Sistema de interação computacional baseado na detecção das mãos utilizando classificadores em cascata

Anne Livia da F. Macedo[1], Igor Ruiz Gomes[1]

[1]Faculdade de Computação – Universidade Federal do Pará (UFPA)
Castanhal – PA – Brazil

{annelivia16, ruiz.igor}@gmail.com

Resumo. *Com o avanço da tecnologia, sistemas que possibilitam uma interação humano-computador mais natural e acessível, terão um impacto muito positivo na sociedade. Diante disso, esta pesquisa propõe o desenvolvimento de um processo de software de baixo custo computacional que permite a interação a distância com o computador através da detecção das mãos via webcam, utilizando o algoritmo de classificação Haar Cascade. Os resultados obtidos foram considerados satisfatórios ao indicar uma precisão de 82%, recall de 80%, F1-score de 81% e um tempo de resposta adequado e eficiente para a interação com diversas aplicações computacionais em tempo real.*

1. Introdução

Em conformidade com [Oudah et al. 2020], os gestos com as mãos são uma das formas de comunicação não verbal que podem ser usadas em diversos campos, como controle de robôs e interação humano-computador (IHC). Para [Sun et al. 2020], a simplificação do processo de interação humano-computador tornou-se um ponto de pesquisa muito relevante quando se trata de sistemas de controle inteligente. Segundo [Cardoso et al. 2015], as soluções baseadas em gestos com as mãos mais eficazes fazem uso de dispositivos eletromecânicos como luvas ou sensores. Contudo, esses equipamentos são normalmente muito custosos, exigem procedimentos de calibração complexos, e não fornecem uma interface totalmente intuitiva.

Com o avanço expressivo da tecnologia da informação nos últimos anos, os esforços para o desenvolvimento de sistemas que permitam uma interação com as máquinas de maneira mais natural e acessível, terão um impacto muito positivo para a população, principalmente no que tange a acessibilidade digital. Diante desse contexto e almejando contribuir com pesquisas na área, este trabalho propõe o desenvolvimento de um sistema de baixo custo computacional que, utilizando apenas a webcam do computador e algoritmos de visão computacional, viabiliza o controle das teclas direcionais através da detecção das mãos do usuário em tempo real.

2. Metodologia

A principal contribuição desta pesquisa consiste na implementação de um sistema de baixo custo computacional, prático e acessível, apto a permitir a interação a distância com o computador através, exclusivamente, da detecção das mãos (gesto estático), sem a necessidade do uso de hardwares e sensores externos, complexos, de alto custo e não acessíveis.

Para concretizar a proposta do projeto, foi implementado um programa que permite que o usuário possa acionar as teclas direcionais através do posicionamento das mãos em determinadas regiões do vídeo que é capturado em tempo real pela webcam do computador, utilizando um algoritmo de aprendizagem de máquina supervisionado e algumas técnicas de processamento digital de imagens. O desenvolvimento foi composto de três etapas principais: (1) treinamento do algoritmo de classificação *Haar Cascade*; (2) aplicação do classificador treinado para a detecção das mãos; (3) definição das regiões responsáveis por representar cada uma das teclas de interesse (cima, baixo, esquerda e direita). O programa foi desenvolvido através da linguagem de programação C++, com o suporte dos algoritmos pré-implementados da biblioteca OpenCV 4.1.0. Além disso, o treinamento do classificador foi realizado baseando-se no trabalho de [Rezaei 2014], que disponibiliza aplicações essenciais para gerar a cascata de classificadores *Haar-Like*.

2.1. Treinamento do classificador Haar Cascade

O *Haar Cascade* é um algoritmo para detecção de objetos que tem como fundamento a abordagem de aprendizagem de máquina proposta por [Viola e Jones 2001], em que uma função em cascata é treinada a partir de muitas imagens positivas e negativas, para posteriormente efetivar a detecção de objetos em outras imagens. Esta abordagem oferece técnicas inteligentes de baixo custo computacional que permite o treinamento para detecção de objetos em circunstâncias adversas como a baixa iluminação [Pereira 2017].

Para efetuar o treino do classificador são necessários um conjunto de amostras que exibe o alvo de interesse (imagens positivas) e um conjunto que não apresenta o alvo (imagens negativas). Portanto, para compor as amostras positivas, foram consideradas 2496 imagens contendo apenas as mãos fechadas com a palma voltada para a frente e o polegar ao lado do indicador, conforme a Fig. 1. Algumas imagens foram coletadas através da webcam utilizada no desenvolvimento e outras são oriundas da internet e das bases de dados *Hand Gesture Database* [Maqueda et al. 2015] e *Significant (ASL) Sign Language Alphabet Dataset* [Muvezwa 2019].

Figura 1. Amostras positivas.

Em relação ao conjunto de negativos, foram empregues 7994 imagens, na qual algumas eram aleatórias e outras apresentavam regiões faciais e do corpo humano para indicar um background tematicamente similar aos que são encontrados nas amostras positivas. Esses dados foram obtidos através da internet e do conjunto de imagens *Real and fake face detection* [Yonsei University 2019]. Após a elaboração das amostras, foi realizado o treinamento do classificador que durou aproximadamente 16 horas.

2.2. Detecção das mãos

A captura das mãos foi efetuada através do método *detectMultiscale* pertencente ao *OpenCV* que implementa a classificação em cascata percorrendo a imagem de entrada

em diferentes escalas e retorna, após todos os procedimentos, uma lista de retângulos, indicando os locais onde as mãos foram encontradas com sucesso.

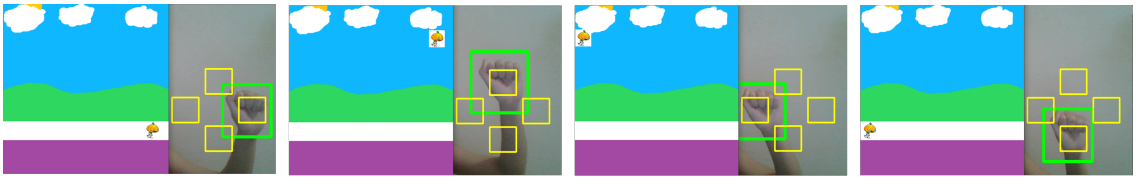
Inicialmente faz-se a conversão do frame obtido em tempo real para a escala de cinza e a equalização do histograma da imagem. Após o pré-processamento, foi especificado como argumentos para os parâmetros do método o valor 1.09 para o fator de escala, que determina o quanto uma imagem será redimensionada em cada uma das escalas. O valor 30 para a quantidade mínima de vizinhos, que objetiva evitar falsas detecções, fazendo com que neste caso, somente os objetos que apresentarem pelo menos 30 detecções sobrepostas sejam considerados, e por fim, foi estabelecido que somente os objetos com dimensões mínimas de 110x110 pixels, serão considerados.

2.3. Interação com o computador

A resolução da webcam integrada é de 640x480 pixels e optou-se pelo espelhamento da imagem de modo que o movimento em tempo real esteja na mesma direção que o movimento realizado pelo usuário. Em seguida, foi mapeada e determinada as regiões de interesse (ROIs) nos frames, de forma que cada uma das áreas indicadas represente uma tecla diferente. Os posicionamentos de cada ROI foram determinados estrategicamente com a intensão de aumentar a agilidade durante os acionamentos sucessivos e permitir uma representação simbólica do que cada área retangular corresponde. O algoritmo avalia qual das teclas deve ser acionada através da detecção da cor do retângulo obtido pelo método de classificação ao localizar as mãos. As coordenadas (x^1, y^1) e (x^2, y^2) das ROIs são: cima (420, 180) e (500, 250); baixo (420, 340) e (500, 410); esquerda (320, 260) e (400, 330); direita (520, 260) e (600, 330).

3. Resultados

Para avaliar os acionamentos sucessivos, foi desenvolvida uma aplicação que permite a movimentação do personagem para as quatro direções. Essa interface possui dimensões de 500x500 pixels e ao especificar uma orientação, o personagem terá um deslocamento de 20px para a direção indicada. A Fig. 2 exibe a interface da aplicação, o resultado da detecção das mãos em tempo real e o processo de acionamento das direções ao posicionar a mão sobre as ROIs referentes as teclas para a direita (a), cima (b), esquerda (c) e para baixo (d), respectivamente.



(a) Direita (b) Cima (c) Esquerda (d) Baixo

Figura 2. Processo de detecção das mãos em tempo real.

Em conformidade com a Fig. 2, para completar todo o percurso do personagem para a direita (a) e para a esquerda (c), são necessários 21 acionamentos cada. Já para cima (b) e para baixo (d) são requeridos 13 acionamentos cada. Baseando-se nesta quantidade de acionamentos das teclas direcionais (68 ao todo), foi apurado o tempo de resposta médio para completar todos os percursos do personagem seguindo a ordem de movimentação da Fig. 2. Três pessoas realizaram o experimento seis vezes e os tempos

médios para o primeiro, segundo e terceiro participante, com precisão de 1/100s, foram de 5.91, 7.18 e 8.27 segundos, respectivamente. O desempenho do detector das mãos foi avaliado com 100 imagens do *dataset* de Maqueda et al. (2015). A Tab. 1 exibe a quantidade de verdadeiros positivos (VP), falsos positivos (FP) e falsos negativos (FN), assim como as taxas de precisão (P), recall (R) e F1-score (F1).

Tabela 1. Desempenho do detector das mãos.

VP	FP	FN	P	R	F1
80	18	20	82%	80%	81%

4. Conclusão

Esta pesquisa propôs o desenvolvimento de um sistema de baixo custo computacional que permite a interação a distância com o computador através da detecção das mãos em tempo real, utilizando um algoritmo de classificação em cascata baseado em características Haar. O modelo proposto alcançou resultados satisfatórios ao propiciar o acionamento das teclas direcionais em um tempo de resposta eficiente e adequado para a execução de inúmeras aplicações computacionais. Ademais, pôde-se concluir que o detector das mãos treinado obteve um desempenho geral apropriado quando considerado os valores obtidos ao computar as métricas de avaliação conhecidas como precisão, recall e F1-score. Para trabalhos futuros, pretende-se aprimorar a performance do sistema e incluir novas funcionalidades como o controle do cursor do mouse.

Referências

Cardoso, T., Delgado, J. and Barata, J. (2015). Hand gesture recognition towards enhancing accessibility. *Procedia Computer Science*, 67:419-429.

Maqueda A. I., et al. (2015). Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns, *Computer Vision and Image Understanding*, Special Issue on Pose & Gesture, 141:126-137.

Muvezwa, K. (2019). Significant (ASL) Sign Language Alphabet Dataset. Disponível em: https://www.kaggle.com/kuzivakwashe/significant-asl-sign-language-alphabet-dataset. Acesso em: 31 ago. 2021.

Oudah, M., Al-Naji, A. and Chahl, J. (2020). Hand gesture recognition based on computer vision: a review of techniques. *Journal of Imaging*, 6(8):73.

Pereira, R. C. (2017). Técnica de rastreamento e perseguição de alvo utilizando o algoritmo Haar Cascade aplicada a robôs terrestres com restrições de movimento. Master's thesis, Universidade Federal do Rio Grande do Norte, Natal, Brasil.

Rezaei, M. (2014). Creating a cascade of haar like classifiers Step by step.

Sun, Y., et al. (2020). Intelligent human computer interaction based on non redundant EMG signal. *Alexandria Engineering Journal,* 59(3):1149-1157.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features, In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition - CVPR 2001*.

Yonsei University, CIPLAB. (2019). Real and Fake Face detection. Disponível em: https://www.kaggle.com/ciplab/real-and-fake-face-detection. Acesso em: 31 ago. 2021.