Provendo um modelo automático de detecção de quedas baseado em rede adversária generativa para assistência de idosos

Allan Costa N. dos Santos, Flávio Luiz Seixas, Natalia Castro Fernandes

[1]Laboratório MídiaCom – Universidade Federal Fluminense (UFF)
Niterói, RJ - Brasil

{allans,fseixas,nataliacf}@id.uff.br

Abstract. *Falls are a serious public health problem and people over 65 are among the most vulnerable to serious injury from a fall. This article proposes and evaluates a model of neural network architecture, called CNN Video Stream Combination (CVSC), to improve the monitoring and safety of the elderly. It is proposed to use a generative neural network to calculate the anomaly score. A model capable of working simultaneously with RGB and infrared cameras is proposed, as falls of older people often occur in low-light environments. The CVSC using an RGB camera presented 97.00% of sensitivity and with an infrared camera it obtained 94.00% of sensitivity.*

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 uma queda. Este artigo propõe e avalia um modelo de arquitetura de rede neural, chamada de CNN Video Stream Combination (CVSC), para melhorar o monitoramento e a segurança de idosos. É proposto o uso de uma rede neural generativa para calcular a pontuação de anomalia. Propõe-se um modelo capaz de trabalhar simultaneamente com câmeras RGB e infravermelho, pois as quedas de idosos frequentemente ocorrem em ambientes com pouca luz. O CVSC utilizando câmera RGB apresentou 97,00% de sensibilidade e com câmera infravermelho obteve 94,00% de sensibilidade.*

1. Introdução

O número de idosos no país está aumentando e, até 2025, o Brasil será o 6º no mundo em quantidade de idosos [Belasco and Okuno 2019]. Isso indica 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% [Quigley et al. 2012].

Quedas são eventos muito graves para os idosos, pois são uma das principais causas de lesões, traumas, internações e óbitos. Além disso, esses eventos contribuem para o declínio funcional e diminuição da autonomia, com consequências diretas na qualidade de vida. Muitos idosos não gostam de ter sempre um acompanhante por perto e os dispositivos e sistemas manuais de alerta de quedas dependem que o idoso acione o alarme. Isso, contudo, nem sempre é possível, pois o idoso pode ter sofrido um acidente que tenha o feito desmaiar ou de alguma forma o impeça de acionar do alarme. Devido a potencial

gravidade da queda, é necessário que o evento seja detectado o quanto antes a fim de evitar riscos ainda piores a saúde do idoso.

Os sistemas de detecção de queda baseados em dispositivos vestíveis possuem a premissa que os idosos sempre estarão portanto os dispositivos, o que não é sempre verdade. Já os sistemas baseados em vídeo possuem limitações na detecção de queda diante de alterações no fundo da imagem, objetos de fundo, alterações de iluminação e movimento da câmera [Mehta et al. 2021]. Algumas abordagens não funcionam em ambientes pouco iluminados, por exemplo, quando o idoso está em seu quarto a noite e precisa se levantar para ir no banheiro ou tomar um remédio. Além disso, a baixa eficácia de alguns algoritmos de detecção de queda se devem às dificuldades de treinamento. Embora as quedas sejam frequentes, nem todos eventos são registrados por câmeras e, por isso, muitos *datasets* são desbalanceados. Com isso, algoritmos tradicionais se limitam a trabalhar com poucas amostras de treinamento. Um outro problema relevante é que muitos desses algoritmos não são próprios para detecção de quedas usando câmeras de infra-vermelho, não sendo capazes de atuar em ambientes escuros.

Este artigo propõe e avalia um modelo de arquitetura de rede neural, chamado de *CNN Video Stream Combination* (CVSC), capaz de usar técnicas de aprendizado de máquina para detecção de quedas de idosos desassistidos. Diferentemente de outras propostas da literatura, o CVSC detecta quedas em ambientes com muita iluminação, com câmeras RGB (Red Green Blue), ou sem iluminação, com câmeras com infravermelho (IR). Assim, o sistema utiliza câmeras domésticas estrategicamente posicionadas em áreas onde idosos com mobilidade reduzida correm maior risco de cair. Esse sistema ajuda a aumentar a independência e, consequentemente, a autoestima do usuário, uma vez que permite que idosos com mobilidade reduzida, mas sem níveis de demência significativa, possam evitar a necessidade de uma assistência 24x7 no seu dia a dia.

Dado a natureza da frequência dos eventos de queda, foi utilizada no CSCV a detecção de anomalias. A detecção de anomalias é um paradigma de classificação no qual o padrão de eventos normais é aprendido a partir dos desvios de distribuição. Esse paradigma é utilizado na tarefa de classificação mesmo em *datasets* desbalanceados. O algoritmo de rede neural convolucional (*Convolutional Neural Network* - CNN) foi usado para calcular o desvio de reconstrução. Quanto maior o erro de reconstrução, maior a chance da presença de uma anomalia e assim a queda é identificada, pois a queda pode ser identificada na imagem como uma mudança espaço-temporal na posição do indivíduo [Mehta et al. 2021]. Dado que há uma mudança repentina, essa mudança gerará um grande erro de reconstrução naquele momento. Com o *Optical Flow Computation* é possível calcular a variação dos *pixels*, e assim a movimentação de uma pessoa [Liu et al. 2018]. O filtro de *Kalman* é utilizado para que o sistema não perca o rastreamento do indivíduo devido as mudanças nos *pixels* provocadas pelas variações de iluminação, e assim, conseguir rastrear a pessoa através do fluxo de imagens de ambos os tipos. A técnica de rastreamento de região de interesse é adaptada para melhorar o desempenho do sistema diante de alterações no fundo da imagem e objetos de fundo. Com base no pré-processamento proposto, o CVSC possibilita utilizar imagens RGB e de infravermelho com a mesma arquitetura. O modelo de rede neural utilizando câmera RGB obteve maior sensibilidade quando comparado com outras técnicas que também utilizam câmera RGB. Assim, a técnica de pontuação de anomalias mostrou-se uma técnica de aprendi-

zado que se adapta bem as variações de iluminação, plano de fundo e movimentação do usuário, o que é ideal para o uso do sistema em situações reais.

Este artigo está organizado da seguinte forma. A Seção 2 descreve os trabalhos relacionados. A Seção 3 descreve a arquitetura proposta e a Seção 4 descreve os resultados obtidos. Por fim, as conclusões são apresentadas na Seção 5.

2. Trabalhos Relacionados

Detectar e responder rapidamente às quedas, principalmente quando o idoso está sozinho, é de suma importância para reduzir as consequências de uma queda. O primeiro sistema de detecção de quedas foi desenvolvido no início da década de 1970 e enviava mensagens de alerta quando um botão de emergência era pressionado [Pannurat et al. 2014]. Os sistemas atuais são muito mais sofisticados e capazes de atuar sem a necessidade de ações por parte da pessoa que caiu. Muitos sistemas de detecção de quedas para idosos são baseados na suposição de que o idoso estará carregando seu *smartphone*. Nesse sentido, sensores de *smartphones*, como acelerômetros, giroscópios e magnetômetros, são usados em algoritmos que detectam quando a queda aconteceu [Mshali et al. 2018]. Outras abordagens usam sensores vestíveis, usando uma lógica semelhante ao uso do telefone celular para detectar e notificar a queda [Ramachandran and Karuppiah 2020]. Uma crítica comum aos sistemas baseados em *smartphones* e/ou detecção de queda de sensores vestíveis em lares de idosos é que não é verdade que os idosos sempre carregam os dispositivos necessários. Outro grupo de propostas discute o monitoramento do ambiente para detecção de quedas, não dependendo do idoso portar qualquer dispositivo. Tais tipos de sistemas utilizam câmeras RGB, câmeras infravermelhas, sensores de videogame, como o *Microsoft Kinect*, sensores de pressão e vibração do piso, microfone, sensores de presença, entre outros [Shojaei-Hashemi et al. 2018, Pannurat et al. 2014].

Shojaei-Hashemi et al 2018 [Shojaei-Hashemi et al. 2018] propõem uma abordagem de aprendizado profundo para detectar quedas humanas, usando rede neural de memória de longo prazo. O modelo não se restringe a nenhuma circunstância específica. Ao contrário dos métodos clássicos cujo desempenho é limitado pelas condições assumidas, a rede neural profunda cuida do próprio padrão de recursos e extrai os recursos mais discriminativos com base nos dados de treinamento e, assim, abrange mais cenários da vida real. Para superar o requisito inerente de abordagens de aprendizado profundo para grandes conjuntos de dados de treinamento, o aprendizado de transferência foi empregado.

Liu et al [Liu et al. 2018] usam restrição de fluxo óptico e imagens reconstruídas para prever quadros futuros. Eles usam um modelo baseado em CNN para estimativa de fluxo, que pode facilitar a retropropagação para perda de fluxo óptico, e realizam o treinamento com imagens RGB. Seguindo essa ideia, é construída uma rede espaço-temporal para reconstrução de fluxo, que recebe como entrada uma janela de quadros do fluxo óptico. Os quadros de fluxo óptico denso são calculados para dois quadros consecutivos. O fluxo é combinado na direção x, y e magnitude para formar uma imagem tridimensional. As imagens de fluxo são delimitadas com sua Região de Interesse (Region Of Interest - ROI) e, para o quadro de diferença, a ROI para imagem de fluxo é a união da ROI dos dois quadros utilizados para calcular o fluxo óptico.

Chen et al[Chen et al. 2021] desenvolvem uma abordagem de detecção de queda

baseada em vídeo usando poses humanas. Primeiro, um estimador de pose leve extrai poses 2D de sequências de vídeo e, em seguida, poses 2D são levantadas para poses 3D. Em segundo lugar, foi introduzida uma rede robusta de detecção de queda que inclui rede convolucional para reconhecer eventos de queda usando poses 3D estimadas, o que aumenta o respectivo campo e mantém baixo custo computacional por convoluções dilatadas.

Xu et al [Xu et al. 2020] desenvolvem uma CNN para detecção de queda através da formação de mapa corporal ósseo 2D da pessoa identificada nas gravações de câmeras RGB. Eles usaram o OPENPOSE para converter a imagem na imagem de esqueleto correspondente. Em seguida, usando a transferência de aprendizado, foi utilizado o conjunto de dados para treinar um novo modelo de detecção de queda. Tanto o trabalho de Chen et al como o de Xu et al foram utilizados para comparação com o modelo proposto nesta pesquisa por serem modelos de CNNs para detectar quedas em específico através de gravações RGB.

Mehta et al.[Mehta et al. 2021] propuseram o 3D Convolutional Autoencoder (3DCAE) para detecção de eventos anormais aplicados 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, o que pode reduzir a taxa de falsos positivos.

Na pesquisa reportada neste artigo, é realizada uma expansão do trabalho de Mehta et al [Mehta et al. 2021] no *ROI Masking* e no pré-processamento para permitir o uso de câmeras RGB e infravermelho, ao invés do uso de câmeras térmicas. Assume-se um cenário, diferentemente dos trabalhos anteriores, aonde falsos-negativos não são toleráveis e no qual as imagens das câmeras podem variar substancialmente com relação às imagens usadas para treino, já que se visa o uso do sistema em casas de repouso com câmeras de monitoramento padrão.

3. CNN *video stream combination* (CVSC)

O CVSC, proposto nesta pesquisa, foi desenvolvido como uma expansão de [Mehta et al. 2021] para dar suporte a idosos, notificando com alta sensibilidade quedas. Visando cenários aonde o idoso tem certa independência, o sistema deve ser capaz de atuar tanto em ambientes iluminados quanto escuros, alarmando quedas sem a necessidade de monitoramento humano. 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. A Figura 1 é a visão geral da arquitetura CVSC, que propõe novos mecanismos de processamento de imagem e *ROI Masking*, estendendo [Mehta et al. 2021] para dar suporte a câmeras RGB e infravermelho. O algorítmo CNN foi usado para calcular o desvio de reconstrução conforme mostrado na Figura 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. Na sequência, são executados dois fluxos em paralelo, como em [Mehta et al. 2021]: um aprendendo explicitamente o movimento na forma de um fluxo óptico e outro recebendo quadros de vídeo brutos como entrada. Para permitir o funcionamento com câmeras RGB e infravermelho, foi necessário propor novos modelos para o pré-processamento e para o rastreamento de indivíduos.

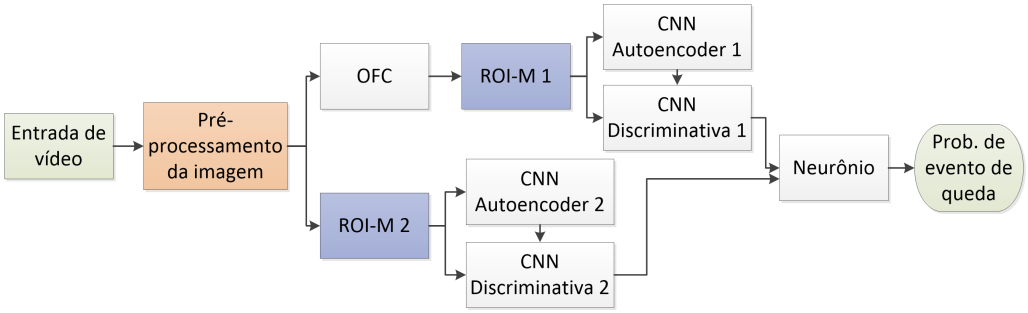

Figura 1. Visão geral da arquitetura CVSC, que propõe novos mecanismos de processamento de imagem e *ROI Masking*, estendendo [Mehta et al. 2021] para dar suporte a câmeras RGB e infravermelho .

3.1. Pré-processamento da imagem

Durante a etapa de pré processamento, a escala da imagem foi redimensionada para 640x480. A taxa de quadros foi atrasada em uma taxa de 1:3 ao longo do vídeo original e em uma taxa de 1:6 no evento de queda. Isso ajuda a estabilizar a pontuação de anomalia por quadro, fazendo com que ela mude menos abruptamente. Todos os quadros do vídeo são divididos em janelas de 8 quadros, utilizando o método de janela deslizante com passo (step) 1.

Em seguida, definiu-se uma filtragem para melhorar a detecção das pessoas nas imagens. Para tanto, com as funções *cv2.dilate e cv2.morphologyEx* da biblioteca *Opencv*, foram realizados testes de filtragem por dilatação, abertura e fechamento [OpenCV 2022]. De acordo com esses testes, o filtro que produziu o melhor resultado foi o *Kalman*, embora seja possível utilizar uma combinação dos filtros anteriores para realçar a diferença na tonalidade. O filtro de *Kalman* também contribui para o rastreamento do idoso e é aplicado nas coordenadas superior esquerda e inferior direita da caixa delimitadora com a velocidade constante.

3.2. *ROI Masking*

O rastreamento de pessoas é realizado para extrair a região de interesse (*Region of Interest* - ROI) dos quadros originais e fluxo óptico para reconstrução baseada em movimento e região. O método baseado em ROI melhora a qualidade do rastreamento, pois o modelo aprende a reconstruir apenas a região de interesse onde a pessoa está. O rastreamento de pessoas é realizado usando a Rede Totalmente Convolucional Baseada em Região (*Region-based Fully Convolutional Networks* - R-FCN) [Dai et al. 2016], treinada no conjunto de dados *Common Objects in Context* (COCO) [Lin et al. 2015]. Essa técnica é bastante eficiente, mas funciona apenas quando há apenas uma pessoa no vídeo. Essa

restrição não é um problema para o CVSC, pois a detecção de queda precisa é necessária quando o idoso está só.

Os blocos *ROI Masking*, ROI-M 1 e 2, da Figura 1 são responsáveis pelo rastreamento de pessoas. Esse rastreamento define caixas delimitadoras, as quais são comparadas entre quadros. A caixa prevista para o quadro atual e a caixa do próximo quadro (se a pessoa for detectada no próximo quadro) são comparadas para verificar se o rastreamento foi perdido. Um contador é necessário para rastrear o número de previsões contínuas do rastreador sem detecção. Quando nenhuma detecção ocorre, o contador é incrementado e quando ultrapassa um limite de 20, o rastreador congela.

Intersection over Union (IoU)[OpenCV 2022] combina as caixas delimitadoras. A IoU é pequena quando o tamanho de uma caixa é grande em comparação com a outra caixa do quadro anterior, o que ocorre quando o detector localiza erroneamente a caixa.Assim, a razão de áreas entre quadros também é utilizada como critério de rastreamento no CVSC. Essa razão pode encontrar os contornos na imagem resultante e selecionar o maior contorno com base na área interna. A menor caixa que contém essa região de contorno é escolhida como candidata à caixa delimitadora da pessoa.

Para melhorar a localização da caixa de contorno, o limite Otsu [OpenCV 2022] é aplicado à imagem para separar o fundo escuro, útil principalmente nas gravações em IR. A imagem resultante ainda pode conter objetos com fundos claros. Para melhorar essa detecção, foram realizados testes com os algoritmos de *Background Subtractor cv2.bgsegm.createBackgroundSubtractorGMG*, *cv2.bgsegm.createBackgroundSubtractorMOG*, *cv2.createBackgroundSubtractorMOG2*, *cv2.createBackgroundSubtractorKNN* e *cv2.bgsegm.createBackgroundSubtractorCNT* [OpenCV 2022]. Eles conseguem melhorar o desempenho do método de detecção diante de alterações no fundo da imagem, objetos de fundo, alterações de iluminação e movimento da câmera. Contudo, caso a pessoa fique parada em uma posição por mais de 10 minutos, por exemplo, os métodos causam falhas no rastreamento do idoso. Portanto, esses métodos são mais recomentados para câmeras em corredores e cozinhas, por exemplo, mas não em quartos e sala de estar, aplicações mais típicas do CVSC.

Na implementação do ROI Masking do CVSC, é utilizado o *checkpoint* da rede *rfcn_resnet101_coco* pré-treinado do modelo de detecção de *tensorflow zoo* para realizar a detecção de pessoas e o filtro *Kalman* para realizar o rastreamento de pessoas. Após a detecção da pessoa, o código de rastreamento de pessoa filtra a pessoa detectada em cada quadro com o filtro *Kalman*. Como o desempenho dos métodos de detecção de queda baseados em vídeo pode ser afetado pelo plano de fundo da imagem e pelos objetos no plano de fundo, apenas a região onde a pessoa está presente foi reconstruída. Portanto, esta região é menos afetada por mudanças nos objetos de fundo e intensidade.

3.3. Fluxo Óptico (OFC)

O fluxo óptico [OpenCV 2022] é o padrão de movimento aparente de objetos da imagem entre quadros consecutivos causados 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 dos *pixels* 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 ambi-

entais que se movem em relação ao plano da imagem da câmera. Tanto a câmera 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 da imagem, assumindo que as mudanças nas intensidades espaço-temporais na sequência são devidas ao movimento relativo da câmera e do ponto ambiente. O fluxo óptico é usado no CVSC para realizar tanto a detecção de movimento quanto a segmentação de objetos.

O movimento e a estrutura 3D podem ser inferidos a partir de campos de velocidade 2D ou campos de deslocamento 2D ou diretamente de derivados de intensidade. O método *Lucas-Kanade* [OpenCV 2022] é uma das abordagens para resolver a equação do fluxo óptico. Assumindo que todos os *pixels* vizinhos terão movimento semelhante, o método de *Lucas-Kanade* considera uma matriz 3x3 ao redor do ponto. A partir de alguns pontos da trilha, são calculados os vetores de fluxo óptico desses pontos. Para pequenos movimentos este método funciona, porém, falha quando há uma grande região de movimento. Para resolver o problema de escala é possível usar o método *pyramids* [OpenCV 2022]. Com esse método, movimentos menores são removidos e movimentos maiores tornam-se pequenos movimentos.

3.4. Arquitetura da rede convolucional

A arquitetura da rede convolucional, mostrada na Figura 1, foi adaptada de [Mehta et al. 2021] e consiste em uma rede adversária generativa. 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 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. 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. Portanto, essa estrutura é capaz de 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 conforme a equação em [Khan et al. 2020].

No segundo caminho, é calculado o fluxo óptico (representado na Figura 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 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 1, recebe tanto

o quadro reconstruído quanto o quadro original como entrada. Ambas as CNN Discriminativas tem arquitetura idêntica às 4 primeiras camadas da CNN *Autoencoder* correspondente. A decodificação opera como codificação, mas ao contrário, usando camadas de convolução 3D. A camada de deconvolução final combina mapas de recursos (*feature maps*) na reconstrução decodificada. Esta camada final usa um passo de $1 \times 1 \times 1$ e preenchimento. Ambas as redes convolucionais discriminativas são unidas por um único neurônio e o resultado é uma probabilidade de evento de queda.

4. Avaliação do Modelo Proposto

4.1. Conjuntos de dados

Os conjuntos de dados usados para treinar e avaliar o modelo foram gravações de câmeras de simulações de quedas e atividade da vida diária (*Activities of Daily Life* - ADL). Dentre os conjuntos de dados disponíveis para tarefas de reconhecimento de atividades humanas (*Human Activity Recognition* - HAR), o *Thermal Simulated Fall* (TSF) [Mehta et al. 2021] foi usado para treinar o modelo de Mehta et al. Este conjunto de dados contém 44 vídeos térmicos de resolução 640x480, entre os quais 9 vídeos com ADL normal e 35 vídeos contendo quedas. 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. Os vídeos ADL contêm um total de 22.116 quadros. Por meio do aprendizado de transferência, o modelo usa os pesos de treinamento da rede *rfcn_resnet101_coco* [Dai et al. 2016] do modelo de detecção de *tensorflow zoo* sobre o conjunto de dados COCO [Lin et al. 2015]. Ele está disponibilizado abertamente e é um conjunto de dados de detecção de objetos, segmentação, reconhecimento no contexto e rotulagem. Tem mais de 200.000 imagens rotuladas, 80 categorias de objetos e 250.000 pessoas. Portanto, o *dataset* COCO é utilizado para treinar a rede para classificar objetos em geral, inclusive detectar pessoas, enquanto que o *dataset* TSF foi utilizado para treinar a rede de forma específica para a detecção de queda.

Para avaliar o modelo proposto nesta pesquisa, foi utilizado o *Dataset "NTU RGB+D"* [Shahroudy et al. 2016] 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. A classificação da ação é dividida em 3 grupos. Um grupo é de ADL, como ler, beber, ouvir músicas, etc. Um grupo é de movimentos relacionados à saúde, como cair, espirrar, etc., e os outros grupos são interações, como pegar coisas, bater, etc. Contém gravações de 2 segundos a 4 segundos usando câmeras de profundidade (resolução 512x424), RGB (resolução 1920x1080) e infravermelho (resolução 512x424). Cada conjunto de dados é capturado por três câmeras *Kinect V2*. Para avaliar o sistema foram utilizados um total de 108 vídeos com ocorrência de quedas em RGB e 100 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.

4.2. Resultados

4.2.1. Rastreamento de usuários

Entre os vídeos RGB, não houve falhas quanto ao rastreamento do indivíduo, ou seja, em todos os testes com a câmera RGB, o código conseguiu rastrear a pessoa. Entre os vídeos em infravermelho (IR), houve 2 falhas nos vídeos com ocorrência de queda, ou seja, nesses testes reprovados, o algoritmo não conseguiu identificar o indivíduo. Assim, em uma câmera IR, o sistema não rastreou o indivíduo em 2% dos testes.

4.2.2. Detecção de queda

A Figura 2 mostra o erro de reconstrução por quadro calculando a média e o desvio padrão em um vídeo com a ocorrência de uma queda. O cálculo da pontuação de anomalia é feito através do desvio padrão ou média do erro entre o quadro de entrada e o quadro reconstruído como em [Nogas et al. 2019]. Cada quadro representa a progressão no vídeo. Um alto erro de reconstrução representa um aumento na pontuação da anomalia, o que indica uma maior chance de ocorrer uma queda. Ambos os gráficos mostram que o sistema foi capaz de identificar corretamente a queda aumentando o erro de reconstrução (um pico acima de 0,0175 para o gráfico da média e um pico acima de 0,0050 para o gráfico do desvio padrão).

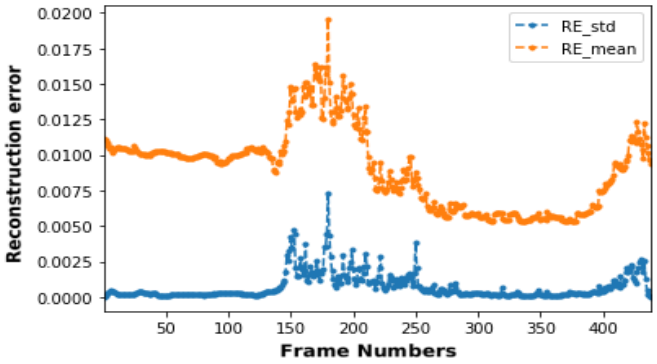

Figura 2. Pontuação de anomalia por quadro para um vídeo de queda

Tabela 1. Resultados do CSCV e comparação com outros modelos, utilizando o mesmo conjunto de dados para teste.

Método	Modalidade	Acurácia	Sensibilidade	Precisão	F1-Score
[Chen et al. 2021]	RGB	99,83%	94,25%	98,73%	96,44%
[Xu et al. 2020]	RGB	91,70%	—	—	—
CVSC	**RGB**	**93,80%**	**97,00%**	**96,00%**	**96,50%**
CVSC	**IR**	**90,48%**	**94,00%**	**96,00%**	**94,99%**

Considerando as amostras em que o sistema conseguiu identificar corretamente o usuário, a Tabela 1 lista as métricas de avaliação do modelo de detecção de quedas mostrando o desempenho de outras propostas de detecção de quedas por meio de visão

computacional, usando o mesmo conjunto de dados de reconhecimento de ação NTU RGB+D. A Tabela 1 mostra os resultados de acurácia, sensibilidade, precisão e *F1-Score* para avaliação com o modelo proposto CVSC e outros modelos de CNN da literatura que também utilizaram gravações RGB como modalidade de detecção de quedas. O trabalho de Chen et al utilizou o COCO *dataset* para treinamento e o NTU RGB+D para ajuste fino e avaliação. O trabalho de Xu et al utilizou o NTU RGB+D e mais outro dois *datasets* para avaliação e treinamento.

Cabe destacar que, pela natureza do problema, falsos positivos, cujo impacto pode ser avaliado pela acurácia, se em pequeno número, são tolerados, mas falsos negativos devem ser evitados a todo custo, o que pode ser avaliado pelas métricas sensibilidade e *F1-Score*. Um falso negativo pode deixar um idoso que caiu desassistido, sendo, portanto, inaceitável. O modelo CVSC supera os modelos da modalidade RGB encontrados na literatura, em especial com relação a sensibilidade.

A sensibilidade considera a relação entre as quedas corretamente classificadas e o total de eventos de quedas. Para um sistema de supervisão de idosos, a sensibilidade se mostra como principal métrica e, idealmente, deve valer 100%. Seguindo esse mesmo objetivo, avalia-se o *F1-Score*, porque essa medida é a média harmônica das métricas de precisão e sensibilidade. A pontuação F1 pode ser uma melhor medida a ser usada quando houver uma distribuição de classes desigual (grande número de ADL), o qual é o caso do problema de detecção de queda, visto que existem mais atividades diárias do que quedas. Cabe destacar que, além de apresentar a melhor sensibilidade e melhor F1-Score, o CVSC é o único modelo capaz de atender a mais de uma modalidade de detecção de queda ao mesmo tempo.

Durante a realização dos experimentos de queda, foram geradas animações mostrando o desempenho do modelo através da pontuação da anomalia por quadro utilizando a média sobre o vídeo em tempo real. A animação mostra a sequência de quadros reconstruídos, do lado esquerdo, os quadros originais e, na parte inferior, o gráfico da pontuação da anomalia por quadro. Esses experimentos estão disponíveis no Google Drive (link) e o código desenvolvido no *GitHub*: ***https://github.com/mestrelan/NEWgreen-SBCAS***.

5. Conclusões

Este artigo propõe e avalia um modelo de arquitetura de rede neural usando técnicas de aprendizado de máquina para melhorar o bem-estar e a segurança de indivíduos em risco de queda, como idosos ou deficientes. Ao analisar as variações de movimentos, o CVSC pode determinar o estado atual da pessoa e enviá-lo ao sistema. Por meio desse método, é possível alcançar um sistema de atenção à saúde que permite maior independência e qualidade de vida para idosos.

O CVSC é o único sistema único capaz de detectar quedas em duas modalidades diferentes de câmera ao mesmo tempo com baixa taxa de falsos negativos, tornando a proposta mais versátil as variações de iluminação. O CVSC é uma nova abordagem que consiste em uma combinação única de algumas das mais recentes técnicas de visão computacional com processamento de imagens, *Deep Learning*, bibliotecas e algoritmos na linguagem de programação *python*.

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 (93,80%), 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, obtendo 96,50% de *F1 Score* para câmeras RGB. O modelo utilizando câmera infravermelha obteve 90,48% de acurácia e 94,99% de *F1 Score*. Esses valores são altos e dão grande fidedignidade ao sistema para operação durante a noite, aonde as quedas de idosos acabam sendo muito frequentes.

As métricas comprovam que o sistema possui uma alta taxa de reconhecimento de quedas e maior pontuação F1 quando comparada ao desempenho de outras propostas de detecção de quedas por meio de visão computacional. A técnica de pontuação de anomalias mostrou-se uma técnica de aprendizado que possui o melhor equilíbrio entre precisão e sensibilidade através da avaliação do *F1 Score*. Essa técnica é capaz de identificar a queda mesmo com a exposição de novos vídeos com variações significativas com relação aos vídeos que foram utilizados para treinamento, o que é ideal para o uso do sistema em situações reais.

Com base nesses resultados, pode-se afirmar que CVSC possui grande potencial para ser implementado em um sistema interno de vigilância e monitoramento por câmeras para monitorar as pessoas de uma casa geriátrica, por exemplo. As limitações de iluminação de modelos anteriores de detecção de queda foram bem solucionadas nesta proposta com a inclusão de infravermelho. Contudo, a arquitetura proposta 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.

5.1. Trabalhos Futuros

Como trabalho futuro, pretende-se adicionar mais camadas de ROI e CNN para que o sistema possa trabalhar com vídeos onde haja a presença de mais de uma pessoa, representando quartos compartilhados por idosos, por exemplo. Pretende-se também realizar testes e treinos com mais amostras de vídeos com a ocorrência de quedas e ADL. Todos os treinos e testes foram realizados em vídeos com a presença de um único indivíduo.

Outra possibilidade é desenvolver mais um bloco de*Deep Learning* na arquitetura para treinar um modelo que possa melhorar o desempenho do método de detecção de queda de vídeo diante de alterações no fundo da imagem, objetos de fundo, alterações de iluminação e movimento da câmera. Pretende-se melhorar o modelo de *Background Subtractor* para que possa ser utilizado em câmeras de diferentes ambientes da casa geriátrica sem perder o rastreamento. Além disso, pretende-se montar um *dataset* real com imagens de uma casa de repouso, para ajustes mais precisos do sistema com base em testes com cenas de quedas mais reais em cenários de atenção ao idoso.

Referências

[Belasco and Okuno 2019] Belasco, A. G. S. and Okuno, M. F. P. (2019). Reality and challenges of ageing.

[Chen et al. 2021] Chen, Z., Wang, Y., and Yang, W. (2021). Video based fall detection using human poses.

[Dai et al. 2016] Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks.

[Khan et al. 2020] Khan, S. S., Nogas, J., and Mihailidis, A. (2020). Spatio-temporal adversarial learning for detecting unseen falls. *Pattern Analysis and Applications*, 24(1):381–391.

[Lin et al. 2015] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

[Liu et al. 2018] Liu, W., Luo, W., Lian, D., and Gao, S. (2018). Future frame prediction for anomaly detection - a new baseline. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6536–6545.

[Mehta et al. 2021] Mehta, V., Dhall, A., Pal, S., and Khan, S. S. (2021). Motion and region aware adversarial learning for fall detection with thermal imaging. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6321–6328.

[Mshali et al. 2018] Mshali, H., Lemlouma, T., and Magoni, D. (2018). Adaptive monitoring system for e-health smart homes. *Pervasive and Mobile Computing*, 43:1 – 19.

[Nogas et al. 2019] Nogas, J., Khan, S., and Mihailidis, A. (EasyChair, 2019). Fall detection from thermal camera using convolutional lstm autoencoder. EasyChair Preprint no. 824.

[OpenCV 2022] OpenCV (2022). OpenCV modules. https://docs.opencv.org/4.x/. Acessado em: 31 de Janeiro de 2022.

[Pannurat et al. 2014] Pannurat, N., Thiemjarus, S., and Nantajeewarawat, E. (2014). Automatic fall monitoring: A review. *Sensors*, 14(7):12900–12936.

[Quigley et al. 2012] Quigley, P. A., Campbell, R. R., Bulat, T., Olney, R. L., Buerhaus, P., and Needleman, J. (2012). Incidence and cost of serious fall-related injuries in nursing homes. *Clinical nursing research*, 21(1):10–23.

[Ramachandran and Karuppiah 2020] Ramachandran, A. and Karuppiah, A. (2020). A survey on recent advances in wearable fall detection systems. *BioMed Research International*, 2020.

[Shahroudy et al. 2016] Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019.

[Shojaei-Hashemi et al. 2018] Shojaei-Hashemi, A., Nasiopoulos, P., Little, J. J., and Pourazad, M. T. (2018). Video-based human fall detection in smart homes using deep learning. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5.

[Xu et al. 2020] Xu, Q., Huang, G., Yu, M., and Guo, Y. (2020). Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications*, 540(C).