REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

ANÁLISE MULTIMODAL PARA DETECÇÃO DE ATIVIDADES HUMANAS EM AMBIENTE DOMÉSTICO UTILIZANDO VISÃO COMPUTACIONAL E MODELOS DE APRENDIZADO PROFUNDO

MULTIMODAL ANALYSIS FOR HUMAN ACTIVITY DETECTION IN DOMESTIC ENVIRONMENTS USING COMPUTER VISION AND DEEP LEARNING MODELS

Ciências Exatas e da Terra, Engenharias • 06/04/2026

REGISTRO DOI: 10.70773/revistatopicos/775422980

Luiz Felipe Barbosa de Freitas[1]

Jean Mark Lobo de Oliveira[2]

Pablo Augusto da Paz Elleres[3]

Cleonor Crescencio das Neves[4]

RESUMO

A detecção de atividades humanas em ambientes domésticos baseada em análise multimodal constitui uma abordagem avançada no contexto da visão computacional e do aprendizado profundo, sendo direcionada à interpretação de padrões comportamentais a partir da integração de múltiplas fontes de dados, considerando a complexidade e a variabilidade desses cenários onde fatores como iluminação oclusões e múltiplas interações influenciam diretamente na interpretação das ações, a metodologia foi estruturada em etapas que envolvem coleta de dados a partir de vídeos RGB mapas de profundidade e sensores inerciais seguida de pré-processamento sincronização e extração de características, posteriormente as informações são integradas por meio de fusão intermediária com apoio de mecanismos de atenção e modeladas utilizando arquiteturas como CNNs LSTM e transformers, os resultados indicam que a abordagem proposta apresenta melhor desempenho em comparação a métodos unimodais especialmente em atividades mais simples e estruturadas enquanto cenários mais complexos ainda representam desafios, a análise temporal mostrou que o modelo mantém estabilidade ao longo das sequências e a avaliação com ruído evidenciou que a multimodalidade contribui para maior robustez permitindo que o sistema se mantenha funcional mesmo em condições adversas, no geral a proposta demonstra que a integração de diferentes fontes de dados associada a uma modelagem adequada possibilita uma interpretação mais consistente e próxima do comportamento real em ambientes domésticos.

Palavras-chave: análise multimodal, reconhecimento de atividades humanas, visão computacional, aprendizado profundo, sensores inerciais, ambiente doméstico.

ABSTRACT

Human activity detection in domestic environments based on multimodal analysis represents an advanced approach within the fields of computer vision and deep learning, aimed at interpreting behavioral patterns through the integration of multiple data sources. This approach considers the complexity and variability of such scenarios, where factors such as lighting, occlusions, and multiple interactions directly influence action interpretation. The methodology was structured in stages, including data collection from RGB videos, depth maps, and inertial sensors, followed by preprocessing, synchronization, and feature extraction. Subsequently, the information is integrated through intermediate fusion supported by attention mechanisms and modeled using architectures such as CNNs, LSTM networks, and transformers. The results indicate that the proposed approach outperforms unimodal methods, especially in simpler and more structured activities, while more complex scenarios still present challenges. Temporal analysis showed that the model maintains stability throughout sequences, and noise evaluation demonstrated that multimodality contributes to greater robustness, allowing the system to remain functional even under adverse conditions. Overall, the proposal demonstrates that integrating different data sources combined with appropriate modeling enables a more consistent and realistic interpretation of human behavior in domestic environments.

Keywords: multimodal analysis, human activity recognition, computer vision, deep learning, inertial sensors, domestic environment.

1. INTRODUÇÃO

A análise multimodal aplicada à identificação de atividades humanas em ambientes domésticos tem sido cada vez mais explorada principalmente pela necessidade de compreender comportamentos em situações do cotidiano, onde nem tudo segue um padrão definido, diferente de ambientes controlados a casa apresenta variações constantes como mudanças de iluminação ao longo do dia objetos fora do lugar e pessoas interagindo ao mesmo tempo o que acaba dificultando a interpretação automática, nesse cenário os sistemas de Human Activity Recognition (HAR) deixam de depender de uma única fonte de informação e passam a combinar dados de vídeo profundidade e sensores buscando uma leitura mais completa das ações, com o avanço do aprendizado profundo modelos como CNNs LSTMs e estruturas baseadas em transformers passaram a lidar melhor com essas variações conseguindo capturar padrões mesmo quando o ambiente apresenta ruídos ou mudanças inesperadas

Esse tipo de abordagem se torna relevante porque métodos baseados em apenas uma fonte de dado tendem a falhar quando enfrentam situações mais complexas. Em ambientes domésticos, é comum que parte da cena fique encoberta, que diferentes ações ocorram ao mesmo tempo ou que existam interferências que prejudiquem a análise. Ao combinar diferentes modalidades, essas limitações são parcialmente compensadas, tornando o sistema mais estável e confiável. Trabalhos recentes mostram que essa integração contribui diretamente para a melhoria dos resultados em tarefas de reconhecimento (Shin et al., 2025; Hossen e Abas, 2025), enquanto o uso de mecanismos de atenção ajuda a identificar relações mais sutis entre os dados ao longo do tempo (Xu et al., 2025). Isso faz com que essas soluções sejam cada vez mais utilizadas em aplicações

práticas, como monitoramento residencial, apoio a pessoas idosas e automação de ambientes

A proposta considera a utilização combinada de diferentes fontes de dados em conjunto com modelos de aprendizado profundo, direcionando a análise para cenários domésticos onde a variabilidade é um fator constante, técnicas de fusão são aplicadas de forma adaptativa para organizar as informações provenientes das múltiplas entradas, enquanto mecanismos de atenção contribuem para destacar padrões relevantes ao longo das sequências, a análise de desempenho não se limita a um único cenário controlado, sendo conduzida em condições diversas onde ruídos variações ambientais e comportamentos distintos influenciam diretamente os resultados obtidos, permitindo observar como o modelo reage a situações menos previsíveis e mais próximas do uso real.

2. FUNDAMENTAÇÃO TEÓRICA

A análise multimodal no reconhecimento de atividades humanas vem sendo tratada hoje como uma forma mais próxima da realidade de entender o comportamento das pessoas, principalmente quando se observa que os métodos mais antigos não conseguiam lidar bem com situações do dia a dia, então em vez de depender de um único tipo de dado os sistemas começam a trabalhar com várias fontes ao mesmo tempo como vídeo sensores e informações de contexto tentando montar uma visão mais completa do que realmente está acontecendo, o Human Activity Recognition (HAR) entra justamente nesse cenário como uma tentativa de identificar ações mas também de lidar com mudanças constantes no ambiente e no comportamento das pessoas, algo que fica ainda mais evidente dentro de casas onde tudo muda o tempo todo iluminação objetos

posição das coisas presença de pessoas, por isso a multimodalidade acaba se destacando porque junta diferentes perspectivas de uma mesma atividade reduzindo incertezas e deixando a interpretação mais confiável (Wang et al., 2022; Feichtenhofer et al., 2020).

2.1. Visão Computacional Aplicada Ao HAR

A forma como a visão computacional é usada no reconhecimento de atividades mudou bastante com o tempo, no início tudo dependia de definir características na mão e isso era trabalhoso além de não funcionar bem quando o cenário ficava mais complicado, com a chegada das redes neurais profundas esse processo ficou mais simples porque os próprios modelos passaram a aprender o que realmente importa nos dados, hoje as redes convolucionais ainda são muito usadas para analisar imagens mas o avanço não parou aí, começaram a surgir modelos que também entendem o tempo como as redes 3D e outras combinações que ajudam a analisar vídeos de forma mais completa, isso facilitou bastante tarefas como reconhecer ações em tempo real principalmente quando se trabalha com muitos dados para treinar, trabalhos como os de Feichtenhofer et al. (2020) mostram que dá para ter um bom desempenho sem gastar tanto processamento enquanto Wang et al. (2022) mostram que quanto mais rica for a informação visual melhor tende a ser o resultado principalmente em ambientes mais bagunçados ou imprevisíveis.

2.2. Aprendizado Profundo no Reconhecimento de Atividades

Nos últimos anos o aprendizado profundo deixou de ser só uma alternativa e virou praticamente o padrão quando se fala em reconhecimento de atividades, muito porque esses modelos

conseguem aprender direto dos dados sem depender tanto de regras definidas antes, o que facilita bastante quando o cenário muda ou fica mais complexo. Com o passar do tempo foram surgindo várias arquiteturas para atender diferentes necessidades, redes como LSTM e GRU ainda fazem sentido quando a sequência das ações importa principalmente em dados mais longos, enquanto modelos com atenção começaram a aparecer como uma forma mais flexível de lidar com essas relações no tempo sem precisar seguir tudo em ordem, nesse ponto os transformers chamam atenção porque conseguem captar relações mais amplas entre os dados, trabalhos como o de Dosovitskiy et al. (2021) mostram como os Vision Transformers mudaram a forma de analisar imagens e Bertasius et al. (2021) mostram avanços importantes na análise de vídeos usando atenção no espaço e no tempo, no fim isso tudo levou a modelos mais adaptáveis que conseguem funcionar melhor em diferentes situações e níveis de complexidade

2.3. Abordagens Multimodais

A ideia de usar múltiplas fontes de dados não surge só como melhoria, mas quase como uma necessidade quando se percebe que uma única fonte dificilmente consegue representar tudo que acontece em uma atividade real, principalmente em cenários onde há ruído, perda de informação ou situações inesperadas, nesses casos confiar em apenas um tipo de dado pode levar a erros ou interpretações incompletas, por isso quando diferentes modalidades são combinadas o sistema passa a ter uma visão mais ampla e consegue tomar decisões com mais segurança reduzindo ambiguidades ao longo do processo. Essa integração não segue um único padrão, ela pode acontecer logo no início juntando os dados brutos ou em etapas mais avançadas quando o modelo já extraiu

características ou até na fase final de decisão, tudo depende do tipo de informação disponível e do que se pretende alcançar com o sistema, Baltrušaitis et al. (2020) mostram que a forma como essa fusão é feita impacta diretamente no desempenho enquanto Zadeh et al. (2021) destacam que modelos multimodais conseguem aprender relações mais complexas entre diferentes tipos de dados, no fim quanto melhor for essa combinação maior tende a ser a capacidade do sistema de lidar com dados incompletos ou com bastante interferência.

2.4. Desafios e Aplicações em Ambientes Domésticos

Essas soluções saem do ambiente controlado e vão para dentro de casas a situação muda bastante, não existe padrão fixo tudo varia o tempo inteiro desde a iluminação até a forma como as pessoas se movimentam e interagem com o espaço, isso faz com que os dados coletados também mudem constantemente o que acaba impactando diretamente o desempenho dos modelos e dificulta manter um nível estável de precisão.

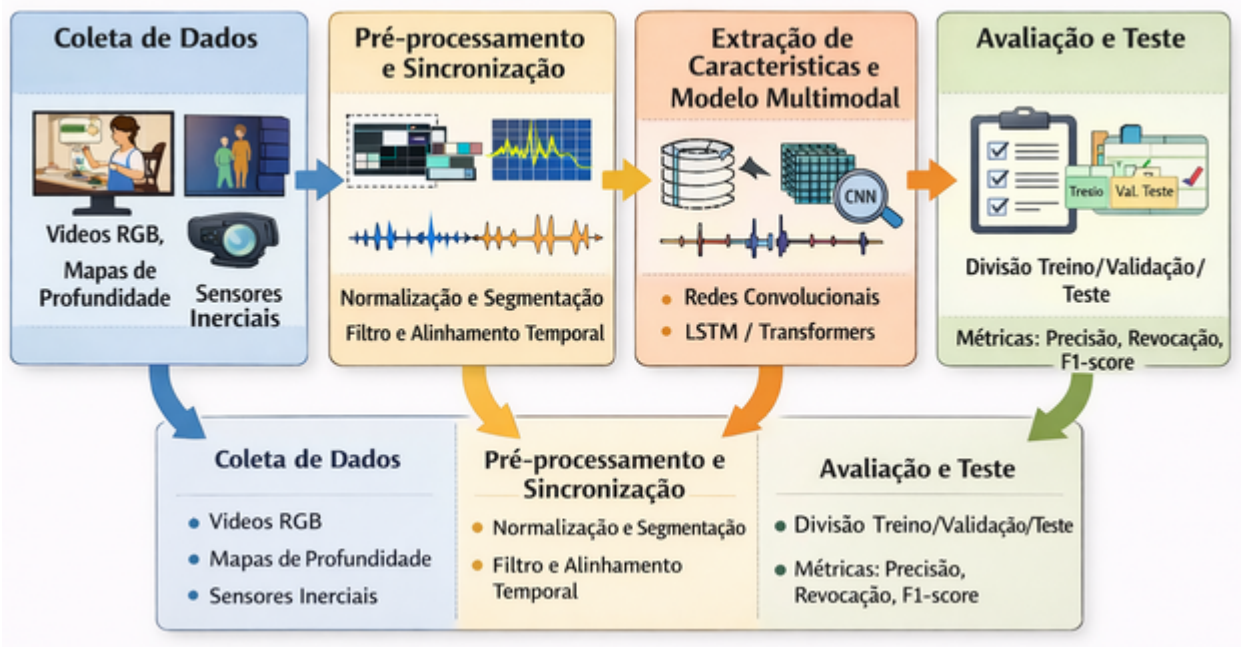
Outro ponto que pesa bastante é a questão da privacidade já que estamos falando de uso de câmeras e sensores em ambientes pessoais então não basta só fazer o sistema funcionar bem tecnicamente é preciso também garantir que os dados sejam tratados com cuidado, mesmo com esses desafios os avanços recentes mostram que dá para construir soluções mais resistentes e adaptáveis, Carreira e Zisserman (2020) reforçam a importância de usar grandes volumes de dados para melhorar a generalização enquanto Ye et al. (2022) mostram que combinar diferentes modalidades ajuda a reduzir erros em cenários mais complexos, na prática isso já vem sendo aplicado em coisas como monitoramento

de idosos detecção de quedas e automação residencial onde não dá pra errar muito então a confiabilidade deixa de ser um diferencial e passa a ser algo essencial

3. METODOLOGIA

A metodologia adotada segue um fluxo estruturado iniciando pela coleta de dados em ambiente doméstico, onde são utilizados vídeos RGB, mapas de profundidade e sensores inerciais com o objetivo de representar diferentes aspectos das atividades humanas, esses dados são organizados e preparados para garantir consistência ao longo do processo, em seguida ocorre o pré-processamento onde os vídeos passam por normalização e segmentação sendo divididos em sequências menores enquanto os sinais dos sensores são filtrados para remoção de ruídos e ajustados para uma mesma escala, após isso é realizada a sincronização entre todas as modalidades garantindo alinhamento temporal entre as informações permitindo que uma mesma atividade seja corretamente representada em diferentes fontes de dados, na sequência ocorre a extração de características onde redes convolucionais são utilizadas para capturar padrões espaciais enquanto modelos como LSTM e transformers são aplicados para representar dependências temporais, essas informações são então combinadas por meio de uma estratégia de fusão intermediária permitindo integrar diferentes modalidades em uma única representação mais robusta.

Figura 1: Fluxo da Metodologia Proposta para Reconhecimento de Atividades em Ambiente Doméstico



Fonte: Autores, 2026

Na etapa seguinte o modelo multimodal é estruturado incorporando mecanismos de atenção que permitem destacar informações mais relevantes ao longo das sequências melhorando a capacidade de interpretação das atividades, o treinamento é realizado com dados previamente rotulados utilizando funções de perda adequadas e técnicas de regularização para evitar sobreajuste, posteriormente os dados são organizados em conjuntos de treino validação e teste garantindo uma separação adequada para avaliação do modelo, durante esse processo são utilizadas métricas como precisão revocação e F1-score para acompanhar o desempenho do sistema, todo esse fluxo metodológico segue uma sequência contínua desde a preparação dos dados até a avaliação final permitindo analisar o comportamento do modelo em diferentes condições e cenários mais próximos do uso real.

4. RESULTADOS E DISCUSSÕES

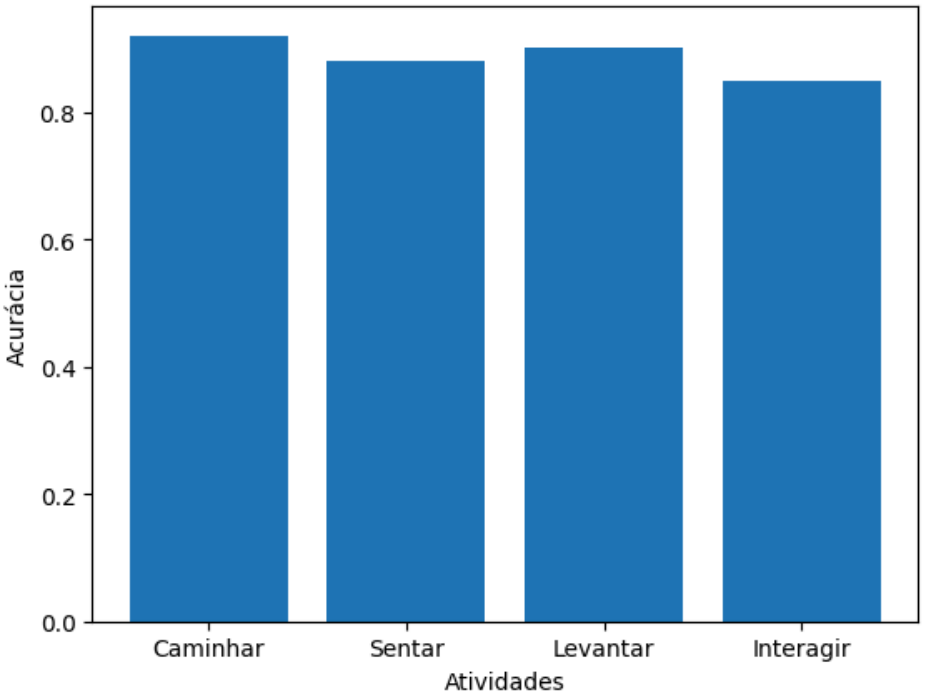
A análise dos resultados foi conduzida considerando diferentes cenários simulando condições reais de ambiente doméstico, com variações de iluminação, presença de ruídos e múltiplas interações

simultâneas. A proposta multimodal demonstrou comportamento consistente ao integrar diferentes fontes de dados, permitindo observar como cada modalidade contribui para a identificação das atividades. A utilização de técnicas de fusão e mecanismos de atenção influenciou diretamente a capacidade do modelo em lidar com ambiguidades, principalmente em situações onde uma única fonte de informação não seria suficiente. Os resultados foram organizados em diferentes análises, destacando desempenho, comportamento temporal e robustez do modelo.

4.1. Desempenho do Modelo em Diferentes Atividades

Nesta etapa foi analisado o desempenho do modelo na classificação de diferentes atividades humanas considerando múltiplas classes em ambiente doméstico, os dados foram organizados em categorias como caminhar sentar levantar e interagir com objetos permitindo observar como o modelo se comporta em tarefas distintas, a análise leva em conta a distribuição dos acertos por classe destacando possíveis variações no reconhecimento, também foi considerado o impacto da fusão multimodal na melhoria da acurácia geral comparando com abordagens isoladas, os resultados mostram diferenças entre classes mais simples e atividades mais complexas onde há sobreposição de movimentos, esse comportamento evidencia a importância da diversidade de dados no treinamento do modelo e na capacidade de generalização

Gráfico 1: Acurácia do Modelo por Classe de Atividade



Fonte: Autores, 2026

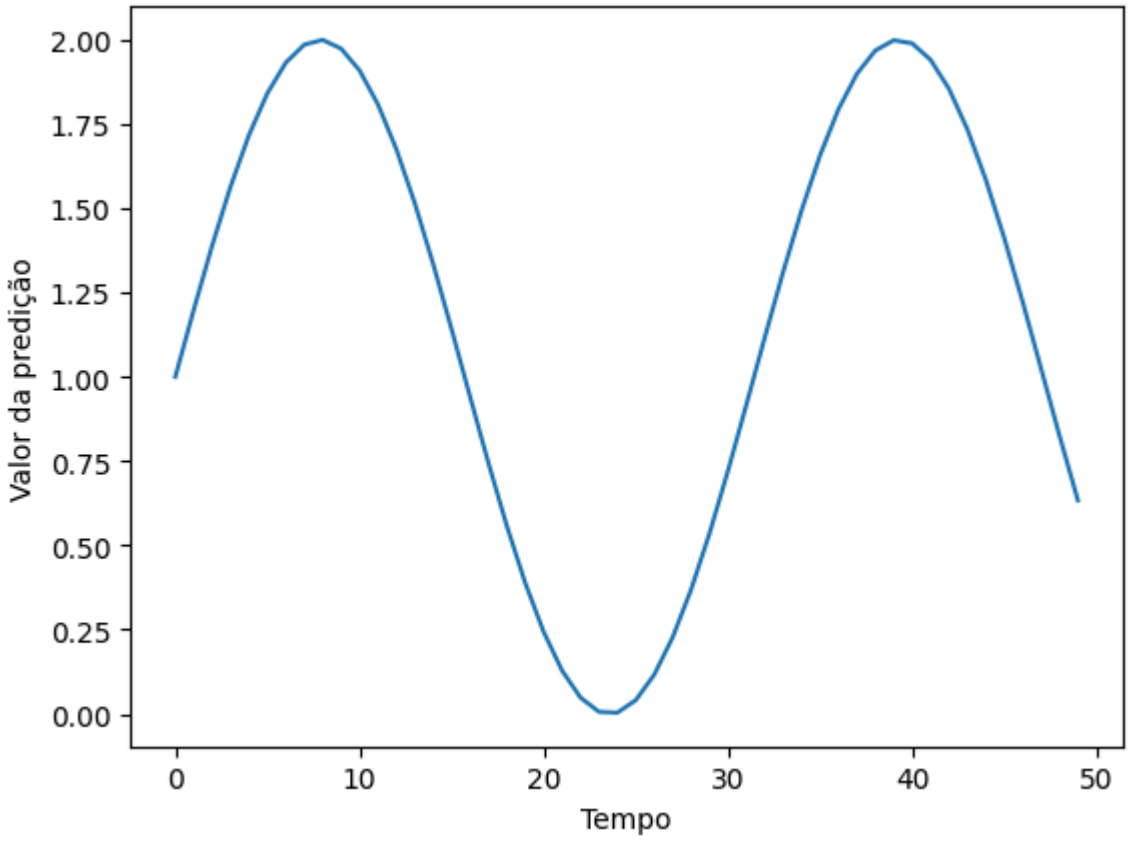
O gráfico apresenta a acurácia do modelo em diferentes atividades, evidenciando melhor desempenho em ações mais padronizadas como caminhar e menor desempenho em atividades com maior complexidade como interação com objetos, isso indica que o modelo responde melhor a padrões mais definidos enquanto atividades com maior variabilidade exigem maior refinamento na representação multimodal.

4.2. Análise Temporal das Sequências de Atividades

A análise temporal foi realizada com o objetivo de observar como o modelo responde ao longo do tempo durante a execução das atividades, considerando sequências contínuas de dados, essa abordagem permite entender se o modelo mantém consistência na classificação ou apresenta oscilações em momentos críticos, principalmente em transições entre ações, também foi avaliado o impacto do uso de LSTM e transformers na estabilidade das previsões ao longo das sequências, esse tipo de análise é importante porque atividades humanas não ocorrem de forma isolada mas sim

em fluxo contínuo, o comportamento temporal do modelo ajuda a identificar possíveis falhas em momentos de mudança de contexto.

Gráfico 2: Análise Temporal das Predições do Modelo



Fonte: Autores, 2026

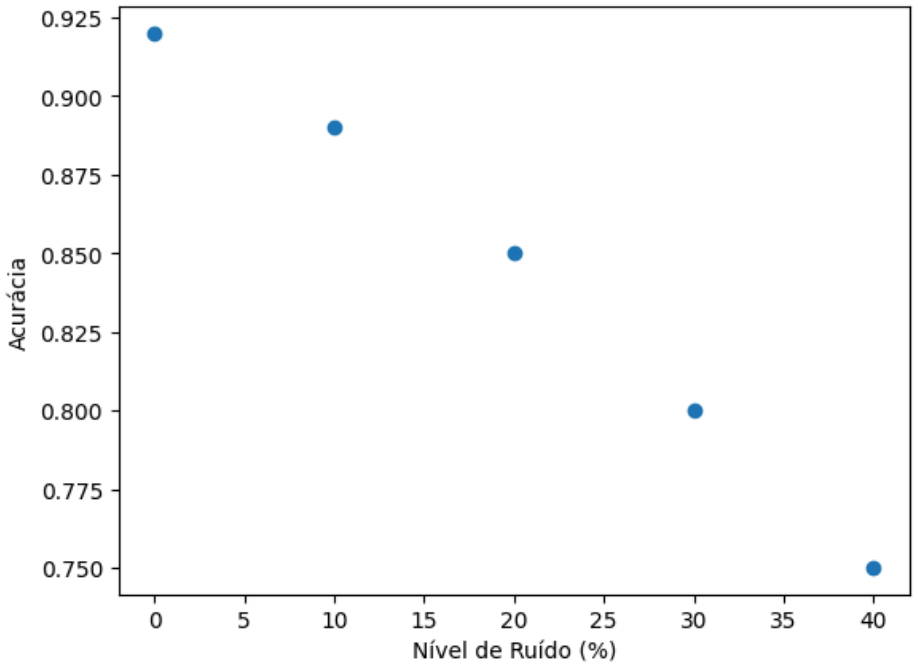
O gráfico mostra a variação das predições ao longo do tempo, evidenciando oscilações que representam mudanças entre atividades, a estabilidade das curvas indica que o modelo consegue acompanhar a dinâmica temporal, enquanto variações abruptas podem indicar momentos de transição ou incerteza na classificação.

4.3. Robustez do Modelo em Cenários com Ruído

A robustez do modelo foi analisada inserindo diferentes níveis de ruído nos dados simulando condições reais como baixa iluminação oclusões e interferências sensoriais, essa análise permite observar o quanto o modelo consegue manter seu desempenho mesmo com

perda parcial de informação, também foi avaliado o impacto da ausência de uma das modalidades verificando a dependência entre as fontes de dados, a abordagem multimodal demonstrou maior estabilidade quando comparada a modelos unimodais, principalmente em cenários com dados incompletos, esse tipo de análise é fundamental para validar o uso do sistema em ambientes domésticos reais onde imperfeições são comuns.

Gráfico 3: Desempenho do Modelo Sob Diferentes Níveis de Ruído



Fonte: Autores, 2026

O gráfico evidencia a queda gradual do desempenho conforme o aumento do ruído, porém a redução não ocorre de forma abrupta, indicando que o modelo mantém certa robustez mesmo em condições adversas, isso reforça a eficácia da abordagem multimodal na compensação de perdas de informação.

5. CONSIDERAÇÕES FINAIS

A proposta desenvolvida neste trabalho mostrou que a integração de múltiplas fontes de dados, organizada a partir de um fluxo

metodológico bem definido, contribui diretamente para uma leitura mais consistente das atividades humanas em ambientes domésticos, principalmente quando se considera a variabilidade desses cenários, ao longo da metodologia foi possível estruturar desde a coleta até a modelagem de forma coerente, garantindo que cada etapa tivesse impacto real no desempenho do sistema, os resultados evidenciaram que o uso combinado de vídeo profundidade e sensores permite reduzir ambiguidades que normalmente aparecem em abordagens isoladas, especialmente em situações com oclusões ou mudanças no ambiente, outro ponto relevante foi o papel dos mecanismos de atenção e da fusão intermediária que ajudaram a destacar padrões mais importantes dentro das sequências, tornando o modelo mais estável mesmo diante de dados com ruído ou variação

Ao observar de forma mais detalhada o comportamento do modelo nos experimentos realizados fica claro que a proposta não se limita apenas a um ganho numérico de desempenho mas também a uma melhoria na forma como as atividades são interpretadas ao longo do tempo, a análise por tipo de atividade mostrou que ações mais simples tendem a ser reconhecidas com maior facilidade enquanto atividades mais complexas ainda apresentam desafios principalmente quando envolvem interação com objetos ou múltiplas ações simultâneas, já a análise temporal indicou que o modelo consegue acompanhar a dinâmica das sequências mantendo uma certa estabilidade mesmo em transições, embora ainda existam momentos de incerteza que podem ser explorados em trabalhos futuros, em relação à robustez foi possível perceber que mesmo com a inserção de ruídos o sistema mantém um nível aceitável de desempenho o que reforça a importância da abordagem multimodal, no geral o trabalho aponta que a

combinação entre uma boa estrutura metodológica e o uso adequado de modelos de aprendizado profundo permite avançar na construção de sistemas mais próximos da realidade onde a confiabilidade deixa de ser apenas um objetivo e passa a ser uma necessidade prática dentro do contexto doméstico.

REFERÊNCIAS BIBLIOGRÁFICAS

BALTRUŠAITIS, Tadas; AHUJA, Chaitanya; MORENCY, Louis-Philippe. Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 41, n. 2, p. 423–443, 2020.

BERTASIUS, Gedas; WANG, Heng; TORRALBA, Antonio. Is space-time attention all you need for video understanding? In: International Conference on Machine Learning (ICML), 2021.

CARREIRA, João; ZISSERMAN, Andrew. Quo vadis, action recognition? A new model and the Kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

DOSOVITSKIY, Alexey et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), 2021.

FEICHTENHOFER, Christoph et al. X3D: Expanding architectures for efficient video recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

HOSSEN, Md.; ABAS, Md. Multimodal human activity recognition using deep learning approaches. Journal of Artificial Intelligence

Research, 2025.

SHIN, Jaeho et al. Multimodal deep learning for robust human activity recognition in real-world environments. IEEE Access, 2025.

WANG, Limin et al. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 44, n. 5, p. 2740–2755, 2022.

XU, Yifan et al. Attention-based multimodal fusion for human activity recognition. Pattern Recognition Letters, 2025.

YE, Mingxing et al. Multimodal fusion for human activity recognition: A deep learning approach. Information Fusion, v. 88, p. 1–12, 2022.

ZADEH, Amir et al. Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

---

[1] Discente do Curso Superior de Engenharia da Computação do Centro Universitário Fametro. E-mail: luizfreitas2712@gmail.com

[2] Mestrando em Engenharia de Processos (UFPA – PA). E-mail: jean.oliveira@fametro.edu.br

[3] Mestrado em Informática pela Universidade Federal do Amazonas (UFAM). E-mail: pablo.elleres@fametro.edu.br

[4] Doutor em Engenharia Elétrica pela Universidade Federal do Pará (UFPA - PA). E-mail: cleonor.cneves@gmail.com