Davi Baechtold Campos

Temporal Transformer-Based Gesture Recognition from Monocular 2D Skeleton Sequences

Curitiba
2025

Davi Baechtold Campos

Temporal Transformer-Based Gesture Recognition from Monocular 2D Skeleton Sequences

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação, pelo Curso de Engenharia de Computação da Pontifícia Universidade Católica do Paraná.

Pontifícia Universidade Católica do Paraná — PUCPR
Campus Curitiba
Curso de Engenharia de Computação

Orientador: Prof. Dr. Alceu de Souza Brito Junior

Curitiba
2025

Aos meus pais, pelo amor, apoio e incentivo incondicional.
Aos meus amigos, pela jornada compartilhada.

Agradecimentos

Agradeço primeiramente

Ao meu orientador, Prof. Dr. Alceu de Souza Brito Junior, pela paciência, conhecimento compartilhado e pela orientação fundamental para a realização deste trabalho.

Aos meus pais e minha família, por todo o suporte, amor e incentivo que foram a base para que eu chegasse até aqui.

Aos meus amigos e colegas de curso, pelas longas noites de estudo, pela ajuda mútua e pelos momentos de descontração que tornaram a caminhada mais leve.

A todos os professores do curso de Engenharia de Computação da PUCPR, pelos ensinamentos que moldaram minha formação profissional e pessoal.

Resumo

O reconhecimento de gestos dinâmicos a partir de sequências de vídeo é um desafio central na área de Interação Humano-Computador (IHC), com aplicações que vão desde interfaces de controle até sistemas de assistência para pessoas com deficiência auditiva. Abordagens tradicionais frequentemente lutam para capturar as complexas dependências temporais inerentes aos gestos. Este trabalho propõe um modelo baseado na arquitetura Transformer para o reconhecimento de gestos a partir de sequências de coordenadas esqueléticas 2D extraídas de vídeos monoculares. O método utiliza o MediaPipe para a extração de pontos-chave das mãos e corpo, que são então processados por um codificador Transformer para classificação. Avaliamos nosso modelo em um dataset de gestos dinâmicos, comparando seu desempenho com baselines como LSTMs e MLPs. Os resultados demonstram a eficácia da arquitetura Transformer em modelar relações temporais, alcançando uma acurácia superior e destacando seu potencial para aplicações em tempo real.

Resumo

Dynamic gesture recognition from video sequences is a central challenge in Human-Computer Interaction (HCI), with applications ranging from control interfaces to assistive systems for the hearing impaired. Traditional approaches often struggle to capture the complex temporal dependencies inherent in gestures. This work proposes a model based on the Transformer architecture for gesture recognition from sequences of 2D skeleton coordinates extracted from monocular videos. The method utilizes MediaPipe for hand and body keypoint extraction, which are then processed by a Transformer encoder for classification. We evaluate our model on a dynamic gesture dataset, comparing its performance against baselines such as LSTMs and MLPs. The results demonstrate the effectiveness of the Transformer architecture in modeling temporal relationships, achieving superior accuracy and highlighting its potential for real-time applications.

Sumário

1 Introdução

O reconhecimento de gestos humanos é uma área de pesquisa fundamental em visão computacional e Interação Humano-Computador (IHC). A capacidade de um sistema computacional interpretar gestos dinâmicos abre um vasto leque de aplicações, desde o controle de dispositivos sem toque até a tradução automática de línguas de sinais, promovendo maior acessibilidade e novas formas de interação.

Contudo, o reconhecimento de gestos a partir de vídeos 2D apresenta desafios significativos. Gestos são, por natureza, sequências temporais de posturas, e a captura das dependências de longo alcance entre os movimentos é crucial para uma classificação precisa. Além disso, variações na velocidade de execução, iluminação, oclusão e pontos de vista podem dificultar a tarefa.

Modelos clássicos baseados em Redes Neurais Recorrentes (RNNs), como LSTMs, têm sido amplamente utilizados para essa tarefa. No entanto, a arquitetura Transformer (VASWANI et al., 2017), originalmente proposta para tarefas de Processamento de Linguagem Natural (PLN), emergiu como uma alternativa poderosa devido ao seu mecanismo de auto-atenção, que permite modelar relações entre todos os pontos de uma sequência, independentemente da distância.

Este trabalho explora o uso da arquitetura Transformer para o reconhecimento de gestos dinâmicos. A abordagem se baseia na extração de coordenadas esqueléticas 2D das mãos e do corpo utilizando a biblioteca MediaPipe (LUGARESI et al., 2019), que oferece uma representação robusta e de baixa dimensionalidade do movimento. Essas sequências de pontos-chave são então alimentadas em um modelo Transformer para classificação.

1.1 Objetivos

O objetivo principal deste trabalho é desenvolver e avaliar um modelo baseado em Transformer para o reconhecimento de gestos a partir de sequências de pontos-chave 2D. Os objetivos específicos são:

- Implementar um pipeline para extração de coordenadas esqueléticas de vídeos usando MediaPipe.
- Desenvolver um modelo de classificação de gestos baseado na arquitetura Transformer.
- Treinar e avaliar o modelo em um dataset de gestos dinâmicos.
- Comparar o desempenho do modelo Transformer com arquiteturas de baseline, como MLP e LSTM.

1.2 Justificativa

A crescente necessidade por interfaces mais naturais e acessíveis impulsiona a pesquisa em reconhecimento de gestos. A arquitetura Transformer, com seu sucesso comprovado em domínios sequenciais como o PLN (DEVLIN et al., 2018), apresenta uma oportunidade promissora para avançar o estado da arte no reconhecimento de gestos, superando algumas das limitações de modelos recorrentes. Este estudo busca validar essa hipótese, fornecendo uma análise comparativa em um contexto prático.

1.3 Estrutura do Documento

Este trabalho está organizado da seguinte forma: O Capítulo 2 apresenta o referencial teórico sobre as tecnologias utilizadas. O Capítulo **??** descreve a metodologia de desenvolvimento. O Capítulo **??** apresenta os resultados obtidos e, finalmente, o Capítulo **??** traz as conclusões e trabalhos futuros.

2 Referencial Teórico

Este capítulo aborda os conceitos fundamentais que sustentam este trabalho. Iniciamos com uma visão geral sobre o reconhecimento de gestos, seguido por uma descrição das principais tecnologias empregadas: a biblioteca MediaPipe para extração de características e a arquitetura Transformer, que constitui o núcleo do nosso modelo.

2.1 Reconhecimento de Gestos

...

2.2 MediaPipe

O MediaPipe é um framework de código aberto desenvolvido pelo Google que permite a construção de pipelines de processamento de dados multimodais, como vídeo, áudio e sensores (LUGARESI et al., 2019). Ele oferece soluções pré-treinadas e otimizadas para tarefas de visão computacional, como detecção de faces, mãos e pose corporal.

Neste trabalho, utilizamos a solução de detecção de mãos e pose do MediaPipe para extrair as coordenadas 2D dos pontos-chave (keypoints) que representam a estrutura esquelética das mãos e do corpo. Essa abordagem transforma os dados de vídeo brutos em uma representação sequencial e estruturada, ideal para ser processada por modelos de aprendizado de máquina.

2.3 Arquitetura Transformer

Proposta originalmente por Vaswani et al. (VASWANI et al., 2017), a arquitetura Transformer revolucionou o campo do Processamento de Linguagem Natural e desde então tem sido adaptada para diversas outras áreas, incluindo visão computacional.

Diferentemente das arquiteturas recorrentes (RNNs) que processam os dados sequencialmente, o Transformer processa a sequência inteira de uma só vez, utilizando um mecanismo chamado de **auto-atenção (self-attention)**. Esse mecanismo permite ao modelo ponderar a importância de cada elemento da sequência em relação a todos os outros, capturando dependências complexas e de longo alcance de forma mais eficaz que as RNNs.

A arquitetura é composta principalmente por blocos de codificador (Encoder) e decodificador (Decoder). Para tarefas de classificação de sequência, como a nossa, apenas a pilha de codificadores é necessária. Cada codificador é composto por uma camada de

auto-atenção multi-cabeça (multi-head self-attention) e uma rede feed-forward totalmente conectada.

Referências

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2018. 7

LUGARESI, C. et al. Mediapipe: A framework for building multimodal applied ml pipelines. In: *Proceedings of the 3rd International Conference on Machine Learning and Computer Science*. [S.l.: s.n.], 2019. 6, 8

VASWANI, A. et al. *Attention Is All You Need*. 2017. 6, 8