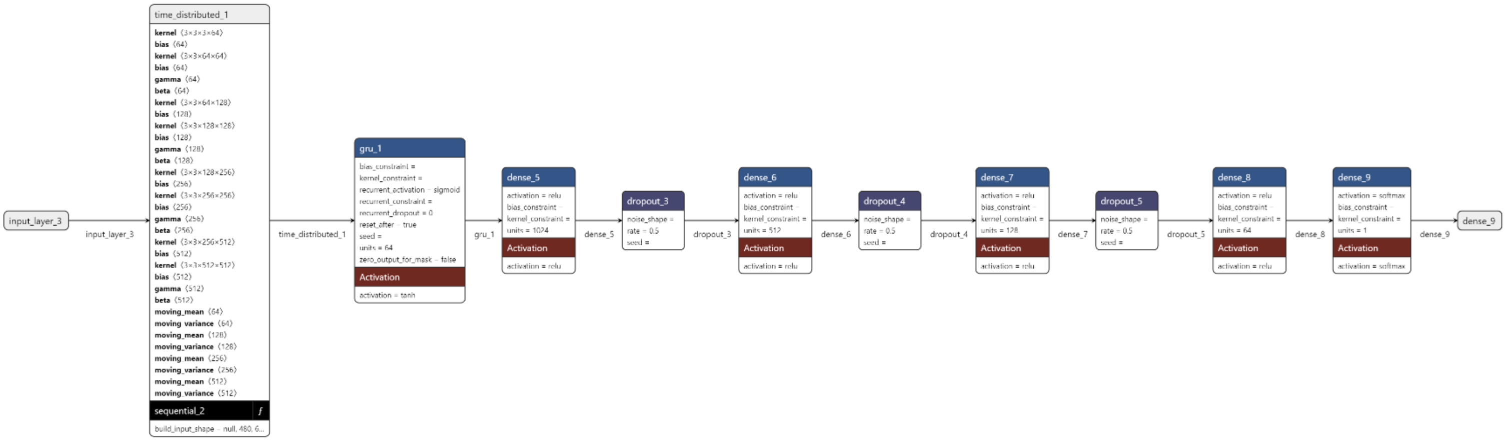
ST-FALL: New spatiotemporal action classification deep learning vision model

This report details SpatioTemporal-FALL, a model designed for fall detection from video sequences. The model integrates convolutional neural networks (CNNs) for spatial feature extraction and recurrent neural networks (RNNs), specifically Gated Recurrent Units (GRUs), for temporal sequence analysis.

Model Architecture



Conv Layers: Initial layers perform spatial feature extraction from each frame independently. These layers capture hierarchical visual features crucial for recognizing actions like falls.

TimeDistributed Layer: Applied after convolutional layers, TimeDistributed layer ensures consistent feature extraction across sequences of frames. This layer is pivotal in processing the temporal dimension of **video data**.

GRU (Gated Recurrent Unit): GRUs are employed to capture temporal dependencies across frames. Unlike traditional RNNs, GRUs mitigate issues of vanishing gradients and are effective in handling long-term dependencies in sequential data.

Dense Layers with Dropout: Fully connected layers interpret extracted features, enhancing the model's ability to classify actions. Dropout regularization is used to prevent overfitting.

Global Max Pooling: This layer aggregates spatial information across feature maps.

Dataset, Training and Evaluation

The dataset used includes:

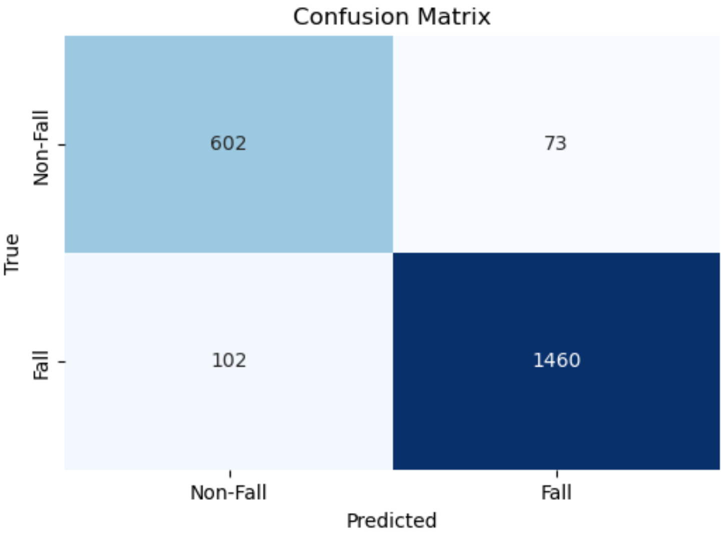
1. Multiple cameras fall dataset: A dataset of 24 fall sequences captured over 8 points of view, totaling 192 videos.
2. Multimodal Fall Dataset: A dataset containing four visual modalities: infra-red, depth, RGB and thermal cameras. These modalities offer benefits such as obfuscated facial features and improved performance in low-light conditions.
3. URFD: This dataset contains 70 (30 falls + 40 activities of daily living) sequences.

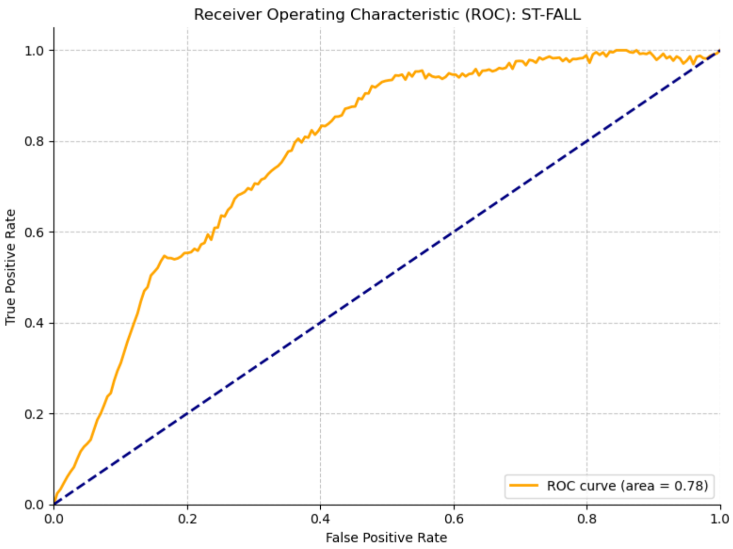
Footage from all three datasets were down sampled and framerates were dropped to reduce compute costs. From each training example (i.e. video), 10-20 intermediate frames that recognized as fall were sampled randomly. From these frames, the bounding box coordinates of the person were extracted by passing them through a YOLOv7 classifier.

The model was trained using binary cross-entropy loss and optimized with the Adam optimizer.

The model was tested on 10% of the original dataset – around 2200 frames.

The model was evaluated against unseen footage, sourced from YouTube (4), Dailymotion(2) and Vimeo(1).





Metrics & Metadata

1. Input Sequence Resolution: 1280x720 (per frame)

2. Sequence Length: 16 frames

Performance Metrics

3. Average Precision (AP) over sequence: 58.3%

4. Temporal Recall: 31.2%

5. Train/Sequence Loss: 2.1%

6. Train/Pose Loss: 0.8%

7. Train/Temporal Consistency Loss: 0.5%

8. Val/Sequence Loss: 3.2%

9. Val/Pose Loss: 1.2%

10. Val/Temporal Consistency Loss: 0.9%

Training Parameters

11. Learning rate: 0.00382

12. Batch size: 4 sequences

IoU Metrics

13. AP50 (50% IoU) over sequence: 76.8%

14. AP75 (75% IoU) over sequence: 63.5%

Efficiency Metrics

15. Average fps (for full sequence): ~6

16. Latency (per sequence): ~267 ms

Additional Metrics

17. Temporal Cohen's Kappa Score: 0.73

18. Sequence-wise Accuracy: 87.2%