Laughter in Motion: Pose-Based Detection Across Annotation Modalities in Natural Social Interactions

Author: Vassil Guenov

Responsible Professor: Hayley Hung

Supervisors: Litian Li, Stephanie Tan

01. Introduction

Laughter is a rich, multimodal social cue that is essential for human interaction. As autonomous agents engage more with real-world social settings, understanding non-verbal signals becomes crucial. While audio and facial cues have been well-studied, body movement remains underexplored. This project investigates laughter detection based on pose estimation and how different annotation modalities influence model performance.

02. Research Questions

RQ1: Can body pose features alone suffice for laughter detection in noisy, in-the-wild settings?

RQ2: How does the annotation modality influence classifier performance?

RQ3: How do we split the data into smaller chunks (segments) to maximize classifier performance?

RQ4: Do labelling modalities bias the importance of specific pose features?

03. Conflab Dataset

- 8 videos around 2 minutes each
- 48 participants
- 4 cameras for each video
- 60 frames per second
- 3 modalities: audio-only, video-only, and audio-visual
- Each video is annotated by 3 people per modality
- Instance - a participant-video-frame triple

04. Data Preprocessing

- Majority rule per modality as ground truth; if tied - consider it a positive instance
- Only the camera view with the highest key points is considered for every agent (the others are discarded)
- 30 % occlusion threshold: if more than 30% of the points are occluded we discard the segment

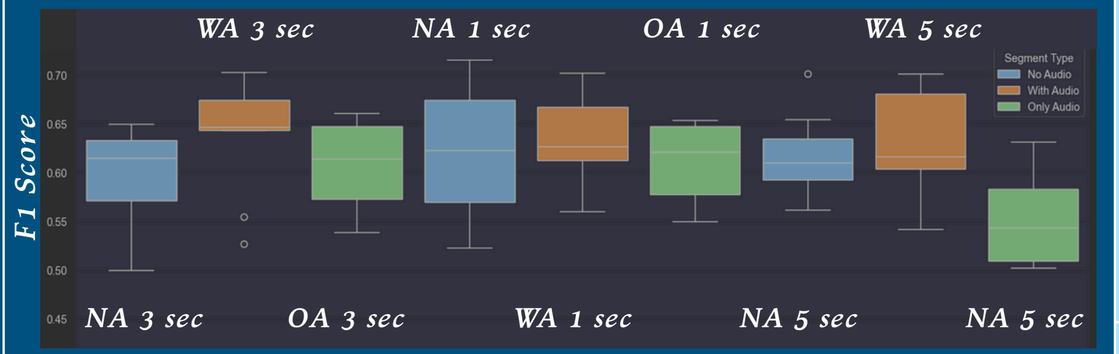
08. Model Performance Results & Discussion

Impact of Segmentation Technique

- Variate of length models overfitted on length of segments
- 3 Second fixed window models performed the best
- Highlight the balance of purity and transition
- Decrease in results of previous studies (0.72 to 0.64) but still comparable

Impact of Annotation Modality

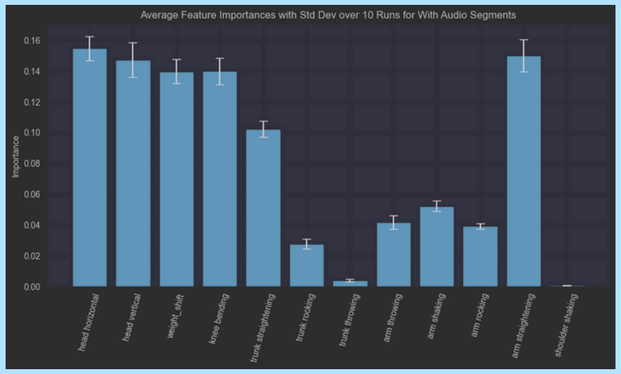
- Audio-visual annotations performed consistently the best
- This is in line with previous research
- Video-only performed significantly better than audio-only



Segmentation-Modality Combination

9. Feature Importance

- Upper body feature set outperforms lower body one; whole feature set beats both.
- This indicates that only upper body models could be trained without hindering too much performance, which is cheaper and faster.
- **Head movements** and **arm straightening** are the most important features.
- Lower body features became more important in audio-only annotation, indicating relations that are hidden from the eye.



10. Answers to RQs

RQ1: Pose-based models reached F1-scores up to 0.64 using only 2D keypoints, confirming their reliability for laughter detection in real-world settings.

RQ2: Audiovisual annotations consistently outperformed others, showing that richer cues lead to better label quality and model performance.

RQ3: Fixed 3-second windows gave the best results by balancing context and label purity, avoiding the overfitting seen in variable-length segments.

RQ4: Feature importance shifted by modality—AV emphasized head and arms, while audio-only focused on lower-body motion, revealing annotation bias.

Methodology

05. Segmentation

- **Separation:** Assure 2 seconds between segments of different labels
- **Noise:** Consider sequences of positive instances of less than 20 frames as noise
- **Delay**: Append 12 frames to compensate for reaction delay
- We apply two segmentation strategies

Continuous Segments

- 100% pure; contain only 1 type of instances
- Variate of length
- Non-laughter segments are much longer than laughter

Fixed Length Segments

- Short laughter segments are padded to a desired length
- Fixed length of 1, 3, 5 seconds
- Offer to also capture transitional behaviour

06. Feature Engineering

We extract motion-based features inspired by Niewiadomski et al.:

- •Kinematics: velocity, acceleration, and displacement of joints
- Body symmetry
- Temporal Rhythmicity: periodic motion via peak detection
- Features adapted to 2D (ConfLab); abdomen features excluded

07. Model Training

- Classifier: Random Forest (RF)
- 12 models: 3 modalities × 4 segment types
- Training protocol:
- Participant-disjoint 75/25 train-test split
- 10-fold cross-validation (stratified)
- Grid search for hyperparameter tuning
- Separate models trained on upper body, lower body, and full-body features