MultiSensor-Home: A Wide-area Multi-modal Multi-view Dataset for Action Recognition and Transformer-based Sensor fusion

Jun Hyung Lee
8/14/25

*Accepted to IEEE International Conference on Automatic Face and Gesture Recognition 2025

The Problem: Limitations of Traditional Action Recognition

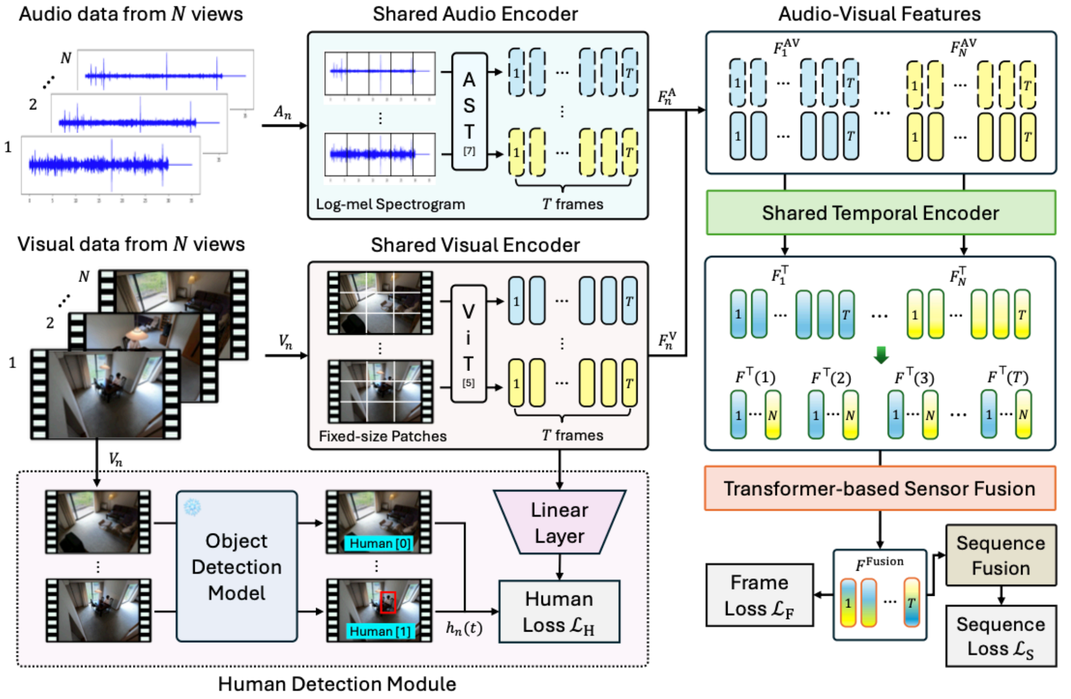
- **Single-View Failures:**
  - A single camera can't capture the full story. Actions can be occluded (blocked) or only partially visible.
- **Narrow vs Wide-Area:**
  - Most research uses Narrow-Area Settings: Multiple cameras all point at the same small spot. This isn't realistic.
  - This paper tackles Wide-Area Distributed Settings: Cameras cover a larger space, and a person moves between views, which is common in real-world environments like homes or offices.
- **Weak Labels vs Strong Labels:**
  - Most datasets have video-level labels (e.g., "this 30-second clip is 'walking'"). This is imprecise.
  - We need frame-level labels (e.g., "'walking' occurs from frame 50 to frame 250"). This allows for much more detailed and accurate models.

Paper Contributions: Two-Part Solution

Contribution 1: Dataset	**Contribution 2: The Method**
MultiSensor-Home	MultiTSF
A new, realistic benchmark dataset for action recogntion	A novel architecture designed for this complex data
Multi-view: 5 Synchronized cameras	Transformer-based Fusion: Combines data from all sensors
Multi-modal: High-resolution RGB video and audio	Human Detection Module: A special component to guide the model to focus on frames with people
Wide-Area: Covers a realistic indoor home environment	State-of-the-Art: Outperforms previous methods on multiple benchmarks
Strongly Labeled: Provides	

Overall Architecture

1. **Multi-modal Feature Extraction:**
   - Takes raw video and audio from all views and extracts initial meaningful features
2. **Human Detection Module:**
   - Helpful task during training. It specifically teaches the model to identify if and where a human is present in the video frames.
3. **Temporal Modeling & Transformer-based fusion:**
   - It first analyzes how an action unfolds over time within each camera view, and then intelligently fuses the information across all views to make a final accurate prediction



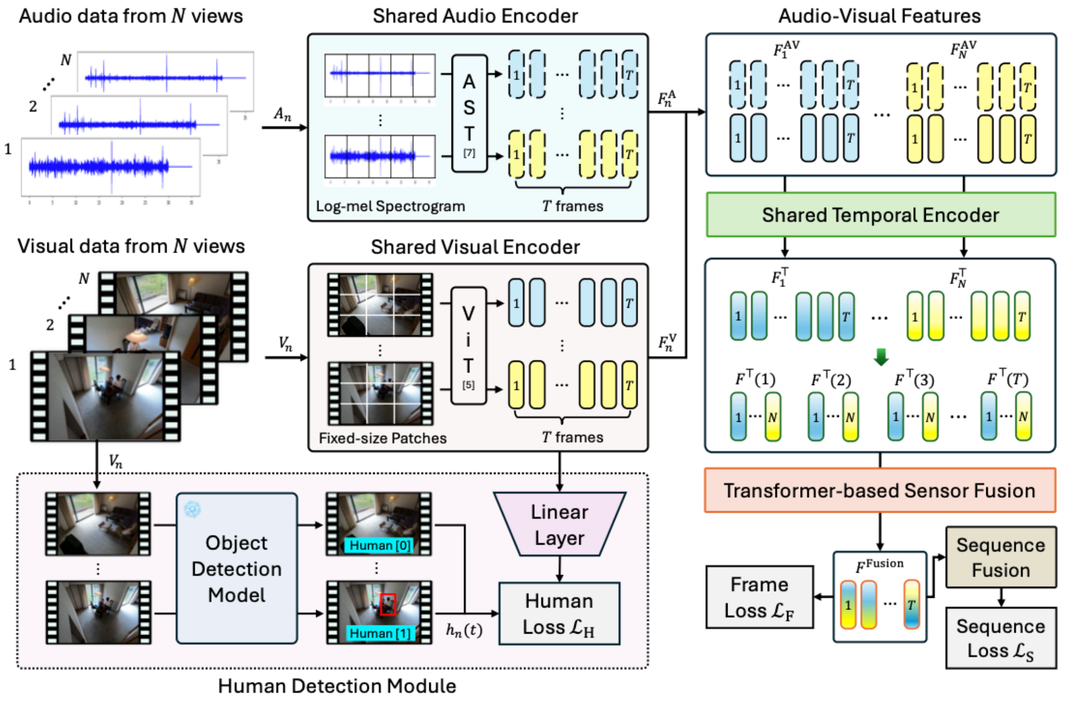
Multi-modal Feature Extraction & Notations

Inputs:

1. $V = \{V_1, V_2, \ldots, V_N\}$**:** The set of video data from all **N** cameras.
   - Each video V_n has dimensions T×D×H×W (Frames, Color Channels, Height, Width).
2. $A = \{A_1, A_2, \ldots, A_N\}$**:** The set of audio data from all **N** cameras.
   - Each audio stream A_n is converted to a spectrogram with dimensions T×F (Time Frames, Frequency Bins).

Process:

1. **Shared Visual Encoder (f_v):** A Vision Transformer (ViT) processes each video Vn .
   - "Shared" means the *exact same* encoder is used for all N views.
   - Output: A sequence of visual feature vectors, F_V^n.
2. **Shared Audio Encoder (f_a):** An Audio Spectrogram Transformer (AST) processes each audio spectrogram A_n .
   - Output: A sequence of audio feature vectors, F_A^n .
3. **Concatenation:** For each view n and each frame t, the audio and visual features are joined together.
   - $F_{AV}^n(t) = [F_A^n(t); F_V^n(t)]$
   - The result is a combined audio-visual feature sequence, F_{AV}^n , for each of the N views.



The Human Detection Module

Goal: Force the model to learn what a human looks like and where they are which helps it focus on “actionable frames”.

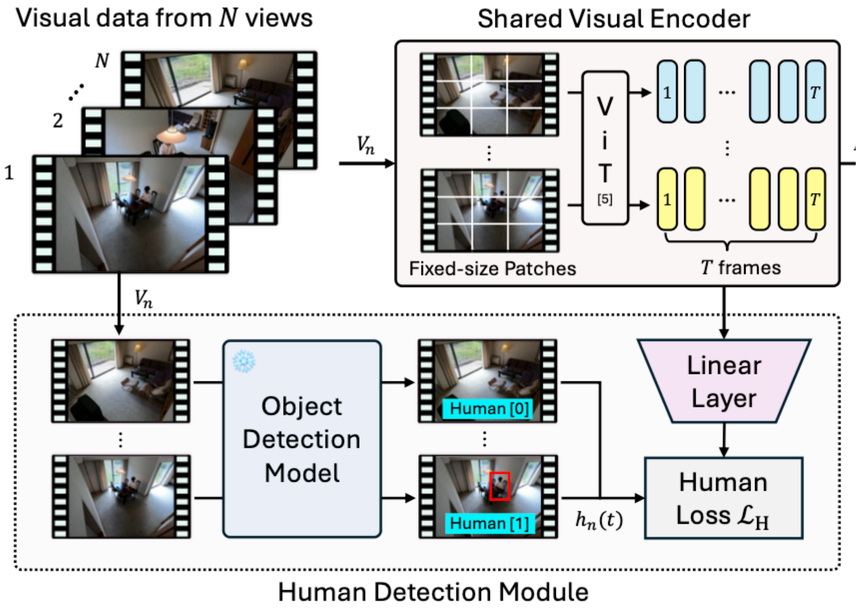
- This is an auxiliary task for better training not part of the final prediction.

Process:

- An off-the-shelf object detector (YOLOv10) is run on all video frames beforehand to create "pseudo-ground-truth" labels, hn (t).
- $h_n(t) = 1$ if a human is detected in frame t of view n.
- $h_n(t) = 0$ otherwise.

Purpose:

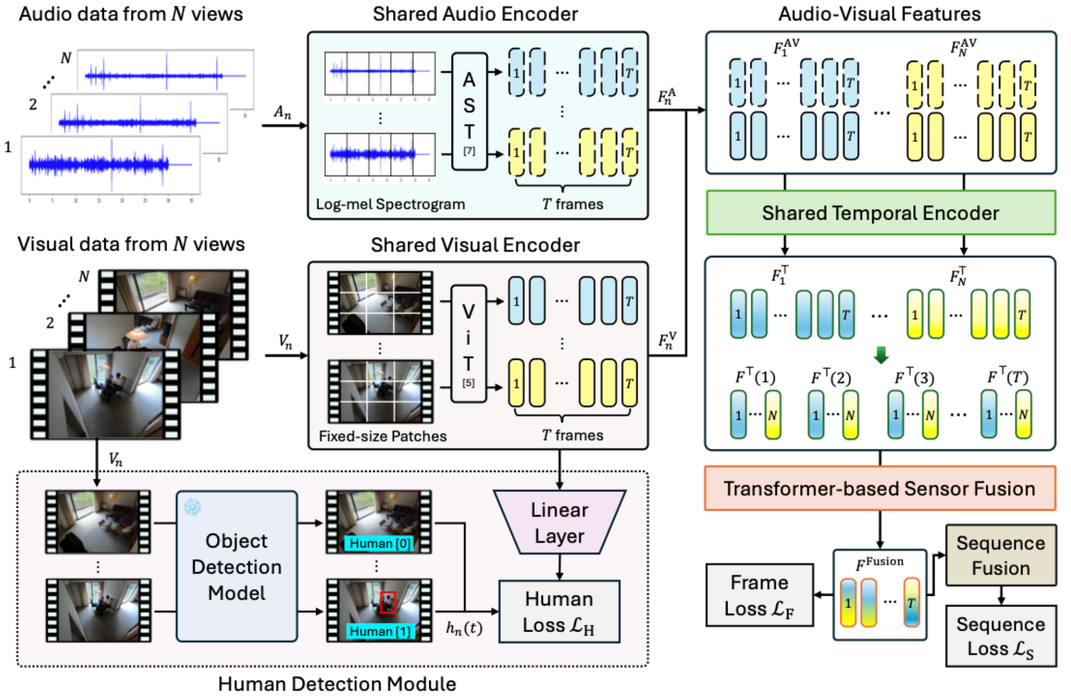
- During training, a special loss function, $\mathscr{L}_H$(Human Loss), is added.
- This loss function penalizes the model if its internal visual features can't distinguish between frames with and without humans.
- This guides the visual encoder (f_v) to produce much more relevant and robust features for the main action recognition task.



Temporal Modeling & Transformer Fusion

Goal: Understand the sequence of an action and combine information from the most relevant camera views.

1. **Shared Temporal Encoder:**
    1. Input: The combined audio-visual features for each view, F_{AV}^{n} .
    2. Purpose: To model temporal relationships. A Transformer learns how features at frame t relate to features at other frames (e.g., 'reaching for a cup' is followed by 'drinking').
    3. Output: A sequence of temporally-aware features, F_{T}^{n}, for each view.
2. **Transformer-based Sensor Fusion:**
    1. This is the key idea. At each single frame t, the model looks at the features from all N views: $\{F_{T}^{1}(t), F_{T}^{2}(t), \ldots, F_{T}^{N}(t)\}$.
    2. Mechanism: It uses a Transformer's self-attention mechanism across the views.
    3. Purpose: To dynamically weigh the importance of each view. If the action "Typing on Laptop" is only clearly visible in View 4, the model learns to assign a high attention score to View 4 and ignore the others for that moment.
    4. Output: A single powerful fused feature vector for each frame $F_{Fusion}(t)$

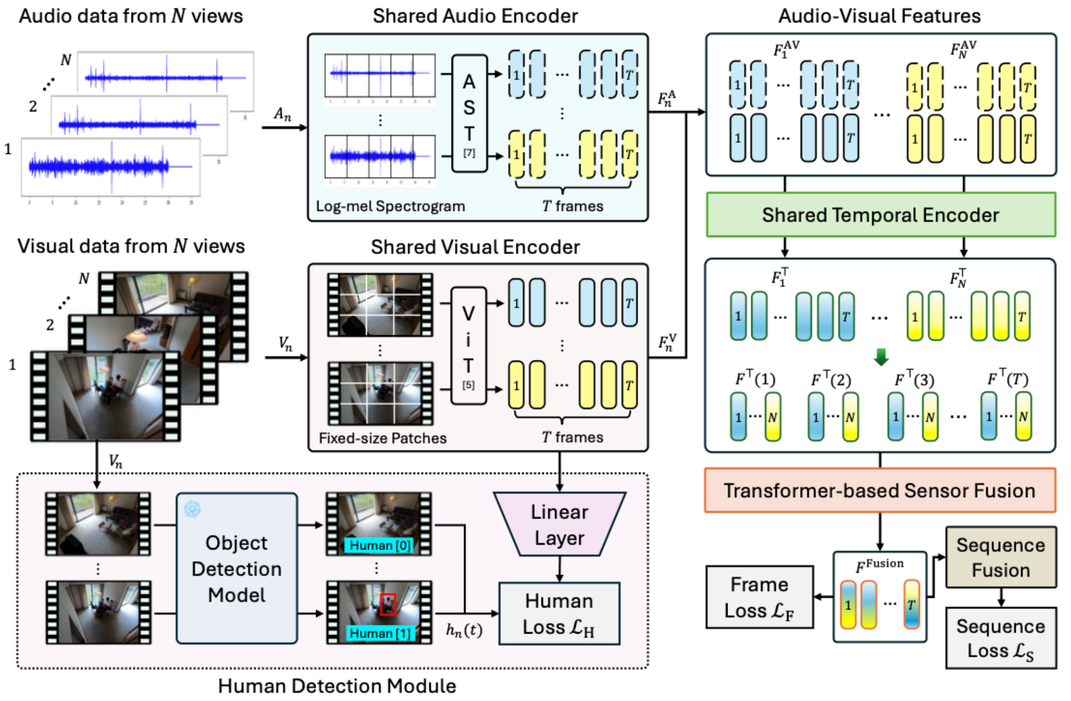


Training the Model: The Learning Objectives (Loss Functions)

Goal: Define the mathematical objectives that guide the model to learn correctly.
The total loss is a weighted sum of three parts.

Total Loss: $\mathscr{L} = \beta_1 \mathscr{L}_H + \beta_2 \mathscr{L}_F + \beta_3 \mathscr{L}_S$ are weights that balance the importance of each term.

1. **Human Loss($\mathscr{L}_H$):**
   - This ensures the visual encoder learns to detect human presence.
2. **Frame Loss($\mathscr{L}_F$):**
   - Purpose: To make correct action predictions at every single frame. This is the primary goal for fine-grained recognition.
   - Input: The final fused features, F_{Fusion}.
3. **Sequence Loss($\mathscr{L}_S$):**
   - Purpose: To make a correct overall prediction for the entire video sequence (the "weak" label).
   - Mechanism: Aggregates the fused features across all frames to make one prediction for the whole clip.



(a) Multi-view inputs (top row) and attention heatmaps (bottom row) highlighting action-relevant regions.

(b) Temporal sequence of video frames (top row) and attention heatmaps (bottom row) highlighting action-relevant regions over time.

Fig. 5. Visualization of multi-view and temporal attention heatmaps from the Shared Visual Encoder on the MultiSensor-Home dataset.

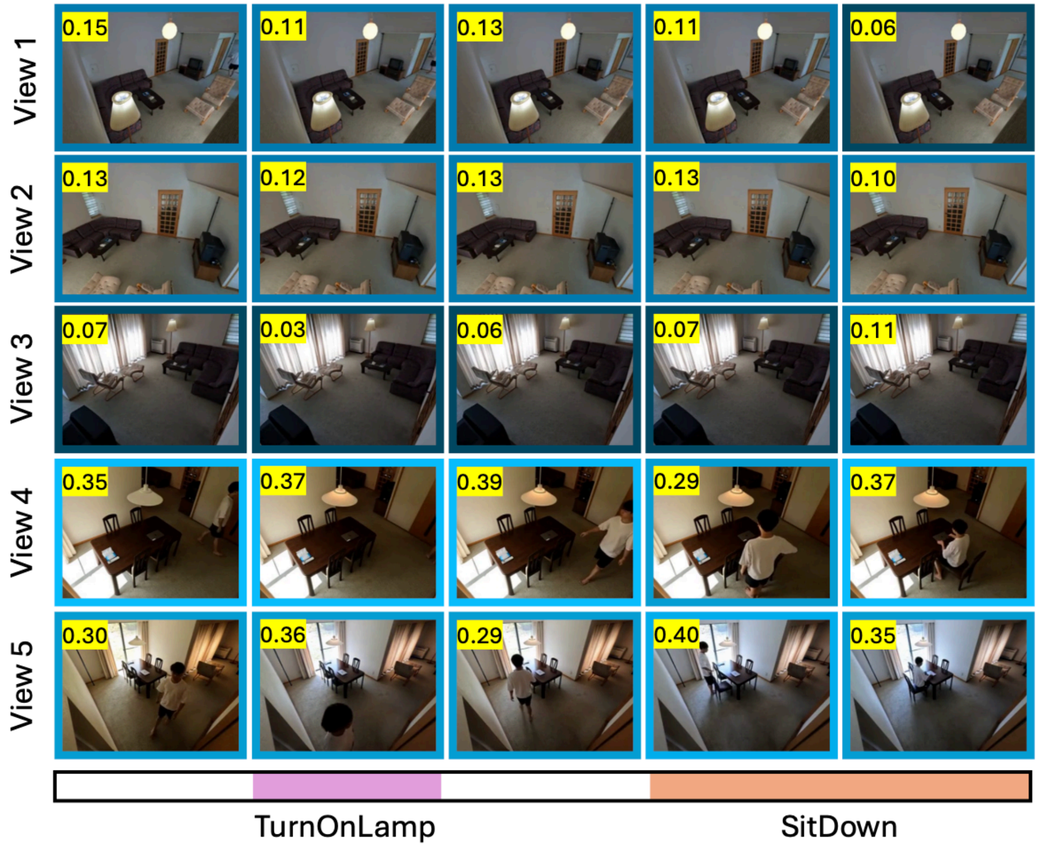


Fig. 6. Attention scores from the Transformer-based Sensor Fusion across multiple views on the MultiSensor-Home dataset.

Research Proposal

Research Proposal Summary: Unsupervised Multi-Camera Audio-Visual Source Localization

1. **Training**:
   - Use **multiTrans** and **multiTSF** dataset
   - The model is trained to predict **Acoustic Source Localization (ASD) pseudo-labels** (the speaker's location) by learning the relationship between these audio cues and the synchronized multi-camera video streams.
   - Through this process the model autonomously learns to associate the directionality of sound with its corresponding visual location in a self-supervised manner.
2. **Testing & Evaluation**:
   - The **audio representation** learned during the training phase is then applied to a **forecasterflexobm** model to serve as a test set.
   - We will evaluate the model's performance by testing its ability to accurately predict **2D Active Speaker Detection (ASD) pseudo-labels** on new unseen data, thereby validating the robustness and effectiveness of our learned representation for a practical downstream task.
     ->forecasterflexobm has provided 2D Active Speaker Detection (ASD) pseudo-labels.

3. Expected Outcome

We expect to deliver a novel framework that significantly improves the accuracy and reliability of unsupervised sound source localization. This technology will enable more robust situational awareness, unlocking advanced capabilities in applications such as autonomous driving, robotics, and smart surveillance systems.