Transformer-based Fusion of 2D-pose and Spatio-temporal Embeddings for Distracted Driver Action Recognition

Erkut Akdag, Zeqi Zhu, Egor Bondarau, Peter H.N. de With

Motivation

- 2023 NVIDIA AI City Challenge.
- Temporal localization of driving actions plays crucial role in advanced driver-assistance systems (ADAS) for ***Smart Cities.***
- Enhance the safety and comfort of road transportation.
- **Aim:** to **localize** and **classify** distracted driving actions from multiple in-vehicle cameras.
- **Challenges:**
  - Demanding localization accuracy.
  - Inter-class similarities of actions.
  - Multi-class task: 16 different classes.

Methodology

- **Problems:**
  - How to handle the data recorded from multiple cameras?
  - The 2D body joint coordinates alone cannot explicitly bring motion information between body parts.
  - How to combine **2D-pose** and **spatio-temporal** features?

- **Solution:** Transformer-based Fusion of 2D-pose and Spatio-temporal Embeddings. **(see Figure 1)**
  - Novel solution for the distracted driver action recognition, based on a transformer model that is **independent** of the number of in-vehicle cameras.
  - Efficient feature extraction from 2D-pose estimation including the key points of the face and hand. (see Figure 2)
  - Fusion of 2D-pose features and video action features by the encoder module with multi-head attention.

Figure 2: Top-Down pose estimation, head pose of a driver during a drinking action, relative distances between hand and face points, set of facial feature points.

Results

Table 1: Experimental results for the proposed solution with two different settings.

Method	*os* score	F_1 score	Precision	Recall
Solution with 2D-Pose skeleton	0.4929	0.6359	0.6591	0.5708
Solution 2D-Pose skeleton&motion	0.5079	0.6452	0.6789	0.5783

Table 2: Experimental results of ablation studies.

Method	*os* score	F_1 score	Precision	Recall
Baseline spatio-temp.	0.3703	0.5126	0.6120	0.4410
+ concat feat. vect. 2D-pose skel.	0.4274	0.5987	0.6364	0.5652
+ concat feat. vect. 2D-pose skel.&mot.	0.4420	0.6330	0.6912	0.5838
Solution w. 2D-pose skel.	0.4322	0.5859	0.6397	0.5403
Solution w. 2D-pose skel.&mot.	0.4493	0.6381	0.6896	0.5846

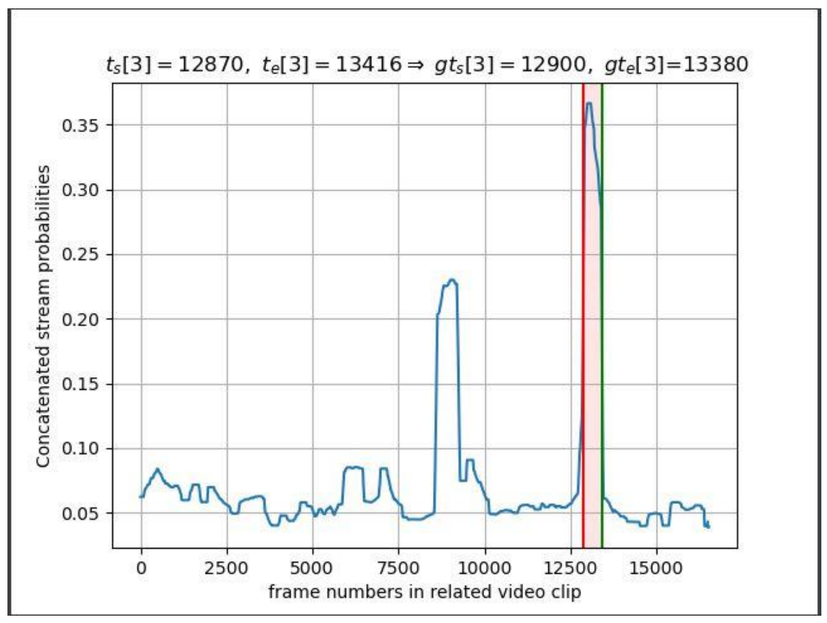

Figure 3: Generated anomaly probabilities for the correct distracted class with the detected peaks denoting localization.

Conclusion

- **The proposed solution has been evaluated on the A2 test set of the 2023 NVIDIA AI City Challenge's Track3, yielding a 0.5079 os score.**
- **Proposed solution is generic and independent of the camera numbers and positions.**

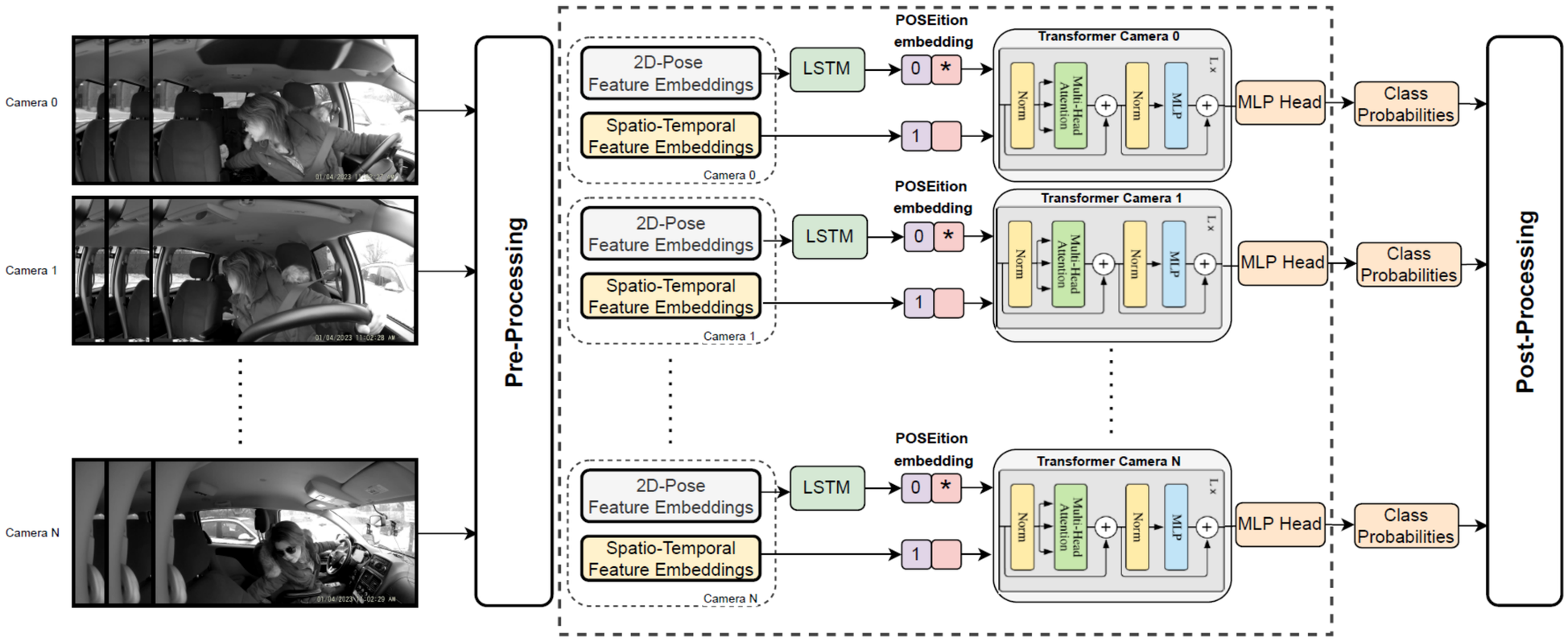

Figure 1: Overview of the proposed architecture. Left: different camera-view inputs to the pre-processing step. Middle: extracted 2D-pose and spatio-temporal embeddings are supplied to the transformer architecture. 2D-Pose embedding is considered as the "POSEition" embedding of the encoder, while the spatio-temporal embedding is the main input of the encoder. Right: 1D-class probabilities obtained after the MLP head per camera view are analyzed for finding the significant peaks for each class in a video. Note that the modules shown with a bold dashed border are used for training only.

Eindhoven University of Technology - Electrical Engineering – Signal Processing Systems (SPS) - VCA Research Group