# Human Emotional Affordance Recognition in Human-Object Interaction using Deep-Learning Models

Hossain, M. , Haque, S. S. , Alam, T. , Rakin, A. A. , Sharife, S. B. , Sharar, S. , Rasel, A. A.
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
{mainul.hossain, shataddru.shyan.haque, tabassum.alam, abdullah.al.rakin, shadman.bin.sharife,
shihab.sharar}@g.bracu.ac.bd, annajiat@bracu.ac.bd

## Abstract

Affordance recognition is the capability to discern the potential action capabilities of objects in an image, which is critical for robot perception and interaction. Emotional affordances, according to the notion, are a recently presented concept that models all of the methods used to collecting & transmitting emotional content in the context of human machine interaction. Emotion recognition, one of several important nonverbal mechanisms by which communication occurs, aids in determining the person's state of mind and condition. In our work, we intend to detect and identify human emotions in various scenarios using various Deep Learning Models, such as identifying what a person, lifting an object such as a knife, may be experiencing, such as happiness/rage/sadness, and predicting the preemptive possible outcome of the scenario based on their emotion. Pre-trained models will be utilized as the project's control basis, while custom trained Deep Learning architectures such as CNN and LSTM will be employed to modify the basic parameters for Video-Image Emotion Recognition. We plan to use 3D AffordanceNet Dataset a benchmark dataset of 23k shapes from 23 semantic Object categories, annotated with 18 visual affordance categories. Classic Feed-Forward CNN models will be used as an architecture for the basis of this input. For emotion processing we will be using Video frames of visual facial expressions selected from Ryerson Audio-Visual Database of Emotional Speech and Song dataset. Basic CNN will be used to extract facial features and for further preservation and processing a LSTM or Transformer based architecture will be used. Both Networks will be jointly trained. Finally, outputs from both benchmarks will be taken and feed into a Recurrent Neural Network to predict dimensional emotions from the video. Initial testing from other research showed a slight problem of overfitting due to the imbalance in the dataset. The problem can be solved by down sampling the dataset. Reducing the number of trained examples should help balance out the performance of the models.

## KEYWORDS

emotional affordance, human-object interaction, emotion recognition, Deep Learning, Pattern Recognition, Video & Image Processing

**References:**

[1] Chang, X., & Skarbek, W. (2021). Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition. Sensors, 21(16), 5452. https://doi.org/10.3390/s21165452

[2] Rangulov, D., &amp; Fahim, M. (2020). Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network. 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS). https://doi.org/10.1109/ipas50080.2020.9334935

[3] Luo, H., Zhai, W., Zhang, J., Cao, Y., &amp; Tao, D. (2021). One-shot affordance detection. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2021/124

[4] Franzoni, V., Milani, A., &amp; Vallverdú, J. (2017). Emotional affordances in human-machine interactive planning and negotiation. Proceedings of the International Conference on Web Intelligence. https://doi.org/10.1145/3106426.3109421