Reconhecimento de Emoções em Vídeos: Uma Análise Comparativa de LSTM, CNNs, YOLO e Vision Transformers no CMU-MOSEI

Daniel Casanova*, Pedro Luiz de Paula Filho*, Kelyn Schenatto*, Alessandra B. G. Hoffmann*
*Universidade Tecnológica Federal do Paraná (UTFPR), Medianeira, Brasil
danielcasanova@alunos.utfpr.edu.br, pedrol@utfpr.edu.br, kelynschenatto@gmail.com, hoffmann@utfpr.edu.br

***Resumo*—This paper presents a comparative study of different deep learning architectures for video analysis using the CMU-MOSEI dataset. We evaluate models that explicitly capture temporal dependencies, such as Long Short-Term Memory (LSTM), against frame-based approaches like ResNet50, Vision Transformer (ViT), and YOLOv11 adapted for classification. The experiments consider accuracy and precision metrics to systematically compare their performance. Results show that the ViT achieved the highest performance (78%), while YOLOv11 and ResNet50 provided competitive alternatives in terms of efficiency and stability. Conversely, the LSTM model obtained the lowest performance (53%), suggesting that explicit temporal modeling was less effective than frame aggregation strategies for this dataset. These findings highlight trade-offs between accuracy, efficiency, and applicability in real-world video-based emotion recognition systems.**

***Keywords*—Video Analysis; Emotion Recognition; Vision Transformer.**

***Resumo*—Este trabalho apresenta um estudo comparativo entre diferentes arquiteturas de aprendizado profundo aplicadas à análise de vídeos utilizando o dataset CMU-MOSEI. Foram avaliados modelos que capturam explicitamente dependências temporais, como a Long Short-Term Memory (LSTM), em comparação com abordagens baseadas em frames, como a ResNet50, o Vision Transformer (ViT) e o YOLOv11 adaptado para classificação. Os experimentos consideraram as métricas de acurácia e precisão para uma análise sistemática de desempenho. Os resultados mostram que o ViT obteve o melhor desempenho (78%), enquanto o YOLOv11 e a ResNet50 se destacaram como alternativas competitivas em termos de eficiência e estabilidade. Por outro lado, o modelo LSTM apresentou o pior resultado (53%), indicando que a modelagem temporal explícita foi menos eficaz que estratégias de agregação de frames neste conjunto de dados. Esses achados evidenciam os trade-offs entre precisão, eficiência e aplicabilidade em sistemas reais de reconhecimento de emoções baseados em vídeo.**

***Palavras-chave*—Análise de vídeo; Reconhecimento de emoções; Vision Transformer.**

I. Introdução

A análise de vídeos tem ganhado crescente relevância em diversas aplicações, tais como vigilância inteligente, análise de comportamento humano, recomendação multimodal e interação homem-máquina. Diferente da análise de imagens estáticas, os vídeos contêm uma dimensão temporal que carrega informações fundamentais sobre dinâmicas e contextos. Ignorar essa dimensão pode levar à perda de padrões essenciais, sobretudo em tarefas de reconhecimento de ações e emoções [1], [2].

Dentre as abordagens que exploram informações temporais, destacam-se as redes recorrentes, em especial as *Long Short-Term Memory* (LSTM) [3]. Em vídeo, essas redes já mostraram ganhos ao modelar a dinâmica de sequências de *frames* [4], [5]. Os fundamentos, vantagens e limitações das LSTM são discutidos na Seção II-A.

Em paralelo, modelos originalmente desenvolvidos para imagens estáticas têm sido adaptados para análise de vídeo por meio da agregação de frames. Redes convolucionais profundas, como a ResNet [6], mostraram-se eficientes na extração de características espaciais, enquanto os *Vision Transformers* (ViT) introduzidos por [7] representam um paradigma baseado em mecanismos de atenção [2], que vêm superando *Convolutional Neural Networks* (CNNs) em múltiplos *benchmarks*. Recentemente, variantes específicas para vídeo, como o ViViT no artigo [8], exploraram a atenção espaço-temporal para capturar dependências de movimento e contexto.

Outro eixo relevante é representado pela família de arquiteturas *You Only Look Once* (YOLO), amplamente empregadas para detecção e classificação em tempo real segundo o estudo de [9]. Desde sua primeira versão, essas arquiteturas evoluíram significativamente, chegando a versões recentes como a YOLOv11, que equilibram velocidade e precisão em cenários complexos [10]. Mais recentemente, o YOLOv12 foi introduzido como uma abordagem *attention-centric*, incorporando mecanismos de atenção (como Area Attention e FlashAttention) e a arquitetura R-ELAN para acelerar o processamento e melhorar a agregação de features, mantendo desempenho em tempo real e superando o YOLOv11 em benchmarks como COCO

[9], [11]. Embora originalmente projetadas para detecção de objetos, adaptações permitem sua aplicação em reconhecimento de ações e classificação baseada em vídeo.

Uma estratégia amplamente empregada para adaptar modelos de imagem ao domínio temporal é a agregação de frames. Técnicas simples, como a média aritmética das predições, demonstraram resultados competitivos em diversos cenários [12], [13], levantando a questão sobre quando métodos complexos de modelagem temporal são de fato necessários [12], [13]. Nesse contexto, surge o problema de investigar se modelos de imagem com estratégias de agregação podem competir, em termos de desempenho e eficiência, com arquiteturas recorrentes como LSTM.

Diante desse cenário, este trabalho tem como objetivo comparar empiricamente diferentes abordagens de análise de vídeo utilizando o dataset multimodal CMU-MOSEI [14]. São exploradas quatro linhas principais: (i) LSTM para modelagem explícita da dependência temporal; (ii) Vision Transformers; (iii) CNNs ResNet aplicadas frame a frame; e (iv) YOLOv11 adaptado para classificação. Todas as abordagens são avaliadas de forma sistemática com base em métricas de desempenho.

A principal contribuição deste estudo é fornecer uma análise comparativa entre modelos que exploram explicitamente a informação temporal (LSTM) e aqueles que dependem de estratégias de agregação de frames (ViT, ResNet e YOLO). O trabalho busca responder se abordagens baseadas em modelos de imagem podem competir em eficácia com arquiteturas temporais, destacando diferenças em precisão e aplicabilidade prática. Além da análise comparativa, os resultados obtidos podem ter implicações práticas em aplicações de monitoramento emocional em ambientes educacionais e terapêuticos, bem como em sistemas de recomendação multimodal. Esses cenários reforçam a relevância do estudo, que busca não apenas avaliar arquiteturas, mas também indicar caminhos para aplicações reais.

II. Revisão da Literatura

A análise de vídeos com técnicas de aprendizado profundo evoluiu de arquiteturas recorrentes para modelos convolucionais, detectores em tempo real e, mais recentemente, transformers. Cada abordagem trouxe contribuições distintas, apresentando também limitações que motivaram o surgimento de novos paradigmas.

A. Redes LSTM para Análise Temporal

As *Long Short-Term Memory* (LSTM) [3] foram concebidas para mitigar o problema do *vanishing gradient* em redes recorrentes, por meio de portas de entrada, saída e esquecimento que permitem capturar dependências de longo prazo em sequências. Essa capacidade mostrou-se essencial em tarefas como reconhecimento de fala [15] e tradução automática [16].

Na análise de vídeo, LSTM são frequentemente combinadas com extratores espaciais convolucionais, compondo modelos híbridos que integram espaço e tempo [4]. Embora eficazes para modelar a dinâmica de ações, apresentam limitações de escalabilidade e alto custo de treinamento [5]. Em síntese, oferecem boa capacidade para relações temporais complexas, mas dificultam a paralelização e podem sobreajustar em *datasets* menores.

B. Redes Convolucionais e ResNet

As *Convolutional Neural Networks* (CNNs) revolucionaram a classificação de imagens após o trabalho de [17]. Posteriormente, [6] introduziram a ResNet, arquitetura baseada em conexões residuais que possibilitam o treinamento de redes muito profundas sem degradação do gradiente. Essa inovação consolidou a ResNet como referência em visão computacional.

Para vídeos, CNNs podem ser aplicadas diretamente em frames individuais [18], ou por meio de convoluções tridimensionais que exploram simultaneamente espaço e tempo [19]. As vantagens das CNNs incluem alta eficiência na extração de padrões espaciais e grande disponibilidade de arquiteturas pré-treinadas. Entretanto, quando aplicadas frame a frame, perdem a dimensão temporal, exigindo mecanismos adicionais de agregação ou modelos híbridos para compensar essa limitação.

C. YOLO e Detecção em Tempo Real

A família de modelos *You Only Look Once* (YOLO), proposta inicialmente por [9], trouxe uma mudança de paradigma ao tratar a detecção de objetos como um problema de regressão unificado. Essa abordagem permitiu realizar predições em tempo real com precisão competitiva, superando métodos anteriores em eficiência. Evoluções posteriores, como YOLOv3 [20] e YOLOv4 [21], refinaram a arquitetura com melhorias em velocidade, profundidade e técnicas de regularização.

Versões mais recentes, como a YOLOv11, destacam-se pela capacidade de equilibrar precisão e desempenho em cenários complexos [10]. Apesar de suas vantagens em tempo de inferência, as arquiteturas YOLO apresentam desvantagens em tarefas que demandam forte modelagem temporal, já que sua concepção original foi voltada para detecção estática.

D. Vision Transformers (ViT)

Inspirados em mecanismos de atenção da área de NLP, os *Vision Transformers* (ViT) foram propostos por [7], demonstrando que, com datasets suficientemente grandes, transformers podem superar CNNs em visão computacional. Diferente das CNNs, o ViT divide a imagem em *patches* que são tratados como

tokens em uma sequência, explorando dependências globais via mecanismos de autoatenção.

Na literatura, diversas variantes estenderam esse paradigma para o domínio de vídeo. O ViViT [8] propôs arquiteturas puramente baseadas em transformers para capturar dependências espaço-temporais, avaliando diferentes esquemas de fatoração da atenção entre dimensões espaciais e temporais. O TimeSformer [12], por sua vez, introduziu a decomposição fatorizada da atenção espaço-temporal, reduzindo o custo computacional e viabilizando o treinamento em larga escala. Outros trabalhos, como Video Swin Transformer [13], exploraram janelas deslocadas para modelar relações hierárquicas e alcançar melhor escalabilidade, enquanto MViT (Multiscale Vision Transformers) [22] incorporaram resoluções múltiplas para capturar contextos em diferentes granularidades.

Esses avanços consolidam os transformers visuais como um dos principais paradigmas emergentes em análise de vídeo, capazes de superar CNNs 3D e LSTM em benchmarks desafiadores. Contudo, seu alto custo computacional e a necessidade de grandes volumes de dados continuam sendo limitações práticas amplamente discutidas na literatura.

E. Estratégias de Agregação Temporal

Uma abordagem alternativa para lidar com a dimensão temporal consiste em aplicar modelos de imagem a cada frame e, posteriormente, agregar as predições. Técnicas simples, como a média aritmética, mostraram-se eficientes em diversos cenários [12], [13]. Embora menos sofisticadas que LSTM ou ViT, essas estratégias apresentam a vantagem de serem computacionalmente leves, facilitando a aplicação em tempo real. Por outro lado, ignoram dependências dinâmicas mais complexas, podendo levar à perda de informações relevantes.

F. Visão Geral e Surveys

Diversos levantamentos na literatura consolidam o progresso das técnicas de análise de vídeo. Como apresentado em [2] foram revisados os métodos e tendências recentes, enquanto em [1] destacam-se avanços em previsão e reconhecimento de ações humanas. Esses *surveys* apontam que não há uma solução única, mas sim um conjunto de abordagens que devem ser escolhidas conforme o equilíbrio desejado entre precisão, custo computacional e aplicabilidade.

III. Metodologia

A metodologia proposta neste trabalho foi organizada em dois fluxos principais: (i) um modelo LSTM robusto com múltiplas melhorias arquiteturais, destinado a capturar explicitamente dependências temporais em vídeos; e (ii) um conjunto de arquiteturas baseadas em processamento frame a frame, cujas predições foram posteriormente agregadas. Ambas as abordagens foram aplicadas sobre o dataset multimodal CMU-MOSEI [14], embora neste estudo tenha sido considerada exclusivamente a modalidade visual (vídeo). Os dados textuais e de áudio, embora disponíveis, não foram utilizados, visto que o objetivo foi avaliar a eficácia das arquiteturas de *deep learning* na classificação baseada apenas em expressões faciais visuais.

A Figura 1 apresenta uma visão geral da metodologia, destacando os dois fluxos comparados: um *pipeline* temporal baseado em LSTM e um *pipeline frame-based*.

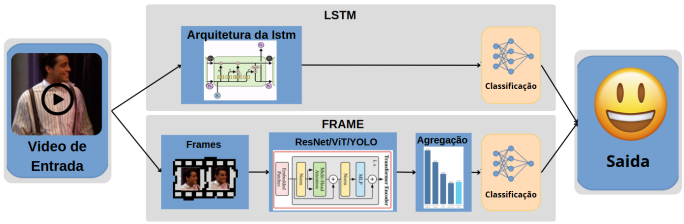

Figura 1: Visão geral da metodologia, comparando o fluxo LSTM e o fluxo *frame-based*. Fonte: Autoria Própria.

A. Dataset

O dataset adotado neste estudo foi o **CMU-MOSEI** [14], amplamente utilizado em tarefas de análise multimodal. Ele contém mais de 23 mil vídeos coletados do YouTube, anotados com múltiplas modalidades: texto, áudio e vídeo. Embora o corpus completo permita investigações multimodais, neste trabalho foi considerada exclusivamente a modalidade **visual**, ou seja, apenas os frames dos vídeos.

A escolha de utilizar somente a modalidade visual teve como objetivo avaliar a eficácia de arquiteturas de *deep learning* na tarefa de reconhecimento de expressões faciais e emoções em vídeos, abstraindo informações textuais e de áudio.

A Figura 2 apresenta exemplos de frames que compõem o CMU-MOSEI, evidenciando a diversidade de indivíduos, expressões e condições de gravação presentes no corpus. Essa variabilidade é crucial para testar a robustez dos modelos propostos.

B. Pré-processamento e Amostragem de Frames

Os vídeos foram decompostos em sequências de frames por amostragem uniforme, mantendo a ordem temporal. Cada frame foi redimensionado para 224×224 pixels e normalizado com estatísticas do ImageNet (média e desvio padrão por canal), de modo a alinhar o domínio visual ao pré-treinamento das arquiteturas avaliadas. Para o fluxo LSTM, as sequências foram construídas com janelas fixas de comprimento T (segmentos não sobrepostos), enquanto para o fluxo *frame-based* os frames foram processados de forma independente, com posterior

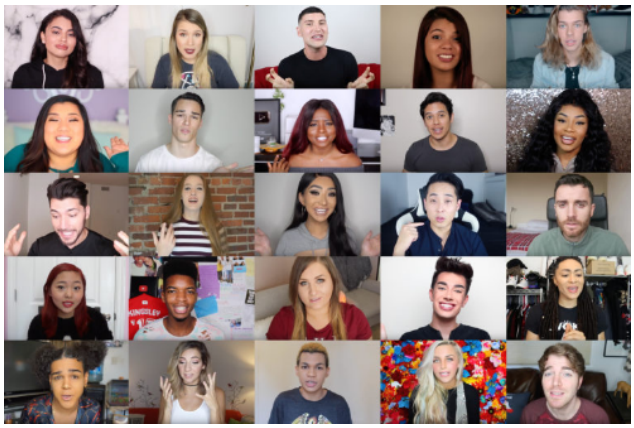

Figura 2: Exemplo de imagens do dataset CMU-MOSEI utilizado no estudo. Fonte: Adaptado de [14].

agregação das predições ao nível de vídeo. Quando necessário, aplicou-se *center-crop* após o redimensionamento para preservar a razão de aspecto. Não foi realizada detecção de rosto: o modelo recebe o *frame* integral, refletindo a variabilidade de enquadramento e iluminação do CMU-MOSEI.

C. Divisão dos Dados e Reprodutibilidade

Os dados foram divididos em partições de treino (70%), validação (15%) e teste (15%), estratificando por classe para preservar a distribuição original. Todas as execuções utilizaram sementes fixas para inicialização pseudoaleatória, garantindo reprodutibilidade dos resultados. O conjunto de validação foi usado para seleção de hiperparâmetros e para o acionamento de *early stopping* quando aplicável, enquanto o conjunto de teste foi reservado exclusivamente para a avaliação final.

D. Transfer Learning e Fine-Tuning

Todos os modelos empregados foram inicializados com pesos pré-treinados em grandes bases de dados de imagens e adaptados ao problema específico por meio de *transfer learning*. Apenas as camadas finais de classificação foram ajustadas, de modo a reduzir o tempo de treinamento e melhorar a generalização em um dataset de tamanho moderado, como o CMU-MOSEI [23], [24]. O processo de *fine-tuning* foi realizado no *framework* PyTorch [25], utilizando os seguintes hiperparâmetros: taxa de aprendizado inicial de 1×10^{-4}, regularização L2 de 1×10^{-4}, *batch size* igual a 16 e 30 épocas de treinamento. Para estabilidade numérica, foi aplicada acumulação de gradientes a cada dois *batches*, e a taxa de aprendizado foi ajustada com o agendador *Cosine Annealing* ($T_{max} = 30$, $\eta_{min} = 1 \times 10^{-6}$).

E. Métricas e Procedimento de Avaliação

A avaliação considerou duas métricas: acurácia e precisão. A acurácia corresponde à proporção de predições corretas sobre o total de amostras, enquanto a precisão é definida, por classe, como

$$\text{Precisão} = \frac{VP}{VP + FP}$$

onde:

- VP: número de instâncias positivas corretamente classificadas (verdadeiros positivos),
- FP: número de instâncias negativas incorretamente classificadas como positivas (falsos positivos).

É reportada a precisão média ponderada pelas frequências das classes, refletindo o desempenho global sob desbalanceamento. Para cada vídeo, o rótulo final foi obtido diretamente (no fluxo LSTM) ou pela agregação das predições por *frame* via média aritmética (no fluxo *frame-based*). Todas as métricas foram computadas no conjunto de teste, após seleção de hiperparâmetros com base no desempenho de validação.

F. Fluxo 1: LSTM Robusto

Implementamos um fluxo temporal híbrido que combina extração espacial profunda com modelagem temporal explícita (fundamentos de LSTM na Seção II-A). O arranjo é:

(i) Extração espacial. *Ensemble* convolucional com ResNet18 (512 *features*) e EfficientNet-B0 (1280 *features*); as saídas são concatenadas para formar uma representação rica de cada *frame*.

(ii) Modelagem temporal. Convoluções temporais 1D (*Temporal Convolutional Networks*, TCN) para padrões de curto prazo, seguidas de LSTM com atenção de múltiplas cabeças (*Multi-Head Attention*) para dependências de longo alcance.

(iii) Robustez e regularização. *Stochastic Depth*, *dropout* adaptativo, L2 adaptativa e *data augmentation* temporal ("*Random Temporal Cropping*", "*Random Frame Dropping*" e "*Temporal Mixup*").

Essa organização segue evidências recentes de arquiteturas híbridas (p.ex., EfficientNet+TCN+LSTM) para vídeo [26], buscando atenuar limitações clássicas de LSTM (paralelização e sensibilidade a hiperparâmetros) e aproveitar sinais em diferentes escalas temporais.

Para controle, treinamos também uma LSTM *baseline* (um único *backbone* convolucional seguido de uma camada LSTM simples), utilizada apenas como referência metodológica.

G. Fluxo 2: Modelagem Baseada em Frames

Cada vídeo foi decomposto em uma sequência de imagens por amostragem uniforme. Em seguida, cada frame foi

processado independentemente por arquiteturas distintas (ResNet50, ViT-Large-Patch16-224 e YOLO11-Style-ResNet50). As predições individuais foram posteriormente agregadas por média aritmética para formar a decisão final do vídeo.
As arquiteturas avaliadas nesse fluxo foram:

- **ResNet50**: arquitetura com 24,6 milhões de parâmetros, empregando conexões residuais profundas. As predições foram obtidas frame a frame e posteriormente agregadas por média aritmética.
- **ViT-Large-Patch16-224**: modelo transformer visual com 307 milhões de parâmetros, capaz de capturar dependências globais por meio de mecanismos de autoatenção, aplicado diretamente sobre *patches* de cada frame.
- **YOLO11-Style-ResNet50**: variante adaptada da família YOLO para classificação, baseada em ResNet50, com aproximadamente 24,7 milhões de parâmetros, destacando-se pelo equilíbrio entre precisão e eficiência em tempo de inferência.

A agregação das predições foi realizada por média aritmética das saídas obtidas em cada frame, método simples, mas eficaz em reduzir variações locais [12], [13]. Apesar de sua eficiência, tal estratégia não modela dependências temporais explícitas, o que pode limitar o desempenho em contextos onde a ordem dos frames é essencial para a correta classificação.

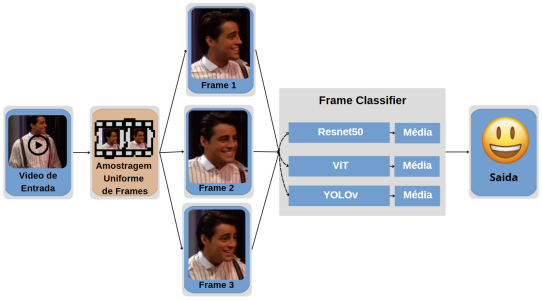

Figura 3: Fluxo detalhado da abordagem baseada em frames, incluindo amostragem uniforme, classificação e etapa de agregação. Fonte: Autoria Própria.

H. Resumo da Abordagem

Dessa forma, a metodologia compara dois paradigmas distintos: (i) um fluxo baseado em sequenciamento temporal robusto, que combina *ensemble* de CNNs, TCN, LSTM e atenção multi-cabeça; e (ii) um fluxo fundamentado em arquiteturas de imagem aplicadas a frames individuais, complementado por agregação estatística. Essa organização permitiu avaliar de forma sistemática as diferenças entre precisão e capacidade de generalização.

IV. Resultados e Discussão

Nesta seção são apresentados e discutidos os resultados obtidos pelos modelos implementados. As métricas utilizadas foram acurácia e precisão, que fornecem uma visão geral da capacidade de classificação correta dos vídeos do dataset CMU-MOSEI. A Tabela I resume os resultados obtidos para cada arquitetura.

Tabela I: Resultados de acurácia e precisão para os modelos avaliados.

Modelo	Precisão	Acurácia
LSTM	0.5354	0.5354
ResNet50	0.6170	0.6271
YOLOv11x-cls	0.7153	0.7178
ViT-Large	**0.7843**	**0.7843**

A. Análise Comparativa

Observa-se que o *Vision Transformer* (ViT-Large) obteve o melhor desempenho em ambas as métricas, com precisão e acurácia de aproximadamente 78%. Esse resultado confirma o potencial dos transformers visuais em capturar dependências globais e padrões complexos nos dados de vídeo [7], [8].

O modelo YOLOv11x-cls alcançou desempenho intermediário, com valores próximos a 72% de precisão e acurácia. Esse resultado mostra que arquiteturas originalmente desenvolvidas para detecção em tempo real podem ser adaptadas para tarefas de classificação de vídeo com competitividade, beneficiando-se de sua eficiência em processamento de imagens. No entanto, a ausência de uma modelagem temporal explícita limita o ganho em cenários mais dependentes da sequência de frames.

A ResNet50 apresentou valores em torno de 62% de acurácia, posicionando-se abaixo do YOLOv11 e do ViT, mas ainda assim superando a LSTM. Isso reflete a robustez das CNNs residuais, que, mesmo sem mecanismos de atenção, continuam relevantes em tarefas de classificação frame a frame devido à sua simplicidade e estabilidade de treinamento.

Por fim, o modelo LSTM obteve o pior desempenho (53%). Apesar das melhorias introduzidas (*ensemble* de CNNs, TCN e atenção multi-cabeças), é observado que: (i) a sinalização temporal útil para emoção no CMU-MOSEI parece ser relativamente fraca frente a pistas estáticas de cada frame; (ii) a integração TCN + LSTM + atenção elevou a complexidade de otimização e a sensibilidade a hiperparâmetros; e (iii) a agregação frame a frame funciona como um forte *baseline* quando as variações temporais são sutis. Importante: também foi treinada uma LSTM mínima (*backbone* único + LSTM de uma camada, sem TCN/atenção), e o resultado foi ainda inferior ao do modelo LSTM robusto. Isso indica que os

aprimoramentos de fato ajudaram a LSTM, embora não o suficiente para superar as abordagens baseadas em frames neste conjunto.

B. Discussão dos Resultados

Os resultados evidenciam que arquiteturas modernas, como ViT e YOLOv11, apresentam maior vantagem em cenários que exigem equilíbrio entre precisão e eficiência. Enquanto o ViT se destaca pela superioridade em métricas de classificação, o YOLOv11 oferece uma alternativa competitiva com menor custo computacional. Já a ResNet50 confirma sua posição como um modelo intermediário estável, ainda amplamente utilizado em tarefas de visão computacional.

O baixo desempenho da LSTM sugere que, para este dataset, a informação temporal explícita não foi suficiente para compensar a ausência de extração espacial sofisticada. Em contrapartida, a simples estratégia de agregação de frames demonstrou ser eficaz, reforçando estudos anteriores que indicam que abordagens baseadas em predições independentes de frames podem ser competitivas em diversos cenários [12], [13].

Em síntese, os experimentos mostram que a escolha do modelo deve considerar principalmente a precisão e o contexto de aplicação. Para cenários que demandam máxima acurácia, o ViT é a melhor escolha; para situações que valorizam maior simplicidade de implementação, o YOLOv11 e a ResNet50 aparecem como alternativas competitivas. Já a LSTM, embora historicamente relevante, mostrou limitações claras neste contexto.

C. Análise Específica do Desempenho da LSTM

Embora a arquitetura LSTM robusta tenha incorporado melhorias relevantes — como representações espaciais enriquecidas via CNNs, TCN para padrões de curto prazo e atenção multi-cabeças para longo alcance — seu desempenho final permaneceu inferior às abordagens baseadas em frames. Alguns fatores ajudam a explicar esse resultado.

Em primeiro lugar, o reconhecimento de expressões e emoções no CMU-MOSEI parece depender majoritariamente de pistas espaciais estáticas, já presentes em um único frame (configuração dos olhos, boca ou sobrancelhas). Nesses casos, a simples agregação das predições individuais tende a ser suficiente, reduzindo a variância e funcionando como um *ensemble* temporal implícito, sem necessidade de modelagem explícita da ordem dos frames.

Outro aspecto é a complexidade de otimização. A combinação TCN+LSTM+atenção aumenta a profundidade e a interdependência entre módulos, tornando o treinamento mais sensível a hiperparâmetros como comprimento da janela temporal ou janela T, taxa de aprendizado e regularização. Pequenas variações nesses elementos afetaram significativamente a convergência, enquanto os modelos baseados em frames mostraram-se mais estáveis.

Além disso, o ruído nas sequências de características — decorrente de variações de enquadramento, iluminação e compressão nos vídeos do YouTube — tende a dificultar a extração de padrões temporais consistentes. Esse ruído pode induzir a LSTM a tentar capturar variações espúrias, reduzindo sua capacidade de generalização.

A definição da janela T também exerce papel crítico: escolhas pouco ajustadas de T podem diluir sinais relevantes de curta duração ou introduzir redundância, diminuindo a eficácia da modelagem sequencial.

Por fim, embora tenham sido aplicadas técnicas de regularização temporal como *stochastic depth*, *dropout* adaptativo e *mixup* temporal, tais mecanismos podem suavizar excessivamente dinâmicas sutis, especialmente em cenários onde o sinal temporal já é limitado.

É importante destacar, entretanto, que o desempenho da LSTM robusta foi superior ao de uma versão *baseline* simplificada (com apenas um *backbone* e uma camada LSTM, sem TCN ou atenção), que apresentou resultados ainda mais baixos. Esse achado sugere que os aprimoramentos de fato contribuíram para mitigar parte das limitações, mas não foram suficientes para superar o forte *baseline* oferecido pelas arquiteturas de processamento frame a frame neste dataset.

V. Comparações com Modelos Existentes no CMU-MOSEI

Além das abordagens testadas neste estudo (LSTM robusta, ResNet50, YOLOv11-cls e ViT-Large), outros modelos relevantes foram avaliados diretamente no dataset CMU-MOSEI. A seguir, destacam-se alguns trabalhos e seus resultados:

- **M3ER** [27]: utiliza fusão multiplicativa de múltiplas modalidades (visão, texto, fala) e alcançou aproximadamente 89,0% de acurácia, cerca de 5 p.p. acima de abordagens anteriores.
- **Attention-based multimodal CNN (Mamieva et al.)** [28]: modelo avaliado também no CMU-MOSEI atingiu $\sim$80,7% de weighted accuracy (WA) e F1 $\approx$ 73,7%.
- **Resultados deste estudo**: entre os modelos avaliados apenas na modalidade visual, o *Vision Transformer* (ViT-Large) obteve o melhor desempenho, com $\sim$78,4% de acurácia e precisão. O YOLOv11-cls atingiu cerca de 71,8%, seguido da ResNet50 com $\sim$62,7%. A LSTM robusta apresentou desempenho inferior, em torno de 53,5%, mesmo após as melhorias propostas.

Esses modelos mostram que estratégias multimodais (especialmente fusão cruzada e pré-treinamento multimodal) têm

potencial para superar métodos baseados apenas na modalidade visual.

Essas comparações situam os resultados atuais dentro do panorama mais amplo da literatura em reconhecimento de emoção por vídeo, ressaltando como as abordagens *frame-based* ainda podem competir de forma eficaz, mas ficam atrás de arquiteturas multimodais avançadas.

VI. Conclusões

Este trabalho apresentou uma comparação entre diferentes abordagens para análise de vídeo no contexto do dataset CMU-MOSEI, explorando tanto modelos que capturam explicitamente a dimensão temporal (LSTM) quanto arquiteturas originalmente desenvolvidas para imagens estáticas (ResNet50, ViT e YOLOv11) adaptadas por meio de estratégias de agregação de frames. Foram utilizadas as métricas de acurácia e precisão como critérios de avaliação.

Os resultados mostraram que o *Vision Transformer* (ViT-Large) obteve o melhor desempenho geral, atingindo valores de aproximadamente 78% em ambas as métricas. É importante ressaltar, entretanto, que neste trabalho o ViT foi aplicado em uma configuração baseada em frames individuais, sem explorar diretamente mecanismos espaço-temporais como os de variantes especializadas (e.g., ViViT, TimeSformer). O ganho observado decorre, portanto, da expressividade espacial das representações extraídas pelos *patches* e da capacidade do modelo em generalizar padrões visuais complexos, mesmo quando as dependências temporais não foram explicitamente modeladas. O modelo YOLOv11 apresentou desempenho intermediário (cerca de 72%), confirmando sua competitividade em cenários que demandam eficiência e menor tempo de inferência. A ResNet50 alcançou valores próximos a 62%, destacando-se como uma alternativa estável de menor complexidade. Por outro lado, a LSTM obteve o pior resultado (cerca de 53%), o que sugere que a simples modelagem temporal não foi suficiente para superar arquiteturas modernas de visão computacional neste dataset específico.

De forma geral, os achados reforçam que a escolha do modelo deve considerar as diferenças de precisão e aplicabilidade. Enquanto o ViT é indicado para aplicações em que a máxima acurácia é essencial, o YOLOv11 e a ResNet50 oferecem alternativas competitivas em cenários que demandam soluções mais simples. A LSTM, embora historicamente relevante, mostrou limitações práticas neste estudo.

A. Trabalhos Futuros

Como desdobramentos naturais deste estudo, propomos as seguintes linhas de continuidade:

- **Integração multimodal (CMU-MOSEI):** incorporar áudio e texto além da visão, testando *early-/late-fusion* e fusão via atenção cruzada (transformers multimodais). Avaliar ganhos em relação ao uso exclusivo da modalidade visual.
- **Métricas e protocolo estatístico:** reportar macro-/weighted F1, recall, balanced accuracy, AUC-ROC/PR (por classe e macro), matriz de confusão, calibração (ECE/Brier). Executar múltiplas rodadas (p. ex., $k \geq 5$ sementes), com ICs de 95% (bootstrap) e testes de significância pareados (McNemar/Bootstrap por vídeo).
- **Ablations da LSTM e variações temporais:** varrer comprimento de janela T, número de camadas/unidades e cabeças de atenção; comparar GRU/TCN puro; testar regularizações temporais (zoneout, layer norm recorrente, dropconnect) e *temporal masking*. Medir o custo/benefício versus a agregação simples de frames.
- **Arquiteturas espaço-temporais:** incluir 3D CNNs (C3D/I3D/R(2+1)D), SlowFast, Video Swin e MViT, bem como transformers específicos para vídeo (ViViT [8], TimeSformer [12]). Comparar com o *frame aggregation* atual em acurácia e custo computacional.
- **Generalização entre domínios:** validar em outros conjuntos de emoção em vídeo (e.g., AFEW, RAVDESS) e/ou *in-the-wild*; realizar *stress tests* (oclusão, iluminação, desfoque) e análise de *domain shift*.
- **Pré-processamento orientado a rosto:** comparar *full-frame* (configuração atual) com detecção/rastreamento facial e alinhamento, quantificando o impacto na robustez e na sensibilidade a enquadramento/iluminação.
- **Eficiência e *deployment*:** medir throughput/latência ponta-a-ponta (por vídeo), uso de memória e energia; aplicar compressão (poda estruturada), quantização (PTQ/QAT) e distilação para *backbones* leves (MobileNetV4 [24], EfficientNetV2 [23]); exportar para ONNX/TensorRT.
- **Novas tarefas e rótulos:** além de classificação discreta, explorar regressão de intensidade e traços contínuos (valência/ativação), e estratégias contra *label noise*.
- **Benchmark de custo:** relatar *wall-clock*, épocas até *early stopping*, custo por amostra e curva de eficiência (Acurácia $\times$ FLOPs/latência) para escolhas informadas em cenários reais.

Essas perspectivas reforçam que ainda há espaço significativo para avanços na área, especialmente em direções que conciliem desempenho, eficiência e integração multimodal.

Referências

[1] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[2] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys*, 2022.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[4] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.
[5] J. Y.-H. Ng *et al.*, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015, pp. 4694–4702.
[6] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.
[8] A. Arnab *et al.*, "Vivit: A video vision transformer," in *ICCV*, 2021, pp. 6836–6846.
[9] M. Rahima, M. Karim, M. Al-Amin, T. Hossain, and M. Uddin, "A comprehensive review of yolo models for object detection," *arXiv preprint arXiv:2501.04665*, 2025. [Online]. Available: https://arxiv.org/abs/2501.04665
[10] Y. Zhang *et al.*, "Comparative performance of yolov8, yolov9, yolov10, yolov11 and faster r-cnn models for detection of multiple weed species," *Smart Agricultural Technology*, vol. 9, p. 100533, 2024.
[11] X. Tian, H. Wang, Z. Li, J. Chen, and Y. Xu, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2501.01599*, 2025. [Online]. Available: https://arxiv.org/abs/2501.01599
[12] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021, pp. 813–824.
[13] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 3202–3211.
[14] A. Zadeh *et al.*, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018, pp. 2236–2246.
[15] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.
[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[18] A. Karpathy *et al.*, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.
[19] D. Tran *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.
[20] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.
[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[22] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 6824–6835.
[23] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning (ICML)*, 2021.
[24] A. Howard *et al.*, "Mobilenetv4: Universal efficient convnets for mobile vision," arXiv preprint, 2023.
[25] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019, pp. 8026–8037.
[26] Autores, "Hybrid efficientnet-b7 with tcn and lstm for video-based engagement detection," *International Journal of Novel Research and Development (IJNRD)*, 2025. [Online]. Available: https://ijnrd.org/papers/IJNRD2503120.pdf
[27] T. Mittal, U. Bhattacharya, R. Chandra, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition," in *Proceedings of AAAI*, 2020, relatório em slides disponível em SlideShare sobre fusão multiplicativa de características multimodais no CMU-MOSEI.
[28] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, p. 5475, 2023.