrições de aparência espacial comumente usadas na intensidade e no gradiente, também foi utilizada uma restrição de movimento temporal na previsão de vídeo, reforçando o fluxo óptico entre os quadros previstos para que a verdade fundamental pelo campo de quadros sejam consistentes. Este trabalho introduz uma restrição temporal na tarefa de previsão de vídeo. Tais restrições espaciais e de movimento facilitam a previsão futura de quadros para eventos normais e, consequentemente, facilitam a identificação de eventos anormais que não atendem à expectativa. Experimentos extensivos em um conjunto de dados disponíveis publicamente validam a eficácia do método em termos de robustez à incerteza em eventos normais e sensibilidade a eventos anormais.

Zahan et al [5] propõem um modelo de rede de convolução de grafos eficiente que explora dependências espaço-temporais e dinâmicas de articulações do esqueleto humano

para detecção precisa de quedas, melhorando os modelos existentes super parametrizados ou avaliados em pequenos conjuntos de dados com muito poucas classes de atividade. As articulações esqueléticas foram exploradas para extração de características com modelos de imagem que ignoram a dependência conjunta entre os quadros, o que é importante para a classificação das ações. O método aproveita a representação dinâmica com características espaço-temporais e simultâneas das articulações esqueléticas. A convolução do grafo usa a matriz de adjacência que incorpora informações de conectividade do nó na operação de convolução para extrair a dependência inerente dos nós. O esqueleto humano também pode ser representado como um grafo com as articulações sendo nós e os ossos sendo arestas. O trabalho é um modelo baseado em *Graph Convolutional Networks* (GCN) que utiliza adjacência de nós físicos para aprender uma melhor representação da estrutura do corpo. A Figura 3.3 mostra uma representação de esqueleto com adjacência de borda. A esquerda a representação do esqueleto dos vizinhos (azul) da articulação da coluna vertebral (vermelho). A direita a representação da matriz N×N para todas as 25 juntas.

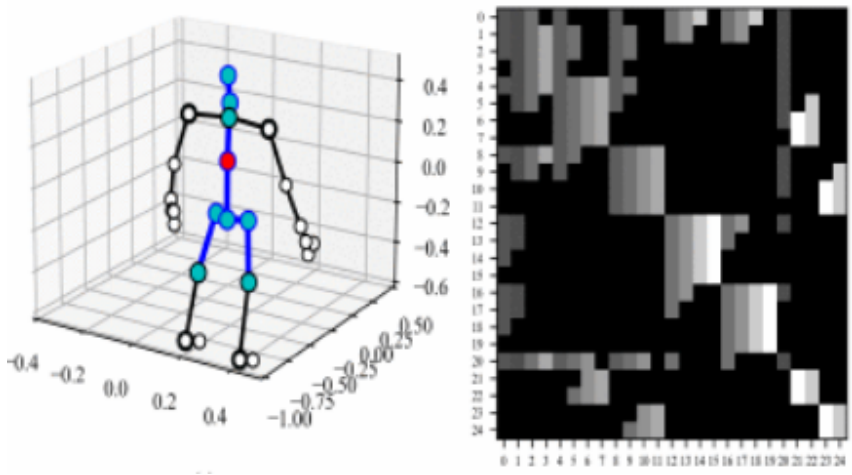


Figura 3.3: Ilustração da estrutura de esqueleto e a representação da matriz para todas as juntas. Fonte: [5].

A convolução 2D tradicional trata a entrada como uma imagem 2D e calcula o mapa de recursos de saída usando o filtro especificado para capturar formas e arestas [5]. A matriz de adjacência atua como uma operação de passagem, permitindo a convolução de grafo para aprender a adjacência do nó. Foi utilizado o particionamento de configuração espacial para gerar a matriz de adjacência sugerida, onde os nós vizinhos são divididos em três subconjuntos: nó raiz, grupo centrípeto: mais próximo do centro de gravidade do nó raiz, e grupo centrífugo: mais longe do nó raiz. Em vez de usar a adjacência usual, esse particionamento de distância aumenta a representatividade do gráfico para convolução espacial [5].

Steven et al [6] propõem um sistema de detecção de queda que usa uma câmera 2D. Assim, como na proposta desta dissertação, *Steven et al* utilizam uma biblioteca de código aberto (OpenCV) e técnicas de visão computacional para detecção de quedas. Eles também usaram uma técnica *Background Subtraction* (BS) que funciona em conjunto com a detecção de movimento para detectar quedas. Eles usaram sinalizadores e cronômetros para melhorar seu sistema. Os sinalizadores são usados para evitar falsos positivos quando a posição inicial é o solo. Os cronômetros determinam o tempo mínimo que uma pessoa deve ficar deitada no chão para detectar uma queda.

Algumas técnicas comuns nos sistemas de detecção de queda baseados em câmeras são: remoção de fundo estático/dinâmico, detecção de esqueleto, operações morfológicas, detecção de membros, etc. *Steven et al* [6] enfoca na remoção dinâmica de fundo e operações morfológicas. A subtração de fundo estático é a técnica de remoção de fundo mais fácil. Ele subtrai uma imagem estática de quadros de vídeo com o mesmo fundo (subtração de matriz). Uma das desvantagens dessa técnica é conhecida como duplicação de objetos. Quando um objeto que já está localizado no fundo é movido e uma segunda imagem do mesmo objeto será criada onde estava localizado ao fundo (criando uma imagem fantasmagórica) e outro no novo local como ilustra a Figura 3.4. Os modelos CVSC3 e CVSC4 propostos nesta dissertação contornam essa deficiência a partir da atualização periódica do fundo e comparação com quadros anteriores. Assim, os pixels que se repetem dentro da sequencia de frames são contados como fundo, contudo, o algoritmo constantemente faz essa verificação dentro de um loop.

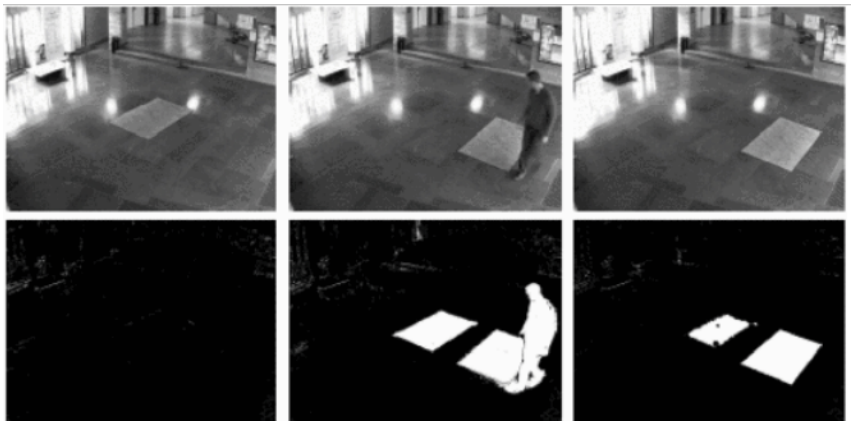


Figura 3.4: Ilustração da falha da técnica de subtração de fundo. Fonte: [6].

A subtração dinâmica de fundo é outra técnica usada em problemas de visão computacional [6]. Ele compara o quadro atual com o anterior e atualiza o fundo. Isso resolve o problema de duplicação experimentado na remoção de fundo estático. Uma das principais desvantagens dos sistemas baseados em câmeras está relacionada à privacidade das

pessoas. Em algumas situações é inadequado colocar uma câmera nos banheiros, embora exista uma alta taxa de quedas no banheiro [80]. Embora o sistema de detecção de queda analise as imagens de forma automática e independente da observação de um humano, algumas pessoas podem se sentir desconfortáveis com a presença de câmeras. Algumas soluções possíveis incluem gravar imagens apenas quando ocorrer uma queda ou usar a câmera no peito/cintura e capturar as imagens da movimentação e não do sujeito [6].

Chen et al [7] desenvolveram uma abordagem de detecção de queda baseada em vídeo usando poses humanas. Em primeiro lugar, um estimador de pose leve extrai poses 2D de sequências de vídeo e, em seguida, as poses 2D são aumentadas para poses 3D. Na segunda fase, uma rede robusta de detecção de queda que inclui uma CNN para reconhecer eventos de queda usando poses 3D estimadas aumenta o respectivo campo e mantém um baixo custo computacional para convoluções. A estimativa da pose humana 3D consiste em localizar a posição dos pontos-chave humanos no espaço 3D a partir de imagens ou vídeos. Assim, a abordagem de detecção de quedas consiste em duas etapas, onde a primeira é estimar poses 3D em sequências de vídeo e a segunda é reconhecer eventos de queda a partir de poses 3D estimadas. O primeiro passo é compatível com qualquer estimador de pose de última geração [7]. Um estimador de pose 2D leve e uma rede de elevação são adotados para reduzir o custo de computação. Na segunda etapa, o artigo apresenta uma rede de detecção de quedas tomando como entrada poses 3D de cada quadro.

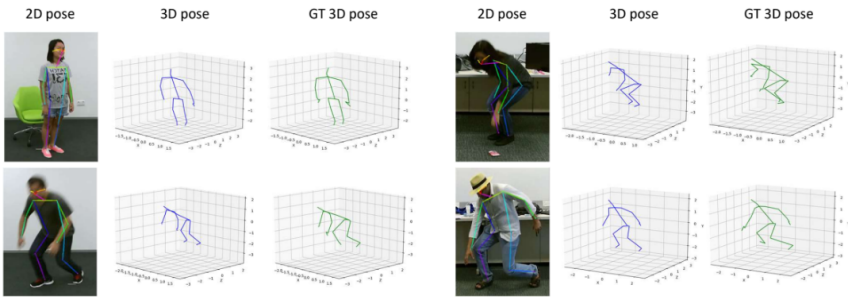


Figura 3.5: Resultados qualitativos de algumas imagens de exemplo. São apresentadas a imagem inicial, poses 2D, poses 3D e poses GT 3D. Fonte: [7].

Para obter uma precisão convincente e, ao mesmo tempo, manter o baixo custo de computação para longas sequências de vídeo, é adotada a convolução temporal dilatada unidimensional [7]. O artigo explora os efeitos de fatores que poderiam contribuir para o desempenho da detecção de quedas, incluindo juntas de entrada e função de perda. A rede de detecção de queda é uma rede totalmente convolucional com conexões residuais que toma uma sequência de poses 3D como entrada e prevê se há um comportamento

de queda. Em redes convolucionais, o caminho do gradiente entre a saída e a entrada tem um comprimento fixo, o que mitiga o desaparecimento e a explosão de gradientes [7]. O número de quadros de *batch* foi definido como 300 para reconhecer quedas em uma sequência de vídeo longa. Além disso, convoluções dilatadas são aplicadas na rede para modelar dependências de longo prazo, mantendo a eficiência. Os primeiros métodos de estimativa de pose humana 3D prevêem diretamente as coordenadas das articulações 3D por meio de redes neurais profundas. Embora um conjunto de recursos adequados para a tarefa possa ser aprendido espontaneamente, esses modelos geralmente possuem grandes recursos computacionais, custo de implantação e alta complexidade. Devido ao desenvolvimento da estimativa de pose humana 2D, os métodos baseados em poses 2D tornaram-se a corrente principal. As poses 2D são concatenadas como entrada para prever as informações de profundidade de cada ponto-chave, reduzindo bastante a complexidade do modelo [7].

Xu et al [8] desenvolveram uma CNN para detecção de queda através da formação de um mapa 2D do corpo ósseo do sujeito identificado nas gravações de câmeras RGBs. Eles usaram OPENPOSE [81] para converter a imagem na imagem do esqueleto correspondente. Em seguida, usando aprendizado de transferência, o conjunto de dados foi usado para treinar um novo modelo de detecção de queda. Tanto o trabalho de *Chen et al* e o de *Xu et al* foram utilizados para a comparação com o modelo proposto nesta pesquisa por serem modelos de CNNs para detectar quedas específicas através de gravações RGBs. OPENPOSE é o código divulgado publicamente da CMU University, e é capaz de converter cada frame em uma imagem de esqueleto correspondente. A imagem do esqueleto de cada quadro foi salva para fornecer o conjunto de dados. Terceiro, foi usado o método de aprendizagem por transferência para obter o conjunto de dados e treinar o mapa do esqueleto correspondente. Assim foi obtido o modelo de detecção de queda. A principal contribuição do trabalho deles é o mapa de esqueleto para detecção de quedas. OPENPOSE é uma biblioteca C++ que pode realizar detecção de pontos-chave de várias pessoas em tempo real, que pode ser compartilhada em uma única imagem. O corpo humano pode ser detectado em conjunto em uma única imagem como mostrado na Figura 3.6. A descrição dos pontos-chave da mão e do rosto (um total de 130 pontos-chave) é muito detalhada, podendo ser descritos dezenas de pontos-chave nos olhos, sobrancelhas, nariz e boca como mostrado na Figura 3.6. OPENPOSE foi originalmente implementado com OpenCV e Caffe, e a versão OpenCV e TensorFlow foi lançada a pouco tempo [8].

OPENPOSE pode inserir a imagem colorida e usar o modelo MobileNet [8] para obter a imagem do esqueleto 2D, conforme mostrado na Figura 3.6. Para imagens de

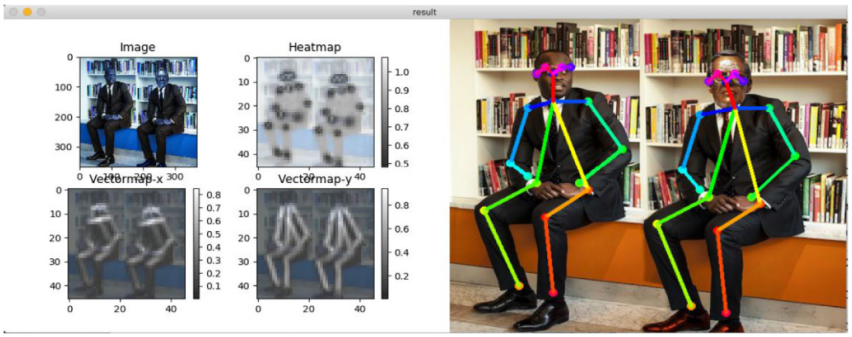


Figura 3.6: Imagem em escala de cinza, mapa de calor, mapa ósseo obtido em imagem de alta definição. Fonte: [8].

alta definição, o efeito será muito bom. Portanto, a resolução da imagem é necessária. OPENPOSE pode fotografar e exibir o mapa do esqueleto em tempo real através da câmera. Esta função possui requisitos de desempenho de hardware mais elevados. São necessárias pelo menos quatro GPUs para ter o efeito em um grande grupo de pessoas. Usando um notebook normal, cada quadro do vídeo terá um atraso de 1 s a 1,5 s [8]. E o mapa ósseo não pode ser exibido corretamente e o efeito em tempo real é muito ruim. Este trabalho tem um bom efeito em locais onde não há obstrução de objetos e a iluminação é forte. Quando houver muitas máscaras nas lentes ou se o ambiente ao estiver muito escuro, os resultados experimentais serão bastante reduzidos [8]. Como exibir pontos corporais ocluídos na presença de obstruções é uma possibilidade de pesquisa futura.

Mehta et al [9] propuseram o 3D Convolutional Autoencoder (3DCAE) em conjunto com o fluxo óptico para detecção de eventos anormais aplicado à detecção de quedas. A estrutura proposta com aprendizagem adversária generativa usa movimento e região para detectar quedas com imagens térmicas. O modelo consiste em uma rede de dois canais, com um canal aprendendo explicitamente o movimento na forma de um fluxo óptico, enquanto o outro recebe quadros de vídeo brutos como entrada. A abordagem pode lidar com situações em que uma pessoa pode não estar presente em um quadro, reduzindo a taxa de falsos positivos. A Figura 3.7 ilustra o modelo proposto. Esse modelo foi utilizado como referência no modelo CVSC.

O modelo tem como base que os recursos espaço-temporais de aprendizagem utilizando a região e movimento em sequências de vídeo melhoram a detecção de quedas quando treinados com a GAN. Para esse fim, a estrutura com reconhecimento de movimento e região consiste em dois canais separados otimizados em conjunto. O primeiro canal recebe a sequência de vídeo térmica e o segundo canal recebe a entrada com o fluxo óptico correspondente. As saídas de ambos os canais são combinadas para fornecer uma

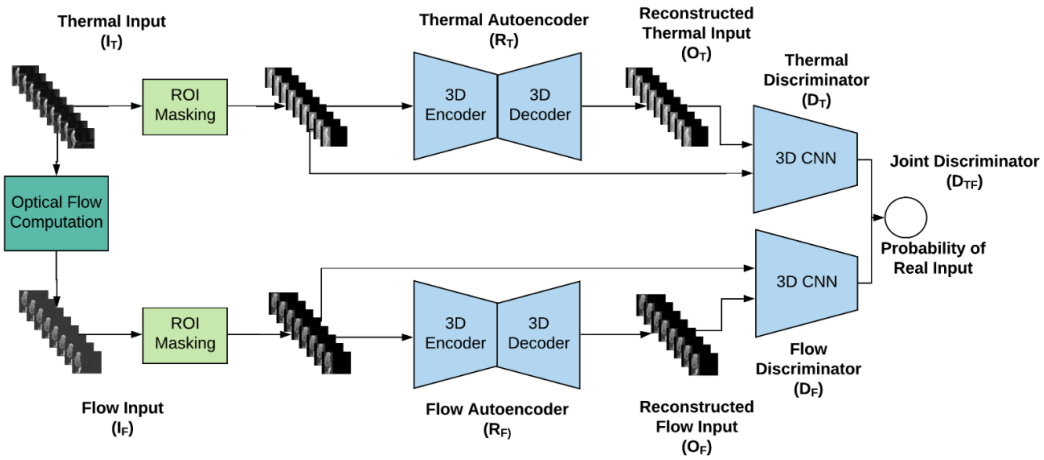


Figura 3.7: Modelo proposto por *Mehta et al*, a qual foi o modelo de base para o CVSC. Fonte: [9]

pontuação discriminativa para o aprendizado da rede GAN. O treinamento conjunto de canais de fluxo térmico e óptico facilita o aprendizado de características discriminatórias baseadas em movimento e região [9]. Eles Realizaram o rastreamento de pessoas em vídeos térmicos de forma que quando uma pessoa não é detectada continuamente por 10 quadros, o algorítmo quebra o vídeo e cria subvídeos, resultando em 22 subvídeos, que contêm ainda 12.454 quadros de vídeos ADL para treinamento. Após a conclusão do treinamento, o bloco 'Thermal Autoencoder', na Figura 3.7, seria capaz de reconstruir sequências ADL de forma eficiente e o bloco '3DCNN', na Figura 3.7, seria capaz de diferenciar entre sequências ADL reais e reconstruídas.

Durante o teste, quando uma sequência de vídeo contendo quadros de queda é mostrada para esta rede, um alto erro de reconstrução e/ou baixa probabilidade do discriminador indicará uma sequência de vídeo anômala. Portanto, este modelo é capaz de identificar quedas com alta precisão. O erro de reconstrução do bloco 'Thermal Autoencoder' ou a saída de probabilidade do bloco '3DCNN' ou sua combinação pode ser usada como uma pontuação de anomalia para identificar quedas invisíveis durante o teste. Assim, o modelo compreende autoencoders convolucionais 3D de dois canais que reconstroem os dados térmicos e as sequências de entrada do fluxo óptico, respectivamente. Portanto, o modelo introduz uma técnica para rastrear a região de interesse, uma restrição de diferença baseada na região e um discriminador conjunto para calcular o erro de reconstrução. Um erro de reconstrução maior indica a ocorrência de uma queda.

Nas imagens térmicas, uma pessoa pode parecer mais brilhante que o fundo devido às diferenças no calor emitido pela pessoa e pelo objeto. No rastreamento da região de interesse, o limite de Otsu [9] é aplicado à imagem térmica para separar o fundo escuro,

Figura 3.8: Quadros resultantes da técnica de rastreamento da região de interesse proposta por *Mehta et al* para câmeras térmicas. Fonte: [9]

como mostrado na Figura 3.8. No canto superior esquerdo da Figura 3.8, mostra a imagem termal original, no canto superior direito a imagem com o limite de Otsu, no canto inferior esquerdo a imagem onde o contorno da pessoa é destacado e no canto inferior direito a imagem com a caixa delimitadora mínima destacada. A imagem com limite ainda pode conter objetos de fundo brilhantes.

Neste trabalho, foi realizado uma expansão de [9] em ROI e pré-processamento. Foi assumido um cenário diferente utilizando cameras RGB e IR ao invés de cameras térmicas e onde falsos negativos não são toleráveis. Os quadros do vídeo foram transformados em escala de cinza e redimensionados. Tanto nas imagens RGB e IR foi realizado a operação de fechamento (closing) morfológico de imagem. O fechamento morfológico é uma técnica utilizada em processamento de imagens para remover pequenos buracos e detalhes escuros (ruídos) nas regiões claras de uma imagem, e também para preencher pequenas lacunas nas regiões escuras. O fechamento é uma combinação de dilatação seguida de erosão. Ele é útil para preencher buracos em regiões claras e remover pequenos objetos escuros como pode ser observado na Figura 3.9.

O pré-processamento proposto possibilita a utilização de imagens RGB e IR. Foi usado o TMF para melhorar a sensibilidade do modelo IR. A técnica de rastreamento da região de interesse melhora o desempenho do sistema na alteração do fundo da imagem e dos objetos de fundo. Esse modelo foi chamado de *Convolutional Video Stream Combination* (CVSC)

Figura 3.9: Quadro resultante após a aplicação da operação de fechamento em uma imagem infra-vermelho.

2. Uma alternativa ao modelo CVSC anterior, alterando a técnica BS para o método de contagem de pixel. Esse modelo foi chamado de CVSC 3. O algoritmo CNT é um método que usa apenas informações dos valores de pixel dos quadros anteriores. CNT BS trabalha com contagem de quadros. Se a cor do pixel permanecer estável por um determinado número de quadros, o algoritmo a considera fundo; caso contrário, é instável, portanto, é removido. O CNT melhorou o tempo de processamento do modelo de [14]. A filtragem de fundo mostrou-se uma técnica importante para diminuir o tempo de processamento e melhorar a sensibilidade do modelo. Assim, a dissertação descreve e avalia um modelo de rede neural chamado de CVSC para melhorar o monitoramento e a segurança de indivíduos com risco de queda, como idosos ou pessoas com mobilidade reduzida. O modelo trabalha com gravações IR, pois quedas também podem ocorrer em ambientes com pouca luz.

Existem basicamente dois tipos de sistemas de detecção de queda: ambientais e vestíveis (wearables). Os dispositivos ambientais são instalados na parede da casa ou mesmo no piso e monitoram apenas aquele ambiente. Então, se a pessoa cair fora de casa, ele não irá detectar. O ponto positivo deles é a pessoa não precisar usar nenhum tipo de equipamento, então não há o risco do esquecimento ou não adaptação a um relógio ou pingente. Além disso, também pode monitorar se a pessoa se levantou da cama, frequência de idas ao banheiro ou se saiu de casa. Esses sistemas podem funcionar por rádio frequência, infravermelho, câmera, sensor de pressão ou até mesmo microfone, mas ainda são muito

Tabela 3.1: Principais características entre os modelos de detecção de queda.

Modelo	**Modalidade de detecção**	**Características específicas**
CVSC	Câmeras RGB e IR	Adaptação com relação a variações de iluminação e objetos de fundo.
[9]	Câmeras térmicas	Usa câmeras térmicas para garantir a privacidade das pessoas.
[8]	Câmeras RGB	Usa mapas de pontos corporais de imagens 2D para fazer detecções de queda.
[7]	Câmeras RGB	Alta precisão de no conjunto de dados NTU RGB+D e desempenho em tempo real em plataforma sem GPU.
[6]	2D camera	A subtração de fundo (dinâmica) e a detecção de movimento são usadas para melhorar o desempenho na detecção de quedas.
[5]	Dados do esqueleto	Usa uma arquitetura leve que é mais fácil de implementar em sistemas embarcados.
[4]	Câmeras RGB	Usa restrição de fluxo óptico e imagens reconstruídas para prever quadros futuros.

difíceis de encontrar no mercado nacional [1].

A Tabela 3.1 mostra as principais vantagens dos principais modelos de visão computacional para detecção de queda. A característica que se destaca no modelo proposto deste trabalho é a adaptação com relação a variações de iluminação e objetos de fundo. O CVSC é o único com mais de uma modalidade de detecção: câmeras RGB e IR. O modelo proposto em [6] se destaca pela subtração de fundo (dinâmica) e a detecção de movimento são usadas para melhorar o desempenho na detecção quedas. Sinalizadores e temporizadores também são empregados para melhorar o sistema. O modelo proposto em [6] Usa uma abordagem baseada na identificação dos pontos corporais em uma estrutura de esqueleto como modalidade de detecção para eliminar preocupações com privacidade e possui uma arquitetura leve que é mais fácil de implementar em sistemas embarcados.

A Tabela 3.2 mostra as principais desvantagens dos modelos de visão computacional para detecção de queda. A desvantagem do modelo de detecção de queda proposto é que

[1] https://www.larpontoi.com/post/qual-detector-de-queda-escolher

Tabela 3.2: Desvantagens entre os modelos de detecção de queda.

Modelo	**Modalidade de detecção**	**Desvantagens**
CVSC	Câmeras RGB e IR	Não detecta queda na presença de grupos de pessoas.
[9]	Câmeras Térmicas	Câmera custosa. Custo de implantação em mais de um ambiente.
[8]	Câmeras RGB	Baixo desempenho na presença de obstruções ou em ambientes escuros.
[7]	Câmeras RGB	Lightweight Pose Network pode ser alterado para outro estimador de pose para maior eficiência e robustez.
[6]	câmeras 2D	O sistema pode ser melhorado através da detecção do piso, eliminando a necessidade de limites de altura.
[5]	Dados do esqueleto	Para aplicações reais, exige o pré-processamento da extração dos dados de esqueleto dos frames.
[4]	Câmeras RGB	Não pode prever o próximo quadro em um momento de movimento abrupto.

ele não detecta queda na presença de várias pessoas no ambiente. Contudo, quando possui mais pessoas no ambiente não seria tão necessário alertar a ocorrência da queda tento em vista que já teria pessoas próximas. Em geral, os modelos de visão computacional tem como desvantagem que em alguns ambientes, como em casa geriátricas ou hospitais, não é adequado colocar uma câmera no banheiro, mas o banheiro costuma ser um ambiente com alto risco de queda. O modelo proposto em [9] é o que lida melhor com a questão da privacidade por utilizar câmeras térmicas, onde a identificação das pessoas é mais difícil. Contudo, ainda assim poderia causar desconforto a presença de câmera no banheiro de um hospital ou casa geriátrica. Além disso as câmeras térmicas possuem um custo mais elevado do que as câmeras RGB. O custo do projeto de monitoramento remoto poderia ser ainda maior proporcionalmente ao número de câmeras térmicas. Além disso, muitos dos hospitais e casas geriatrias já possuem câmeras de monitoramento com RGB e IR, então o ideal seria aproveitar o sistema de monitoramento já instalado.

Os modelos de visão computacional poderiam de fato contar também com um pré-processamento para desidentificar as pessoas no vídeo, mas ainda assim não mudaria o

desconforto da presença de câmera no banheiro. O modelo proposto em [6] usa uma abordagem baseada na identificação em vídeo dos pontos corporais em uma estrutura de esqueleto como modalidade de detecção para diminuir as preocupações com privacidade. Contudo, ainda assim o idoso poderia se sentir desconfortável com a presença de câmera em alguns ambientes e pode não entender que o monitoramento possui desidentificação de pessoas no vídeo. É importante ressaltar que a prevenção de quedas é sempre importante e que nenhum dispositivo trabalha com prevenção. Para prevenir quedas, é importante falar com profissionais de saúde, incorporar atividade física no dia-a-dia e deixar o ambiente adequado a fim de reduzir riscos de quedas.

Capítulo 4

Modelo Proposto

Os sistemas baseados em vídeo têm limitações na detecção de quedas devido a mudanças no fundo da imagem, objetos de fundo, iluminação e movimento da câmera [9]. Algumas abordagens não funcionam em ambientes pouco iluminados e não são adequadas para detectar quedas usando câmeras infravermelhas, podendo não funcionar em ambientes escuros [6]. A proposta do modelo CVSC é a adaptar o pré-processamento de imagem para que o modelo de detecção de queda possa ter alta sensibilidade com relação a variações de iluminação e objetos de fundo. Assim, o CVSC pode ser capaz de lidar com mais de uma modalidade de detecção: câmeras RGB e IR. O CVSC expande o modelo de Metha et al [9] com diferentes técnicas de pré-processamento visando melhorar a qualidade da detecção, considerando os tipos mais usuais de câmeras, visando o cenário de detecção de queda de idosos em casas de repouso. O modelo de Metha et al utiliza câmeras térmicas. Assim, o pré-processamento do CVSC visa melhorar o desempenho do modelo para câmeras RGB e IR pois muitos dos hospitais e casas geriatrias já possuem câmeras de monitoramento com RGB e IR. Além disso o modelo CVSC deve alternar entre o modo de operação junto com a câmera, ou seja, quando a iluminação diminuir, o sensor da câmera identifica a baixa luminosidade e troca para o modo de operação infra-vermelho e passa a enviar um stream de video IR ao invés de RGB. Assim, que o modelo CVSC passar a receber esse novo stream de vídeo IR. Ele identifica o stream IR e trocar também o seu modo de operação de RGB para IR. O CVSC pode realizar esse identificação mudar o seu modo de operação porque a resolução das gravações IR e RGB são diferentes.

4.1 Convolutional Video Stream Combination - CVSC

O CVSC [14] é um modelo baseado em um modelo de rede neural proposto para a detecção de quedas. Esse modelo inicial serviu como base para o estudo de mais técnicas de pré-processamento, que levaram ao desenvolvimento dos modelos CVSCs 2, 3 e 4. O objetivo dos modelos propostos e analisados é apoiar a detecção de quedas de idosos com alta sensibilidade, ou seja, impedindo com alta probabilidade que uma queda seja erroneamente classificada como uma não-queda. É importante notar que esse é o principal requisito do sistema. Além disso, é importante que não sejam gerados muitos alarmes falsos de queda, para que os alarmes não venham a ser desconsiderados pelos cuidadores e familiares.

O CVSC é um sistema de monitoramento de detecção de queda para idosos em ambientes fechados; porém, visando cenários onde o idoso tenha alguma independência, o sistema deve ser capaz de atuar tanto em ambientes claros quanto escuros, alertando quedas sem a necessidade de monitoramento humano. Cabe destaque que, mesmo em casas de repouso, usualmente os idosos não são assistidos 100% do tempo, muito embora existam pessoas no local que podem auxiliar em casos de emergência. Assim, o sistema auxilia gerando alarmes para essas pessoas próximas quando quedas ocorrem nos momentos que os idosos estão desassistidos.

No cenário alvo tratado, usualmente, a pessoa estará sozinha na cena da câmera, uma vez que é esse o caso no qual uma queda precisa ser notificada a terceiros que não estão no local. Diferentemente de outras propostas, entende-se que o sistema deve ser capaz de atuar nas situações de claro e escuro para uso real com idosos. Como base para o CVSC, foi utilizado o método de detecção de anomalia, de forma que o sistema identifique a queda considerando a distribuição desbalanceada de eventos das atividades cotidianas [14]. É um paradigma de classificação no qual o padrão de eventos normais é aprendido a partir dos desvios de distribuição.

A Figura 4.1 é a visão geral do modelo CVSC, que propõe novos mecanismos de processamento de imagem e *ROI Masking*, estendendo [9] para dar suporte a câmeras RGB e infravermelho, além de aumentar a sensibilidade do algoritmo classificador. O algoritmo CNN foi usado para calcular o desvio de reconstrução conforme mostrado na Figura 4.1. Após as gravações, o bloco 'Pré-Processamento de Imagem' executa o redimensionamento da imagem, a transformação em tons de cinza e diminui a taxa de quadros de entrada.

A CNN calcula o desvio da reconstrução [14]. Em seguida, dois fluxos são executados em paralelo, como em [9]: um explicitamente aprendendo o movimento na forma de um

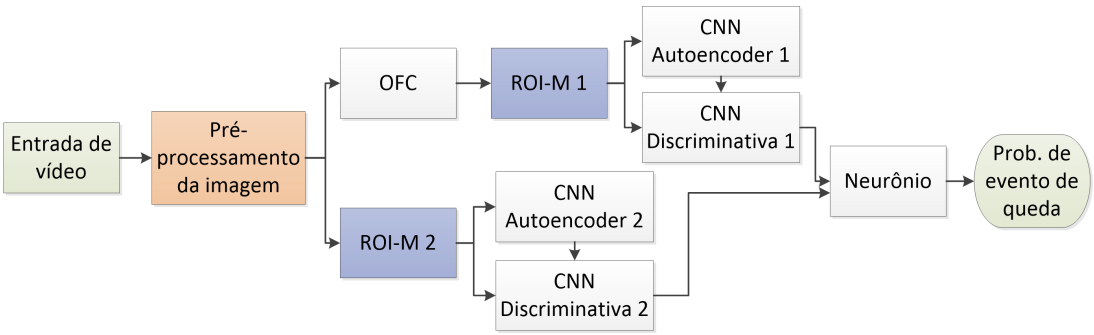


Figura 4.1: Visão geral do modelo CVSC, a qual é baseada na proposta de Mehta et al [9].

fluxo óptico [15] e o outro recebendo quadros de vídeo brutos como entrada.

Na implementação proposta, o CVSC redimensiona a escala da imagem para 640×480 e divide todos os quadros de vídeo em janelas de 8 quadros usando o método de janela deslizante com passo (step) 1 [14]. Em seguida, são aplicados filtros para melhorar a detecção de pessoas nas imagens. É possível usar uma combinação dos filtros para destacar a diferença de tonalidade. O filtro de Kalman contribui para rastrear o idoso. O algoritmo aplica com velocidade constante o filtro de Kalman nas coordenadas superior esquerda e inferior direita da caixa delimitadora. O código e detalhes de implementação estão disponíveis no *GitHub* em [82].

4.1.1 Seleção de Filtros

O filtro de limiarização realiza uma operação de limiarização em um quadro de imagem (*frame*). A limiarização é uma técnica de processamento de imagem que converte uma imagem em escala de cinza em uma imagem binária, onde cada pixel é convertido em preto ou branco com base em um valor limite (threshold) que é aplicado a todos os pixels da imagem [2]. O valor máximo que pode ser atribuído a um pixel da imagem binária é definido como 255 neste trabalho. Foi utilizada a técnica de limiarização binária com um valor limiar otimizado automaticamente usando o algoritmo Otsu no quadro de imagem (frame) [19]. O resultado é uma imagem binária com valores de pixel de 0 ou 255. O método de Otsu é utilizado para buscar um *threshold* ideal para separação dos elementos na frente e no fundo de uma imagem.

O filtro de fechamento realiza uma operação de fechamento morfológico em uma imagem. O fechamento morfológico é uma técnica de processamento de imagem que é usada para preencher pequenos buracos e descontinuidades em objetos de uma imagem binária ou em escala de cinza [2]. O fechamento morfológico é uma combinação de uma operação

(a) Frame final após duas iterações da operação.

(b) Frame após 10 iterações da operação.

(c) Frame após 50 iterações da operação.

Figura 4.2: Efeito do número de iterações da operação morfológica.

de dilatação seguida de uma operação de erosão. A operação de dilatação expande as áreas dos objetos na imagem, enquanto a operação de erosão remove pequenas lacunas dentro dessas áreas [19]. O kernel estruturante é uma matriz que define a forma e o tamanho da vizinhança que será considerada em cada pixel da imagem durante a operação morfológica [16].

A Figura 4.2 mostra o efeito do filtro de fechamento sobre o frame final. Quando maior o número de iterações maior é a área em torno da pessoa identificada pelo algorítimo. A Figura 4.2(a) mostra que o frame final após duas iterações da operação morfológica de fechamento. Na Figura 4.2(a) o contorno da pessoa foi detectado melhor. A Figura 4.2(b) mostra que o frame final após 10 iterações da operação morfológica de fechamento e a Figura 4.2(b) após 50 iterações da operação.

Nesse trabalho, o filtro de fechamento foi aplicado sobre o frame de máscara. O resultado é uma imagem com pequenos buracos e descontinuidades preenchidos e com uma aparência mais suave e uniforme como mostra a figura Figura 4.2. Como o objetivo do pré-processamento é diminuir a influência de objetos de fundo e iluminação, no trabalho foi utilizado duas iterações da operação morfológica para que o algoritmo fosse capaz de identificar a pessoa com a menor área de contorno e ainda assim garantir uma imagem suavizada. Assim, o filtro de fechamento aplica uma operação de fechamento morfológico na imagem usando um kernel estruturante e repetindo a operação duas vezes. O resultado é uma imagem com pequenos buracos e descontinuidades preenchidos e com uma aparência mais suave e uniforme.

O filtro de abertura realiza uma operação de abertura morfológica em uma imagem. A abertura morfológica é uma técnica de processamento de imagem usada para remover ruídos e outras pequenas irregularidades de uma imagem binária ou em escala de cinza [2]. A abertura morfológica é uma combinação de uma operação de erosão seguida de uma operação de dilatação. A operação de erosão remove pequenos objetos da imagem, enquanto a operação de dilatação expande as áreas restantes [19]. O kernel estruturante é uma matriz que define a forma e o tamanho da vizinhança que será considerada em cada pixel da imagem durante a operação morfológica [16]. Neste caso, a operação será aplicada duas vezes. Foi realizado apenas duas interações para garantir que o filtro aplicasse uma operação morfológica suave. A partir de duas iterações, a alteração, principalmente nos contornos, começa a apresenta pontas e traços grossos. Esse filtro aplica uma operação de abertura morfológica na imagem com o kernel estruturante e repetindo a operação duas vezes. O resultado é uma imagem com ruídos e outras pequenas irregularidades removidas.

O filtro de dilatação realiza uma operação de dilatação morfológica em uma imagem. A dilatação morfológica é uma técnica de processamento de imagem que é usada para expandir as áreas brancas dos objetos de uma imagem binária ou em escala de cinza [2]. O kernel estruturante usado aqui foi um kernel estruturante na forma de elipse com

tamanho 3 x 3 [19]. O filtro de dilatação aplica uma operação de dilatação morfológica na imagem usando um kernel estruturante. O resultado é uma imagem com as áreas brancas dos objetos expandidas, tornando-os maiores e mais conectados como se pode observar na Figura 4.3. A Figura 4.3 mostra a comparação entre os frames com e sem o filtro de dilatação. A Figura 4.3(a) mostra que o frame sem o filtro de dilatação apresenta alguns ruídos na pessoa, contudo a Figura 4.3(b) mostra que o frame de dilatação corrigiu os ruídos presentes na pessoa, mas apresenta ruídos brancos no fundo da imagem, que podem ser corrigidos com a aplicação do filtro de fechamento, como mostrado na Figura 4.2.

(a) Frame sem o filtro de dilatação.

(b) Frame após o filtro de dilatação.

Figura 4.3: Comparação entre os frames com e sem o filtro de dilatação.

O filtro de desfoque (ou mediana) é uma técnica de processamento de imagem que substitui o valor de cada pixel da imagem pela mediana dos valores dos pixels na janela definida em torno dele [2]. Esse filtro ajuda a remover o ruído de alta frequência da imagem, preservando os limites dos objetos. Portanto, foi aplicado um filtro de mediana na imagem com uma janela de tamanho 5x5 pixels, removendo o ruído de alta frequência da imagem e preservando os limites dos objetos. O resultado é uma imagem suavizada e

com menos ruído, o que pode ajudar em análises subsequentes da imagem [19].

Para descobrir a melhor combinação de filtros necessários para melhorar o modelo de detecção de queda com relação a iluminação e garantir alta sensibilidade, foram realizados testes entre combinações dos filtros. As combinações de filtros testadas nesse trabalho foram: o filtro de fechamento [17], o filtro de abertura [17], o filtro de dilatação [18], o filtro de desfoque [19], o filtro de limiarização [20]. Também foram utilizadas as técnicas de subtração de fundo: TMF [21], CNT [2] e MOG2 [22]. Durante os testes foram realizadas diversas combinações de filtros e para facilitar a identificação a Tabela 4.1 mostra a nomenclatura utilizada. Foram testadas um total de 126 combinações diferentes e o Capítulo 5 descreve as diferentes combinações analisadas e os resultados obtidos.

Tabela 4.1: Filtros usados nesse trabalho e a nomenclatura utilizada.

Filtro	**Identificador**
Operações Morfológicas	
Sem filtro	0
Abertura	O
Fechamento	C
Dilatação	D
Limiarização	T
Desfoque	B
Subtração de Fundo	
Sem Subtração de Fundo	1
FMT	2
CNT	3
MOG2	4

4.1.2 *ROI Masking*

Os blocos *ROI Masking*, ROI-M 1 e 2, da Figura 4.1 são responsáveis pelo rastreamento de pessoas. Esse rastreamento define caixas delimitadoras, as quais são comparadas entre quadros. Foi realizado o rastreamento de pessoas para extrair a região de interesse ROI dos quadros originais e fluxo óptico para reconstrução baseada na região, conforme descrito em [9]. O método baseado em ROI melhora a qualidade do rastreamento à medida que o modelo aprende a reconstruir apenas a região de interesse onde a pessoa está. Conforme proposto em [9], foi realizado o rastreamento de pessoas usando as redes totalmente convolucionais baseadas em região - *Region-based Fully Convolutional Networks* (R-FCN) [83], treinadas no conjunto de dados *Common Objects in Context* (COCO) [84]. Essa técnica é bastante eficiente, mas só funciona quando apenas uma pessoa está no vídeo. Essa

restrição não é um problema, pois a detecção precisa de quedas é mais necessária quando o idoso está sozinho.

Esse rastreamento define caixas delimitadoras, que o algoritmo compara entre os quadros. Conforme indicado por Mehta et al [9], foi usado um contador para rastrear o número de quadros contínuos sem detecção. Quando nenhuma detecção ocorre, o algoritmo incrementa o contador. Quando excede um limite de 20, o rastreador congela. 20 é suficiente para permitir que o algoritmo identifique que nenhum indivíduo está presente no ambiente. Com um contador acima de 20, o algoritmo processa mais quadros desnecessários pelas camadas da CNN, conforme comprovado por [9], aumentando o tempo de processamento.

O *Intersection over Union* (IoU)[16][85] corresponde às caixas delimitadoras. O IoU é menor quando o tamanho da caixa é maior que a caixa do quadro anterior, o que ocorre quando o detector localiza a caixa erroneamente. Assim, utiliza-se a razão de área entre quadros como critério de rastreamento. Essa proporção pode encontrar os contornos na imagem resultante e selecionar o contorno mais proeminente com base na área interna. A menor caixa que contém essa região de contorno é candidata a caixa delimitadora da pessoa.

Foi aplicado o limite Otsu [16] à imagem para separar o fundo escuro, também conforme especificação do trabalho de Mehta et al [9], que aplica esse método para imagens térmicas. Observou-se nos testes do CVSC que esse método é particularmente útil em imagens IR para melhorar a localização da caixa de contorno.

Dado que os dados de entrada do CVSC (câmeras RGB e IR) são diferentes dos dados do trabalho de Mehta el al (câmeras térmicas), observou-se que a saída após a aplicação do límite de Otsu era uma imagem que frequentemente ainda continha objetos com fundos claros. Para melhorar essa detecção, foram realizados testes com o algoritmo de BS [6] [19]. Esse algortimo pode melhorar o desempenho do método de detecção diante de mudanças de fundo, objetos, iluminação e movimento da câmera. Contudo, se a pessoa estiver na mesma posição por muito tempo, os métodos baseados em BS geram falhas no rastreamento do idoso, porque o algoritmo reconhece a pessoa como pano de fundo. Assim, esses métodos com BS são adequados para câmeras em locais onde a pessoa se move, tais como corredores e cozinhas, mas não em quartos e salas de estar.

Na implementação do *ROI Masking*, foi usado o *checkpoint* da rede *rfcn_ resnet101_ coco* pré-treinada a partir do modelo de detecção do *tensorflow zoo* [83] para realizar a detecção de pessoas, e o filtro *Kalman* [86] para realizar o rastreamento de pessoas. O código

de rastreamento de pessoa funciona filtrando a pessoa em cada quadro com o filtro *Kalman*. Objetos e imagens de fundo podem afetar o desempenho dos métodos de detecção de queda baseados em vídeo. O algoritmo apenas reconstrói a região onde a pessoa está presente. Portanto, mudanças na intensidade de objetos e fundo afetam menos o ROI.

4.1.3 Processamento do Fluxo Óptico

O fluxo óptico [76] [67] é o padrão de movimento aparente de objetos de imagem e foi utilizado por *Mehta et al* [9] para o cálculo da estimativa de movimento aparente entre quadros consecutivos causado pelo movimento da pessoa. No trabalho, o OFC calcula o padrão de movimento aparente de objetos, superfícies e bordas em uma cena causado pelo movimento relativo. O OFC ajuda na detecção de queda final por permitir que o modelo identifique o padrão de movimento.

4.1.4 A rede convolucional

As redes neurais convolucionais são redes biologicamente inspiradas que são usadas em visão computacional para classificação de imagens e detecção de objetos [66] [87]. No modelo de rede neural convolucional, cada camada da rede é tridimensional, possuindo uma extensão espacial e uma profundidade correspondente ao número de características. A noção de profundidade de uma única camada em uma rede neural convolucional é distinta da noção de profundidade em termos do número de camadas [87].

A rede convolucional usada nesse trabalho, com pesos de treinamento importados do trabalho de Mehta et al [9], é mostrada na Figura 4.1. Essa rede apresenta dois caminhos compostos por redes CNN que processam o vídeo de entrada. No primeiro caminho, a entrada para a primeira rede CNN (chamada CNN *Autoencoder 1* na Figura 4.1) é uma janela de *frames* da reposta do bloco *ROI Masking*. Nessa rede, foram utilizados filtros 3D 3×3 com profundidade temporal de 5 em todas as camadas. As camadas convolucionais da rede recebem apenas a região de interesse com a caixa que delimita a pessoa como entrada [14]. A saída de ambos os caminhos de rede convolucional é conectada por uma camada totalmente conectada de um único neurônio com uma função *sigmóide* para gerar uma probabilidade de a sequência de quadros ser original ou reconstruída. Alto erro de reconstrução e/ou baixa probabilidade na resposta do neurônio indicarão uma sequência de vídeo anormal. Essa estrutura visa identificar quedas com alta acurácia. O erro de reconstrução ou saída de probabilidade ou sua combinação pode ser usada como uma

pontuação de anomalia para identificar quedas durante o teste. O erro de reconstrução e a pontuação de anomalia são calculados como em [88] e [14] usando a média e o desvio padrão.

No segundo caminho, é calculado o fluxo óptico (representado na Figura 4.1 pelo bloco OFC) e posteriormente o bloco *ROI Masking* destaca a pessoa da imagem. A entrada da segunda rede CNN (chamada de CNN *Autoencoder 2* na Figura 4.1) é uma janela de quadros de fluxo óptico. Os filtros 3D 3×3 foram usados com profundidade temporal de 5 na primeira camada. Na segunda camada de convolução, filtros 2 × 2 com profundidade temporal de 4 foram usados para reconstruir a profundidade temporal de comprimento ímpar.

Cada caminho consiste em uma CNN para reconstruir a janela de entrada. Ambos os caminhos seguem, após o processamento pela CNN *Autoencoder*, para uma CNN para discriminar os quadros reconstruídos da janela do quadro original. Em outras palavras, essa rede neural, chamada CNN Discriminativa 1 e 2 na Figura 4.1, recebe tanto o quadro reconstruído quanto o quadro original como entrada. Ambas as CNN Discriminativas têm camadas idêntica às 4 primeiras camadas da CNN *Autoencoder* correspondente. A primeira camada de convolução 3D usa convoluções 3D com passo (*stride*) de 1 × 2 × 2 e preenchimento (*padding*), e o restante usa passo de 2 × 2 × 2 e preenchimento, arranjo importado do trabalho de *Mehta et al* [9]. Cada dimensão (profundidade temporal, altura e largura) é reduzida por um fator de 2 com cada camada de convolução 3D, exceto a primeira, que reduz apenas a dimensão espacial, permitindo assim uma sequência de cálculos entre as camadas das redes neurais sem degradar completamente a dimensão temporal. A decodificação opera como a codificação, mas ao contrário, usando camadas de convolução 3D. A camada de deconvolução final combina mapas de características (*feature maps*) na reconstrução decodificada. Esta camada final usa um passo de 1 × 1 × 1 e preenchimento, arranjo importado do trabalho de *Mehta et al* [9]. Ambas as redes convolucionais discriminativas são unidas por um único neurônio e o resultado é uma probabilidade de evento de queda. A normalização em lote (*batch normalization*) é usada em todas as camadas da CNN discriminadora, exceto na camada de entrada. A função de ativação *LeakyRelu* é usada em todas as camadas ocultas dessa rede, com um coeficiente de inclinação negativo definido com 0,2 [88]. Na última rede *feed-forward*, as redes CNN individuais são conectadas por um único neurônio *sigmóide* para gerar uma probabilidade de evento de queda.

A rede convolucional é alimentada pela janela de quadro mascarada e a perda de

reconstrução (*reconstruction loss*) baseada em região é usada. Os quadros nos quais uma pessoa não está localizada são removidos após o rastreamento. Os quadros são combinados por suas caixas delimitadoras (máscara ROI). A rede utiliza o otimizador *Stochastic Gradient Descent* (SGD) com taxa de aprendizado de 0,0002 [9] para o discriminador 1 e 2 e otimizador *AdaDelta* para o *autoencoder* 1 e 2 em todos os modelos [9]. SGD e *AdaDelta* são algoritmos de otimização usados para treinar redes neurais em aprendizado de máquina. O SGD é frequentemente escolhido para problemas de classificação binária, especialmente quando o conjunto de dados é grande [89]. Esse é o senário deste trabalho tento em vista que deseja-se classificar queda e não queda dentro de um conjunto grande de gravações. *AdaDelta* é um método de taxa de aprendizagem adaptativa que tem como principal vantagem que não é necessário definir uma taxa de aprendizagem padrão [1]. Todos os modelos foram treinados por 300 épocas. Os pesos da rede foram definidos conforme o modelo disponibilizado por [9], sendo usados como base para os testes das combinações de modelos de pré-processamento de imagem e detecção de movimento propostas e analisadas.

No mapeamento proposto no CVSC, na camada de entrada da CNN, as características correspondem aos canais de cores RGB e, nos canais ocultos, essas características representam mapas de características ocultos que codificam vários tipos de formas na imagem.

O modelo contém camadas de convolução, que realizam a operação de convolução, na qual um filtro é utilizado para mapear as ativações de uma camada para a próxima [9]. Uma operação de convolução usa um filtro tridimensional de pesos com a mesma profundidade da camada atual, mas com uma extensão espacial menor. O produto escalar entre todos os pesos no filtro e qualquer região espacial (do mesmo tamanho do filtro) em uma camada define o valor do estado oculto na próxima camada após a aplicação de uma função de ativação *LeakyRelu* [87]. A operação entre o filtro e as regiões espaciais em uma camada é realizada em todas as posições possíveis para definir a próxima camada (na qual as ativações mantêm suas relações espaciais da camada anterior). As conexões em uma rede neural convolucional são muito esparsas, pois qualquer ativação em uma determinada camada é função de apenas uma pequena região espacial na camada anterior [87]. É possível visualizar espacialmente quais partes da imagem afetam porções específicas das ativações em uma camada. As características nas camadas de nível inferior capturam linhas e outras formas primitivas, enquanto as características nas camadas de nível superior capturam formas mais complexas, como *loops* (que geralmente ocorrem em formas

[1] https://paperswithcode.com/method/adadelta

geométricas). Portanto, as camadas posteriores podem criar formas novas compondo as formas nessas características calculadas.

As CNNs pré-treinadas de recursos publicamente disponíveis, como ImageNet[90] e a utilizada nesse trabalho, que advém de [9], geralmente estão disponíveis para uso de maneira pronta para outras aplicações e conjuntos de dados. Isso é obtido usando a maioria dos pesos pré-treinados na rede convolucional [87] [66]. Os pesos da camada de classificação final podem ser aprendidos a partir do conjunto de dados disponível. Os pesos nas primeiras camadas são úteis porque aprendem vários tipos de formas nas imagens que podem ser úteis para praticamente qualquer tipo de aplicação de classificação. Além disso, as ativações de recursos na penúltima camada podem até ser usadas para aplicações não supervisionados [87]. Por exemplo, pode-se criar uma representação multidimensional de um conjunto de dados de uma imagem arbitrária passando cada imagem pela rede neural convolucional e extraindo as ativações da penúltima camada [87]a [66]. Posteriormente, qualquer tipo de indexação pode ser aplicada a essa representação para recuperar imagens semelhantes a uma imagem de destino específica. Essa abordagem geralmente oferece resultados bons na recuperação de imagens devido à natureza semântica dos recursos aprendidos pela rede. Vale ressaltar que o uso de redes convolucionais pré-treinadas é tão popular que o treinamento raramente é iniciado do zero [87].

Dessa forma, o modelo CVSC possui uma GAN de [9] com a importação de pesos de treinamento. A entrada para a rede CNN é uma janela de quadro da resposta do *ROI Masking*. As camadas convolucionais da rede recebem apenas a região de interesse, com a caixa delimitando a pessoa como entrada. A saída da rede convolucional prossegue para uma camada totalmente conectada de um único neurônio com uma função sigmoide. O neurônio gera uma probabilidade de que a sequência do quadro seja original ou reconstruída. Alto erro de reconstrução e baixa probabilidade na resposta do neurônio indicam uma sequência de vídeo anormal. O algoritmo usa o erro de reconstrução como uma pontuação de anomalia para identificar quedas durante o teste. O algoritmo calcula o erro de reconstrução e pontuação de anomalia pelo erro médio entre a entrada e os quadros reconstruídos como em [91]. O algoritmo calcula o fluxo óptico para identificar o movimento em uma sequência de quadros. O *ROI Masking* destaca a pessoa na imagem.

4.2 CVSC 1, 2, 3 e 4

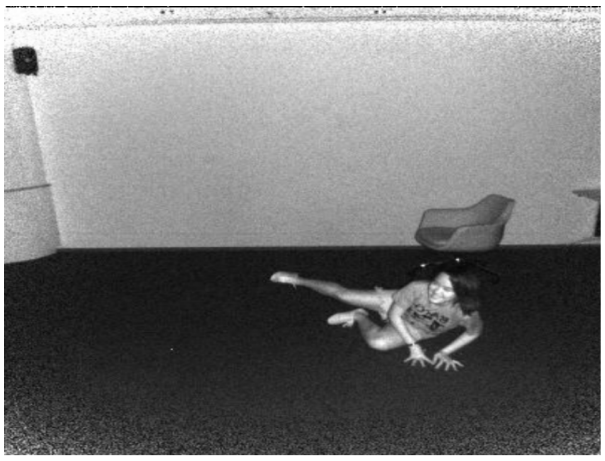
Essa seção descreve as diferentes variações entre os modelos CVSC. Foram desenvolvidas novas abordagens de pré-processamento e combinações de filtros com o objetivo de avaliar a sensibilidade do sistema de detecção de queda com relação as variações de iluminação e objetos de fundo. Cada modelo é composto de uma combinação de filtros que segue a nomenclatura da Tabela 4.1. Cada modelo CVSC possui diferentes abordagens de remoção de fundo que exploram a vizinhança dos pixels, o histórico dos pixels e a estabilidade dos pixels de formas diferente permitindo desenvolver filtros com a seletividade ideal para o problema em questão. Assim, cada modelo foi desenvolvido visando processar vídeos RGB e IR, gravações sobre altas variações de iluminação e fundo estático. Por fim, cada modelo foi avaliado para definir aquele que possui a mais alta sensibilidade em detecção de queda mesmo em ambientes pouco iluminados. Assim, os modelos são variações no bloco 'Pré-Processamento de Imagem' na Figura 4.1 com a visão geral do modelo CVSC. Esse é o bloco de entrada do modelo e impacta diretamente em todo o processamento posterior do modelo.

4.2.1 CVSC 1

CVSC 1 é o modelo sem adição de filtro de BS. Foram realizados testes em um pequeno subconjunto de dados de 20 vídeos de queda do *dataset* para definir a melhor combinação de filtros para esse modelo. Os modelos avaliados possuem uma sigla, conforme a Tabela 4.1. A letra "C"representa que foi utilizado o filtro de fechamento, a letra "O", o filtro de abertura, a letra "D", o filtro de dilatação, a letra "B", o filtro de desfoque, a letra "T"o filtro de limiarização e "0"indica que não foi utilizado nenhum dos filtros anteriores. O Capítulo 5 descreve os resultados obtidos para as diferentes combinações de pré-processamento analisadas.

4.2.2 CVSC 2

CVSC FMT CODB, chamado neste trabalho de CVSC 2, é uma alternativa ao uso do modelo CVSC com a adição do método BS de TMF e os filtros de abertura e fechamento, filtro de dilatação e filtro de desfoque [21]. Para desenvolver o filtro TMF é necessário calcular o valor mediano dos pixels do vídeo para determinar o plano de fundo. O valor mediano dos pixels forma uma nova imagem que é utilizada como máscara. O algoritmo percorre os *frames* do vídeo realizando uma sobreposição com a máscara e calculando a



(a) Frame de um momento da queda no vídeo em IR.

(b) Plano de fundo encontrado pelo CSVC 2.

(c) Máscara correspondente utilizada para a filtragem do frame no CVSC 2.

Figura 4.4: Figura dos frames do pré-processamento do CVSC 2.

diferença para formar uma nova imagem considerada como primeiro plano. Cada frame do video original possui a sua correspondente máscara e a imagem resultante da subtração.

A Figura 4.4(a) mostra o frame de um momento da queda no vídeo em IR. A Figura 4.4(b) mostra o plano de fundo encontrado pelo CSVC 2. A Figura 4.4(c) mostra a máscara correspondente utilizada para a filtragem do frame no CVSC 2 e formação da imagem resultante, a Figura 4.5. A Figura 4.5 mostra que o método de filtro TMF conseguiu destacar corretamente a pessoa no frame, contudo a Figura 4.4(b) mostra um sombreamento no centro da imagem que não foi filtrado corretamente bem como alguns ruídos causados pelo movimento corpo próximo ao solo.

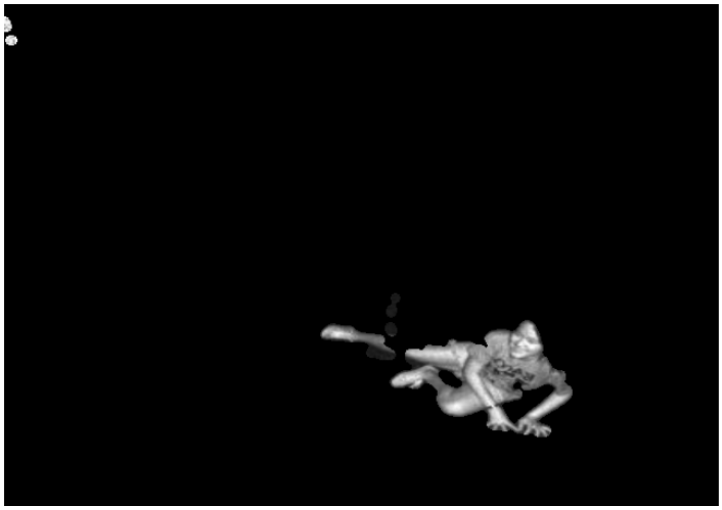


Figura 4.5: Frame resultante do processo de TMF do CVSC 2

O modelo aplica uma operação bitwise "AND"em cada pixel da imagem usando uma máscara. A função realiza a operação lógica "AND"entre os valores dos pixels do frame e os valores correspondentes na máscara. O resultado dessa operação é armazenado novamente na variável frame, substituindo a imagem original pelos pixels resultantes da operação "AND". Isso significa que apenas os pixels onde a máscara que contém valores diferentes de zero serão mantidos, enquanto os demais pixels serão definidos como zero, resultando em uma espécie de recorte da imagem original.

4.2.3 CVSC 3

O CVSC CNT CODB, chamado neste trabalho de CVSC 3, é uma alternativa ao modelo CVSC anterior ao trocar a técnica BS pelo método Count (CNT) com os mesmos filtros usados no CVSC2 [2]. O método CNT é um método que usa apenas informações dos valores de pixel dos quadros anteriores. CNT BS trabalha com contagem de frames. O processo de subtração é semelhante ao do método TMF. É um algoritmo duas vezes mais rápido que o MOG2 (Mistura de Gaussianos 2) em hardware barato (comparado ao Raspberry Pi3) [16]. Se a intensidade do pixel permanecer estável (*PixelStability*) por um

Figura 4.6: Máscara correspondente utilizada para a filtragem do frame no CVSC 3

determinado número de quadros, o algoritmo a considera fundo; caso contrário, é instável, portanto, é considerado como fundo. O CNT é muito mais rápido do que qualquer outra solução BS no OpenCV (otimizado com a ferramenta Valgrind) [16]. Nesse trabalho foram realizado dois projetos de filtro CNT. O primeiro, chamado de CNT 1 com menor estabilidade de pixel. O segundo, chamado de CNT 2, com maior estabilidade de pixel, obteve maiores métricas que o anterior e portanto o outro foi desconsiderado nos resultados e comparações entre os modelos apresentados no Capítulo 5.

A Figura 4.6 mostra a máscara correspondente utilizada para a filtragem do frame no CVSC 3 e formação da imagem resultante, a Figura 4.7. A Figura 4.6 mostra que a área da mascara correspondente no CVSC 3 é maior do que no CVSC 2. Isso evita que o algoritmo acabe filtrando de mais e acabe cortando partes da pessoa do frame. Contudo, também existe a presença de um contorno no centro da imagem que não foi filtrado bem como o cuido do IR nas bordas da imagem.

4.2.4 CVSC 4

O CVSC MOG2 CODB, chamado neste trabalho de CVSC 4, é uma alternativa ao modelo CVSC anterior ao trocar a técnica BS pelo método MOG2 [2][67] [16]. O método MOG2 é um método que agrupa os valores de intensidade de cada pixel no vídeo selecionando um número apropriado da distribuição Gaussiana para cada pixel [92]. O primeiro passo da técnica consiste em caracterizar cada pixel por sua intensidade no espaço de cores RGB ou no espaço IR bidimensional. O MOG usa várias distribuições gaussianas para modelar o plano de fundo de um pixel, cada gaussiano usando três parâmetros:

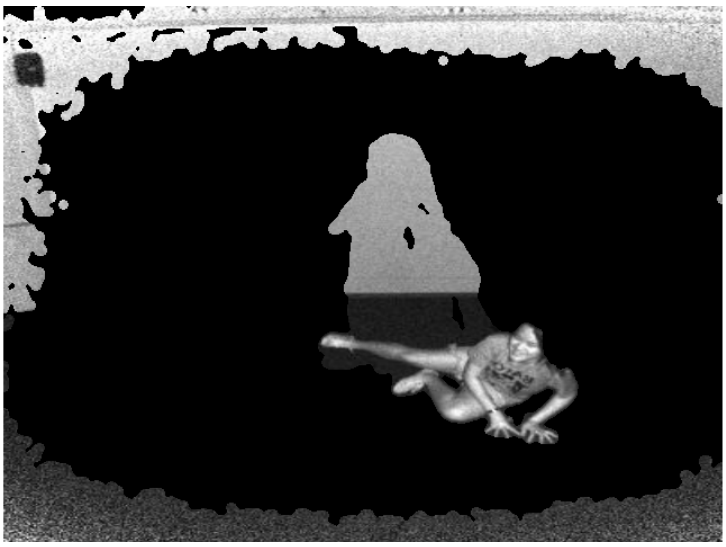


Figura 4.7: Frame resultante do processo de contagem do CVSC 3

- Mean($\mu i, t$): são as estimativas das médias de intensidade de cor.
- Variância($\sigma i, t$): são as estimativas da variância da média.
- Weight(k): número de gaussianos por pixel, representando o tempo que essas cores estiveram presentes na cena.

Figura 4.8: Máscara correspondente utilizada para a filtragem do frame no CVSC 4

A segunda etapa do método consiste em classificar cada pixel. O algoritmo atualiza os parâmetros gaussianos do pixel a cada novo pixel (independente de outros pixels) com base na taxa de aprendizado para ajustar a cena, considerando o valor do novo pixel de entrada para rastrear as mudanças de fundo. O algoritmo classifica os pixels que não

correspondem aos "Gaussianos de fundo"como primeiro plano. O algoritmo agrupa pixels de primeiro plano se a média dos gaussianos corresponder ao novo valor de pixel.

A Figura 4.8 mostra a máscara correspondente utilizada para a filtragem do frame no CVSC 4 e formação da imagem resultante, a Figura 4.9. A Figura 4.8 mostra que a área da mascara correspondente no CVSC 4 contorna melhor o corpo da pessoa realizando assim uma segmentação mais precisa. Ao realizar uma filtragem mais precisa, o algoritmo evita de enviar características da imagem desnecessárias para a rede CNN.

Figura 4.9: Frame resultante do processo de mistura gaussiana do CVSC 4

As Figura 4.9, Figura 4.7 e Figura 4.5 são as imagens resultantes do pré-processamento de imagem utilizado nos modelos CVSC4, CVSC3, CVSC2 respectivamente. Comparando essas figuras pode-se perceber que o modelo que melhor conseguiu destacar a pessoa na imagem foi o CVSC4. Na Figura 4.7 o fundo não foi totalmente removido, enquanto que na Figura 4.5, uma pequena parte da pessoa foi removida com o fundo da imagem.

Capítulo 5

Avaliação do Modelo Proposto

Esse capítulo trata da avaliação dos modelos CVSC, apresentar os resultados do processamento de imagens utilizado em cada um deles e comparar com outros modelos CNN na literatura que também usaram gravações RGB como modalidade de detecção de queda. O objetivo dos testes é para avaliar o desempenho dos modelos tanto na detecção de queda quanto em tempo de processamento. Para tal, foram utilizados gravações de simulações de queda em um conjunto de dados disponível abertamente. Os conjuntos de dados usados para treinar e avaliar o modelo foram gravações de câmeras de simulações de quedas e atividade da vida diária ADL. Assim, serão apresentados os resultados de avaliação, as técnicas de processamento de imagem mais rápidas e o modelo CVSC que obteve maior sensibilidade.

5.1 Ambiente de testes

Os testes foram processados na máquina com sistema operacional Debian GNU/Linux 11 bullseye (x86-64), Cinnamon Version 4.8.6, Linux Kernel 5.10.0-18-amd64, processador 11th Gen Intel Core i7-11700F 2.50 GHz x 8, memória de 15.5Gb, HD de 1256.7 GB e placa de vídeo NVIDIA Comporation TU116 [GeForce GTX 1660]. Os testes foram processados utilizando a linguagem de programação Python 3.7.6 em conjunto principalmente com as bibliotecas Keras 2.3.1, Tensorflow 1.14.0 e Opencv 4.2.0.

5.2 Conjunto de dados

5.2.1 COCO *dataset*

O *dataset* COCO [84] é utilizado para treinar a rede de Mehta et al [9], usado como base do modelo CVSC, para classificar objetos em geral, inclusive detectar pessoas. As categorias do conjunto de dados COCO formam um conjunto representativo relevante para aplicações práticas e elas ocorrerem com frequência suficiente para permitir a coleta em um grande conjunto de dados [84]. As categorias de partes de objetos foram incluídas na categoria "coisas". As “coisas” incluem objetos para os quais instâncias individuais podem ser facilmente rotuladas, por exemplo, pessoa, cadeira, carro, enquanto as categorias incluem materiais e objetos sem limites claros, como por exemplo, céu, rua e grama. O COCO *dataset* fornece uma localização precisa de instâncias de objetos [84]. A especificidade das categorias de objetos pode variar significativamente. Por exemplo, um cachorro pode ser membro das categorias “mamífero”, “cachorro” ou “pastor alemão”. Para permitir a coleta prática de um número significativo de instâncias por categoria, foi usado um limitar no conjunto de dados para categorias básicas, ou seja, rótulos de categoria que são comumente usados por humanos ao descrever objetos (cachorro, cadeira, pessoa).

Também existem algumas categorias de objetos partes de outras categorias de objetos [84]. Por exemplo, um rosto pode ser parte de uma pessoa. A inclusão de categorias de partes de objetos (rosto, mãos, rodas) é benéfica para muitas aplicações de visão computacional. Os autores usaram várias fontes para coletar categorias básicas de objetos de “coisas”. Primeiro foi compilada uma lista de categorias combinando categorias e um subconjunto das 1200 palavras usadas com mais frequência que denotam objetos visualmente identificáveis [84].

As imagens foram agrupadas em três tipos básicos: imagens de objetos icônicos, imagens de cenas icônicas e imagens não icônicas [84]. As imagens típicas de objetos icônicos têm um único objeto grande em uma perspectiva canônica centrada na imagem. Imagens de cenas icônicas são tiradas de pontos de vista canônicos e geralmente não têm pessoas. As imagens icônicas têm a vantagem de poderem ser facilmente encontradas pesquisando diretamente por categorias específicas usando a pesquisa de imagens do *Google* ou do *Bing*. Embora as imagens icônicas geralmente forneçam instâncias de objetos de alta qualidade, elas podem carecer de características importantes, informações contextuais e pontos de vista não canônicos, mas o objetivo um conjunto de dados de modo que a maioria das imagens fossem não icônicas [84]. Foi demonstrado que conjuntos de dados contendo mais

imagens não icônicas são melhores em generalização [84].

Para permitir um cronograma de lançamento mais rápido, o COCO *dataset* foi dividido em duas partes aproximadamente iguais [84]. A primeira metade do conjunto de dados foi lançada em 2014, a segunda metade lançada em 2015. A versão de 2014 contém 82.783 amostras treinamentos, 40.504 amostras de validações e 40.775 imagens de teste (aproximadamente 1/2 treinamento, 1/4 validação e 1/4 teste) [84]. Há cerca de 270.000 pessoas segmentadas e um total de 886.000 instâncias de objetos segmentados somente nos dados de treinamento avaliação de 2014. A versão cumulativa de 2015 contém um total de 165.482 imagens de treinamento, 81.208 de validação e 81.434 de teste.

O COCO *dataset* está disponível abertamente e é um conjunto de dados de detecção de objetos, segmentação, reconhecimento no contexto e rotulagem. Tem mais de 200.000 imagens rotuladas e 80 categorias de objetos.

5.2.2 TSF *dataset*

Dentre os conjuntos de dados disponíveis para tarefas de reconhecimento de atividades humanas *Human Activity Recognition* (HAR), o *Thermal Simulated Fall* (TSF) [93] foi usado para treinar o modelo de Mehta et al [9], usado como base do modelo CVSC, para detectar quedas. Este conjunto de dados consiste em vídeos capturados por uma câmera termográfica *FLIR ONE* montada em um telefone *Android* em uma sala com uma única visualização. Os autores gravaram um total de 44 vídeos, dos quais 35 vídeos contêm um evento de queda (36.391 quadros no total, 828 quadros de queda) e 9 vídeos (22.116 quadros) contêm apenas ADL. A resolução das imagens térmicas é 640×480. Os vídeos de ADL incluem diferentes cenários, como uma sala vazia, uma pessoa entrando em uma sala, sentada em uma cadeira ou deitada em uma cama, enquanto os vídeos de queda incluem uma pessoa caindo de uma cadeira, cama ou caindo enquanto caminha. Por meio do aprendizado de transferência, o modelo usa os pesos de treinamento da rede *rfcn_resnet101_coco* [83] do modelo de detecção de *tensorflow zoo* sobre o conjunto de dados COCO [84].

5.2.3 *NTU RGB+D dataset*

O *dataset "NTU RGB+D"* [94] da *Nanyang Technological University* (NTU) foi utilizado apenas para avaliar os modelos propostos nesta pesquisa de forma a garantir a independência entre os *datasets* de treinamento e teste. Ele é um *dataset* disponibilizado

abertamente e muito utilizado em reconhecimento de atividades humanas. O conjunto de dados contém um total de 60 modalidades de ação (classes) para tarefas de reconhecimento de atividades e um total de 56.880 vídeos. Para avaliar o sistema foram utilizados um total de 132 vídeos com ocorrência de quedas em RGB e 155 vídeos de quedas em IR. As câmeras de profundida e as outras modalidades que o *dataset* inclui, como *3D Skeletons*, *Masked Depth Maps*, *Full Depth Maps*, não foram utilizadas no experimento porque fogem ao objetivo de detectar quedas em lares de idosos.

As amostras do NTU *dataset* são capturadas em 80 pontos de vista distintos da câmera. A faixa etária dos participantes no conjunto de dados é de 10 a 35 anos [94]. Embora não tenham sido realizadas simulações de quedas com idosos para não ocorrer em riscos de acidentes, a faixa etária selecionada traz uma variação realista para a qualidade das ações e quedas simuladas. Embora o conjunto de dados seja limitado a cenas internas, devido à limitação operacional o sensor, foi usado a inconstância do ambiente capturando em várias condições de fundo [94]. Essa grande quantidade de variação em pessoas e visualizações torna possível ter avaliações de modalidades e visualizações cruzadas mais precisas para vários métodos de análise de ação baseados.

Para coletar este conjunto de dados, foi utilizado câmeras *Microsoft Kinect v2*. Foi usado quatro modalidades principais de dados fornecidas por este sensor: mapas de profundidade, informações de juntas 3D, quadros RGB e sequências de infravermelho. Os mapas de profundidade são sequências de valores de profundidade bidimensionais em milímetros. Para manter todas as informações, foi aplicada a compactação sem perdas para cada quadro individual. A resolução de cada quadro de profundidade é de 512 $\times$ 424. A informação da junta consiste em localizações tridimensionais das 25 juntas principais do corpo para corpos humanos detectados e rastreados na cena. Os pixels correspondentes em quadros RGB e mapas de profundidade também são fornecidos para cada junta e cada quadro. Os vídeos RGB são gravados na resolução fornecida de 1920 $\times$ 1080. As sequências infravermelhas também são coletadas e armazenadas quadro a quadro em 512 $\times$ 424. Foram gravadas 60 classes de ação no total, divididas em três grandes grupos: 40 ações diárias ADL, como por exemplo beber, comer e ler, 9 ações relacionadas à saúde, como por exemplo espirrar, cambalear e cair e 11 ações mútuas, como por exemplo socar, chutar e abraçar.

No total, 40 pessoas participaram da coleta de dados. As idades dos sujeitos estão entre 10 e 35 anos com variedade em idade, gênero e altura. Cada pessoa recebe um número de ID único em todo o conjunto de dados. Foram usadas três câmeras ao mesmo

tempo para capturar três visualizações horizontais diferentes da mesma ação. Para cada configuração, as três câmeras foram posicionadas na mesma altura, mas em três ângulos horizontais diferentes: $-45°$, $0°$, $+45°$. Cada pessoa foi solicitada a realizar cada ação duas vezes, uma vez em direção à câmera esquerda e outra em direção à câmera direita. Desta forma, foi capturada duas vistas frontais, uma vista lateral esquerda, uma vista lateral direita, uma vista lateral esquerda a 45 graus e uma vista lateral direita a 45 graus. As três câmeras recebem identificadores de câmera únicos. A câmera 1 sempre observa as visualizações de 45 graus, enquanto as câmeras 2 e 3 observam as visualizações frontal e lateral. Para aumentar ainda mais as visualizações das câmeras, em cada configuração foi mudada a altura e as distâncias das câmeras aos participantes. Todos os números de câmera e configuração são fornecidos para cada amostra de vídeo [94].

5.3 Resultados

O modelo foi avaliado usando gravações de câmeras RGB e IR.Também é relevante avaliar o modelo com registros de IR, pois as quedas também podem ocorrer em ambientes com pouca luz, por exemplo, se o idoso acordar assustado no meio da noite, precisar de algum medicamento ou tiver alguma confusão mental. Cabe destacar que muitas câmeras de monitoramento RGB já incluem infravermelho, dando ao sistema outro modo de detecção sem custo adicional. As pontuações de anomalia foram calculadas usando a média e outra usando o desvio padrão.

Para avaliar o desempenho da detecção de quedas como uma anomalia, a precisão, a sensibilidade (ou *Recall*) e a pontuação F1 (*F1 - Score*) foram calculadas [95] e comparadas com outras propostas recentes de detecção de quedas no mesmo conjunto de dados.

Em nossa pesquisa, a condição positiva (P) é a detecção da queda e a condição negativa (N) representa a ausência de queda ($P = FN + TP$, $N = FP + TN$). Portanto, um verdadeiro positivo (TP) é um resultado de teste que indica corretamente a detecção da queda de uma pessoa. Um verdadeiro negativo (TN) é um resultado de teste que indica corretamente a ausência de uma queda. Um falso positivo (FP) é um resultado de teste que indica erroneamente que uma pessoa caiu no vídeo. Um falso negativo (FN) é um resultado de teste que indica erroneamente a ausência da queda uma pessoa em um vídeo.No link do GitHub em [96] está disponível o arquivo .csv com os resultados da avaliação dos modelos CVSCs.

- População ($Pop = TP + TN + FP + FN$): Representa o número total de casos avaliados na pesquisa.
- Prevalência ($Prev = (TP + FN)/Pop$)[97][98]: Serve para medir o equilíbrio dos dados dentro da população total.
- Acurácia ($Acc = (TP + TN)/Pop$)[99][97][100]: É a medida do sucesso do sistema na identificação de quedas. O total de conquistas positivas e negativas sobre a população total indica o grau de sucesso do modelo.
- Precisão (ou Valor Preditivo Positivo) ($Prec = TP/(TP + FP) = 1 - FDR$) [99][100][98]: Precisão expressa a proporção de unidades que o modelo diz serem Positivas e, na verdade são Positivas. Em outras palavras, a Precisão nos diz o quanto se pode confiar no modelo quando ele prevê um teste como uma queda de uma pessoa.
- sensibilidade (ou *Recall*, taxa de acerto ou taxa positiva verdadeira) ($Sen = TP/(TP + FN) = 1 - FNR$) [99][98]: Esta é uma das principais métricas em nossa pesquisa sobre identificar quedas. Ele avalia o quanto o modelo não identificou as quedas que realmente ocorreram. A pesquisa trata de uma situação que, se não identificada, poderia resultar em danos críticos de saúde. Idealmente, esta métrica deve ser tão próxima de 100 %.
- F1 Score (or *F measure* or *Dice Similarity Coefficient*)($F1 = 2*((Prec*Sen)/(Prec+Sen))$)[100][98]: F1 é a média harmônica de precisão e sensibilidade. É uma medida da precisão de um teste para análise estatística de um classificador binário. É calculado a partir da precisão e sensibilidade do teste. O valor mais alto possível de um F-score é 1,0, indicando precisão e sensibilidade perfeitos, e o menor valor possível é 0, se a precisão ou sensibilidade for zero. A métrica F1 Score indica o balanceamento entre a precisão e a sensibilidade.
- Valor Preditivo Negativo ($NPV = TN/(TN + FN) = 1 - FOR$)[100][98]: A fração de casos verdadeiramente não queda de todos os casos que o modelo previu como não-queda.
- Taxa de descoberta falsa ($FDR = FP/(TP + FP) = 1 - PPV$)[101][98]: FDR é um método para conceituar a taxa de erros de FP em condições positivas previstas durante o teste múltiplas comparações. Esta medida avalia a taxa do modelo identificando incorretamente uma queda entre todos os casos em que o modelo identificou uma queda.

- Taxa de Falsa Omissão ($FOR = FN/(TN+FN) = 1-NPV$)[98] FOR é a medida da taxa de erro FN pelo total de condições negativas previstas. Esta medida avalia a taxa com que os modelos identificam erroneamente uma não queda em todos os casos em que o modelo identificou uma não queda.

- Taxa de falsos positivos (ou queda ou probabilidade de falso alarme) ($FPR = FP/(TN+FP) = 1-TNR$)[100][97]: FPR é a probabilidade de rejeitar falsamente a hipótese negativa (não queda) para um teste específico.

- Taxa de falsos negativos (ou taxa de erros) ($FNR = FN/(TP+FN) = 1-TPR$)[100][98]: A taxa de falsos negativos é a proporção de positivos que testaram negativo com o teste. A probabilidade condicional de um resultado de teste negativo, dado que a condição procurada está presente. Esta é também uma das principais métricas na nossa investigação sobre identificação de quedas.

- Taxa de Negativos Verdadeiros (ou Especificidade ou Seletividade) ($TNR = TN/(TN+FP) = 1-FPR$) [100][98]: TNR mede a proporção de negativos reais que nossos modelos corretamente identificado como tal. É a percentagem de não quedas que os modelos identificaram corretamente em relação ao número real de não quedas.

- Razão de verossimilhança positiva (ou razão de verossimilhança positiva, razão de verossimilhança para resultados positivos) ($LR+ = TPR/FPR$)[102][103]: As razões de verossimilhança usam a sensibilidade e a especificidade do teste para determinar se um resultado de teste altera de forma útil a probabilidade de existência de uma condição. LR+ representa a probabilidadede uma pessoa que teve o teste positivo dividido pela probabilidade de uma pessoa que não teve o teste positivo. Bons classificadores têm LR+ maior que 1 porque possuem taxas de TPR mais altas.

- Razão de verossimilhança negativa (ou razão de verossimilhança negativa, razão de verossimilhança para resultados negativos)($LR- = FNR/TNR$)[102][103]: LR- representa a probabilidade de uma pessoa ter o teste negativo (não queda) dividida pela probabilidade de uma pessoa não ter o teste negativo. Bons classificadores têm LR- tendendo a 0 porque possuem taxas de TNR mais altas.

- Razão de chances de diagnóstico ($DOR = LR+/LR-$)[103]: Esta medida é utilizada para estimar a capacidade discriminativa do modelo e também para comparação entre dois classificadores. É uma medida da eficácia de um classificador. É definido como a razão entre as chances de o teste ser positivo, se o sujeito tiver caído, e as chances de o teste ser positivo, se o sujeito não tiver caído.

- Pontuação F-Beta ($\beta = 0,4$ e $FBeta = (1+\beta^2) * ((Prec * Recall)/((\beta^2 * Prec) + Recall))$) [103]: Esta medida é outra variante das medidas F. Esta variante representa a média harmônica ponderada entre precisão e Recall. Esta métrica é sensível a mudanças nas distribuições de dados. β é um parâmetro para análise de pontuação, se usar β igual a 1, temos a média harmônica, então temos a pontuação F1. $\beta > 1$ dá mais peso para Recall, enquanto $\beta < 1$ favorece a precisão. Por exemplo, beta = 2 torna Recall duas vezes mais importante que a precisão, enquanto beta = 0,5 faz o oposto. Assintoticamente, β tendendo ao $+\infty$ a pontuação considera apenas recall e beta tendendo a 0, considera apenas precisão [1].

- Coeficiente de correlação de Matthews (MCC) (ou Coeficiente Phi(ϕ ou $r\Phi$))[99][103]: faixas de MCC entre -1 e +1, sendo -1 para classificação totalmente errada ($TP = TN = 0$) e 1 para classificação perfeita ($FP = FN = 0$). $MCC = 0$ indica classificação aleatória ($TP * TN = FP * FN$). Esta métrica representa a correlação entre as classificações observadas e previstas e é calculada diretamente a partir da matriz de confusão. Esta métrica é sensível a dados desequilibrados.

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{1}$$

- Índice de Informação (ou Bookmaker, Informedness ou índice de Youden) ($BM = Sen + TNR - 1$) [103]: Esta métrica avalia o poder discriminativo do classificador. A fórmula do índice de Youden combina sensibilidade e especificidade como na métrica DOR. A métrica BM varia de zero quando o classificador é ruim até 1, representando um classificador perfeito. Também é adequado com dados não balanceados.

- Marcação (ou deltaP (Δp)) ($MK = Prec + NPV - 1$)[98][103]: A marcação quantifica o quão marcada uma condição é para o classificador e especifica o probabilidade de que uma condição seja marcada pelo classificador (versus acaso). Nesse casso de pesquisa, a condição é a presença da queda. Esta métrica é uma codificação daquilo que é incomum e isso se reflete em definições probabilísticas formais de marcação e informação como componentes unidirecionais com correção de chance do coeficiente de correlação de Matthews (MCC) correspondente ao Δp. Assim, a Marcação é uma medida de confiabilidade de previsões positivas e negativas do modelo. A métrica MK varia de zero quando o classificador é ruim até 1, representando um classificador perfeito.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html

- Índice Fowlkes–Mallows (FM) [104]: O índice Fowlkes-Mallows é um método de avaliação usado para determinar a similaridade entre dois clusters obtidos após um algoritmo de clustering e também uma métrica para medir matrizes de confusão. Esta medida de similaridade pode ser entre dois agrupamentos hierárquicos ou um agrupamento e uma classificação de referência. Um valor mais alto para o índice indica maior similaridade entre os clusters e os rankings de referência. O valor mínimo possível do índice de Fowlkes-Mallows é 0, que corresponde à pior classificação binária possível, onde todos os elementos foram classificados incorretamente. E o valor máximo possível do índice de Fowlkes-Mallows é 1, que corresponde à melhor classificação binária possível, onde todos os elementos foram perfeitamente classificados.

$$FM = \sqrt{Prec^{*}Sen} \tag{2}$$

- Pontuação de Ameaça (TS) (ou Threat Score, Índice de Sucesso Crítico *Critical Success Index* (CSI), índice de Jaccard ou coeficiente de similaridade de Jaccard) ($TS = TP/(TP+FN+FP)$)[105] [106]: Esta métrica é definida como a proporção de acertos pela soma de acertos, alarmes falsos e erros. TS varia de 0 a 1 (onde 1 representa previsão perfeita). TS não é imparcial, atribui pontuações mais baixas para eventos mais raros. TS ignora verdadeiros negativos. A métrica parte da justificativa que para o desenvolvimento inicial do índice de sucesso crítico é evitar a inflação excessiva (ou seja, otimismo indevido) das métricas de teste como consequência de um número muito grande de verdadeiros negativos (ou seja, desequilíbrio de classe).

- Acurácia balanceada ou taxa de classificação balanceada ($BA = (Recall + TNR)/2$) [103][107][99]: Este métrica combina as métricas de sensibilidade e especificidade. Se o conjunto de dados for balanceado, ou seja, se as classes tiverem quase o mesmo tamanho, a acurácia e a exatidão balanceada tendem a convergir para o mesmo valor. Na verdade, a principal diferença entre Acurácia e Acurácia Balanceada surge quando o conjunto de dados inicial mostra uma distribuição desigual para as classes. Além disso, taxa de erro de equilíbrio (*Balance error rate* (BER)) ou taxa de metade de erro total (*Half total error rate* (HTER)) representa $1 - BA$. As métricas BA e BER podem ser usadas com conjuntos de dados desequilibrados. A precisão balanceada dá peso igual a cada classe e sua indiferença à distribuição das classes ajuda a identificar possíveis problemas preditivos também para classes raras e sub-representadas. Portanto, é uma métrica importante para identificação de pessoas e quedas devido à tendência de conjuntos de dados desequilibrados.

- Pontuação Hamming ($Ham = (TN + TP)/(TP + FP + TN + FN)$ [108]: Se o classificador obtiver previsões mais corretas, isso resultará em uma pontuação de Hamming mais alta. Quanto maior o valor medido, melhor. O melhor valor possível é 1 (se um modelo acertou todas as previsões) e o pior é 0 (se um modelo não fez uma única previsão correta).

A AUC é a área sob a curva *Receiver Operating Characteristic* (ROC), que mede a capacidade do modelo de distinguir entre exemplos positivos e negativos. A curva ROC é uma curva que mostra a taxa de verdadeiros positivos em relação à taxa de falsos positivos para diferentes limiares de classificação. A taxa de verdadeiros positivos é a proporção de exemplos positivos que foram corretamente classificados como positivos pelo modelo, enquanto a taxa de falsos positivos é a proporção de exemplos negativos que foram erroneamente classificados como positivos pelo modelo.

A AUC é um valor numérico que representa a área total sob a curva ROC. Ela é usada como uma medida de desempenho do modelo em problemas de classificação binária. Quanto maior o valor da AUC, melhor é o desempenho do modelo, indicando uma melhor capacidade de discriminação. Um modelo com AUC de 1,0 é perfeito, enquanto um modelo com AUC de 0,5 é equivalente a uma escolha aleatória.

Entre as métricas, a pesquisa está interessada principalmente em uma boa avaliação da sensibilidade, pois está relacionada à taxa de falsos negativos, que é muito importante para o sistema de detecção de quedas. Um Falso Negativo é quando o indivíduo caiu, mas o sistema não identificou a queda. Essa é exatamente a situação de risco à saúde que o sistema mais quer evitar, pois o indivíduo não pode ficar no chão sem acionar um alerta do sistema. Assim, quanto maior a taxa de falsos negativos, menor a sensibilidade do sistema.

Durante a execução dos experimentos de queda, animações mostram o desempenho através da pontuação de anomalia por quadro usando a média sobre o vídeo em tempo real. A animação mostra a sequência de quadros reconstruídos e originais e o gráfico da pontuação de anomalia por quadro. Essas experiências estão disponíveis no *Google Drive* em [109]. A Figura 5.1 mostra o quadro após a técnica BS à esquerda, o quadro reconstruído à direita e o gráfico de pontuação de anomalia abaixo. Este gráfico mostra o erro de reconstrução por quadro calculando a média em um vídeo com queda. A média do erro entre o quadro de entrada e o quadro reconstruído calcula a pontuação da anomalia como em [91]. Cada quadro representa a progressão no vídeo. Um alto erro de reconstrução corresponde a um aumento na pontuação da anomalia e a uma maior chance

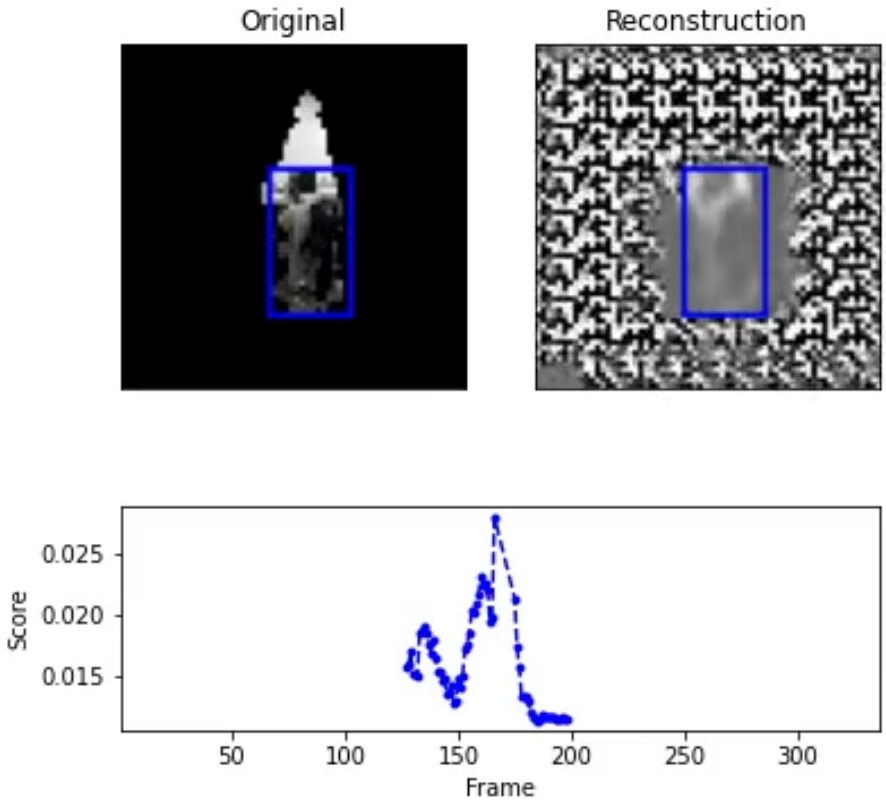


Figura 5.1: Animação da pontuação de anomalia por quadro no vídeo em que ocorre uma queda.

de queda. O gráfico mostra que o sistema consegue identificar a queda com aumento do erro de reconstrução.

A Figura 5.2 mostra o erro de reconstrução por quadro calculando a média e o desvio padrão em um vídeo com a ocorrência de uma queda. Cada quadro representa a progressão no vídeo. Um alto erro de reconstrução representa um aumento na pontuação da anomalia, o que indica maior chance de ocorrência de queda. Ambos os gráficos mostram que o sistema conseguiu identificar corretamente a queda aumentando o erro de reconstrução (um pico acima de 0,0175 para o gráfico de média e um pico acima de 0,0050 para o gráfico de desvio padrão). Considerando as amostras onde o sistema conseguiu identificar corretamente o usuário, a Tabela 5.3 lista as métricas de avaliação do modelo de detecção de queda e mostra o desempenho de outras propostas de detecção de queda por meio de visão computacional usando o mesmo conjunto de dados de reconhecimento de ação NTU RGB+D. Avaliando os experimentos, pode-se perceber que existem algumas variações quanto ao momento exato da queda e o momento em que a pontuação da anomalia ultrapassa o limiar.

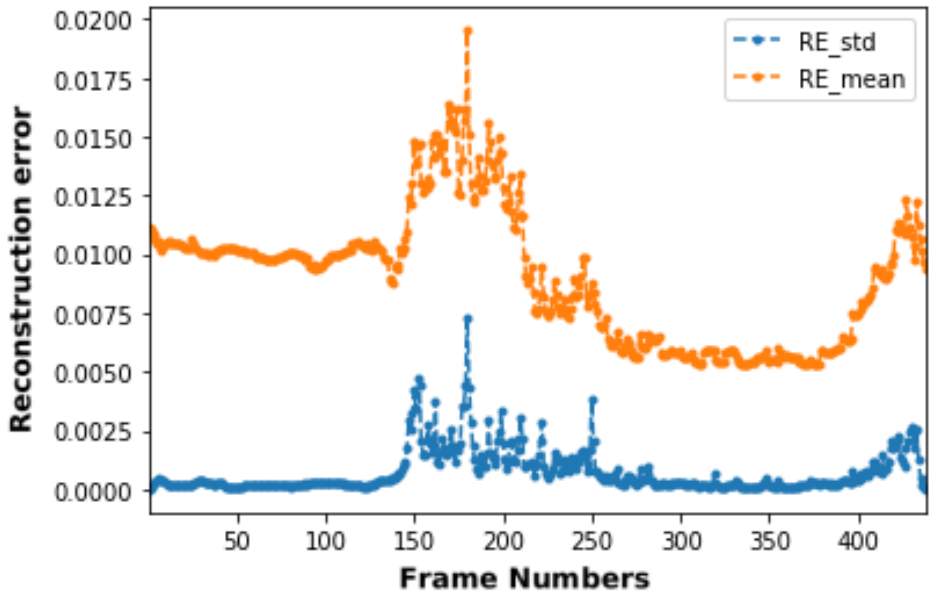


Figura 5.2: Gráfico da pontuação de anomalia por quadro para um vídeo em que ocorre uma queda.

5.3.1 Resultados do CVSC 1

CVSC 1 é o modelo sem adição de filtro de BS. Foram realizados testes em um pequeno subconjunto de dados de 20 vídeos de queda do *dataset* para definir a melhor combinação de filtros para o modelo. A Figura 5.3 apresenta os modelos CVSC1 de maior acurácia, Recall e F1-Score. Os modelos avaliados são referenciados no eixo x do gráfico, cada um com uma sigla específica conforme a Tabela 4.1. Os resultados são apresentados em termos de métricas de desempenho de modelos de aprendizado de máquina.

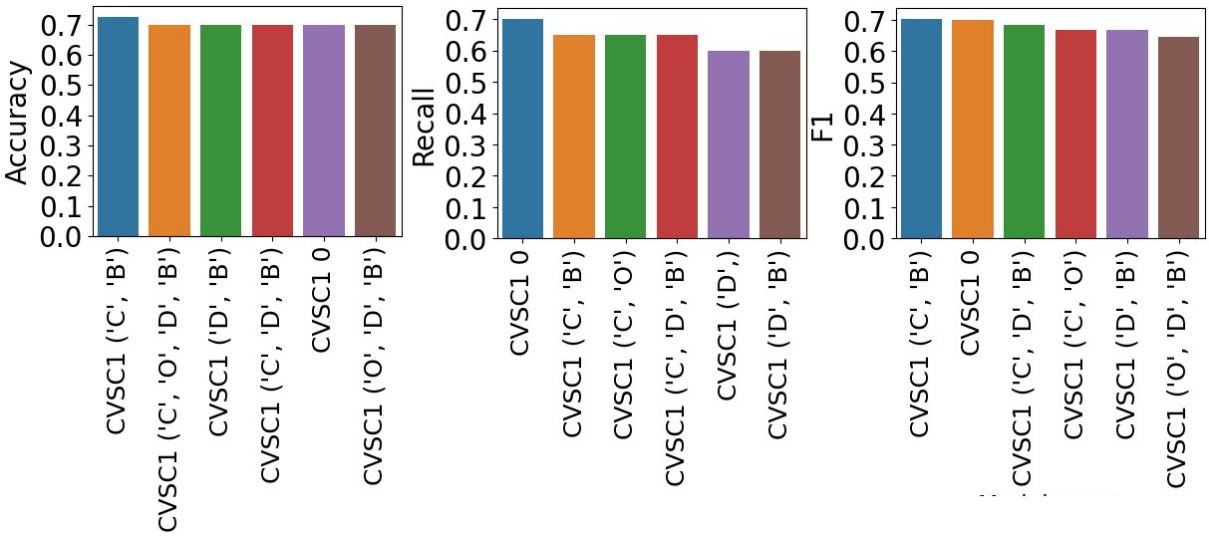


Figura 5.3: Acurácia, Recall e F1-Score do modelo CVSC1 com diferentes filtros

O modelo CVSC1 (C,B) é o modelo que possui a maior acurácia e F1-Score entre os modelos CVSC1. Esse mesmo modelo fica em segundo lugar com relação ao Recall. O modelo que possui o maio Recall foi o CVSC1 sem a adição de filtros. Os quatro modelos de maior acurácia utilizaram o filtro de desfoque.

5.3.2 Resultados do CVSC 2

No CVSC2 foram realizados testes em um pequeno subconjunto de dados de queda do *dataset* para definir a melhor combinação de filtros para esse modelo. A Figura 5.4 apresenta os modelos CVSC2 de maior acurácia, Recall e F1-Score.

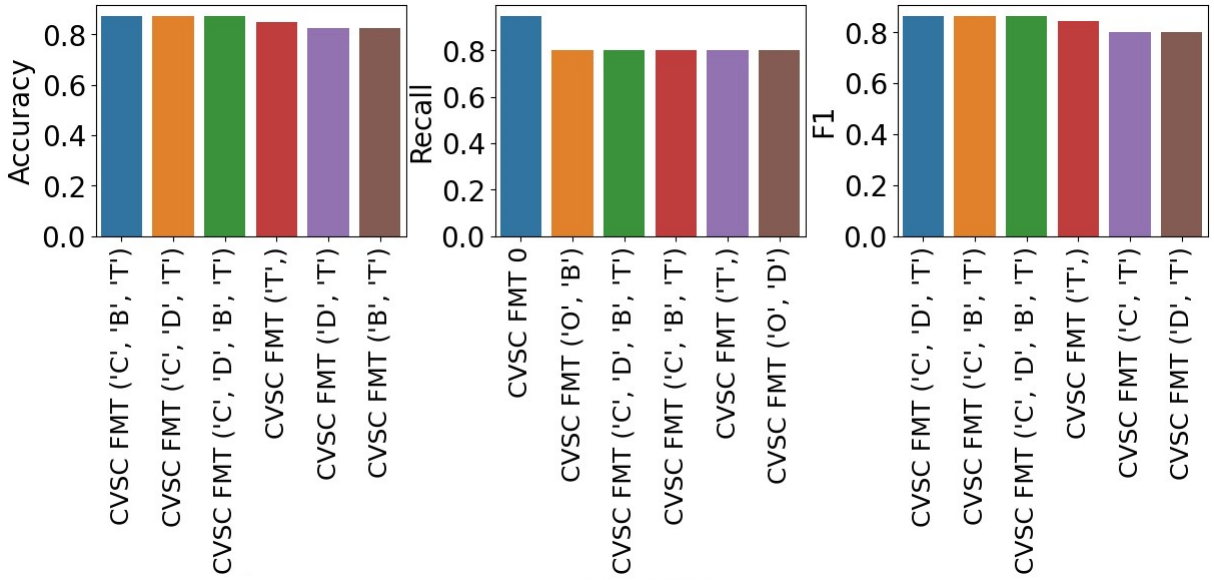


Figura 5.4: Acurácia, Recall e F1-Score do modelo CVSC 2 com diferentes filtros.

O modelo CVSC2 (C,B,T) é o modelo que possui a maior acurácia entre os modelos CVSC 2. Os filtros de fechamento e desfoque são os mesmos filtros utilizados no modelo CVSC 1 de maior acurácia demonstrando a eficácia deles no modelo de detecção de queda melhorando as métricas do modelo através do pré-processamento. As métricas de Acurácia, Recall e F1-Score aumentaram com relação aos modelos CVSC 1, demonstrando a melhora do modelo com filtro de BS em comparação com o modelo sem BS. Assim, o pré-processamento da imagem permitiu melhorar a acurácia da detecção de queda pelo modelo de visão computacional. Todos os 6 modelos de melhor acurácia e melhor F1-Score utilizam o filtro de limiarização demonstrando que a combinação da filtragem mediana temporal e o filtro de limiarização melhora as métricas de acurácia e F1-Score do modelo de detecção de queda.

5.3.3 Resultados do CVSC 3

No CVSC3 foram realizados testes em um pequeno subconjunto de dados de queda do *dataset* para definir a melhor combinação de filtros para esse modelo. A Figura 5.5 apresenta os modelos CVSC3 de maior acurácia, Recall e F1-Score.

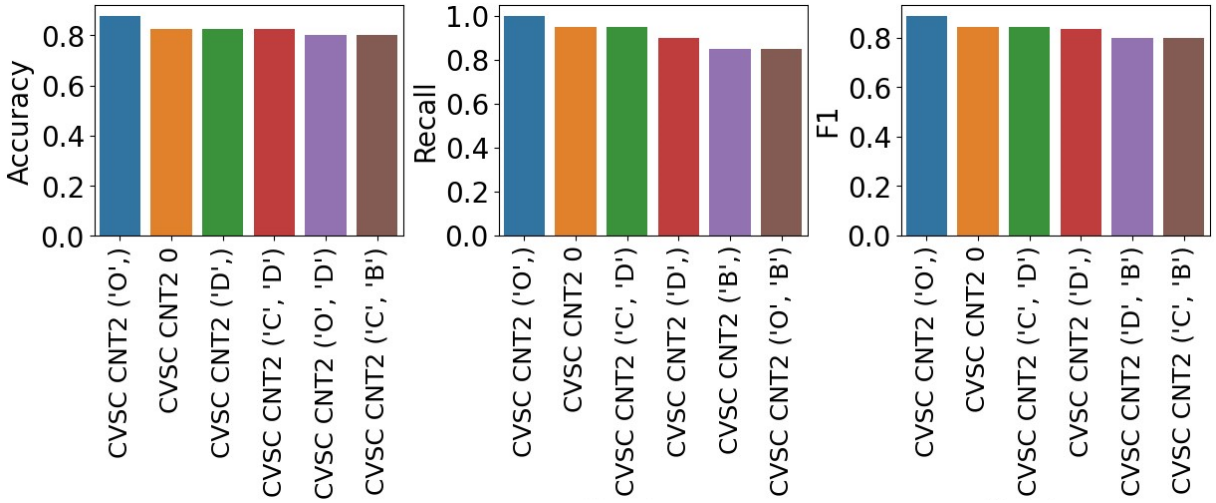


Figura 5.5: Acurácia, Recall e F1-Score do modelo CVSC 3 com diferentes filtros.

O modelo CVSC3 (O) é o modelo que possui a maior acurácia entre os modelos CVSC3. Os filtros de fechamento e desfoque não obtiveram desempenhos altos como nos modelos anteriores, mostrando que não é uma boa combinação a técnica de BS CNT com os filtros de fechamento e desfoque. Para a técnica de subtração de fundo CNT, o filtro que produz um desempenho melhor na detecção de queda é o filtro de abertura. O modelo com o filtro de abertura obteve a melhor métrica em acurácia, Recall e F1-Score. A Figura 5.5 também mostra que nenhum dos modelos CVSC2 com melhores métricas utilizou o filtro de limiarização demonstrando que não é uma boa combinação o filtro de limiarização com a técnica CNT. O modelo CVSC 3 tém maior Recall que o modelo CVSC 2 que por sua vez tem maior Recall que o modelo CVSC 1 o que demostra que melhorar o pré-processamento da imagem, o filtro e reduzir a influencia da iluminação e objetos de fundo deixa o modelo de visão computacional com maior sensibilidade para detectar a queda.

5.3.4 Resultados do CVSC 4

No CVSC4 foram realizados testes em um pequeno subconjunto de dados de queda do *dataset* para definir a melhor combinação de filtros para esse modelo. A Figura 5.6 apresenta os modelos CVSC4 de maior acurácia, Recall e F1-Score.

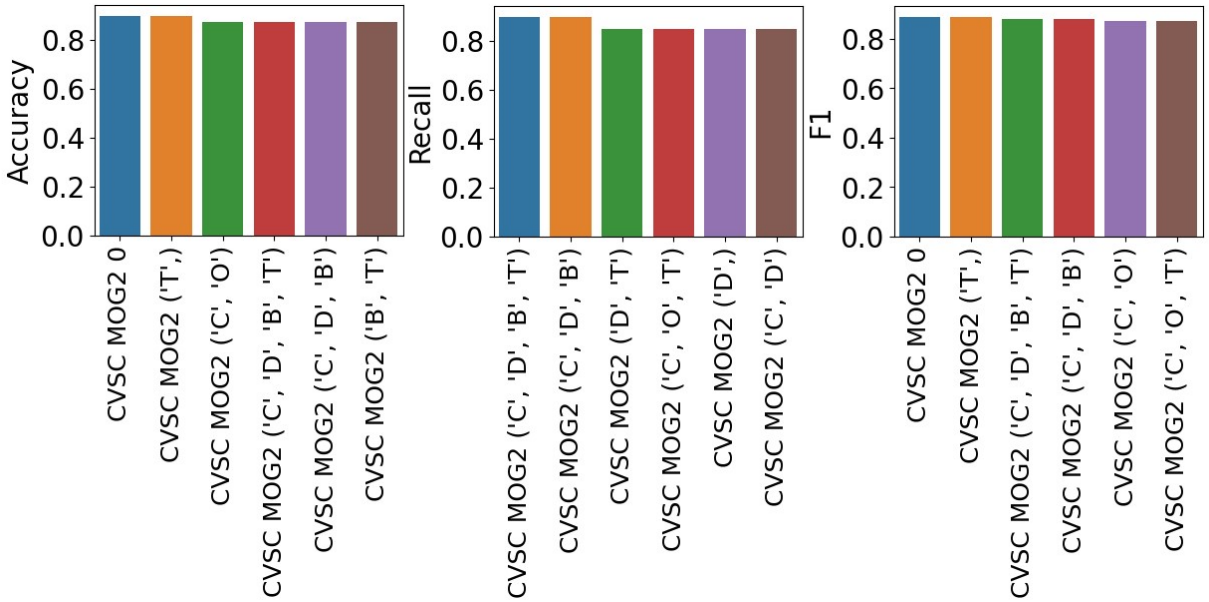


Figura 5.6: Acurácia, Recall e F1-Score do modelo CVSC 4 com diferentes filtros.

Comparando os modelos, os modelos com maior pontuação foram os CVSC MOG2 0 e CVSC MOG2 ('T',) com 0.9 de acurácia. A técnica de subtração de fundo MOG2 é a técnica de subtração de fundo que produz um desempenho melhor em termos de acurácia dos modelos CVSC.

5.3.5 Comparação dos modelos em termos das métricas de desempenho

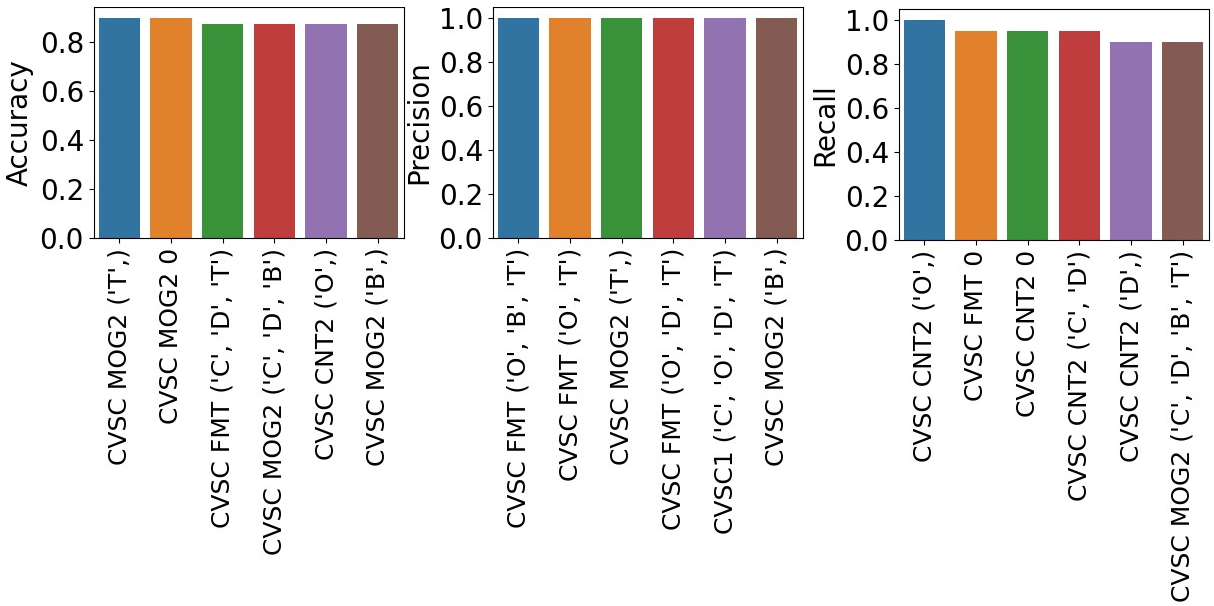


Figura 5.7: Maiores métricas de Acurácia, Precisão e Recall dos modelos CVSCs.

Os modelos CVSC foram comparados em termos das métricas de aprendizado de

máquina mencionadas. A Figura 5.7 apresenta os resultados dos 6 melhores modelos em termos de acurácia, sensibilidade e precisão. A Figura 5.8 mostra os gráficos dos modelos em ordem de *F1 - Score*, AUC e NPV. Comparando os modelos, os modelos com maior pontuação foram os CVSC MOG2 0 e CVSC MOG2 ('T',) com 0,9 de AUC e 0,8 de Recall. Portanto, o CVSC4 obteve maior pontuação e assim o filtro de remoção de fundo usando mistura de gaussianas melhorou o desempenho do modelo CVSC sem remoção de fundo.

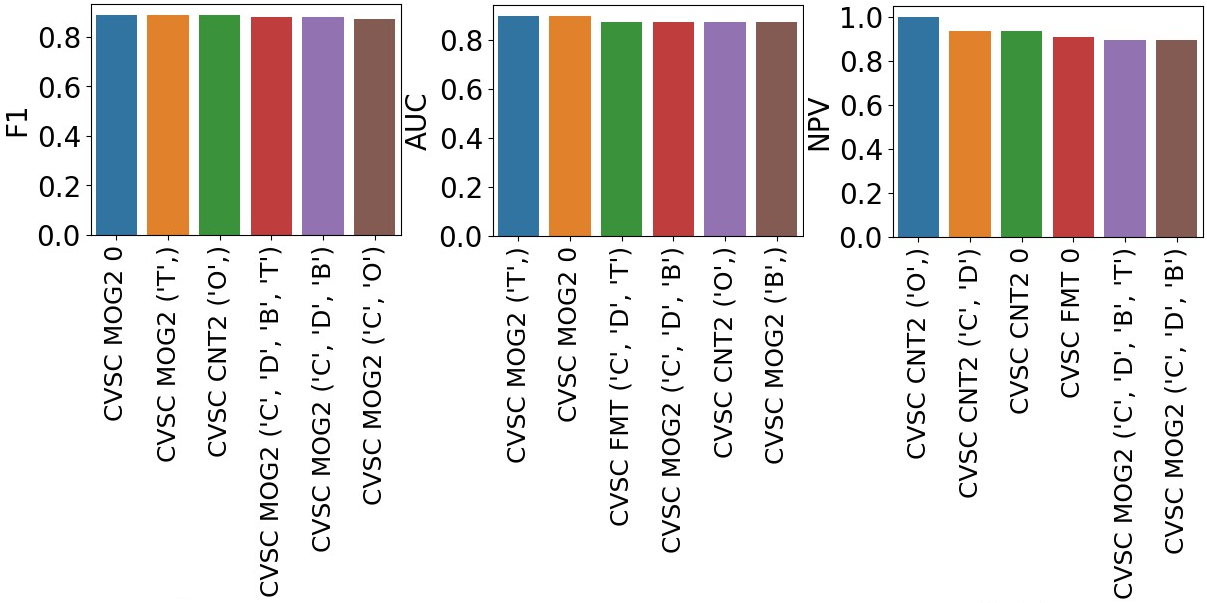


Figura 5.8: Maiores métricas de F1-Score, AUC e NPV dos modelos CVSCs.

A Figura 5.7 mostra os gráficos dos modelos em ordem de acurácia, precisão e Recall. Comparando os modelos, os modelos com maior pontuação foram os CVSC MOG2 0 e CVSC MOG2 ('T',) com 0,9 de acurácia. Portanto, o CVSC4 obteve maior pontuação também em termos de acurácia e assim provando que vale a pena incluir um estagio a mais no modelo em pré-processamento de imagem para melhorar a acurácia do modelo. O modelo de maior Recall foi o CVSC CNT2 ('O'), foi o terceiro modelo de maior F1-Score e quinto de maior AUC. Os modelos de maior precisam possuem a métrica próxima de 1. Os dois modelos de maior AUC também são os modelos CVSC MOG2 0 e CVSC MOG2 ('T',). Assim como o modelo CVSC CNT2 ('O') tem o maior Recall ele também tem o maior NPV.

A Figura 5.10 mostra o gráfico dos modelos em ordem das métricas FOR, FPR e FNR. A Figura 5.10 mostra que todos os modelos de maior FOR e FNR utilizaram o filtro de limiarização. A Figura 5.10 mostra que o modelo de maior FPR foi o modelo CVSC FMT 0.

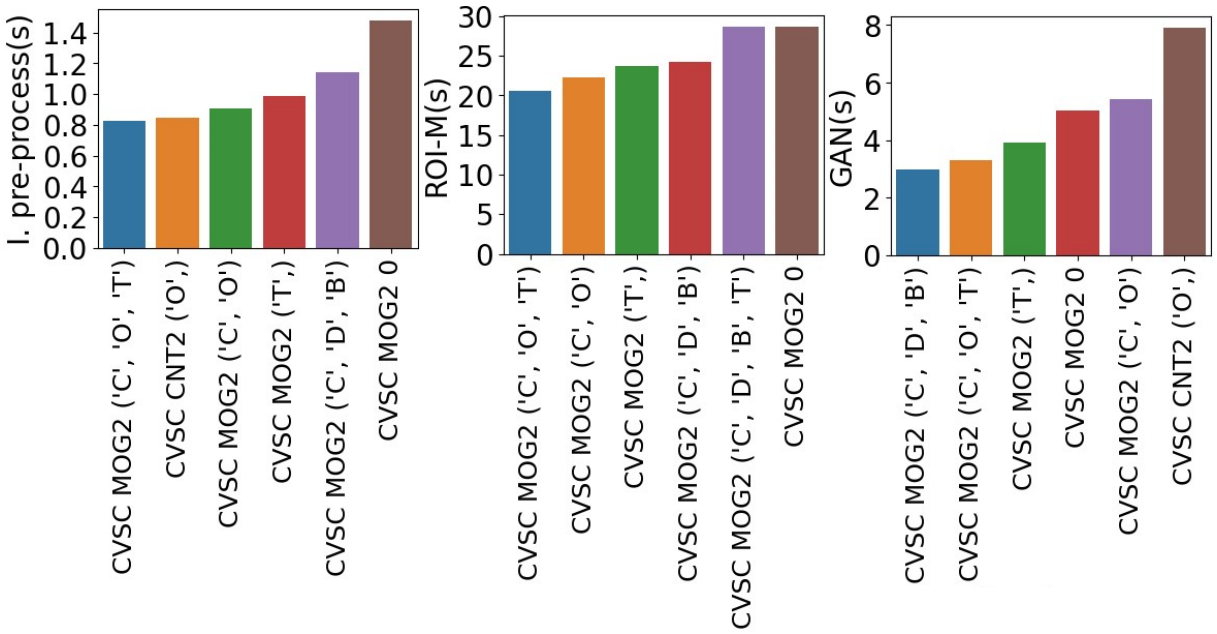


Figura 5.9: Tempo de pré processamento de imagem, mascaramento ROI e GAN dos modelos CVSCs de maior F1-Score.

A Figura 5.11 mostra o gráfico dos modelos em ordem das métricas TNR, MCC e BM. A Figura 5.11 mostra que cinco primeiros modelos de maior TNR utilizaram o filtro de limiarização. A Figura 5.11 mostra que os modelos de maior MCC e BM foi o modelo CVSC MOG2 0 e CVSC MOG2 ('T').

A Figura 5.8 mostra o gráfico dos modelos em ordem de pontuação F1. A Figura 5.9 mostram os gráficos do tempo de pré-processamento de imagem, mascaramento ROI e GAN dos modelos desses modelos. A Figura 5.12 mostra o tempo de processamento do OFC e de processamento total dos modelos de maior F1 -Score e gráfico dos modelos de maior FDR. FDR representa a taxa de descoberta falsa então quanto maior essa métrica maior é a taxa de falso positivo do modelo e portanto, pior é o classificador.

Comparando os modelos, os modelos com maior pontuação foram os CVSC MOG2 0 e CVSC MOG2 ('T',) com 0.8889 de pontuação F1. Ao comparar o tempo total de processamento do modelo CVSC MOG2 0 de 58.60100s e do modelo CVSC MOG2 ('T',) de 47.30136s conclui-se portanto que o modelo de melhor desempenho foi o CVSC MOG2 ('T',), assim o filtro de limiarização ajudou a diminuir o tempo de processamento e consequentemente é o filtro ideal para ser utilizado no modelo CVSC 4. O modelo de menor tempo total de processamento foi o CVSC1 ('C', 'O', 'T') com 26.21476s, mas obteve métricas baixas de acurácia (0.55) e *recall* (0.0). O modelo CVSC CNT2 ('O', 'B') possui o maior tempo total de processamento de todos os modelos com tempo de 109.53201s e obteve medidas de acurácia (0.75), *recall* (0.85) e pontuação F1 (0.7727). Esse modelo

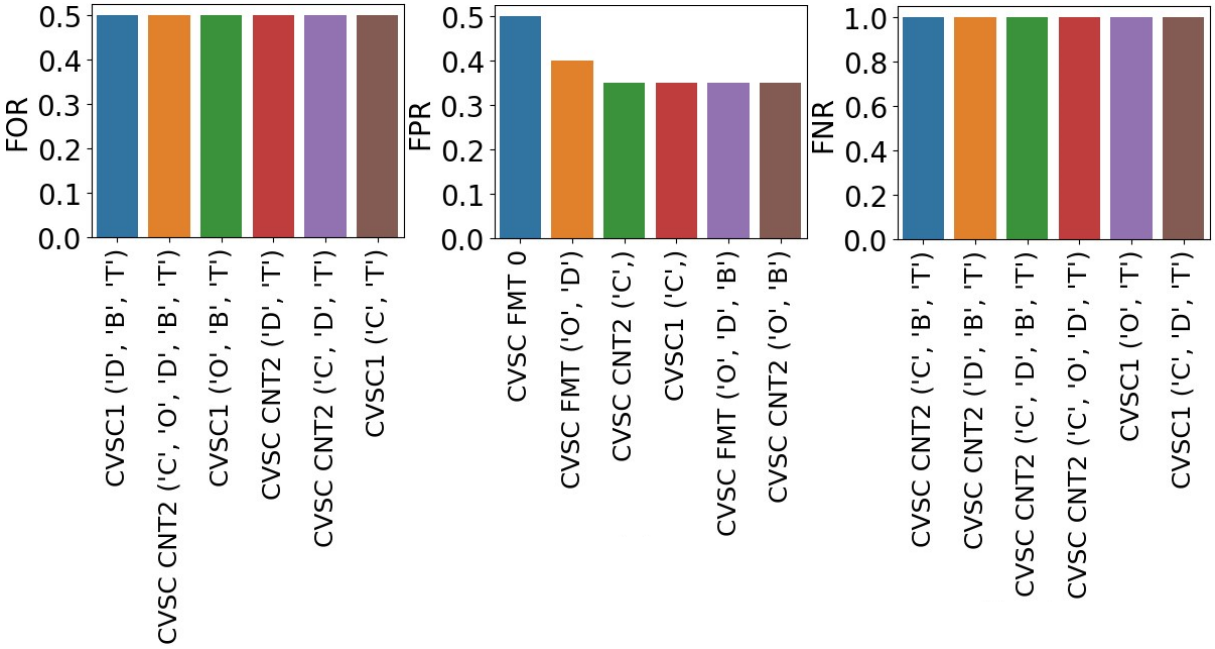


Figura 5.10: Maiores métricas de FOR, FPR e FNR dos modelos CVSCs.

possui tempo total de processamento 4.177 vezes maior que o modelo de menor tempo. O modelo CVSC MOG2 ('T',) possui tempo total de processamento 1.802 vezes maior que o modelo de menor tempo. Conclui-se que de fato, após aplicar os filtros de subtração de fundo, o modelo pode processar mais rápido e com melhor acurácia, sensibilidade e pontuação F1. A subtração de fundo deixa o modelo menos suscetivo a falsos negativos devido a objetos de fundo ou variações de iluminação.

5.3.6 Comparação dos modelos em termos de tempo de processamento

Os modelos CVSC foram comparados em termos do tempo de processamento de cada etapa do modelo. A Figura 5.13 mostra o gráfico dos modelos CVSC em ordem de menores tempos de pré-processamento de imagem e tempo total de processamento do modelo. O modelo de menor pré-processamento de imagem (com 0.55744s) foi o CVSC1 ('D', 'T'), mas obteve métricas baixas de acurácia (0.55), *recall* (0.1) e pontuação F1 (0.1818). O modelo CVSC MOG2 0 obteve tempo de pré-processamento de imagem (com 1.48093s) e é um dos modelos que obteve métricas mais altas de acurácia (0.9), *recall* (0.8) e pontuação F1 (0.8889). O modelo CVSC MOG2 0 possui tempo de pré-processamento de imagem 8.13 vezes maior que o modelo de menor tempo. O modelo CVSC MOG2 ('T',) obteve 0.99052s de tempo de pré-processamento de imagem. Esse modelo possui tempo de pré-processamento de imagem 5.44 vezes maior que o modelo de menor tempo.

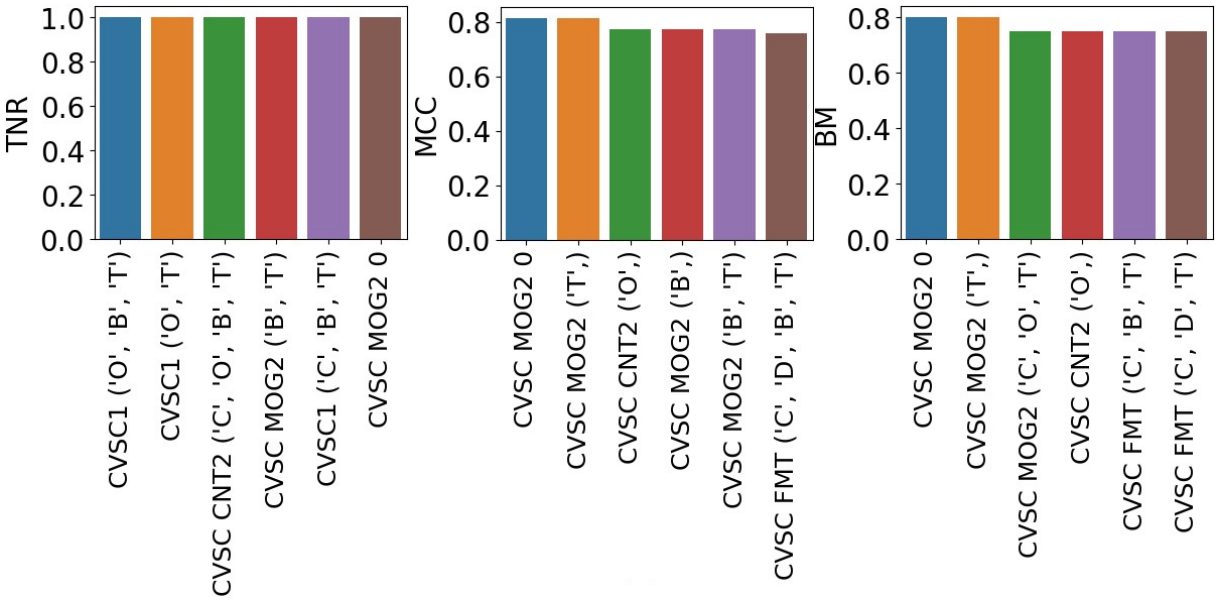


Figura 5.11: Maiores métricas de TNR, MCC e BM dos modelos CVSCs.

Após a etapa de pré-processamento da imagem, as etapas seguintes recebem como entrada uma imagem suavizada e com menos ruído, o que pode ajudar em análises subsequentes da imagem. Como a queda é uma variação abrupta do movimento de um indivíduo, os filtros de subtração de fundo na etapa de pré-processamento foram projetados para filtrar da imagem o que estivesse parado. Assim, filtro é capaz de remover o fundo da imagem, mas também remove a própria pessoa caso ela não esteja em movimento por exemplo. Nesse caso, resulta em uma sequência de quadros na cor preta que são identificados e não são processados pelas etapas posteriores de mascaramento da região de interesse, cálculo do fluxo ótico e da rede CNN, mas mesmo assim esse é um exemplo que o modelo pode concluir que não houve uma queda visto que não houve movimentação no vídeo. Portanto, os menores tempos de processamento na etapa de mascaramento da região de interesse, cálculo do fluxo ótico e da rede CNN foram de 0s. Esses são os casos onde o modelo concluiu mais vezes se ouve queda ou não apenas com a etapa de pré-processamento de imagem.

O modelo que obteve maior tempo de processamento na etapa de mascaramento da região de interesse foi o CVSC CNT2 0 com 59.05879s (acurácia de 0.825, *recall* de 0.95 e pontuação F1 de 0.8444). O modelo que obteve maior tempo de processamento na etapa do cálculo do fluxo ótico foi o CVSC1 0 com 0.27946s (acurácia de 0.7, *recall* de 0.7 e pontuação F1 de 0.7). O modelo que obteve maior tempo de processamento na etapa da rede CNN foi o CVSC CNT2 (’O’, ’B’) com 11.75652s (acurácia de 0.75, *recall* de 0.85

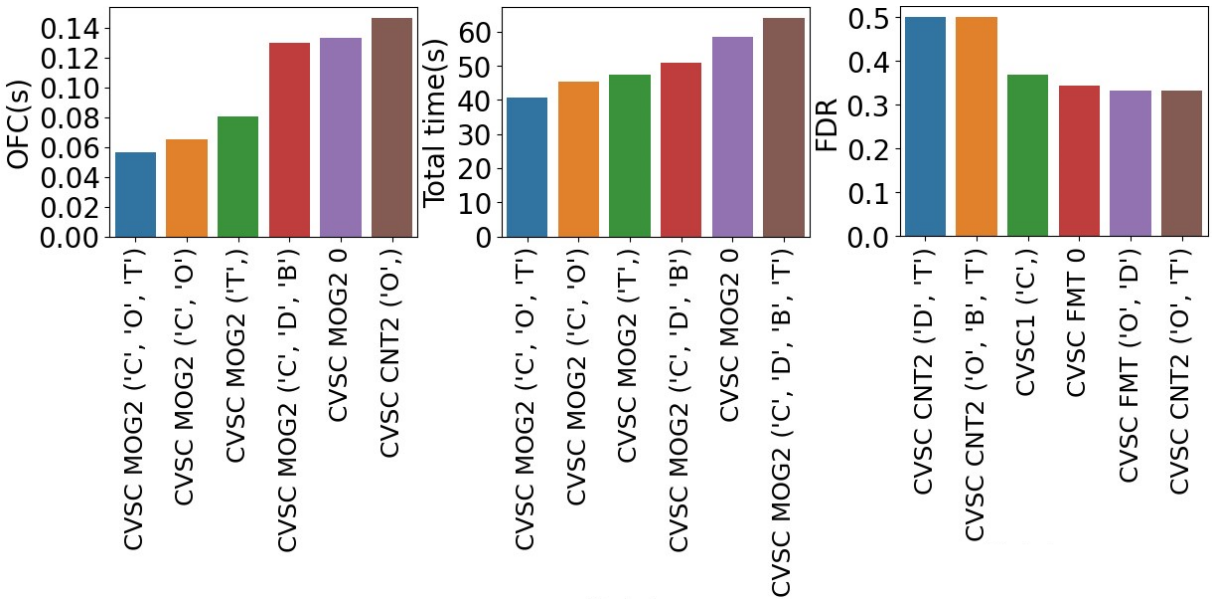


Figura 5.12: Tempo de processamento do OFC e de processamento total dos modelos de maior F1 -Score e gráfico dos modelos de maior FDR.

e pontuação F1 de 0.7727). Os modelos de menores tempos nas etapas de cálculo do fluxo ótico, mascaramento ROI e GAN, ou seja, que concluíram mais vezes se ocorreu uma queda ou não apenas com a etapa de pré-processamento de imagem, obtiveram desempenho inferior aos modelos que obtiveram tempos maiores de processamento nas etapas posteriores. Isso mostra que embora em alguns casos pode-se concluir se a pessoa caiu ou não apenas com a etapa de pré-processamento, ainda existe alguns casos que não foram identificados corretamente na etapa de pré-processamento de imagem, mas que foram identificados corretamente com a ajuda das etapas posteriores. Portanto, cada etapa do modelo é importante para identificar quedas com alta sensibilidade e a etapa de pré-processamento pode ser usada sim de maneira equilibrada para reduzir o tempo total de processamento do modelo.

5.3.7 Tabela dos modelos de maior pontuação e dos modelos de menor tempo

A Tabela 5.1 mostra as métricas dos modelos CVSC de maior pontuação F1. A Tabela 5.2 mostra as métricas dos modelos CVSC de menor tempo total de processamento. A Tabela 5.2 mostra que todos os modelos de menor tempo total de processamento utilizam o filtro de limiarização. A Tabela 5.1 mostra que dentre os modelos de melhores métricas, o modelo que possui menor tempo total de processamento foi o modelo que utiliza o filtro de limiarização. Os modelos que utilizam o filtro de remoção de fundo MOG2 obtiveram

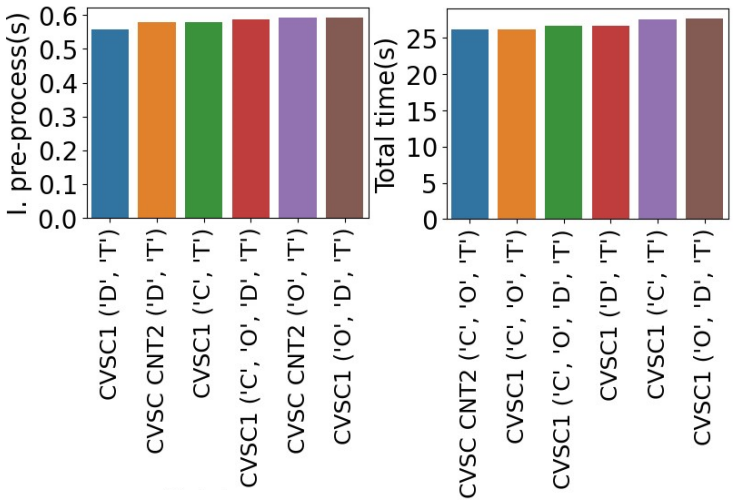


Figura 5.13: Modelos CVSC de menores tempos de pré-processamento e tempo total de processamento do modelo.

as melhores métricas.

Tabela 5.1: Tabela dos modelos CVSC de maior pontuação F1

Model name	CVSC MOG2 0	**CVSC MOG2 ('T')**	CVSC CNT ('O')
Accuracy	0.9	**0.9**	0.875
Precision	1.0	**1.0**	0.8
Recall	0.8	**0.8**	1.0
F1	0.8889	**0.8889**	0.8889
AUC	0.9	**0.9**	0.875
I. pre-proc.(s)	1.48093	**0.99052**	0.84817
ROI-M(s)	28.72982	**23.76672**	54.69319
OFC(s)	0.13296	**0.08086**	0.14679
CNN(s)	5.02954	**3.90141**	7.91803
Total time(s)	58.60100	**47.30136**	98.77532

Tabela 5.2: Tabela dos modelos CVSC de menor tempo de processamento total

Model name	CVSC CNT2 (C,O,T)	CVSC1 (C,O,T)	CVSC1 (C,O,D,T)
Accuracy	0.5	0.5	0.525
Precision	0.0	0.0	1.0
Recall	0.0	0.0	0.05
F1	0.0	0.0	0.0952
AUC	0.5	0.5	0.525
I. pre-proc.(s)	0.64673	0.62164	0.58816
ROI-M(s)	0.0	0.0	0.69898
OFC(s)	0.0	0.0	0.00207
CNN(s)	0.0	0.0	0.14254
Total time(s)	26.17540	26.21476	26.68316

O modelo CVSC1 ('C', 'B'), que utiliza apenas os filtros de fechamento e desfoque

em sequencia, obteve acurácia de 0.725, o qual e maior que a acurácia do modelo CVSC1 ('C') de 0.625 e do modelo CVSC1 ('B') de 0.65, modelos que utilizaram somente os filtros de fechamento e desfoque respectivamente. Além disso, o modelo CVSC1 ('D',), o qual é o modelo que utiliza apenas o filtro de dilatação, obteve 0.65 de acurácia que é igual a acurácia do modelo CVSC1 ('B') e do modelo CVSC1 ('O',), o qual é o modelo que utiliza apenas o filtro de abertura. Mas utilizando esses filtros em conjunto no modelo CVSC1 ('O', 'D', 'B'), a acurácia é de 0.7, a qual é maior que a acurácia dos modelos utilizando os filtros individuais. Pode-se pensar então que quanto mais filtros melhor a acurácia? Isso não é verdade porque o modelo CVSC1 ('C', 'D', 'B', 'T'), que utiliza os filtros de fechamento, dilatação, desfoque e limiarização, foi o modelo dentre os modelos que não utilizaram subtração de fundo, com a pior acurácia (0.5).

Dentre os modelos que não utilizaram técnicas de subtração de fundo, o modelo CVSC1 ('O', 'D', 'B'), que utiliza os filtro de abertura, dilatação e desfoque, foi o modelo mais lento, com tempo de processamento total de 108.49960s e acurácia de 0.7. O modelo mais lento, ou seja, com maior tempo de processamento total foi o CVSC FMT 0, modelo que utilizou apenas a técnica de subtração de fundo TMF, com tempo de 128.16381s. Além disso, o modelo que obteve o melhor *trade-off* entre o tempo total de processamento e acurácia foi o CVSC MOG2 ('T',),modelo que utiliza a técnica de subtração de fundo MOG2 e o filtro de limiarização, com tempo de processamento 47.30136s e acurácia de 0.9. Os resultados mostram que em geral as técnicas de subtração de fundo melhoram a acurácia e podem diminuir o tempo de processamento total dependendo da combinação de filtros usados.

5.3.8 Comparação com outros modelos na literatura que também usaram gravações RGB

A Tabela 5.3 mostra os resultados de acurácia, sensibilidade, precisão e *F1-Score* para uma avaliação dos modelos CVSCs apenas com outros modelos CNN na literatura que também usaram gravações RGB como modalidade de detecção de queda. *Chen et al* [7] usou o conjunto de dados COCO para treinamento e NTU RGB+D para ajuste fino e avaliação.*Xu et al* [8] usou o NTU RGB +D e outros dois conjuntos de dados para avaliação e treinamento. Foram usados 132 vídeos de queda RGB e 155 IR para avaliar os modelos CVSC. A tabela 5.3 mostra que a sensibilidade do modelo CVSC 3 RGB é de 99,23% superando outras propostas que utilizam o mesmo *dataset* sobre a mesma modalidade de detecção RGB. A sensibilidade do modelo CVSC 4 IR é de 100% porque o

modelo não obteve nenhum falso positivo. Portanto, nossos modelos são muito confiáveis para identificar corretamente uma pessoa. A tabela 5.3 mostra que o F1 Score do modelo CVSC 3 RGB é de 98,09% superando outras propostas que utilizam o mesmo *dataset* sobre a mesma modalidade de detecção RGB. O F1 Score do modelo CVSC 4 IR é de 98,69%.

Tabela 5.3: Resultados do CVSC em comparação com outros modelos, usando o mesmo conjunto de dados para teste.

Método	**Modalidade**	**Sensibilidade**	**Acurácia**	**Precisão**	**F1-Score**
[7]	RGB	94,25%	99,83%	98,73%	0,96
[8]	RGB	—	91,70%	—	—
CVSC	**RGB**	**96,92%**	**94,07%**	**96,92%**	**0,96**
CVSC 2	**RGB**	**77,97%**	**75,61%**	**95,83%**	**0,85**
CVSC 3	**RGB**	**99,23%**	**96,29%**	**96,99%**	**0,98**
CVSC 4	**RGB**	**99,21%**	**96,24%**	**96,95%**	**0,98**
CVSC	**IR**	**95,74%**	**93,78%**	**97,83%**	**0,96**
CVSC 2	**IR**	**99,31%**	**96,71%**	**97,33%**	**0,98**
CVSC 3	**IR**	**94,11%**	**91,77%**	**97,29%**	**0,95**
CVSC 4	**IR**	**100%**	**97,43%**	**97,42%**	**0,98**

Devido à natureza do problema, falsos positivos, cujo impacto pode ser avaliado pela precisão, se em pequeno número, são tolerados. Um resultado falso negativo em uma queda não percebida, e pode ser estimado por sensibilidade e métricas F1-Score. Um falso negativo pode deixar uma pessoa idosa caída desacompanhada, portanto inaceitável. O modelo supera os modelos de modalidade RGB encontrados na literatura, principalmente no que diz respeito à sensibilidade.

A sensibilidade considera a relação entre quedas classificadas corretamente e o número total de eventos de queda. Para um sistema de detecção de quedas para idosos, a sensibilidade é a métrica principal e, idealmente, deve ser de 100%. O F1-Score é avaliado porque esta medida é a média harmônica das métricas de precisão e sensibilidade. A pontuação F1 pode ser uma medida melhor a ser utilizada quando há uma distribuição desequilibrada das classes (grande número de ADL), como é o caso do problema de detecção de quedas, pois há mais atividades diárias do que quedas. Além disso, o CVSC apresenta a melhor sensibilidade, o melhor F1-Score, podendo lidar com mais de uma modalidade de detecção de quedas ao mesmo tempo. Além de apresentar a melhor sensibilidade e o melhor F1-Score, o CVSC é capaz de lidar com mais de um tipo de detecção de queda ao mesmo tempo.

A Tabela 5.3 mostra que o desempenho do CVSC2 IR foi melhor que o CVSC IR. Os

resultados CVSC2 IR mostram que o FMT para remover o fundo e os filtros aplicados funcionam bem para as imagens IR. As imagens IR são mais ruidosas do que as imagens RGB, então os filtros são mais recomendados em imagens IR. O algoritmo BS melhora o desempenho do modelo CVSC2 IR contra falsos negativos resultantes de objetos de fundo e variações de iluminação. O modelo CVSC 2 IR melhorou em quase todas as métricas em relação ao CVSC IR. No entanto, o modelo CVSC 2 RGB tem uma sensibilidade menor do que o modelo CVSC RGB. Para melhorar o resultado CVSC 2 RGB, foi desenvolvido o modelo CVSC 3 com a técnica CNT, que permite um melhor controle da filtragem de fundo pela estabilidade dos pixels do que a técnica FMT, que filtra apenas em relação à mediana temporal. A partir dos testes, foi utilizado a técnica de contagem de pixels para obter uma filtragem mais estável que o FMT e assim melhorar o resultado geral entre as modalidades CVSC 2 RGB. Portanto, a Tabela 5.3 mostra que a sensibilidade do modelo CVSC 3 RGB é maior que a dos modelos anteriores.

O CSVC 4 possui boas métricas em RGB e IR. O CSVC 4 IR tem a melhor sensibilidade entre os modelos da tabela. O CSVC 4 tem sensibilidade de 100%, o que significa que o algoritmo detectou todas as ocorrências de queda. O que é um importante parâmetro para o escopo da pesquisa em detecção de queda. Portanto, o modelo proposto pode atender ao objetivo. O sistema não deixou nenhum idoso desassistido em caso de queda. Observou-se que o método baseado em ROI melhora a qualidade da reconstrução na região ROI, pois o modelo aprende a reconstruir apenas a região de interesse.

5.3.9 Comparação dos modelos em termos de tempo de processamento utilizando GPU

Nesta seção são apresentados os resultados do tempo de processamento no mesmo computador, mas agora utilizando GPU. As especificações da máquina são mostradas na Figura 5.14. Foram realizados testes nesse computador com sistema operacional Debian sem utilizara a GPU e utilizando a GPU para avaliar a variação no tempo de processamento dos modelos.

Na Figura 5.15, o terminal superior à esquerda mostra o processamento da detecção de queda em vídeo, sendo executada pelo código utilizando a GPU NVIDIA GeForce GTX 1660. O termina inferior à esquerda representa as métricas de uso da GPU, de uma forma simplificada e direta, enquanto a direita representa as métricas detalhadas de uso da GPU. As métricas apresentadas no quadro inferior esquerdo, marcam 30,75 Watts de potência, com 19% de utilização da GPU, 32% da velocidade máxima da ventoinha, 44ºC

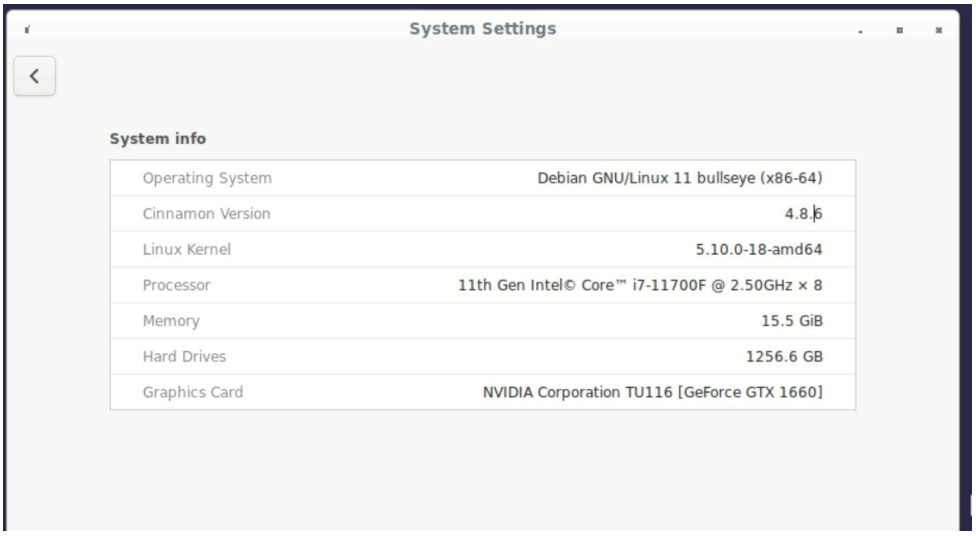


Figura 5.14: Especificações da máquina com sistema operacional Debian.

no interior da placa de vídeo com 5412 MB de memória em uso. No terminal da direita, as métricas mais detalhadas mostram que a versão do Nvidia SMI é 5420.61.05, com versão 520.61.05 do driver, e versão 11.8 do CUDA. Este mesmo terminal, mostra em detalhes os processos utilizando a GPU.

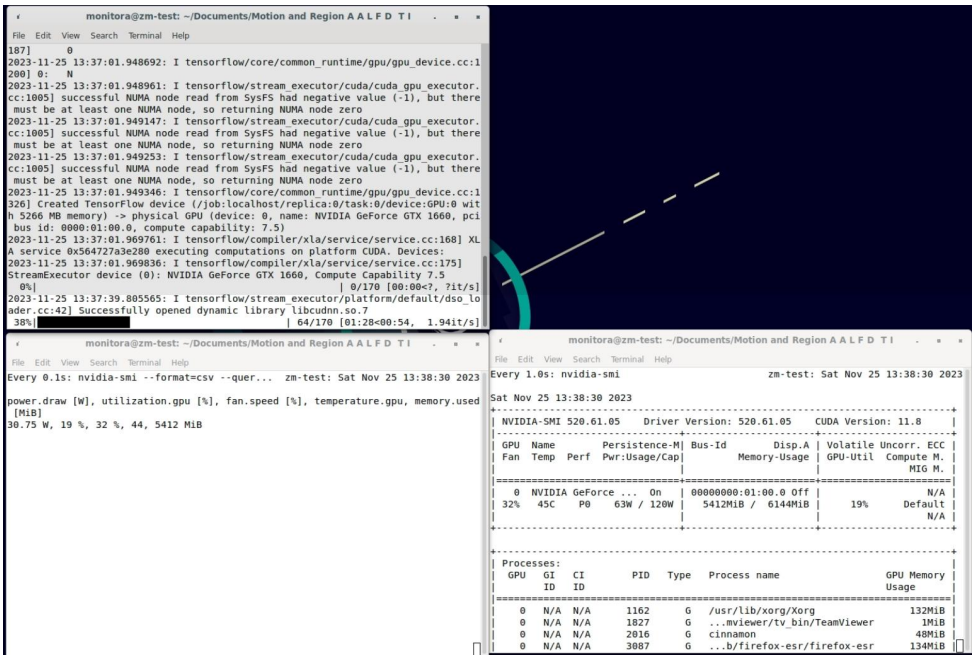


Figura 5.15: Métricas da GPU durante o processamento do modelo de visão computacional.

A Figura 5.15 mostra o gráfico da utilização da GPU (em azul) e do uso de memória da mesma (em marrom) durante o processamento do modelo ao longo do tempo, na mesma máquina anterior. A Figura 5.15 mostra que o uso da memória da GPU antes de inciar o processamento do modelo estava abaixo de 25% de uso e aumenta para acima de 75% de

uso ao longo do processamento do modelo. A Figura 5.15 também mostra que o gráfico da utilização da GPU estava por volta de 0% antes de iniciar o processamento do modelo. Durante o processamento do modelo, a utilização da GPU sobe para acima de 75% no início, mas depois segue por volta de 25% com algumas variações. A Figura 5.15 também mostra a utilização da CPU, da GPU, o consumo de memória principal e da memória da GPU por cada processo utilizando a GPU. A Figura 5.15 mostra o PID do processo e o tipo de processo.

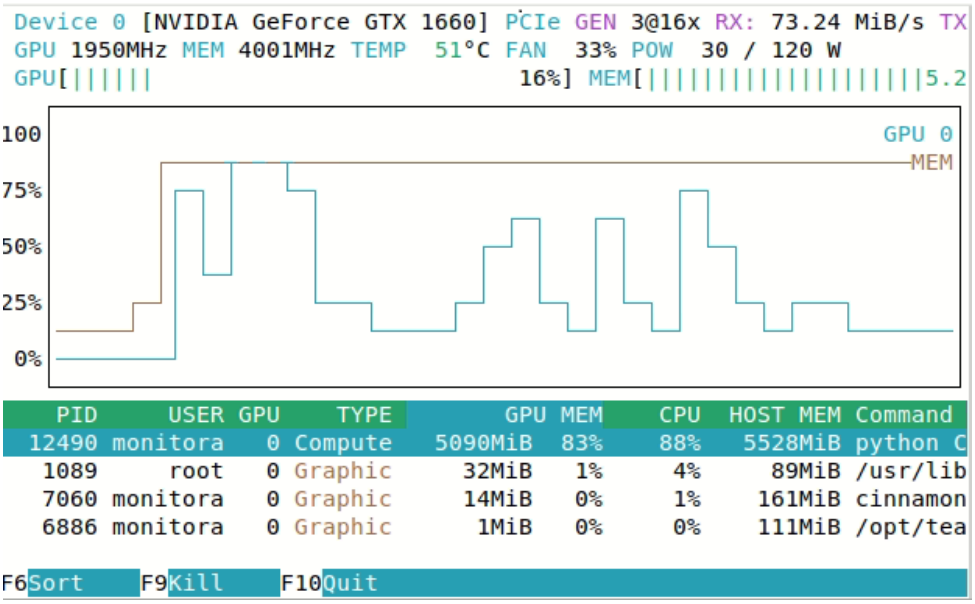


Figura 5.16: Gráfico da utilização da GPU (em azul) e do uso de memória (em marrom) durante o processamento do modelo ao longo do tempo.

Também foram realizados os mesmos testes com o mesmo dataset, no entanto, em um computador diferente. Agora utilizando uma CPU Intel Xeon E-2356G, com 128GB de memória RAM, e uma GPU Nvidia Quadro RTX A4000 com Windows. A avaliação de desempenho foi realizado em uma máquina com a configuração especificada a seguir. Trata-se de um Servidor Monoprocessado Xeon E-2356G | 128GB | SSD960| Rack.

- Processador: Intel® Xeon® E-2356G (6C/12T @ 3.20 GHz - Rocket Lake)
- Placa Mãe: Gigabyte® Server Board Xeon, Modelo MX33-BS0.
- Memoria: 128 GB UDIMM DDR4-3200.
- RAID-1: SSD de 960 GB PNY® CS900. RAID suportado em Windows 10, Server 2016/2019, Red Hat e SUSE Linux. Sem Unidades Óticas (CD, DVD).
- Saídas de Rede Gigabit: 02 (duas) Portas Gigabit Intel® i210 incorporadas.

- Dual Port 10 GbE Adicional: Placa Intel® X520-DA2 (2x 10GbE SFP+)
- Placa de Vídeo NVIDIA Quadro RTX A4000 16GB GDDR6 ECC, PCI Express 4.0 x16
    - Núcleos CUDA: 6144
    - Núcleos tensores: 192
    - Desempenho do núcleo RT: 37.4 TFLOPS
    - Desempenho do tensor: 153.4 TFLOPS
    - Conector de energia: 1x 6-pin PCIe
    - Conectores de display: 4x DisplayPort 1.4a
- Gabinete: Rack Mount 4U, sem trilhos
- Fonte: REAL CORSAIR RM1000X. Fonte de alimentação 1000W de potência, ATX12V v2.31 e EPS12V v2.92, Bivolt Automático, PFC Ativo, ventoinha de 135mm com controle automático de velocidade, uma linha de +12V, ATX20 / 24, 2x ATX12V (4+4 pinos), 8x PCI Express (6+2 pinos), 11x SATA, 12x Molex, 1x Floppy, CABEAMENTO MODULAR, cabos com malha de proteção, certificações 80 PLUS GOLD e NVIDIA SLI-Ready ,Corretor do fator de potência ativo (PFC) com valor de PF de 0,99.
- Cabeamento e Amarração: "Origami Design"para otimização de fluxo de ar.

A Figura 5.17 mostra, no terminal à esquerda o processamento da detecção de queda em vídeo, sendo executada pelo código utilizando a GPU NVIDIA RTX A4000. A Figura 5.17 mostra à direita do terminal o gráfico da utilização da GPU em % pelo tempo. O código esta processando uma série de vídeos de quedas simuladas do dataset. Entre o processamento dos vídeos o gráfico de utilização da GPU fica por volta de 0%. Quando o modelo esta processando o vídeo, a utilização da GPU fica um pouco abaixo de 50%.

A Figura 5.18 mostra gráfico da utilização da GPU pelo tempo utilizando o software da NVIDIA. Durante o início do processamento, o gráfico da utilização da GPU oscila mais, contudo depois que o código está processando a séroe de vídeos do dataset, o gráfico da utilização fica pouco abaixo de 50% durante o processamento do vídeo do dataset.

A Figura 5.19 mostra o tempo de pré-processamento de imagem dos modelos CVSCs de maior F1-Score sem utilizar GPU no PC Debian (esquerda), utilizando GPU no PC Debian (meio) e no PC Windows com GPU (direita). A Figura 5.20 mostra o tempo de

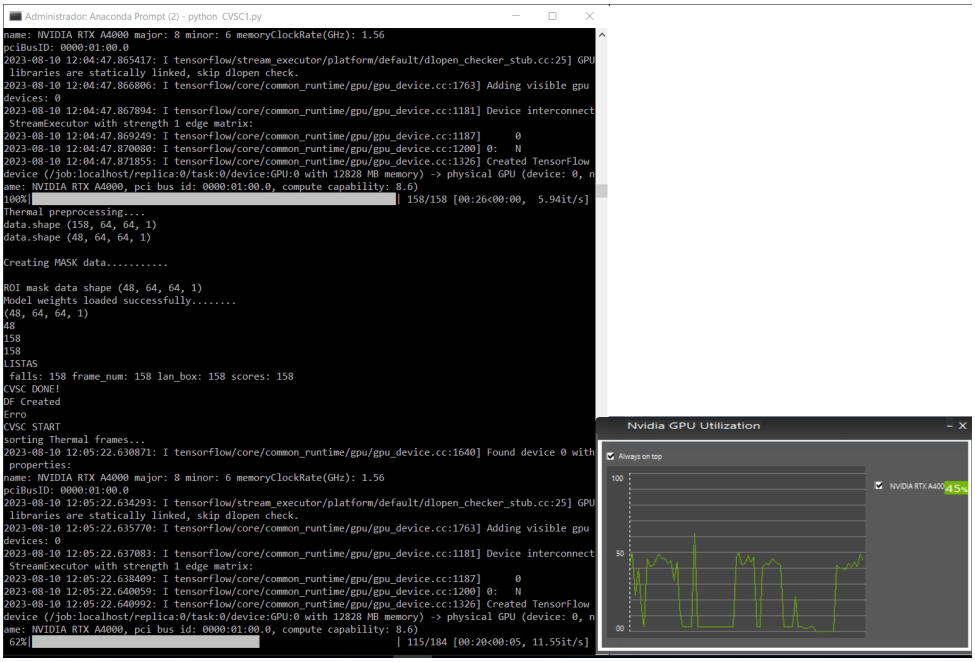


Figura 5.17: Terminal mostrando o processamento da detecção de queda em vídeo, sendo executada pelo código utilizando a GPU NVIDIA RTX A4000.

processamento total dos modelos CVSCs de maior F1-Score sem utilizar GPU (esquerda), utilizando GPU no PC Debian (meio) e no PC Windows com GPU (direita).

Usando a GPU Nvidia GeForce GTX 1660, os tempos de processamento total foram reduzidos em todos os casos. O CVSC MOG2 0, no PC Debian sem GPU, durou 58,6010 segundos, enquanto, com assistência da GPU, reduziu seu tempo para 51,7508 segundos. O CVSC MOG2 (‘T’), no PC Debiam sem GPU, durou 47,3014 segundos, enquanto, com assistência da GPU, reduziu seu tempo para 43,5603 segundos. O CVSC CNT (‘O’), no PC Debian sem GPU, durou 98,7753 segundos, enquanto, com assistência da GPU, reduziu seu tempo para 82,8209 segundos.

O modelo CVSC MOG2 0, no servidor, reduziu seu tempo de processamento total para 33,5222 segundos. O CVSC MOG2 (’T’), no servidor, reduziu seu tempo de processamento para 44,6915 segundos. O CVSC CNT (’O’), reduziu seu tempo de processamento para 63,2107 segundos. Essa redução no tempo de processamento é importante tendo em vista a gravidade do evento de quedae portanto, quanto mais rapido for detectado, mais rapido o idoso poderá ser acudido.

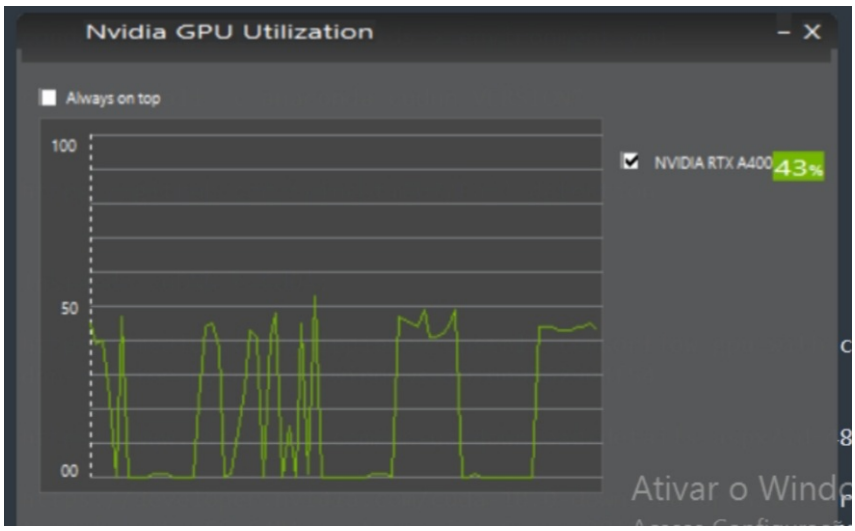


Figura 5.18: gráfico da utilização da GPU pelo tempo utilizando o software da NVIDIA.

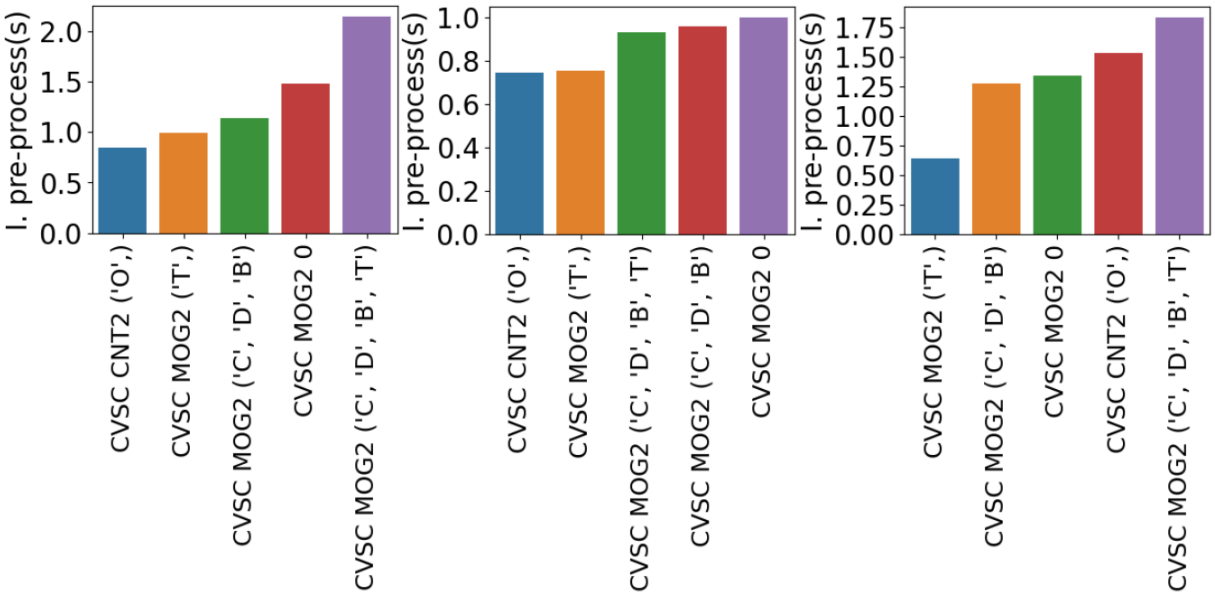


Figura 5.19: Tempo de pré-processamento de imagem dos modelos CVSCs de maior F1-Score sem utilizar GPU (esquerda), utilizando GPU no PC Debian (meio) e no PC Windows com GPU (direita).

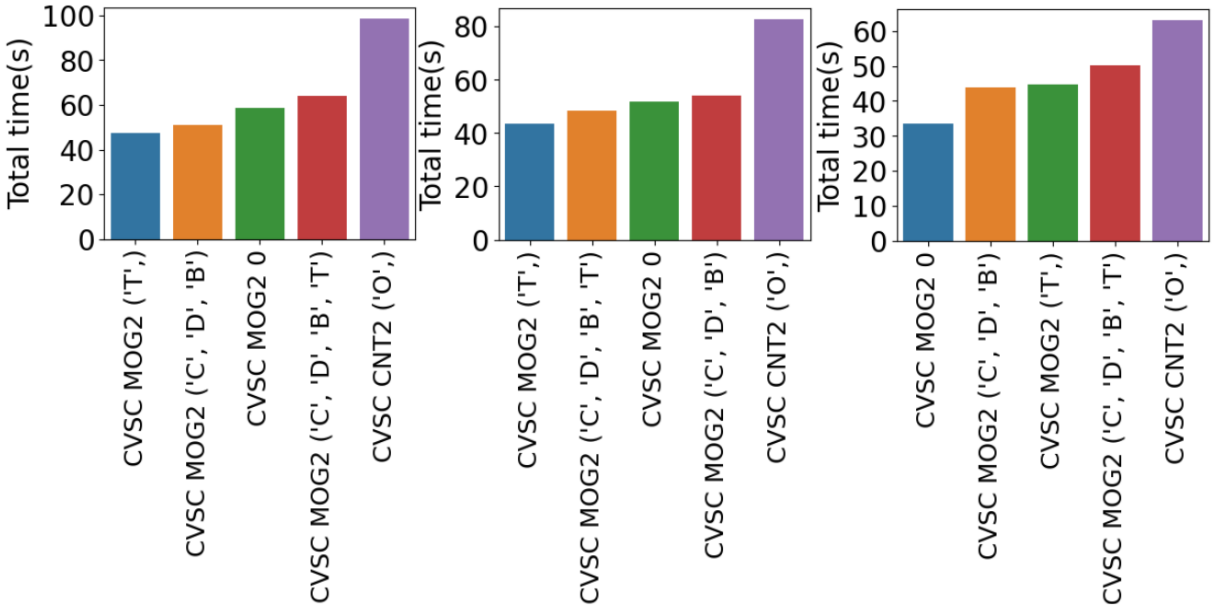


Figura 5.20: Tempo de processamento total dos modelos CVSCs de maior F1-Score sem utilizar GPU (esquerda), utilizando GPU no PC Debian (meio) e no PC Windows com GPU (direita).

Capítulo 6

Conclusões

Portanto, este trabalho propõe e avalia um modelo de rede neural usando técnicas de aprendizado de máquina e visão computacional para melhorar o bem-estar e a segurança de indivíduos em risco de queda, como idosos ou pessoas com mobilidade reduzido. Considerando que as quedas são um grave problema de saúde pública e as pessoas com mais de 65 anos estão entre as mais vulneráveis a lesões graves decorrentes de uma queda e que as quedas podem afetar negativamente a mentalidade do idoso, resultando em baixa autoestima, pois se tornam dependentes de uma pessoa que o acompanha constantemente, além do desprazer das constantes idas ao hospital. Uma abordagem natural e prática para idosos ou pessoas com mobilidade limitada requer um sistema eficaz para verificar remotamente seu bem-estar onde quer que estejam. de reconstrução maior indica que ocorreu uma queda.

Neste trabalho foi coberto uma ampla gama de tópicos de visão computacional. Desde a formação da imagem, como as imagens podem ser pré-processadas para remover ruído ou desfoque, segmentadas em regiões ou convertidas em descritores de características. Várias imagens podem ser combinadas e registradas, com os resultados usados para estimar movimento, rastrear pessoas e mesclar imagens em composições e renderizações mais atraentes e interessantes. As imagens também podem ser analisadas para produzir descrições semânticas de seu conteúdo.

O trabalho também expôs algumas técnicas matemáticas. Isso inclui matemática contínua, como processamento de sinal, abordagens variacionais, geometria tridimensional e projetiva, álgebra linear e mínimos quadrados. Também expôs tópicos em matemática discreta e ciência da computação, como algoritmos de grafos e otimização combinatória [67]. Como muitos problemas em visão computacional são problemas inversos que envolvem a estimativa de quantidades desconhecidas a partir de dados de entrada ruidosos, também

foi examinado as técnicas aprendizado de máquina para aprender modelos probabilísticos de grandes quantidades de dados de treinamento.

Como a disponibilidade de imagens visuais parcialmente rotuladas na *Internet* continua a aumentar exponencialmente, esta última abordagem continuará a ter um grande impacto em nesse campo de pesquisa [2]. O campo de pesquisa é tão amplo e princípios unificadores que possam ser usados para simplificar o estudo de visão computacional na área da saúde estão sendo desenvolvidos [39] [2]. Parte do trabalho está na ampla definição de visão computacional, que é a análise de imagens e vídeos, bem como na incrível complexidade inerente à formação de imagens visuais [67]. Da mesma forma, a visão computacional se baseia em uma ampla variedade de subdisciplinas, o que torna difícil cobrir um curso de um semestre. Por outro lado, a incrível amplitude e complexidade técnica dos problemas de visão computacional é o que atrai muitas pessoas para esse campo de pesquisa.

Ao analisar as movimentações de comportamento, o sistema pode determinar o estado atual da pessoa e enviá-lo ao sistema. Através desse método é possível alcançar um sistema de saúde para idosos através de meios técnicos. O modelo CVSC 4 utilizando câmera RGB obteve 96,24% de acurácia, 96,95% de precisão, 99,21% de sensibilidade e 98,07% de F1 - score. O modelo utilizando câmeras de infra-vermelho obteve 97,43% de acurácia, 97,42% de precisão, 100% de sensibilidade e 98,69% de F1 - score, conforme mostrado na Tabela 5.3. As métricas provam que o sistema tem uma alta taxa de reconhecimento de quedas e 5.3 mostra o desempenho de outras propostas de detecção de quedas através de visão computacional utilizando o mesmo *NTU RGB+D Action Recognition Dataset* para comparação.

A capacidade de tratar simultaneamente imagens claras (RGB) e escuras (infravermelho) é essencial para cenários de cuidado a idosos. Embora o sistema não tenha obtido a melhor acurácia entre os sistemas existentes, essa métrica foi suficientemente alta utilizando câmera RGB, evitando a ocorrência frequente de falsos-positivos. Em contrapartida, considerando tanto a sensibilidade quanto o F1-Score, as quais são métricas relacionadas ao impacto dos falsos negativos, o sistema apresentou melhores resultados do que as demais propostas que utilizaram o mesmo *dataset* de teste. Esses valores são altos e dão grande fidedignidade ao sistema para operação durante a noite, aonde as quedas de idosos acabam sendo muito frequentes. Contudo, o modelo proposto apresenta limitações quanto a detecção de queda na presença de várias pessoas ou a detecção de quedas simultâneas, situações que serão tratadas em pesquisas futuras.

Portanto, comparando os resultados de outras propostas, o método tem um desempenho competitivo com propostas de detecção recentes. A proposta do artigo é uma nova abordagem que consiste em uma combinação única de algumas das mais recentes técnicas de visão computacional com processamento de imagem, *Deep Learning*, bibliotecas e algoritmos na linguagem de programação *python*. Como dito em [13], o desenvolvimento de filtros foi fundamental para o desempenho dos modelos em câmeras RGB e IR. Esse modelo que pode ser implementada em um sistema interno de vigilância e monitoramento por câmeras, também chamado de *Circuito Fechado de Televisão* (CFTV), para monitorar pessoas e pode contribuir para a segurança contra o risco de acidentes com quedas. A técnica de pontuação de anomalias mostrou-se uma técnica de aprendizado que se adapta bem e é capaz de identificar a queda mesmo com a exposição de novos cenários de vídeo e quedas com grandes variações dos vídeos que foram utilizados para treinamento, sendo ideal para o uso do sistema em situações reais. O modelo foi treinado usando TSF e COCO *dataset*, mas avaliada no conjunto de dados NTU RGB+D.

Capítulo 7

Trabalhos futuros

Por último, como muito sobre o processo de formação da imagem é inerentemente incerto e ambíguo, uma abordagem estatística que modela a incerteza no mundo (por exemplo, o número e os tipos de animais em uma imagem) e o ruído no processo de formação da imagem geralmente é essencial [67]. As técnicas de inferência bayesiana podem então ser usadas para combinar modelos anteriores para estimar as incógnitas e modelar sua incerteza. As técnicas de aprendizado de máquina podem ser usadas para criar os modelos probabilísticos em primeiro lugar. Aprendizagem e inferência com algorítimos eficientes, como programação dinâmica, cortes de grafos e propagação de características, podem desempenhar um papel crucial nesse processo [67] [2].

Dada a amplitude de conceitos cobertos nesse trabalho, que novos desenvolvimentos serão aplicados no futuro? Uma das tendências recentes em visão computacional é usar grandes quantidades de dados visuais parcialmente rotulados na *Internet* como fontes para aprender modelos visuais de cenas e objetos. Tendo em vista as abordagens baseadas em dados assim bem-sucedidas em campos relacionados, como reconhecimento de fala, tradução automática, síntese de fala e música e até computação gráfica (tanto na renderização baseada em imagem quanto na animação da captura de movimento). Um processo semelhante está ocorrendo na visão computacional, com alguns dos novos trabalhos ocorrendo na interseção dos campos de reconhecimento de objetos e aprendizado de máquina. Técnicas quantitativas mais tradicionais em visão computacional, como estimativa de movimento, correspondência estéreo e aprimoramento de imagem, se beneficiam de modelos anteriores melhores para imagens, movimentos e disparidades, bem como técnicas de inferência estatística eficientes, como aquelas para Markov não homogêneo e de campos aleatórios ordem superior [67] [2]. Algumas técnicas, como a correspondência de características e estrutura de movimento, amadureceram a ponto de poderem ser aplicadas

a coleções quase arbitrárias de imagens de cenas estáticas. São técnicas estão relacionadas ao reconhecimento visual de grandes quantidades de dados que amadureceram e podem contribuir a pesquisa documentada nessa dissertação.

No entanto, a lacuna entre o computador e o desempenho humano nessa área ainda é grande e provavelmente permanecerá assim por muitos anos [67] [2]. Embora todos esses sejam desenvolvimentos encorajadores, a lacuna entre o desempenho humano e da máquina na compreensão semântica da cena permanece grande. Pode levar muitos anos até que os computadores possam nomear e delinear todos os objetos em uma fotografia com a mesma habilidade de uma criança de dois anos [67] [2]. Contudo, o estado da arte atual permite a criação de modelos robustos e suficientemente precisos, como os modelos documentados nessa pesquisa. No entanto, deve-se lembrar que o desempenho humano geralmente é o resultado de muitos anos de treinamento e familiaridade e geralmente funciona melhor em situações ecologicamente importantes. Por exemplo, enquanto os humanos parecem ser especialistas em reconhecimento facial, nosso desempenho real quando mostrado pessoas que não se conhece bem não é tão bom [67].

A combinação de algoritmos de visão com técnicas gerais de inferência que raciocinam sobre o mundo real provavelmente levará a mais avanços, embora alguns dos problemas possam vir a ser relacionados a AI completas, no sentido de uma emulação completa da experiência e inteligência humana. Seja qual for o resultado desses esforços de pesquisa, a visão computacional já está tendo um tremendo impacto em muitas áreas, incluindo fotografia digital, efeitos visuais, imagens médicas, segurança e vigilância e pesquisa baseada na Web [67] [2]. A amplitude dos problemas e técnicas inerentes a este campo, combinada com a riqueza da matemática e a utilidade dos algoritmos resultantes, garantirá que esta continue sendo uma área de estudo empolgante nos próximos anos.

Sendo assim, em trabalhos futuros, será adicionado mais camadas de ROI e CNN para que o sistema possa trabalhar com vídeos com a presença de mais de uma pessoa no quadro, representando quartos compartilhados por idosos, por exemplo. Outra possibilidade é desenvolver outro bloco de Deep Learning para treinar um modelo que possa melhorar o desempenho do método de detecção de queda diante de mudanças de fundo de imagem, objetos de fundo, mudanças de iluminação e movimento de câmera. Pretende-se aprimorar o modelo BS com câmeras em vários ambientes diferentes em uma casa de repouso. Além disso, planejamos montar um *dataset* real com imagens de uma casa de repouso para ajustes mais precisos do sistema a partir de testes com cenas realistas.

Referências

[1] HENRY, J.; PYLYPCHUK, Y.; SEARCY, T.; PATEL, V. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, Office of the National Coordinator for Health Information Technology ..., v. 35, n. 35, p. 2008–2015, 2016.

[2] BURGER, W.; BURGE, M. J. *Digital image processing: an algorithmic introduction using Java*. [S.l.]: Springer, 2016.

[3] CARDENAS, J.; GUTIERREZ, C. A.; AGUILAR-PONCE, R. Effects of antenna orientation in fall detection systems based on wifi signals. In: *2020 IEEE Latin-American Conference on Communications (LATINCOM)*. [S.l.: s.n.], 2020. p. 1–6.

[4] LIU, W.; LUO, W.; LIAN, D.; GAO, S. Future frame prediction for anomaly detection - a new baseline. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 6536–6545.

[5] ZAHAN, S.; HASSAN, G. M.; MIAN, A. Modeling human skeleton joint dynamics for fall detection. In: *2021 Digital Image Computing: Techniques and Applications (DICTA)*. [S.l.: s.n.], 2021. p. 01–07.

[6] HERNANDEZ, S. D.; DELAHOZ, Y.; LABRADOR, M. Dynamic background subtraction for fall detection system using a 2D camera. In: *2014 IEEE Latin-America Conference on Communications (LATINCOM)*. [S.l.: s.n.], 2014. p. 1–6.

[7] CHEN, Z.; WANG, Y.; YANG, W. *Video Based Fall Detection Using Human Poses*. 2021.

[8] XU, Q.; HUANG, G.; YU, M.; GUO, Y. Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications*, v. 540, n. C, 2020.

[9] MEHTA, V.; DHALL, A.; PAL, S.; KHAN, S. S. Motion and region aware adversarial learning for fall detection with thermal imaging. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. [S.l.: s.n.], 2021. p. 6321–6328.

[10] HO, J. Y.; HENDI, A. S. Recent trends in life expectancy across high income countries: retrospective observational study. *BMJ*, British Medical Journal Publishing Group, v. 362, 2018.

[11] QUIGLEY, P. A.; CAMPBELL, R. R.; BULAT, T.; OLNEY, R. L.; BUERHAUS, P.; NEEDLEMAN, J. Incidence and cost of serious fall-related injuries in nursing homes. *Clinical Nursing Research*, Sage Publications Sage CA: Los Angeles, CA, v. 21, n. 1, p. 10–23, 2012.

[12] BELASCO, A. G. S.; OKUNO, M. F. P. *Reality and challenges of ageing*. [S.l.]: SciELO Brasil, 2019. 1–2 p.

[13] SKORKA, O.; ISPASOIU, R. Tradeoffs with rgb-ir image sensors. *IEEE Transactions on Electron Devices*, v. 69, n. 6, p. 2915–2923, 2022.

[14] SANTOS, A.; SEIXAS, F.; FERNANDES, N. Provendo um modelo automático de detecção de quedas baseado em rede adversária generativa para assistência de idosos. In: *Anais do XXII Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2022. p. 120–131. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/21625>.

[15] GUO, Z.; YANG, M.; CHEN, N.; XIAO, Z.; YAN, B.; LIN, S.; ZHOU, L. Lightvo: Lightweight inertial-assisted monocular visual odometry with dense neural networks. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. [S.l.: s.n.], 2019. p. 1–6.

[16] Open Source Computer Vision Documentation. *OpenCV modules*. 2022. `https://docs.opencv.org/4.x/`. Accessed on January 31.

[17] VINCENT, L. Morphological area openings and closings for grey-scale images. In: SPRINGER. *Shape in picture: mathematical description of shape in grey-level images*. [S.l.], 1994. p. 197–208.

[18] SERRA, J.; VINCENT, L. An overview of morphological filtering. *Circuits, Systems and Signal Processing*, Springer, v. 11, p. 47–108, 1992.

[19] WU, Y.; LEE, T. Time-frequency feature decomposition based on sound duration for acoustic scene classification. In: IEEE. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2020. p. 716–720.

[20] BRADLEY, D.; ROTH, G. Adaptive thresholding using the integral image. *Journal of graphics tools*, Taylor & Francis, v. 12, n. 2, p. 13–21, 2007.

[21] LIU, W.; CAI, Y.; ZHANG, M.; LI, H.; GU, H. Scene background estimation based on temporal median filter with gaussian filtering. In: IEEE. *2016 23rd International Conference on Pattern Recognition (ICPR)*. [S.l.], 2016. p. 132–136.

[22] 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.

[23] BABUTA, A.; OSWALD, M.; RINIK, C. Machine learning algorithms and police decision-making: legal, ethical and regulatory challenges. Royal United Services Institute, 2018.

[24] TZIMAS, T. *Legal and Ethical Challenges of Artificial Intelligence from an International Law Perspective*. [S.l.]: Springer Nature, 2021.

[25] GICHOYA, J. W.; MCCOY, L. G.; CELI, L. A.; GHASSEMI, M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ health & care informatics*, BMJ Publishing Group, v. 28, n. 1, 2021.

[26] GEIS, J. R.; BRADY, A. P.; WU, C. C.; SPENCER, J.; RANSCHAERT, E.; JAREMKO, J. L.; LANGER, S. G.; KITTS, A. B.; BIRCH, J.; SHIELDS, W. F. Ethics of artificial intelligence in radiology: summary of the joint european and north american multisociety statement. *Canadian Association of Radiologists Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 70, n. 4, p. 329–334, 2019.

[27] CIRILLO, D.; CATUARA-SOLARZ, S.; MOREY, C.; GUNEY, E.; SUBIRATS, L.; MELLINO, S.; GIGANTE, A.; VALENCIA, A.; REMENTERIA, M. J.; CHADHA, A. S. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, Nature Publishing Group UK London, v. 3, n. 1, p. 81, 2020.

[28] BITTERMAN, D. S.; AERTS, H. J.; MAK, R. H. Approaching autonomy in medical artificial intelligence. *The Lancet Digital Health*, Elsevier, v. 2, n. 9, p. e447–e449, 2020.

[29] LARSSON, S.; HEINTZ, F. Transparency in artificial intelligence. *Internet Policy Review*, v. 9, n. 2, 2020.

[30] KUBAT, M.; KUBAT. *An introduction to machine learning*. [S.l.]: Springer, 2017.

[31] PATI, J. Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach. *IEEE Access*, IEEE, v. 7, p. 4232–4238, 2018.

[32] YU, Z.; WONG, H.-S.; WANG, H. Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, Oxford University Press, v. 23, n. 21, p. 2888–2896, 2007.

[33] TUR, C.; KALINCIK, T.; OH, J.; SORMANI, M. P.; TINTORÉ, M.; BUTZKUEVEN, H.; MONTALBAN, X. Head-to-head drug comparisons in multiple sclerosis: urgent action needed. *Neurology*, AAN Enterprises, v. 93, n. 18, p. 793–809, 2019.

[34] YUN, T.; COSENTINO, J.; BEHSAZ, B.; MCCAW, Z. R.; HILL, D.; LUBEN, R.; LAI, D.; BATES, J.; YANG, H.; SCHWANTES-AN, T.-H. Unsupervised representation learning improves genomic discovery for lung function and respiratory disease prediction. *medRxiv*, Cold Spring Harbor Laboratory Press, p. 2023–04, 2023.

[35] JENSEN, P. B.; JENSEN, L. J.; BRUNAK, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, Nature Publishing Group UK London, v. 13, n. 6, p. 395–405, 2012.

[36] LONCARIC, F.; CAMARA, O.; PIELLA, G.; BIJNENS, B. Integration of artificial intelligence into clinical patient management: focus on cardiac imaging. *Revista Española de Cardiología (English Edition)*, Elsevier, v. 74, n. 1, p. 72–80, 2021.

[37] CERCHIONE, R.; CENTOBELLI, P.; RICCIO, E.; ABBATE, S.; OROPALLO, E. Blockchain's coming to hospital to digitalize healthcare services: Designing a distributed electronic health record ecosystem. *Technovation*, Elsevier, v. 120, p. 102480, 2023.

[38] KONONENKO, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, Elsevier, v. 23, n. 1, p. 89–109, 2001.

[39] CLEOPHAS, T. J.; ZWINDERMAN, A. H. Machine learning in medicine-a complete overview. Springer, 2015.

[40] SUN, T. Applying deep learning to audit procedures: An illustrative framework. *Accounting Horizons*, American Accounting Association, v. 33, n. 3, p. 89–109, 2019.

[41] OBERSTE, L.; FINZE, N.; HOFFMANN, P.; HEINZL, A. Supporting the billing process in outpatient medical care: Automated medical coding through machine learning. In: AISEL. *Conference of the European Colloid and Interface Society (ECIS): Research Papers*. [S.l.], 2022. v. 2022, p. 1–18.

[42] CABALLERO, E.; FERREIRA, V. C.; LIMA, R. A.; ALBUQUERQUE, C.; MUCHALUAT-SAADE, D. C. Lator: Link-quality aware and thermal aware on-demand routing protocol for wban. In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. [S.l.: s.n.], 2020. p. 337–342.

[43] LIMA, R. A.; FERREIRA, V. C.; CABALLERO, E.; ALBUQUERQUE, C. V.; SAADE, D. C. M. Simulation of iso/ieee 11073 personal health devices in wbans. In: *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. [S.l.: s.n.], 2019. p. 221–224.

[44] FERREIRA, V. C.; BALBI, H. D.; SEIXAS, F. L.; ALBUQUERQUE, C. V. N. D.; SAADE, D. C. M. Wireless body area networks: An overview. *Minicurso do XXXV Simpósio Brasileiro de Telecomunicações*, SBRT, 2017.

[45] CABALLERO, E.; FERREIRA, V.; LIMA, R. A.; SOTO, J. C. H.; MUCHALUAT-SAADE, D.; ALBUQUERQUE, C. Bns: a framework for wireless body area network realistic simulations. *Sensors*, MDPI, v. 21, n. 16, p. 5504, 2021.

[46] SEIXAS, F. L.; CONCI, A.; SAADE, D. C. M. Sistema de apoio à decisão aplicado ao diagnóstico de demência, doença de alzheimer e transtorno cognitivo leve. *Jornal Brasileiro de TeleSSaúde*, v. 2, n. 4, p. 143–144, 2013.

[47] SWEENEY, C.; BERNARD, P. S.; FACTOR, R. E.; KWAN, M. L.; HABEL, L. A.; JR, C. P. Q.; SHAKESPEAR, K.; WELTZIEN, E. K.; STIJLEMAN, I. J.; DAVIS, C. A. Intrinsic subtypes from pam50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics. *Cancer epidemiology, biomarkers & prevention*, AACR, v. 23, n. 5, p. 714–724, 2014.

[48] LIAO, Z.; YOU, R.; HUANG, X.; YAO, X.; HUANG, T.; ZHU, S. Deepdock: enhancing ligand-protein interaction prediction by a combination of ligand and structure information. In: IEEE. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.], 2019. p. 311–317.

[49] CIMA, I.; SCHIESS, R.; WILD, P.; KAELIN, M.; SCHÜFFLER, P.; LANGE, V.; PICOTTI, P.; OSSOLA, R.; TEMPLETON, A.; SCHUBERT, O. Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 108, n. 8, p. 3342–3347, 2011.

[50] CHEN, H.; LAN, X.; YU, T.; LI, L.; TANG, S.; LIU, S.; JIANG, F.; WANG, L.; HUANG, Y.; CAO, Y. Development and validation of a radiogenomics model to predict axillary lymph node metastasis in breast cancer integrating mri with transcriptome data: A multicohort study. *Frontiers in Oncology*, Frontiers Media SA, v. 12, p. 1076267, 2022.

[51] SANTOS, A. C.; FIRMINO, R. M.; SOTO, J. C.; MEDEIROS, D. S.; MATTOS, D. M.; ALBUQUERQUE, C. V.; SEIXAS, F.; MUCHALUAT-SAADE, D. C.; FERNANDES, N. C. Aplicações em redes de sensores na área da saúde e gerenciamento de dados médicos: tecnologias em ascensão. *Sociedade Brasileira de Computação*, 2020.

[52] PIKE NASSER M. MUSTAFA, D. T. M.; BRUSIC, V. Sensor networks and data management in healthcare: Emerging technologies and new challenges. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, Milwaukee, WI, USA, USA, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8753998/authors>.

[53] MILAN, A. A.; FERNANDES, N. C.; MEDEIROS, D. S. V. A monte carlo approach for antenna blocking probability estimation in mobile networks. In: *2022 25th Conference on Innovation in Clouds, Internet and Networks (ICIN)*. [S.l.: s.n.], 2022. p. 146–150.

[54] BADIDI, E.; MOUMANE, K. Enhancing the processing of healthcare data streams using fog computing. *2019 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, Barcelona, Spain, Spain, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8969736/authors>.

[55] BAEK, W.-S.; KIM, D.-M.; BASHIR, F.; PYUN, J.-Y. Real life applicable fall detection system based on wireless body area network. In: IEEE. *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC)*. [S.l.], 2013. p. 62–67.

[56] KEPSKI, M.; KWOLEK, B. Embedded system for fall detection using body-worn accelerometer and depth sensor. In: IEEE. *2015 IEEE 8th International conference on intelligent data acquisition and advanced computing systems: technology and applications (IDAACS)*. [S.l.], 2015. v. 2, p. 755–759.

[57] FORTI, M. M. H. . A. G. . G. A. . G.; ZHOU, M. A smartphone-enabled fall detection framework for elderly people in connected home healthcare. *IEEE Network*, IEEE, Milwaukee, WI, USA, USA, v. 33, p. 58 – 63, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8933560>.

[58] AJERLA, D.; MAHFUZ, S.; ZULKERNINE, F. A real-time patient monitoring framework for fall detection. *Wireless Communications and Mobile Computing*, Hindawi Limited, v. 2019, p. 1–13, 2019.

[59] MSHALI, H.; LEMLOUMA, T.; MAGONI, D. Adaptive monitoring system for e-health smart homes. *Pervasive and Mobile Computing*, v. 43, p. 1 – 19, 2018. ISSN 1574-1192. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574119217305370>.

[60] RAMACHANDRAN, A.; KARUPPIAH, A. A survey on recent advances in wearable fall detection systems. *BioMed Research International*, v. 2020, 2020.

[61] ALMOHAMMADI, N.; SEN, A. A. A.; BORIE, H.; ALMUHAMMADI, A.; ALKHODRE, A.; YAMIN, M. A framework for enhancing relief system of health domain by iot. In: *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. [S.l.: s.n.], 2019. p. 1326–1330.

[62] COLóN, L. N. V.; DELAHOZ, Y.; LABRADOR, M. Human fall detection with smartphones. In: *2014 IEEE Latin-America Conference on Communications (LATINCOM)*. [S.l.: s.n.], 2014. p. 1–7.

[63] DIAS, P. V. G. F.; COSTA, E. D. M.; TCHEOU, M. P.; LOVISOLO, L. Fall detection monitoring system with position detection for elderly at indoor environments under supervision. In: *2016 8th IEEE Latin-American Conference on Communications (LATINCOM)*. [S.l.: s.n.], 2016. p. 1–6.

[64] VERGüTZ, A.; SILVA, R. da; NACIF, J. A. M.; VIEIRA, A. B.; NOGUEIRA, M. Mapping critical illness early signs to priority alert transmission on wireless networks. In: *2017 IEEE 9th Latin-American Conference on Communications (LATINCOM)*. [S.l.: s.n.], 2017. p. 1–6.

[65] GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial networks. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 11, p. 139–144, 2020.

[66] SKANSI, S. *Introduction to Deep Learning: from logical calculus to artificial intelligence*. [S.l.]: Springer, 2018.

[67] SZELISKI, R. *Computer vision: algorithms and applications*. [S.l.]: Springer Nature, 2022.

[68] SOLOMON, C.; BRECKON, T. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. [S.l.]: John Wiley & Sons, 2011.

[69] BERGH, F. van D.; LALIOTI, V. Software chroma keying in an immersive virtual environment. *South African Computer Journal*, n. 24, 1999.

[70] HARALICK, R. M.; SHAPIRO, L. G. Image segmentation techniques. *Computer vision, graphics, and image processing*, Elsevier, v. 29, n. 1, p. 100–132, 1985.

[71] ISLAM, K. Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing*, Elsevier, v. 101, p. 103970, 2020.

[72] SAMBOLEK, S.; IVASIC-KOS, M. Automatic person detection in search and rescue operations using deep cnn detectors. *Ieee Access*, IEEE, v. 9, p. 37905–37922, 2021.

[73] MING, Z.; ZHU, M.; WANG, X.; ZHU, J.; CHENG, J.; GAO, C.; YANG, Y.; WEI, X. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing*, Elsevier, v. 119, p. 104394, 2022.

[74] WANG, K.; WANG, H.; LIU, M.; XING, X.; HAN, T. Survey on person re-identification based on deep learning. *CAAI Transactions on Intelligence Technology*, Wiley Online Library, v. 3, n. 4, p. 219–227, 2018.

[75] WU, D.; ZHENG, S.-J.; ZHANG, X.-P.; YUAN, C.-A.; CHENG, F.; ZHAO, Y.; LIN, Y.-J.; ZHAO, Z.-Q.; JIANG, Y.-L.; HUANG, D.-S. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, Elsevier, v. 337, p. 354–371, 2019.

[76] OpenCV. *Optical Flow*. 2022. `https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html`. Accessed on January 31.

[77] SOTO, J. C.; GALDINO, I.; CABALLERO, E.; FERREIRA, V.; MUCHALUAT-SAADE, D.; ALBUQUERQUE, C. A survey on vital signs monitoring based on Wi-Fi CSI data. *Computer Communications*, Elsevier, v. 195, p. 99–110, 2022.

[78] ALMEIDA, G. C. de; SANTOS, A. C. dos; SOARES, C. L. d. A.; PINTO, P. C. A.; BELLO, F. d. S. D.; BOECHAT, Y. E. M.; SEIXAS, F. L.; SANTOS, A. A. S. dos; MESQUITA, C. T.; MESQUITA, E. T. Nova geração da telessaúde: Oportunidades, tendências e desafios. *Sociedade Brasileira de Computação*, 2023.

[79] SOTO, J. C.; GALDINO, I.; CABALLERO, E.; FERREIRA, V.; MUCHALUAT-SAADE, D.; ALBUQUERQUE, C. Monitoramento de sinais vitais utilizando redes wi-fi. In: CAMPISTA, M.; DUARTE, F. (Ed.). *Livro de Minicursos do SBRC 2022*. 1. ed. [S.l.]: SBC, 2022. cap. 5.

[80] HITCHO, E. B.; KRAUSS, M. J.; BIRGE, S.; DUNAGAN, W. C.; FISCHER, I.; JOHNSON, S.; NAST, P. A.; COSTANTINOU, E.; FRASER, V. J. Characteristics and circumstances of falls in a hospital setting: a prospective analysis. *Journal of general internal medicine*, Wiley Online Library, v. 19, n. 7, p. 732–739, 2004.

[81] CHEN, W.; JIANG, Z.; GUO, H.; NI, X. Fall detection based on key points of human-skeleton using openpose. *Symmetry*, MDPI, v. 12, n. 5, p. 744, 2020.

[82] Allan. *Algorítmo e detalhes de implementação*. 2023. `https://github.com/mestrelan/frothy-lilac/`. Accessed on May 23.

[83] DAI, J.; LI, Y.; HE, K.; SUN, J. *R-FCN: Object Detection via Region-based Fully Convolutional Networks*. 2016.

[84] LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; BOURDEV, L.; GIRSHICK, R.; HAYS, J.; PERONA, P.; RAMANAN, D.; ZITNICK, C. L.; DOLLáR, P. *Microsoft COCO: Common Objects in Context*. 2015.

[85] Learn OpenCV. *Intersection over Union for Object Detection*. 2022. `https://learnopencv.com/intersection-over-union-iou-in-object-detection`. Accessed on January 31.

[86] PNEVMATIKAKIS, A.; POLYMENAKOS, L. 2d person tracking using kalman filtering and adaptive background learning in a feedback loop. In: SPRINGER. *Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006, Southampton, UK, April 6-7, 2006, Revised Selected Papers 1*. [S.l.], 2007. p. 151–160.

[87] AGGARWAL, C. C. Neural networks and deep learning. *Springer*, Springer, v. 10, p. 978–3, 2018.

[88] KHAN, S. S.; NOGAS, J.; MIHAILIDIS, A. Spatio-temporal adversarial learning for detecting unseen falls. *Pattern Analysis and Applications*, Springer Science and Business Media LLC, v. 24, n. 1, p. 381–391, Jul 2020. ISSN 1433-755X.

[89] DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, v. 12, n. 7, 2011.

[90] HUH, M.; AGRAWAL, P.; EFROS, A. A. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[91] NOGAS, J.; KHAN, S.; MIHAILIDIS, A. *Fall Detection from Thermal Camera Using Convolutional LSTM Autoencoder*. EasyChair, 2019.

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

[93] VADIVELU, S.; GANESAN, S.; MURTHY, O.; DHALL, A. Thermal imaging based elderly fall detection. In: SPRINGER. *Asian conference on computer vision*. [S.l.], 2017. p. 541–553.

[94] SHAHROUDY, A.; LIU, J.; NG, T.-T.; WANG, G. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 1010–1019.

[95] PANNURAT, N.; THIEMJARUS, S.; NANTAJEEWARAWAT, E. Automatic fall monitoring: A review. *Sensors*, MDPI AG, v. 14, n. 7, p. 12900–12936, Jul 2014. ISSN 1424-8220.

[96] Allan. *Arquivo .CSV com os resultados dos modelos CVSCs*. 2023. `https://github.com/mestrelan/frothy-lilac/blob/main/metricas.csv`. Accessed on 2023.

[97] MAZUROWSKI, M. A.; HABAS, P. A.; ZURADA, J. M.; LO, J. Y.; BAKER, J. A.; TOURASSI, G. D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, v. 21, n. 2, p. 427–436, 2008. ISSN 0893-6080. Advances in Neural Networks Research: IJCNN â€™07. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608007002407>.

[98] MUSHTAQ, Z.; YAQUB, A.; HASSAN, A.; SU, S. F. Performance analysis of supervised classifiers using pca based techniques on breast cancer. In: *2019 International Conference on Engineering and Emerging Technologies (ICEET)*. [S.l.: s.n.], 2019. p. 1–6.

[99] GRANDINI, M.; BAGLI, E.; VISANI, G. *Metrics for Multi-Class Classification: an Overview*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2008.05756>.

[100] ERICKSON, B. J.; KITAMURA, F. Magician's corner: 9. performance metrics for machine learning models. *Radiology: Artificial Intelligence*, Radiological Society of North America, v. 3, n. 3, 2021.

[101] MEHRALIVAND, S.; YANG, D.; HARMON, S. A.; XU, D.; XU, Z.; ROTH, H.; MASOUDI, S.; SANFORD, T. H.; KESANI, D.; LAY, N. S.; MERINO, M. J.; WOOD, B. J.; PINTO, P. A.; CHOYKE, P. L.; TURKBEY, B. A cascaded deep learningâ€“based artificial intelligence algorithm for automated lesion detection and classification on biparametric prostate magnetic resonance imaging. *Academic Radiology*, v. 29, n. 8, p. 1159–1168, 2022. ISSN 1076-6332. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1076633221003779>.

[102] CUBUK, C.; GARRETT, A.; CHOI, S.; KING, L.; LOVEDAY, C.; TORR, B.; BURGHEL, G.; DURKIE, M.; CALLAWAY, A.; ROBINSON, R. Clinical likelihood ratios and balanced accuracy for 44 in silico tools against multiple large-scale functional assays of cancer susceptibility genes. *Genetics in Medicine*, Nature Publishing Group, v. 23, n. 11, p. 2096–2104, 2021.

[103] THARWAT, A. Classification assessment methods. *Applied Computing and Informatics*, Emerald Publishing Limited, 2020.

[104] FOWLKES, E. B.; MALLOWS, C. L. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, Taylor & Francis, v. 78, n. 383, p. 553–569, 1983.

[105] MCINROY, B.; FENG, W.; PAN, Y. An empirical study on performance measures for online advertising. In: *2015 12th International Conference on Information Technology - New Generations*. [S.l.: s.n.], 2015. p. 38–43.

[106] BERTELS, J.; EELBODE, T.; BERMAN, M.; VANDERMEULEN, D.; MAES, F.; BISSCHOPS, R.; BLASCHKO, M. B. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In: SPRINGER. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. [S.l.], 2019. p. 92–100.

[107] HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

[108] SINGLA, K.; BISWAS, S. Machine learning explanability method for the multi-label classification model. In: IEEE. *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. [S.l.], 2021. p. 337–340.

[109] Allan. *Google Drive link with the experiments animations*. 2023. `https://drive.google.com/drive/folders/1HcP7nzmm856qcjIREyKlFWHYc19xjX0C?usp=sharing`. Accessed on May 23.