RECONHECIMENTO DE GESTOS DE MÃOS OFENSIVOS EM CONTEXTOS MULTICULTURAIS UTILIZANDO LLM

GUILHERME GOMES LUCCAS RODRIGUES[1],
FABRICIO B. NARCIZO[2,3], MARIO T. SHIMANUKI[4]

[1] Cursando Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP, Câmpus Caraguatatuba, rodrigues.luccas@aluno.ifsp.edu.br
[2] AI Research Scientist, GN Advanced Science, GN One, fbnarcizo@jabra.com
[3] Part-time Lecturer, Computer Science Department, IT University of Copenhagen, København S, Denmark, narcizo@itu.dk
[4] Professor no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, IFSP, Campus Caraguatatuba, mario@ifspcaragua.net

Área de conhecimento (Tabela CNPq): Computabilidade e Modelos de Computação 1.03.01.01-1

RESUMO: Com a evolução dos meios digitais de comunicação, diversas tecnologias surgiram para prevenir a propagação de mensagens ofensivas tanto textuais como corporais. Ambientes como redes sociais e grandes corporações frequentemente envolvem diversas culturas, onde gestos específicos podem ser interpretados de modo diferente entre indivíduos. Diante disso, esse estudo visa criar um algoritmo de reconhecimento de gestos ofensivos das mãos considerando aspectos culturais, por exemplo o flexionamento dos dedos ou a direção do polegar, e a inserção dessas informações em um Modelo de Linguagem em Grande Escala (LLM) para sua análise e identificação do gesto como ofensivo ou não considerando os aspectos culturais envolvendo o país do gesto analisado, para que, gestos como o "joinha" que podem ser ofensivos em países como Iraque, sejam analisados levando em consideração sua cultura. Com isso criando um modelo que possa moderar gestos ofensivos em plataformas virtuais criando um ambiente mais seguro para seus usuários.

PALAVRAS-CHAVE: gestos ofensivos; pontos; detecção; LLM.

1 INTRODUÇÃO

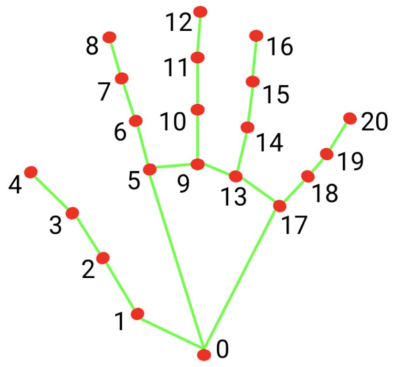
Com o crescente uso das tecnologias de comunicação por vídeos, como redes sociais e videoconferências, diversas questões sobre o seu uso surgiram, sendo uma dessas questões a moderação de conteúdos ofensivos dado a grande quantidade de usuários de diversas culturas que utilizam ambientes virtuais todos os dias. Embora as recentes tecnologias possuam algoritmos capazes de identificar conteúdos ofensivos textuais, a identificação de gestos ofensivos ainda apresenta certa dificuldade por conta da natureza dos gestos humanos, em que podem variar de posição, forma e ainda por cima a cultura pode influenciar todo o significado de um gesto [1], já que um gesto ofensivo em uma região pode ser considerado não ofensivo ou ter um significado completamente diferente em outra, por exemplo o gesto "polegar para cima" (thumb up em inglês) mostrado na figura 1, amplamente conhecido como um gesto positivo em muitos lugares, pode ser interpretado de forma ofensivos em certos países do Oriente Médio como o Iraque.

Figura 1: Exemplo de uma das imagens representando o gesto *thumbs up*

Em meio a essa grande variedade de gestos de mãos ofensivas, em locais como grandes redes sociais ou empresas com alcance mundial, é possível que existam confusões em certos momentos por conta da diferença de cultura entre essas pessoas. Levando em consideração esse problema na detecção de gestos de mão ofensivos, podemos utilizar Modelos de Linguagem de Larga Escala (*Large Language Model* – LLM), devido a sua capacidade de interpretar dados como textos e imagens simultaneamente. As LLMs seriam uma ferramenta de grande ajuda para detectar gestos ofensivos [2 e 3], considerando o contexto cultural de cada país.

2 TEORIA

Para identificar gestos de mão, é necessário o uso de tecnologias que ajudem no mapeamento dos pontos de referência da mão. Esses pontos representam diferentes regiões da mão, que são categorizadas de 0 a 20, sendo no total 21 pontos [4], como ilustrado na Figura 2. Esse mapeamento pode ser feito pelo uso da ferramenta Media Pipe que utiliza diversas técnicas de Machine Learning e IA para implementar diversas soluções, sendo uma delas a função Hands, que possibilita a detecção das mãos, possibilitando o desenvolvimento deste estudo.



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figura 2: Categorização dos 21 pontos, seus respectivos nomes e sua localização na mão humana pelo media pipe
Fonte: https://mediapipe.readthedocs.io/en/latest/solutions/hands.html

A partir da identificação dos pontos nas mãos humanas, é possível desenvolver algoritmos para a detecção de gestos ou ações específicas das mãos, por exemplo: se uma pessoa está com um dedo flexionado, ou com a mão fechada, a orientação da palma

da mão, etc. É possível desenvolver modelos para aprimorar ferramentas existentes e criar novas tecnologias com objetivos específicos, como, por exemplo, a detecção de gestos da Língua Brasileira de Sinais (LIBRAS) [5] e [6], ou, como abordado nessa pesquisa, o reconhecimento de gestos [7] ofensivos.

3 MATERIAL E MÉTODOS

A metodologia utilizada constitui-se em duas etapas: 1) criação de um algoritmo para detectar movimentações de pontos de interesse localizados nas mãos – como flexão dos dedos, ou orientação da palma; e 2) a inserção dos resultados obtidos em uma LLM, para analisar e identificar os gestos como ofensivo ou não ofensivo, levando em consideração o país e contexto cultural daquele gesto, a Figura 3 demonstra um fluxograma contendo todos os passos necessários para a elaboração do projeto.

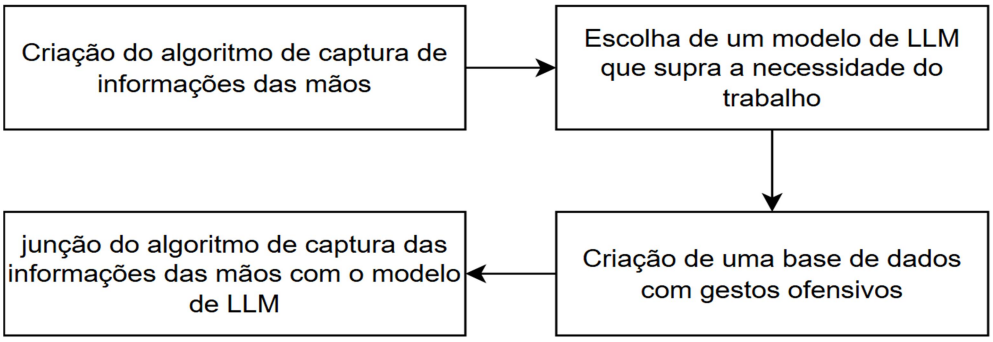


Figura 3: Fluxograma contendo todas as etapas feitas durante o projeto

Para a criação do algoritmo de detecção de padrões das mãos, foram implementadas seis regras de detecção [8] de acordo com a Tabela 1, sendo estas: (1) a verificação dos dedos flexionados, aplicando em todos os dedos individualmente e com seu retorno sendo se o alvo está flexionado, estendido ou está em um meio termo; (2) verificação da proximidade entre os dedos, com um retorno dizendo se os dedos estão juntos, separados ou quase juntos; (3) verificação individual do contato dos dedos com o polegar, retornando se algum dedo está em contato, em um meio termo, ou totalmente separado; (4) detecção da direção para qual o polegar está sendo apontado, retornando se o polegar está direcionado para cima, para baixo, e, caso não esteja em nenhuma das duas direções, será considerado que está flexionado ou em outro sentido; (5) orientação da palma, em que é verificado a direção para qual a palma da mão está direcionada, sendo: esquerda, direita, baixo, cima, dentro, ou fora; e (6) retorno da posição da mão com base nas coordenadas dos pontos de interesse.

Tabela 1: Regras de detecção implementadas durante o projeto
fonte:GestureGPT: Toward Zero-shot Free-form Hand Gesture Understanding with Large Language Model Agents [8]

Regra	Aplicável a	Retorno
Flexão dos Dedos	Polegar, Indicador, Médio, Anelar, Mínimo	1: Estendido 0: Entre -1: Curvado
Proximidade dos Dedos	Indicador-Médio, Médio-Anelar, Anelar-Mínimo	1: Juntos 0: Entre -1: Separados
Contato do Polegar	Polegar-Indicador, Polegar-Médio, Polegar-Anelar, Polegar-Mínimo	1: Contato 0: Entre -1: Sem Contato
Direção do Polegar	Polegar	1: Para Cima -1: Para Baixo 0: Outras Direções/Curvado
Orientação da Palma	Palma	Codificação: [Esquerda, Direita, Para Baixo, Para Cima, Para Dentro, Para Fora] Zeros: Desconhecido
Posição da Mão	Mão	Coordenadas em Ponto Flutuante

A partir das regras definidas, o projeto foi desenvolvido usando-as como funções. Cada função utiliza suas entradas e retorna os resultados correspondentes, como ilustrado na tabela 1 nas colunas "*applicable to*" e "*value*". Para construir essas funções, comparações entre pontos da mão são realizadas, utilizando fórmulas como a distância euclidiana para calcular a distância entre dois pontos, e operações envolvendo ângulos para determinar direções específicas dos dedos. A tabela 2 apresenta um exemplo do resultado obtido ao analisar uma imagem do gesto *"thumbs up"*

Tabela 2: Demonstração dos resultados obtidos por uma das imagens

Informação captada	Lado da mão	Resultado
Flexão dos Dedos	Esquerdo	[1, -1, -1, -1, -1]
Proximidade dos Dedos	Esquerdo	[-1, -1, 0]
Contato do Polegar	Esquerdo	[-1, -1, -1, -1]
Direção do Polegar	Esquerdo	1
Orientação da Palma	Esquerdo	[-1, -1, -1, -1]

Além da criação das funções para obtenção de informações das mãos, utilizamos um modelo de Linguagem de Grande Escala (LLM) para identificar gestos. Foram testadas diversas LLMs focadas em visão computacional usando o aplicativo *LMStudio*, que facilita a procura e utilização de modelos LLM para diferentes finalidades Durante os testes optou-se na utilização do modelo *Llava:13b* do Ollama [9] devida sua capacidade em trabalhar com a análise de imagens. Após a escolha da LLM foi iniciado o desenvolvimento do sistema, onde, primeiramente, foram adquiridas diversas imagens disponíveis na internet em sites de imagens de uso livre como Pexels [10] e Unsplash [11] e por conta de uma pouca quantidade de imagens somente com os gestos, imagens de autoria própria foram inseridas para a criação de uma base de dados com alguns gestos ofensivos e não ofensivos totalizando 280 imagens, para a realização dos testes iniciais.

Com a procura das imagens concluída, os testes com o modelo de LLM foram iniciados. O primeiro passo foi criar um código para definir um *prompt* capaz de interpretar a imagem enviada, considerando o contexto cultural do país inserido junto com a imagem, a Figura 4 mostra algumas imagens do gesto *"thumbs up"* utilizadas como teste durante o estudo.

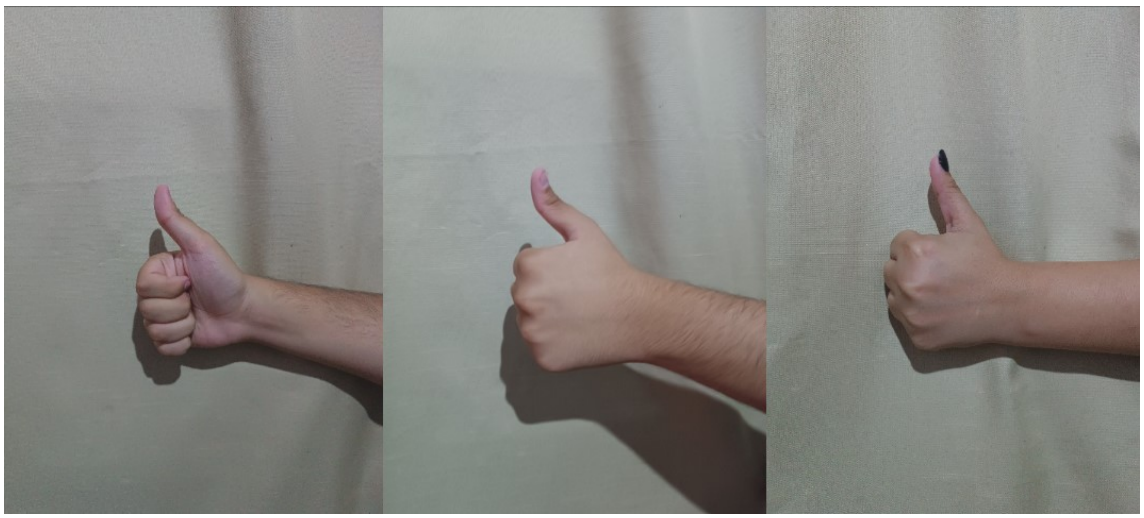


Figura 4: Imagens do gesto "*thumbs up*" utilizadas durante o projeto

Como mostrado na Tabela 2, o prompt de sistema foi usado para contextualizar a tarefa, enquanto o prompt de usuário incluía a variável 'country' para especificar o país. O modelo então classificava a imagem com base no contexto cultural local. A Tabela 3 apresenta os resultados obtidos com os prompts iniciais para o gesto 'joinha' nos contextos culturais do Brasil e do Iraque usando o gesto "*thumbs up*".

Tabela 3: Resultados obtidos pela LLM do gesto "*thumbs up*" no contexto cultural do Brasil e Iraque

Prompt de Sistema	Prompt de usuário	País	Resultado
"You are an image classifier. Your job is to classify images as either 'offensive' or 'not offensive'. Respond ONLY with one of these words"	"In this {country}, the hand gesture is: offensive or not offensive"	Brasil	"Not Offensive"
"You are an image classifier. Your job is to classify images as either 'offensive' or 'not offensive'. Respond ONLY with one of these words"	"In this {country}, the hand gesture is: offensive or not offensive"	Iraque	"Not offensive"

Após os primeiros testes com os *prompts*, notamos que a LLM tinha uma dificuldade em interpretar o país e o retorno esperado, por conta disso foi elaborado um conjunto de parâmetros para garantir que o modelo consiga interpretar tanto a imagem como o país em que ela deve levar em consideração ao analisar, com isso, chegamos nos parâmetros: (1) seu trabalho, onde é inserido na LLM o que ela é e o que deve fazer, nesse caso sendo uma LLM focada em analisar imagens levando em consideração o contexto multicultural; (2) o que receber, nesse caso a imagem e o país de origem; (3) como interpretar, sendo esse parâmetro de grande importância para que a LLM entenda que ela deva utilizar o parâmetro do país para interpretar a imagem; e (4) como retornar o resultado, permitindo uma análise estruturada e formatação adequada para posterior

avaliação. Na Tabela 4, é mostrado o prompt revisado, com uma explicação mais detalhada tanto no prompt de sistema quanto no de usuário, usando as variáveis 'country' e 'gesture' para melhorar a interpretação da LLM."

Tabela 4: Resultados obtidos pela LLM do gesto “thumbs up” no contexto cultural do Brasil e Iraque usando um prompt mais elaborado

Prompt de Sistema	Prompt de usuário	País	Resultado
"You are an expert in cultural hand gestures from different countries. In the context of {country}, classify the hand gesture as either 'offensive' or 'not offensive'. Respond ONLY with one word: either 'offensive' or 'not offensive'. Provide no explanations."	in {country} the hand gesture {gesture} is: 'offensive' or 'not offensive'? Respond ONLY with one word: either 'offensive' or 'not offensive'"	Brasil	”Not offensive”
"You are an expert in cultural hand gestures from different countries. In the context of {country}, classify the hand gesture as either 'offensive' or 'not offensive'. Respond ONLY with one word: either 'offensive' or 'not offensive'. Provide no explanations."	in {country} the hand gesture {gesture} is: 'offensive' or 'not offensive'? Respond ONLY with one word: either 'offensive' or 'not offensive'"	Iraque	”Offensive”

Com o prompt final finalizado, começamos os teste iniciais utilizando como base os países Iraque e Brasil, focando inicialmente em quatro gestos: *“thumbs up”*, *“OK sign”*, *“middle finger”* e *“rock and roll”*. O principal foco foi o gesto *“thumbs up”,* que é positivo no Brasil, mas ofensivo no Iraque[1]. A Tabela 5 ilustra os resultados iniciais usando o gesto *“thumbs up”* e *“middle finger”* no contexto dos países Brasil e Iraque usando exatamente 40 imagens de cada gesto, onde anteriormente houve muitos erros de interpretação do prompt envolvendo o país Iraque. No entanto, com o prompt aprimorado, os resultados melhoraram significativamente.

[1] https://www.businessinsider.nl/hand-gestures-offensive-different-countries-2018-6

Tabela 5: Respostas do modelo referente à análise
Dos gestos “thumbs up” e “middle finger”

	Iraque Thumbs up	Brasil Thumbs up	Iraque Middle Finger	Brasil Middle finger
Total de dados:	40			
Porcentagem de ofensivos	100%	20%	100%	100%
Porcentagem de não ofensivo	0%	80%	0%	0%

5 CONSIDERAÇÕES FINAIS

Com a utilização do prompt mais detalhado, o modelo demonstrou grande eficiência na maior parte dos testes com os gestos utilizados durante todo o processo de desenvolvimento, promovendo uma inovação na análise de gestos, pois mesmo em trabalhos que envolvam o reconhecimento de gestos, poucos trabalham especificamente com gestos ofensivos, como mostrado na tabela 6 que compara esse trabalho com o trabalho de XIN, Z. et al [7]

Tabela 6: Comparação entre este trabalho e o artigo GestureGPT de XIN, Z. et al

Atividade relacionadas	Trabalhos comparados			
	XIN, Z (2023)	Das (2023)	Silva (2022)	Guilherme (2024)
Reconhecimento de gestos	X	X	X	X
Classificação de ofensividade				X
Contexto multicultural				X

Em trabalhos futuros, espera-se que o algoritmo de informações das mãos seja integrado com a LLM capaz de identificar os gestos, dispensando a análise visual da imagem e focando exclusivamente nas informações e o envio de variáveis culturais selecionadas, como PIB, média de idade ou proficiência em inglês, pois no momento o contexto do países está sendo considerado pelo próprio modelo, com o algoritmo criado, ele terá a capacidade de auxiliar na identificação de gestos ofensivos e prevenir propagação de ofensas em vídeos em redes sociais, videochamadas, streamings em canais de entretenimento ou até em noticiários ao vivo, em alguns casos, sendo possível criar ambientes menos agressivos para seus usuários evitando situações que gestos poderiam ser usados para ofender outras pessoas, com isso, cumprindo o objetivo deste estudo, criando um modelo de reconhecimento de gestos ofensivos levando em consideração o contexto multicultural.

REFERÊNCIAS

[1] ARCHER, D. Unspoken diversity: Cultural differences in gestures. Disponível em: <https://qualquant.org/wp-content/uploads/video/1997%20Archer79-105.pdf>. Acesso em: 8 out. 2024.

[2] Large language models in textual analysis for gesture selection. [s.d.]. Disponível em: <https://ar5iv.labs.arxiv.org/html/2310.13705>. Acesso em: 8 out. 2024.

[3] Probing language models' gesture understanding for enhanced human-AI interaction. [s.d.]. Disponível em: <https://ar5iv.labs.arxiv.org/html/2401.17858>. Acesso em: 8 out. 2024.

[4] Guia de detecção de pontos de referência do rosto. Disponível em: <https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker?hl=pt-br>. Acesso em: 4 out. 2024.

[5] SILVA, R. P. DA. Visão computacional: um estudo de caso aplicado à língua brasileira de sinais (LIBRAS). 2022.

[6] DAS, A. et al. Development of a real time vision-based hand gesture recognition system for human-computer interaction. 2023 IEEE 3rd Applied Signal Processing Conference (ASPCON). Anais...IEEE, 2023.

[7] INDRIANI; HARRIS, M.; AGOES, A. S. Applying hand gesture recognition for user guide application using MediaPipe. Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021). Anais...Paris, France: Atlantis Press, 2021.

[8] XIN, Z. et al. GestureGPT: Toward zero-shot interactive gesture understanding and grounding with large language model agents. 2023. Disponível em: <http://arxiv.org/abs/2310.12821>. Acesso em: 8 out. 2024

[9] LIU, H et. al. LLaVA: Large Language and Vision Assistant Visual Instruction Tuning. In: NEURAL INFORMATION PROCESSING SYSTEMS, 2023, [Local da Conferência]. University of Wisconsin-Madison, Microsoft Research, Columbia University. Disponível em: <https://llava-vl.github.io>. Acesso em: 7 out. 2024.

[10] UNSPLASH. Disponível em: https://www.unsplash.com. Acesso em: 21 nov. 2024.

[11] PEXELS. Disponível em: https://www.pexels.com. Acesso em: 21 nov. 2024.