RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Cross-Modal Inference of Acceleration Readings from Visual Inputs

Taqiya Ehsan
Advisor: **Jorge Ortiz**
Dept. of Electrical & Computer Engineering, Rutgers University

Abstract

This research tackles the challenges in person re-identification (ReID) by proposing a cross-modal inference pipeline that integrates visual and inertial measurement unit (IMU) sensor data. Traditional ReID methods relying solely on visual features face limitations in diverse environmental conditions. The introduced approach demonstrates increased resilience to variations in appearance by fusing data from multiple modalities. The study focuses on person-mobile device ReID, mapping individuals in video streams to IMU data from mobile devices. Rigorous testing, starting from a basic Sequence-to-Sequence Long-Short Term Memory network, achieves up to 100% matching accuracy, emphasizing the method's effectiveness. The proposed pipeline holds promise for real-world applications, particularly in assistive technologies, showcasing the potential of cross-modal inference for enhanced accuracy and efficiency.

Objectives

- Person-mobile device re-identification using cross-modal inference integrating visual and IMU sensor data from mobile devices
- Infer individual acceleration from video data and synchronize with mobile phone data
- Compare performance of Stacked as well as Sequence-to-Sequence LSTM networks
- Evaluate the impact of the attention mechanism on the sequence-to-sequence LSTM network

Methods Pipeline

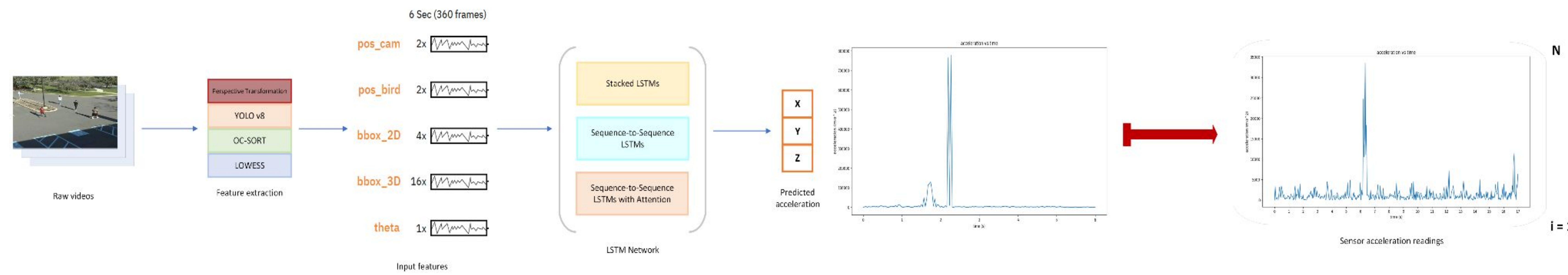

Figure: Method Pipeline

Data Collection & Processing

- A custom dataset with four (4) individuals was collected in a controlled environment, simulating a city streetscape
- Synchronized data was collected from RTSP and GoPro cameras, and Sensor Logger in participants' mobile phones
- The data collection process lasted for 1 hour, during which participants are recorded performing various actions and walking patterns

Figure: Single frame of collected video stream

- Perspective transformation was done to obtain the bird's eye view of scene using homography
- YOLO V8 + OC-SORT were used to assign track IDs to the individuals and track their trajectories
- 3D bounding boxes were generated by applying Locally Weighted Least Squares Regression Smoother (LOWESS) to the camera and bird's eye views
- The video was segmented into approx. 6 second slices ensuring sliced scenes capture all individuals
- Video inferences and sensor readings are synchronized

Figure: **top:** 3D bounding boxes around the individuals
bottom: Bird's eye view of the individuals walking

Extracted Features

Input Data:

1. pos_cam_x
2. pos_cam_y
3. pos_bird_x
4. pos_bird_y
5. bbox_2D_x1
6. bbox_2D_y1
7. bbox_2D_x2
8. bbox_2D_y2
9. bbox_3D_x1
10. bbox_3D_y1
11. bbox_3D_x2
12. bbox_3D_y2
13. bbox_3D_x3
14. bbox_3D_y3
15. bbox_3D_x4
16. bbox_3D_y4
17. bbox_3D_x5
18. bbox_3D_y5
19. bbox_3D_x6
20. bbox_3D_y6
21. bbox_3D_x7
22. bbox_3D_y7
23. bbox_3D_x8
24. bbox_3D_y8
25. Theta

Output Data:

1. X
2. Y
3. Z

Key Takeaways & Future Work

- LSTM models perform progressively better as more complexity is added to the architecture
- Addition of an attention layer makes the pre-existing sequence-to-sequence models more efficient
  Investigate the impact of using contrastive loss with DTW distance on the sequence-to-sequence LSTM architecture with attention
- Train the overall best performing model with additional data and finetune architecture parameters for most accurate user mapping
- Progressively add more complexities to the architecture to obtain best possible accurate person ReID

Training & Evaluation

- Training data is split into sequences of (92, 180, 25) for the input and (92, 180, 3) for labels to mirror the format (samples, time steps, features)
- Training is done following the leave-one-out method
- All models are implemented with Keras backend with ReLU activation
- Training occurs for 50 epochs with Adam optimizer and a batch size of 1, without resetting the states for every epoch
- The model outputs are compared with the ground truth IMU data using Dynamic Time Warping (DTW) distance
  - objective is to map an output to the track ID whose ground truth has minimum DTW distance

Model Performance

Average user mapping accuracy for a threshold of **k** lowest DTW distance values with leave-one-out training:

Model	k = 1	k = 2	k = 3
LSTM A	0%	75%	75%
LSTM B	0%	50%	75%
LSTM C	0%	50%	75%
LSTM D	50%	75%	75%

Vanilla LSTMs

Model	k = 1	k = 2	k = 3
Seq2Seq A	50%	50%	75%
Seq2Seq B	0%	25%	100%
Seq2Seq C	25%	50%	75%
Seq2Seq D	25%	50%	75%

Sequence-to-Sequence LSTMs

Model	k = 1	k = 2	k = 3
Seq2Seq_attn A	50%	75%	100%
Seq2Seq_attn B	25%	50%	75%
Seq2Seq_attn C	25%	50%	75%
Seq2Seq_attn D	25%	50%	100%

Sequence-to-Sequence LSTMs with Attention

- The best performing **Vanilla LSTM** model is *LSTM D* with test loss of 3.44 and **75%** correct matching
- The best performing **Sequence-to-Sequence (Seq2Seq) LSTM** model is *Seq2Seq B* with test loss of 1.18 and **100%** correct matching
- The best performing **Sequence-to-sequence LSTM** model with **attention** is *Seq2Seq_attn A* with test loss 2.33 and **100%** correct matching

WINLAB