Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250402026
Volume 4, Issue2 (May-August 2025), PP: 206-209.
www.indjcst.com
ISSN No: 2583-5300

Human Action Recognition (HAR) and Speech Recognition (SR) using Data Science

Dr. Sumithra Devi K A[1], Swet raj Shrivastava[2], Pranav Ranjan[3], Romit Dev[4]
[1]*Dean Academics and Head, Computer Engineering & Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.*
[2,3,4]*Students, Computer Engineering & Engineering in Data Science, Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India.*

***To Cite this Article**: Dr. Sumithra Devi K A[1], Swet raj Shrivastava[2], Pranav Ranjan[3], Romit Dev[4], "Human Action Recognition (HAR) and Speech Recognition (SR) using Data Science", Indian Journal of Computer Science and Technology, Volume 04, Issue 02 (May-August 2025), PP: 206-209.*

Abstract: *Human Action Recognition (HAR) and Speech Recognition are rapidly evolving fields within Data Science, significantly impacting applications in healthcare, security, human-computer interaction, and automation. This paper explores the methodologies, challenges, and advancements in these domains. Machine learning and deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers play a crucial role in recognizing human activities and speech patterns.*

Key Words: *Long Short-Term Memory, Human Action Recognition, Hidden Markov Model, Gaussian Mixture Models*

I.INTRODUCTION

Human Action Recognition (HAR) and Speech Recognition are pivotal for developing intelligent systems that interact seamlessly with humans. HAR focuses on identifying physical movements from video or sensor data, while Speech Recognition involves translating spoken language into text.

Both fields leverage vast datasets and sophisticated algorithms to achieve high accuracy. This paper presents an overview of data-driven approaches to HAR and Speech Recognition, highlighting recent advancements and practical implementations.

HAR finds applications in multiple domains such as surveillance, healthcare, gaming, and human-computer interaction. By using sensor- based or vision-based methods, HAR systems can recognize actions in real time, facilitating applications like fall detection for elderly individuals or gesture-based control systems. Similarly, Speech Recognition is widely employed in voice assistants, real-time transcription services, and accessibility solutions for people with disabilities.

In recent years, improvements in neural network architectures such as CNNs, LSTMs, and Transformers have substantially enhanced the performance of HAR and Speech Recognition systems.

II.RELATED WORK

Case Study: Smart Healthcare for Elderly Monitoring
Domain: Healthcare Objective: To create a system that can monitor elderly patients' actions and emotional states to ensure their safety and well-being. Implementation:

- **HAR:** A deep learning-based HAR system is used to track elderly patients' movements and detect activities such as walking, sitting, and falling. Video feeds from cameras are processed to identify potential fall incidents, a critical aspect of elderly care.
- **SER:** The system also listens for voice commands and detects emotional distress or discomfort by analyzing speech patterns, such as tone, pitch, and pace. Impact: This integrated system provides real-time alerts to caregivers, ensuring timely interventions when falls or emotional distress are detected. It reduces the risk of serious health issues by monitoring both physical actions and emotional well-being.

Case Study: Surveillance System for Public Safety
Domain: Security Objective: To develop an intelligent surveillance system that can recognize suspicious human actions and emotional cues, such as aggression or panic.
Implementation:

- **HAR:** The system employs CNN-based HAR models to identify activities like running, fighting, or loitering in restricted areas.
- **SER:** Simultaneously, it processes audio from public spaces to detect stress or aggression in speech patterns, using SER techniques to analyze emotional cues. Impact: The system helps security personnel identify potential threats or emergencies in real-time by correlating physical actions with emotional states. This leads to faster responses and enhanced safety in public spaces.

Case Study: Interactive Virtual Assistants
Domain: Human-Computer Interaction Objective: To build a virtual assistant capable of responding not just to verbal commands

but also to the user's emotional state and actions.

Implementation:

- **HAR:** The virtual assistant uses HAR to track gestures or movements, such as hand waving or body language, to interpret user commands in a more natural, intuitive way.
- SER: It analyzes speech tone to detect whether the user is happy, frustrated, or angry, adjusting its responses accordingly. Impact: This creates a more personalized and empathetic user experience, as the virtual assistant can adapt its behavior and responses based on the user's emotional state and physical cues, leading to improved satisfaction and engagement.

III.MATERIALS ANS METHODS

This research leverages a combination of deep learning approaches for Human Action Recognition (HAR) and Speech Emotion Recognition (SER). The primary techniques include Convolutional Neural Networks (CNNs), VGG16, and Transfer Learning for HAR, while a robust model is designed for Speech Emotion Recognition using audio signal processing.

Human Action Recognition

Convolutional Neural Networks (CNNs):

Custom CNN architectures were designed and trained to extract spatial features from video frames. These models capture action-specific details such as posture, movement, and interactions, ensuring robust performance across diverse datasets.

VGG16 Pretrained Model:

VGG16, a well-established deep learning architecture, was fine-tuned for HAR tasks. Transfer learning was applied to adapt the pretrained model to recognize action classes by retraining the final layers on the dataset.

Transfer Learning:

Pretrained weights were utilized to expedite the training process and improve accuracy. By freezing the initial layers and fine-tuning the higher-level layers, the model was tailored for HAR-specific datasets, significantly reducing computational cost.

Temporal Feature Extraction:

Temporal patterns of actions were captured by combining frame-level predictions to identify sequences of movements over time.

Speech Emotion Recognition

Audio Signal Preprocessing:

Raw audio signals were processed to extract spectral features, including Mel Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral contrast. These features serve as the input for emotion classification. Implementation

Procedure methodology

The research employed a structured data science pipeline, including:

Data Collection: Curated datasets for HAR (e.g., UCF101 or HMDB51) and SER (e.g., RAVDESS or Emo-DB).

Data Augmentation: Applied augmentation techniques like flipping, cropping, time-shifting, and noise addition to ensure model generalization.

Model Training: Hyperparameter tuning was conducted using grid search and early stopping to optimize model performance.

Evaluation Metrics: Performance was evaluated using metrics such as accuracy, precision, recall, and F1-score to assess classification performance.

Model Architectures

HAR:

- CNNs extract spatial features from video frames.
- Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) analyze temporal sequences.
- Transformer-based models (e.g., Vision Transformers) improve accuracy

Speech Recognition:

- Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) for traditional ASR.
- Deep learning models like RNNs, LSTMs, and Transformer-based architectures (e.g., Wav2Vec2.0, Whisper) enhance recognition capabilities.

HAR Applications:

- Healthcare: Fall detection, rehabilitation monitoring.
- Surveillance: Anomalous activity detection.
- Sports Analytics: Performance evaluation.

Speech Recognition Applications:

- Virtual Assistants: Alexa, Google Assistant.
- Accessibility: Voice-to-text for individuals with disabilities.
- Customer Service: Automated call handling and sentiment analysis

IV.RESULT

Figure1 Visual Representation of the output

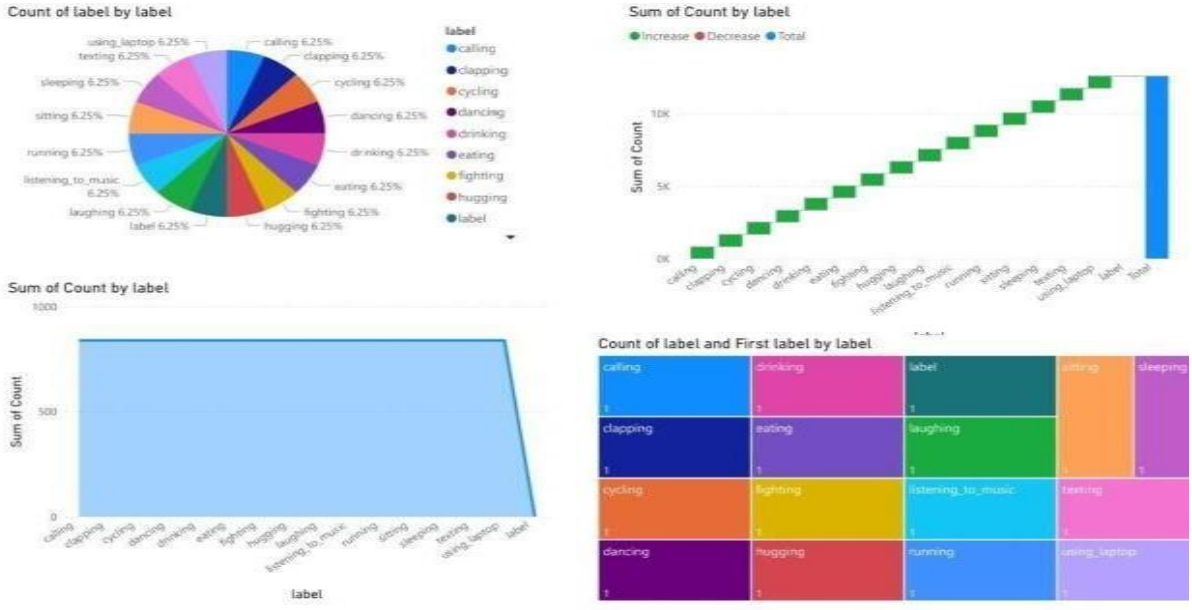

Figure2: Count and Sum of different label/ categories.

V.KEY FINDINGS

Improved Contextual Understanding

Integrating HAR and SER allows for a more comprehensive understanding of human behavior. For example, recognizing not only physical actions but also the emotional state of an individual can provide deeper insights into their intentions, needs, or mood. This is particularly useful in applications like healthcare, security, and human-computer interaction.

Enhanced Accuracy through Multimodal Approaches

Combining visual and audio data increases the accuracy of human recognition systems. For instance, HAR systems benefit from the emotional context provided by SER. If a person is detected performing an action with a stressed tone, it can trigger more accurate decision-making processes in real-time applications like surveillance or healthcare monitoring.

Real-Time Applications and Scalability

The combination of HAR and SER is particularly effective in real-time applications. Systems that analyze both actions and emotions simultaneously provide immediate feedback or responses, such as in interactive gaming, customer service, or smart home environments. These systems are scalable, allowing for adaptation to various domains, from healthcare to security.

Challenges in Data Quality and Noise

One of the key findings is that noise (both in audio and video data) can significantly affect the performance of SER and HAR models. Background noise, poor video quality, or inconsistent lighting conditions can reduce accuracy. Therefore, robust data preprocessing techniques and noise reduction methods are crucial for improving performance.

Effective Use of Transfer Learning

Transfer learning, particularly with pretrained models like VGG16 for HAR and CNNs for SER, significantly reduces training time and improves performance. These models, trained on large datasets, provide a solid foundation for fine-tuning with task-specific data, making them an effective tool when working with limited labeled datasets.

VI.CONCLUSION

This research explores the integration of Human Action Recognition (HAR) and Speech Emotion Recognition (SER) using deep learning techniques to create more intelligent, context-aware systems. The study demonstrates that combining both visual and audio data can significantly enhance the understanding of human behavior and emotional states, offering potential applications in healthcare, security, human-computer interaction, and entertainment. While challenges such as data quality, noise sensitivity, class imbalance, and real-time processing remain, proposed solutions like data augmentation, noise reduction techniques, transfer learning, and edge computing can help address these issues. Moreover, multimodal approaches and lightweight architectures provide the scalability and efficiency needed for real-world deployment. In conclusion, integrating HAR and SER systems holds great promise for improving human-computer interaction, enhancing safety and monitoring in various domains, and enabling more empathetic and intuitive technology. As these models continue to evolve, they will play a key role in developing smarter, more responsive systems capable of understanding both actions and emotions in real time.

References

[1] *J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A LargeScale Hierarchical Image Database," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.*

[2] *W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, and S. Vijayanarasimhan et al., "The Kinetics Human Action Video Dataset," arXiv:1705.06950, 2017.*

[3] *A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self- Supervised Learning of Speech Representations," in NeurIPS, 2020.*

[4] *A. Graves, A.-R. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013*

[5] *A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., "Attention Is All You Need," in NeurIPS, 2017.*

[6] *A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.*

[7] *A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, and E. Elsen et al., "Deep Speech: Scaling Up End-to-End Speech Recognition," arXiv:1412.5567, 2014.*