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

Investigating modality annotation impact for detecting laughter in the wild

Vassil Guenov[1]

Supervisor(s): Hayley Hung[1]**, Litian Li**[1]**, Stephanie Tan** [1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Vassil Guenov
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Litian Li, Stephanie Tan, Julian Urbano Merino

Abstract

Laughter is a complex multimodal behavior and one of the most essential aspects of social interactions. Although previous research has used both auditory and facial cues for laughter detection, these approaches are commonly afflicted with difficulties in noisy, occluded, and privacy-sensitive settings. This paper explores the potential of using body posture alone—captured through 2D keypoint estimation as a robust signal for automatic laughter detection in naturalistic settings. We create a machine learning pipeline using the ConfLab dataset, which segments pose data, extracts motion-based features, and trains Random Forest classifiers on various annotation modalities (audio-only, video-only, and audiovisual) and segmentation methods (fixed and variable length). We show that, while variable-length segmentation yields optimal performance, it leads to overfitting. On the other hand, fixed-duration segmentation with three-second windows and audiovisual annotations achieves a pragmatic compromise and reaches F1-scores (65%) comparable to earlier efforts in ideal environments. Upper-body movement, especially head and arm motion, is seen to be salient cues to laughter via feature importance analysis. Annotation modality is also found to significantly affect both classification performance and relative pose feature importance. These findings demonstrate the viability of pose-based laughter detection and reveal how annotation choices shape model behavior, offering insights for affective computing in the wild.

1 Introduction

Laughter is a universal behavior and a fundamental building block of human interactions. Apart from its most well-recognized function as a response to humor, laughter can also occur in response to embarrassment, awkwardness, or agreement, thus facilitating nuanced aspects of human nature [15]. Furthermore, laughter is a multimodal signal that combines its characteristic sound, facial expressions, and body motions. Darwin described excessive laughter as: "the whole body is often thrown backwards and shakes, or is almost convulsed; the respiration is much disturbed; the head and face become gorged with blood; with the veins distended; and the orbicular muscles are spasmodically contracted in order to protect the eyes. Tears are freely shed."[2][14] This richness in modality makes laughter an important non-verbal social cue that has attracted researchers from various disciplines [4].

With the growing development in fields such as cognitive robotics and artificial intelligence, interactions between humans and machines are becoming increasingly common[7]. To facilitate natural communication, it is essential for automated systems to understand not only linguistic content but also non-verbal cues like laughter. From here then arises the need for reliable laughter detection systems[13].

Significant effort has already been devoted to detecting laughter using its acoustic properties and facial expressions [3][6][9]. While effective in controlled environments, these approaches are often impractical in real-world ("in-the-wild") social interactions, where conversations are noisy, spontaneous, and participants may be partially occluded[5]. In such settings, audio signals can be compromised and multiple camera angles may be required to maintain consistent facial visibility. Pose estimation, however, offers a promising alternative. It enables laughter detection based on full-body motion, which is often observable even when faces and voices are not clearly captured. As a single camera can capture multiple individuals, and occluded keypoints can be inferred from visible ones, pose estimation reduces system complexity and increases robustness [9][6]. Moreover, pose-based methods are more privacy-preserving, as body posture is generally considered less sensitive than facial imagery or voice recordings.

Another major challenge in laughter detection is its subjectivity. While some instances of laughter are clearly recognizable, others—such as smirks or restrained expressions—may be more ambiguous.[1] This ambiguity complicates annotation and often necessitates multiple annotators per instance. However, the presentation modality during annotation (e.g., audio-only, visual-only, or audiovisual) can significantly influence labeling outcomes [14], [8]. Given that most annotations are performed with audio or audiovisual input, it remains unclear how visual-only annotations might differ—both in terms of label distribution and downstream model performance—especially when the features are derived exclusively from pose data.[10]

This study seeks to answer the following research question: **How reliable is pose estimation in correctly identifying laughter in the wild?** In order to do that we drew inspirations from three different studies. First, we extend work presented by Griffin et al. and Niewiadomski et al. in laughter detection based solely on pose estimation to in-the-wild social contexts.[6][9] Second, we leverage the methodological framework of Quiros et al. to analyze how annotation modality affects model performance and feature importance in pose-based laughter detection. [14] In doing so, this study aims to uncover modality-specific biases in annotation and assess whether pose estimation alone can serve as a reliable signal for laughter detection.

The aim of this paper can be further split into answering the following sub questions:

1. Can body pose features alone suffice for laughter detection in in-the-wild settings?
2. How does the annotation modality influence classifier performance?
3. How do we split the data in smaller chunks (segments) to maximize classifier performance?
4. Do labeling modalities bias the importance of specific pose features in the best performing model?

2 Related Work

The rich multimodal nature of laughter, as a social signal, has attracted the attention of researchers across various fields such as psychology, affective computing, and

human-computer interaction. It serves various social functions, including affiliative bonding, tension diffusion, and norm enforcement.[15] The social function and characteristics of laughter have been rigorously studied highlighting its vocal aspects as the most prevalent, but also recognizing that laughter can affect facial expression, physiological changes and distinctive body movements. [4] [3]

Traditional approaches to automatic laughter detection have heavily relied on vocal cues and facial expressions. For example, Schroder et al. emphasize the need for artificial agents to recognize non-verbal cues like laughter for meaningful social interaction.[13] Several systems have been proposed to classify laughter based on acoustic features in controlled environments.[12] However, these systems often fail in real-world settings where occlusions, noise, and camera limitations compromise data fidelity, as shown by Gillick et all. when comparing the performance of audio-based laughter detection models in controlled and spontanious environments. [5]

To address these limitations, recent research has shifted toward using body movement as a modality for laughter detection. Griffin et al. first showed that laughter could be detected using pose estimation with accuracy relative to that of human guess [6]. Similarly, Niewiadomski et al. expanded on their work giving new settings and feature sets, training classifiers like Random Forest and SVM and again achieving F1-scores close to those of human annotators[9]. Those studies however take place in a controlled environment with perfect visibility, and as such we aim to translate their methods in more natural settings.

Complementary to these efforts, Di Lascio et al. explored laughter recognition through physiological and movement data, collected via a non-invasive wrost-worn devises, collected via non-invasive wrist-worn devices[3]. This provides an alternative to pose estimation in laughter detection in terms of a privacy-safe evaluation.

Dupont et al. provide a comprehensive review of multimodal laughter research, highlighting the interdisciplinary nature and technological challenges of detecting laughter across modalities[4]. Building on this, the ConfLab dataset represents a major step toward privacy-conscious and ecologically valid data collection in social settings. It enables robust pose estimation even from overhead perspectives, thus facilitating the study of unscripted social behavior[11].

A key gap identified in the field is the subjectivity and ambiguity in labeling laughter, especially when cues are subtle or mixed with other emotional expressions. Quiros et al. investigated how annotation modality—audio-only, video-only, and audio-visual—impacts the consistency of laughter labels and subsequent model performance. They found significant differences in inter-rater agreement and model accuracy depending on the annotation source [14]. These findings suggest that the modality presented to annotators can bias the labeling process and potentially confound model training and evaluation. Their study however does not investigate those effects on pose estimation data and how different modalities affect the accuracy of those methods - a question we aim to answer in this paper.

3 Methodology

To investigate the reliability of laughter detection using pose estimation, we construct a complete machine learning pipeline comprising the following stages:

1. **Preprocessing:** We begin with pose estimation data from various participants in several videos.
2. **Segmentation:** The data set is divided into sub-parts; *segments*.Each segment contains:
   - 2D keypoints for each frame,
   - video id,
   - participant's id,
   - start and end frame indices,
   - a binary label indicating if any of the frames in the segment are annotated as laughter.

   To achieve this result we utilize two segmentation approaches:
   - *Continuous segmentation*: Generates variable-length segments that are label-pure (all frames share the same label).
   - *Fixed-length segmentation*: Generates segments of fixed duration (1s, 3s, or 5s); shorter segments are padded to desired length.
3. **Feature Extraction:**: From each segment, we extract a 12-dimensional feature vector of the keypoints across its frames.
4. **Classification:** We use a Random Forest classifier for training. One model for all combinations of:
   - segmentation strategy (variable-length, 1s, 3s, 5s), and
   - modality (no audio (video-only), audio only, audio-visual),

   resulting in 12 unique model configurations. Appendix A contains a more detailed diagram of the methodology.

3.1 Data Preprocessing

The experiments draw upon the *ConfLab*[1] dataset that captures natural social interactions in a relaxed environment with multi-camera over-head video. The data we are concerned with consists of eight video segments with length of around two minutes[11]. Each segment is shot by 4 different cameras at 60 fps and contains up to 17 body keypoints for every participant at every frame. We select the optimal camera angle for each video-participant-frame triple (we refer to this combination as an instance from now on) to avoid copying of the same behavior. Each instance is independently labeled by three annotators for each modality (audio-only, visual-only, and audiovisual), totaling to 9 annotations per instance. To construct a ground truth label, we use a majority labeling strategy: whenever two out of three annotators from any modality marks the instance as laughter, it is considered a positive sample instance for this modality. In case of a tie we

[1] https://data.4tu.nl/collections/6034313

consider the instance as positive. This approach increases the amount of labeled laughter instances and reduces noise.

Despite removing noise there is still a stark difference in inter-agreement between modalities. 83.5% of laughter instances have only one modality for which the participant in the frame is considered laughing (positive instance); 14.2% have two modalities labeling them as positive, while only 2.3% have all three modalities consider the instance positive. The figures below illustrate this trend - Figure 1 showcases the variety of modality-unique instances (instances that are positive only for a single modality), while Figure 2) highlights the lack of agreement between any two modalities, and to even lesser degree between all three. As we can see both from the data and figures, the amount of instances, which only one modality considers as laughing is significantly greater than the amount in which there is agreement between any of the modalities. To that extend we can take this data as support to the claim that modality of annotation would affect the performance of the models and as such is worth investigating. After filtering and aligning all of the pose data, it was time to segment it.

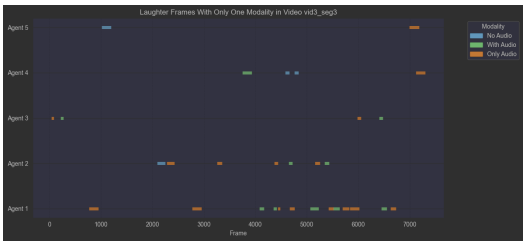

Figure 1: Annotation agreeness across modalities. Highlights frames labeled as containing laughter by only one of the modalities. Blue for no-audio segment, green for with audio segments, and orange - only audio segments.

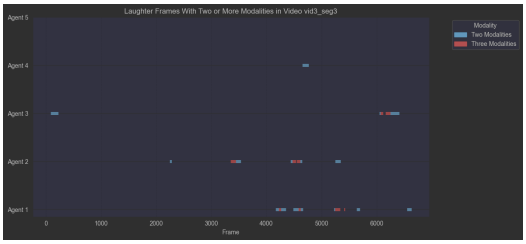

Figure 2: Annotation agreeness across modalities. Shows segments where there are two (blue) or three (red) modalities that agree that a given frame contains laughter.

3.2 Segmentation Strategies

To train temporal classifiers, we segmented the data into smaller chunks to construct a final dataset containing both positive and negative samples. Since the goal is to identify laughter, each sample must be sufficiently descriptive, capturing not only the laughter itself but also the transitional states before and after. With this objective—and considering the characteristics of our dataset—we made several key decisions.

First, because human reactions are not instantaneous, we introduced a delay to the segments. Prior research estimates the average human reaction time at approximately 200 ms [16]. Accordingly, all laughing segments were padded with an additional 12 frames (1/5 of a second) at the beginning, compensating for delayed human responses to laughter.

Second, to reduce noise caused by brief and potentially ambiguous events, we excluded sequences of 20 frames (excluding the delay) or fewer from the set of positive samples. Such short instances were treated as noise rather than valid positive episodes in the segmentation techniques described below.

Third, we introduced a two second gap between any positive and negative segments. By doing so we ensure maximal separability between different types of annotation. Moreover we ensure all transitional behavior is associated with identifying laughter, allowing the model to learn not only the expected behavior once laughter is present but also the preceding and residual motion in case the latter is more prominent than the former.

Finally, to maintain segment purity and avoid mixtures of positive and negative instances, we chose not to use sliding windows, as they tend to introduce impurity by blending different classes within a segment. Instead, we focused on identifying positive instances and build our segments from them. To that extent we explored two segmentation strategies, each trying to find a balance between interpretability, data coverage, and annotation purity

Variable-Length Continuous Uniform Segmentation is the first segmentation technique focuses entirely on purity of the segment. Segments are formed by grouping sequential positive instances of the same label. This ensures all instances in a given segment belong to the same type of behavior making labeling the segment itself trivial. On one hand, this preserves natural episode lengths and is especially suitable for learning organic behavior. On the other it creates a dataset with non-uniform segments. Moreover as shown by Figure 3) the laughter segments are much shorter then their non-laughter counterparts.

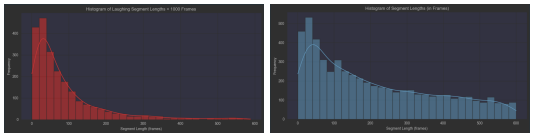

Figure 3: Length distribution of laughter episodes (left) and the whole dataset(right) using sequential laughter segmentation.

Table 1 shows more in depth statistics into the laughing segments only. As we see there is a big discrepancy between the lengths across modalities but also in single modalities as well.

In order to better address this issue we decided along with the uniform segmentation to introduce **fixed window length segmentation**, which ensures consistency by splitting the dataset into segments of equal length. Our second segmentation techniques builds off of the first by grouping all sequential positive instances, and then pad them to a number of frames that is divisible by the chosen segment length. For example, if participant x is annotated as laughing from frames

Table 1: Laughter Segment Length Statistics (in frames) Across Annotation Modalities

Modality	Mean	50%	75%	100%
No Audio	81.43	55.00	96.00	606.00
With Audio	69.16	46.00	78.00	550.00
Only Audio	58.36	45.00	71.00	429.00

100 to 145, we obtain an initial segment of 46 frames. After adding the 12-frame delay, the segment becomes 58 frames long. To reach the fixed window length of 60 frames, we pad one frame on each side. In case the original length exceeds the window length we pad it and then divide into smaller multiple smaller segments.

For our negative sample set we follow a similar strategy. We remove all frames that are part of laughter segments, introduced the gap, mentioned eariler, and divide the remainder of the space into window length chunks. Figure 4 depicts the segmentation of three participants for a part of a video and lustrates our two main goals in this strategy. First, different label segments are far not close by ensuring separability. Second all transitional behavior is noted as laughter preserving as much purity as possible.

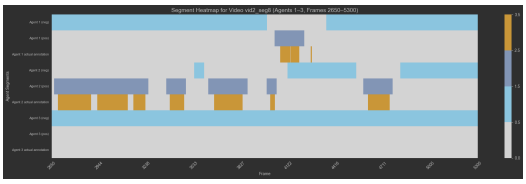

Figure 4: Illustration of segmentation strategy for participants(y-axis) 1,2,3 and frames (x-axis) 2650-5300 of vid2seg8 for no audio annotation and a window size of 60. The light blue show segments eligible to be in the negative dataset away from laughing segments noted in gray. Yellow show the actuall laughing frames annotated as laughing in the dataset.

The window lengths we chose for this experiments:

- **60 Frames** - One second segments; the closest to the mean of all modalities and still bigger than the median ensuring most laughter segments are unbroken
- **180 Frames** - Three second segments; Bigger than 75% of the segments, enough to capture some transitive behavior but still close to the mean
- **300 frames** - Five second segments; Bigger than 90% of the segments so almost no segments will be broken into smaller chunks ensuring preservability of laughter, while also capturing more transitive behavior. Quite larger than the mean, however which can also lead to some non-laughter behavior being caught as well.

Each approach is a compromise between temporal alignment and segment consistency. Before continuing on all segments are stored inside a pandas dataframe with the following columns: Video id, participant id, modality, start frame id, end frame id, label. For each of those segments we then took the keypoints for each frame and performed feature extraction.

3.3 Feature Engineering

With the work of Niewiadomski et al. [9] as a foundation, we extracted a collection of kinematic and postural features based on the 2D pose estimates. Our final feature collection includes:

- **Kinematic Features:** Velocities, accelerations, and displacements of major joints per frame.
- **Postural Dynamics:** Trunk rocking, shoulder shaking, and symmetrical features.
- **Temporal Rhythmicity:** Frequency patterns identified through peak detection techniques.

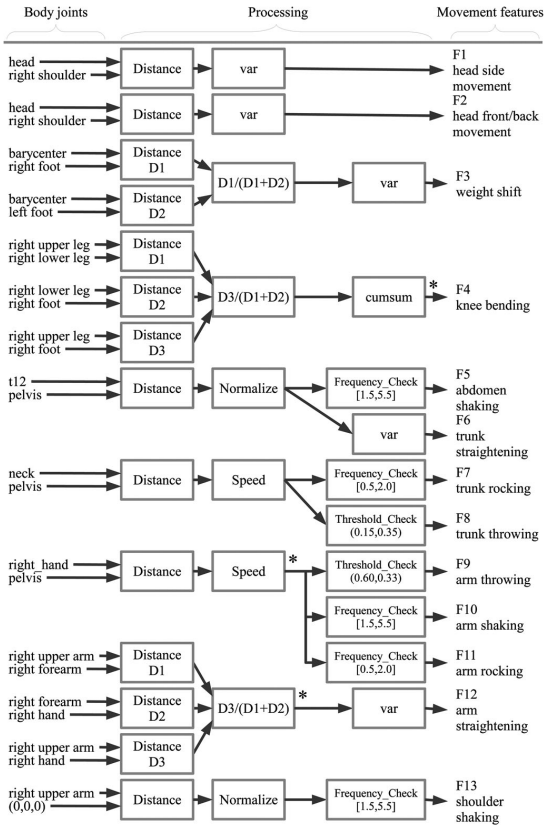

Figure 5: Feature extraction pipeline adapted from Niewiadomski et al.[9]. Each feature corresponds to a particular body motion consisting of multiple joints.

While our feature engineering is inspired by the work of Niewiadomski et al., shown in Figure 5, there are important adaptations required due to the constraints of the *ConfLab* dataset. First of all, our features are annotated in a 2d plain requiring small adjustments to the calculations mainly in f1 and f2 - head movements. Second, as our dataset consists of only 17 keypoints which are less then the ones used by Niewiadomski et al. we removed entirely f5 abdomen shaking, as there were no sufficient keypoints to calculate. Finally,

as a point of reference, we also introduced a statistical based feature vector. For every keypoint we recorded the total distance traveled, mean speed, max speed, and speed variance. This enables us to study whether motion variability in itself is capable of distinguishing laughter from neutral behavior.

3.4 Classification and Evaluation

We trained a Random Forest (RF) classifier from *scikit-learn*[2] to distinguish laughter from non-laughter segments based on the features that we extracted. The use of RF classifier is particular fitting as not only it has proven to be the best performer so far in other laughter detection experiments[6] []niewiadomski2016automated[14], but it is also scale-invariant, efficient with limited data, and provides insight into feature importance. We used the following protocol:

Participant-Disjoint Splits: Ensured that no participant is present in both train and test sets (75-25 split), preventing identity leakage. Figure 6 showcases the distribution of laughter to non-laughter segment based on each participant. From it we can deduce that there are enough subsets of participants to form different test sets for each model, Furthermore each participant's segments will only be present in one of the two sets, ensuring variability and better generalization of the model.

Stratified Cross-Validation: Done to deal with class imbalance and preserve the laughter/non-laughter ratio.

Hyperparameter Tuning: Grid search over RF parameters such as tree numbers, max depth, minimum sample split, etc. with 10-fold cross-validation.

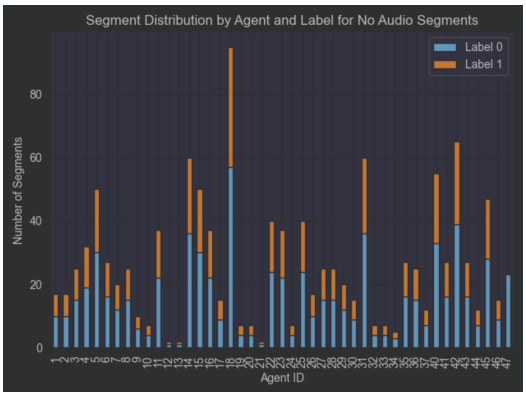

Figure 6: Amount of segment distribution per participant for three second with audio annotated segments.

We trained and evaluated for every labeling modality and segmentation strategy combination ten times and provided average accuracy, F1-score, and feature importance (via mean decrease in Gini impurity). Additionally as our aim was also to distinguish between different feature importance we also trained models on two subsets of the features: upper body (f1, f2, f9, f10, f11, f12, f13) and lower body(f3, f4, f6, f7, f8) as labeled in Figure 5. We can then determine if there is any discrepancy in combining many features or if there is benefit in focusing only on certain parts of the body, saving efforts in keypoint extraction.

[2]https://scikit-learn.org

4 Results

4.1 Performance of Classification

The main part of the research aims to investigate wether pose estimation is really suitable for in the wild laughter detection and as such the performance of the classifiers is of biggest concern. Table 2 summarizes the average classification statistics over 10 runs for each of our twelve models (each trained on different segmentation-modality pair). The results are sorted according to their F1 score. We have added Niewiadomski's RF classifiers as a baseline to which we can compare our results. The relative performance of Random Forest classifiers across segmentation strategies and annotation modalities reveals a clear trend: continuous variable-length segmentation performs considerably better to fixed-length segmentation.

It is immediately obvious from the table that the variable length windows have incredibly high results compare to the rest of the models. In fact, they are the only segmentation technique that achieved higher F1 score than our baseline. This however is immediately undermined by the observed correlation between feature movement and labeling. Most of the features show weak negative correlation but some in the case of arm rocking ($r = -0.299$) and weight shift ($r = -0.273$), which are usually associated with laughter. This suggests overfitting on the length of the segments and thus we will focus on the other model for performance for the rest of this section.

The primary performance metric considered is the F1 score, which balances precision and recall. Among all configurations, the best F1 score of **0.638** was achieved with *Fixed Three Second* segmentation and *With Audio* annotations, closely followed by the *Fixed Five Second With Audio* setup with an F1 score of **0.630**. These results suggest that moderate-length segments (three seconds) combined with richer multimodal annotations (audio and video) yield the most effective laughter detection performance in terms of balancing precision and recall.

tab:performance does not paint the full picture however in terms of performance. Figure 7 shows that there are two extreme outliers in this particular model which decrease the mean significantly. Since other models do not have such big discrepancies we can conclude that the Fixed three second audio-visual model performs considerably the best if we exclude the outliers.

Comparing annotation modalities more broadly, the *With Audio* annotation consistently outperformed both *No Audio* and *Only Audio* across all segmentation lengths. Moreover, among those using fixed length segmentation, the best are precisely the one using *With Audio* annotation, proven by Table 2. We can further see that there is a significant difference in the performance of models using that annotation and the other ones. For example, in the *Fixed One Second* setting, *With Audio* achieved an F1 score of **0.634**, outperforming *No Audio* (**0.622**) and *Only Audio* (**0.610**). A similar trend is observed in the *Three* and *Five Second* segmenta-

Table 2: Average classification performance across segmentation strategies and annotation modalities.

Segmentation	Modality	Accuracy	F1 Score	Precision	Recall
Variable Length Windows	With Audio	0.880	0.860	0.850	0.860
	Only Audio	0.840	0.830	0.820	0.820
	No Audio	0.830	0.810	0.790	0.800
Niewiadomski's RF	-	-	0.72	0.73	0.73
Three Seconds	With Audio	**0.642**	**0.638**	**0.644**	**0.642**
One Second	With Audio	0.645	0.634	0.655	0.620
Five Seconds	With Audio	0.639	0.630	0.644	0.623
One Second	No Audio	0.651	0.622	0.677	0.588
Five Seconds	No Audio	0.635	0.618	0.648	0.593
Three Seconds	No Audio	0.632	0.598	0.658	0.553
One Second	Only Audio	0.616	0.610	0.619	0.607
Three Seconds	Only Audio	0.612	0.608	0.615	0.607
Five Seconds	Only Audio	0.611	0.540	0.658	0.468

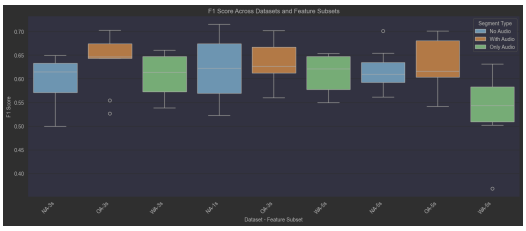

Figure 7: F1 score metric for all the fix sized window models across 10 runs. Three second *with audio* model (second from lef to right) performs the best and with significant outliers which drop its mean further down.

tions. This suggests that the inclusion of both audio and visual cues in annotations provides more informative training signals for the model, resulting in improved generalization. In contrast, *Only Audio* annotations generally performed the worst, particularly in longer segments (e.g., F1 of **0.540** in the *Five Second* condition), indicating that *No Audio* labeling may not adequately capture the nuanced nonverbal expressions of laughter visible in body pose dynamics.

These trends highlight the importance of both annotation richness and temporal context in training effective laughter detection models based on pose data. This is of particular importance in order for us to better understand how annotation modalities can really impact the performance of models. The results also indicate that extending segment duration beyond one second can enhance performance—especially when annotations are derived from multimodal sources—though gains may plateau or diminish if the segment length becomes too long.

4.2 Feature Importance

To further understand how pose estimation models behave we narrow our focus to the specific movements that matter. In this particular section we focus solely on the best performing model to keep results digestible. Adding to the already trained *three second, with audio* model, we also evaluated the performance of two other models trained on the subsets of the whole feature set - one only on upper body, the other only on lower. Figure 8 highlights the the difference in performance. Note that in this diagram we have removed the two outliers we had before on the model, containing the whole feature set, which increased it's average f1 score to around 0.66. This is still better than the upper body model - 0.65 and considerably better than the lower body one - 0.625.

Figure 8: Performance of model when split into upper and lower as well.

Diving a level deeper we also performed feature importance analysis of the top-performing model, which revealed many strong features responsible for detecting laughter. Figure 9 depicts our findings:

- **head horizontal**, **head vertical**, and **arm straightening**—indicating dominant directional movements in upper-body articulation patterns.
- **weight shift**, **knee bending**, and **trunk straightening**—capturing posture modulation behaviors that differentiate laughter from neutral stance.
- **Temporal rhythmicity** features—derived through peak detection, showing periodic bodily movement in synchronization with laughter.

Visual-only and audio-only training models generated importance shifts: visual-only heavily favored `arm straightening` and `arm throwing`, while audio-only emphasized more nuanced body signals like `knee bending` as shown in Figure 10. Comparing all three however paints a clear picture: the most important features are `arm`

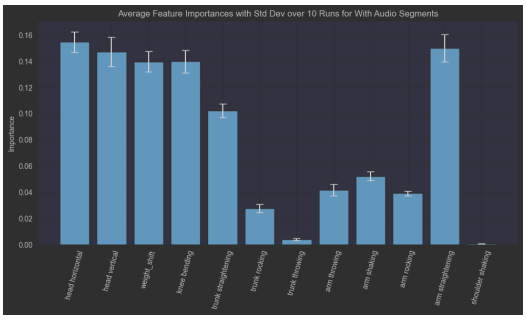

Figure 9: Feature importance of *With Audio* annotation.

`straightening` and `head movement`. While other features vary in their specific importance we can still categorize `head horizontal`, `head vertical`, `weight shift`, `knee bending`, `trunk straightening`, `trunk rocking`, `arm shaking` and `arm straightening` as significantly important. The remaining features hold less importance to the overall model.

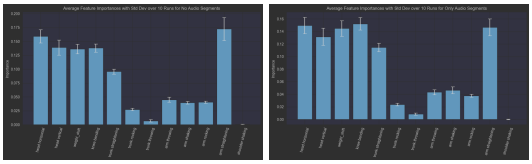

Figure 10: Feature importance of other two annotations.

5 Discussion

The results of this study reveal an important caveat in the design of laughter detection systems based solely on body pose data. While initial metrics suggest that variable-length segmentation yields significantly higher classification scores, closer inspection of the data, feature correlations, and model behavior points to substantial overfitting.

5.1 Overfitting in Variable-Length Segmentation

Initial results showed that variable-length segment models were achieving high F1-scores (0.86), surpassing even models from previous works, but this was subsequently recognized as overfitting. All laughter segments although 100% pure were considerably shorter than the non-laughter segments which lead to the duration of the segment itself becoming a proxy for the label. This is reflected in the strong negative correlation between some features (e.g., arm rocking, weight shift) and laughter categorization. This comes as a surprise, as laughter is typically categorized by a dynamic burst of movements. Therefore due to discrepancy in their length, laughter segments do not have enough time to accumulate high enough values to counter for natural movement. This combined with the unusually high results can lead us to determine that the model picks up characteristics of the data that are not in the feature set, leading to overfitting. As such we cannot determine its results of accurately representing the ability of pose estimation to be used in laughter detection.

5.2 Fixed-Length Segmentation and Model Realism

On the other hand, fixed-length segmentation better resembles the real-world inference scenario in which laughter must be extracted from a stream of continuous input with noisy temporal boundaries. Though lower on raw performance (e.g., F1-scores of 0.61–0.64), the fixed-length models were more balanced in precision and recall and less sensitive to motion amplitude or duration. This suggests that they are learning more generalizable features.

The best-performing model was a realistic trade-off between sufficient context and noise reduction. Moreover this duration offers good balance in capturing transition behavior. However, the performance difference between this model and the variable-length one, although large in size, is less meaningful in light of the overfitting risks inherent to the latter.

Niewiadomski et al. obtained F1-scores comparable to the level of human performance (0.73) on rich 3D motion capture data with dense keypoint coverage.[9] Contrastingly, this study operated using 2D pose data with sparse 17 keypoints and required some adaptations as removing features and dealing with non-available joints. Our model obtained relative score of 0.64 is comparable to that of previous research. It has deteriorated in performance as expected due to the imperfect conditions and lack of ground truth. The results are still considerably better than blind guessing and point to the conclusion that even in imperfect conditions pose estimation is still a viable technique for detecting laughter. This results also support fixed length segmentation as the more adequate segmentation technique to the continuous variable one answering another of our questions.

5.3 Impact of Annotation Modality

Shifting focus from segmentation techniques and purely answering if pose estimation is reliable enough to detect laughter to the more nuance topic of modality's affect on the performance. Annotation modality had a consistent, though secondary, impact on performance. Audiovisual-annotated models performed best overall for all types of segmentation, with no-audio second, and only-audio annotations performing worst. These results are consistent with prior work [14], which suggests that multimodal annotation improves label quality and inter-rater agreement.

5.4 Feature Importance

The final part of our aim to better understand how pose estimation can be used in laughter detection led us to examine how much influence each feature had on the final output. The first step to that consisted of breaking the feature set into two parts - upper and lower body. The lower body provided the worst performance, but more interesting is comparing the other two. Even though the model trained on the full feature set still performed better than the one which was trained on only the upper body features, the difference attributes to 1%. This suggest that lower body features are still important too and can't be easily discarded for maximum performance. Such a small difference however also means that in more extreme and congested situations, especially with overhead

cameras when lower body features might be harder to catch, capturing the movement only of the upper body (head and arms) might not compromise the performance of the model.

The results of comparing the models are further supported by a more in depth analysis of the feature importance of the model adopting the whole feature set. The features bearing the highest significance are all upper body features - head movement and arm straightening. Furthermore if we look down the next most significant features they are all lower body features - weight shift, knee bending and trunk straightening. This further compliments the above results and assumptions as it shows that there are some lower body features that contribute and are more correlated with laughter and we can't undermine their importance if we want to maximize performance. Other features are more neglectable with none of them reaching a higher mean decrease of impurity of more than 5%.

Combining our focus in modality attention and feature importance also offers interesting insights. To summarize the results, audiovisual models prefers more upper body movement with noticeable difference between the top three features and the following two. Visual features however move past this differences and while still prefering the upper body more we can see that difference is a lot smaller while audio-only features are completely opposite-expressing preference in lower body movements. This indicates that there is a correlation between feature importance and annotation modality in pose estimation. Since pose estimation is more connected to the visual aspect of laughter detection, audio-annotations provide interesting insights into movement that can be correlated to laughter but not particularly well visible. We often associate laughter with the facial features and upper body, thus we can neglect the movements of the lower boddy, and the preference of the audio only modality towards exactly those features proves that there are hidden for the human eye connections.

6 Conclusions and Future Work

This study set out to determine if full-body pose estimation alone can reliably detect laughter in natural, in-the-wild social interactions.To that extend, we needed to understand how accurate pose estimation is in detecting laughter. Delving a layer deeper, we set out to understand more in depth what circumstances in terms of modality annotation and segmentation techniques allowed for the best performance. Finally, in order to fully understand the reliability of pose estimation as a method we needed to understand what features contributed the most. In tackling all these problems we reached the following conclusions:

First, to answer the question simply: pose estimation **can** be used to effectively detect laughter even imperfect in the wild conditions. Despite a drop of classification performance of 8% from previous research we can still use pose estimation, even one collected in crowded rooms with sometimes limited visibility and annotated in 2d coordinates instead of 3d ones.

Second, **segmentation strategy has a strong effect on the model performance**.Various length segments although best on paper have a tendency to overfit. Fixed length segmentation, on the other hand, although more reliable still provide a dilemma in choosing the specific length. In our particular cases choosing the three second window worked the best. This is significant as it proves that choosing a length that captures most whole laughter segments without breaking them, but also capturing some transitional behavior works better then going in either end of the spectrum.

Third, **annotation modality has a large impact on both model performance and representations learned**. Audio-visual annotations consistently yielded higher quality model results than audio-only or vision-only labels. This indicates richer sensory cues, improve labeling quality and thus model training. It also cemented results from previous researches and provided insights into ways to cheaply collect data without hindering performances.

Finally, **feature importance provided us with core insights into understanding pose estimation models in general**. It showed that, although all data is important we could still use only the upper part of the human body to annotate and would still get comparable results. This opens new questions and allows for more detailed investigations into only that part of the body, which may increase the accuracy of the model further. On top of that combining modality effect with feature importance opened the door to some interesting discoveries. Despite pose estimation being mostly a visible way of storing information without containing any audio in itself there was some strong correlation between audio-only annotation and lower body features.

Collectively, these results establish that pose estimation is an effective modality for detection of laughter, especially when paired with high-quality annotation and realistic segmentation methods. The trained models adequately identified laughter-relevant patterns in movement, in accordance with the social and physiological understanding of laughter as a full-body phenomenon.

Future Work

Building on such findings, future research would be able to create adaptive segmentation techniques that preserve temporal realism while suppressing label leakage. Moreover, expansion of the dataset in size to cover a broader range of social settings and individual behavior would facilitate greater generalizability. On that topic, choosing a different feature set or combining different features would also be beneficial as it would provide even more clearance on the nuance of body movement in laughter. Finally, integration of temporal modeling structures like RNNs or transformers may enhance the ability of the model to learn complex dynamics over time and better capture the true structure of laughter during conversation.

7 Responsible Research

In this section we will go over matters related to the integrity of the conducted research such as proper data handling, reproducibility and use of generative AI.

In our research we used a third party dataset. We have the approval of the original authors and owners of the dataset to

use the dataset for the required purpose of investigating the possibility of laughter detection in in-the wild social settings. For the duration of the project the dataset has always been present on the remote server, on which it originates from, and whenever we accessed it, it was from the TU Delft's grounds connected to a VPN. Additionally privacy of the participants in the original data collection has been kept both by the owners of the dataset and by us. Each participant is referred by a special id which is only used as a link to the pose estimation itself. The data itself also cannot be used to identify any of the participants by itself.

The research and methodology have both been clearly outlined in our paper. Using those instructions it should be clear to a reader how if they have been allowed access to the data set to conduct the exact same research, or given a different dataset how they can replicate the pipeline to check the generalizability of our methods. The code itself is posted to a public GitHub[3] repository, although for privacy concerns we removed all display of data from the dataset.

In this project use of generative AI was primarily done in order to format the code and paper structure. In the paper We predominantly used AI for structuring sections in LaTeX format. Occasionally, we also used AI to paraphrase sentences that seemed long or convoluted, although such occurrences were rare. We never used AI to write original content and even in the events which we used it to paraphrase the final product was also edited by us. In the code base we predominantly relied on such tools to generate us plots for the already computed results and for formatting code and documentation. All the ideas were based on literature reviews and our own reasoning. We also used it to quickly translate results from csv files to LaTeX-ready tables. In summary AI was predominately used for formatting and slight improvements, but never for novel and significant contributions. I have uploaded example prompts in Appendix B.

A Methodology Diagram

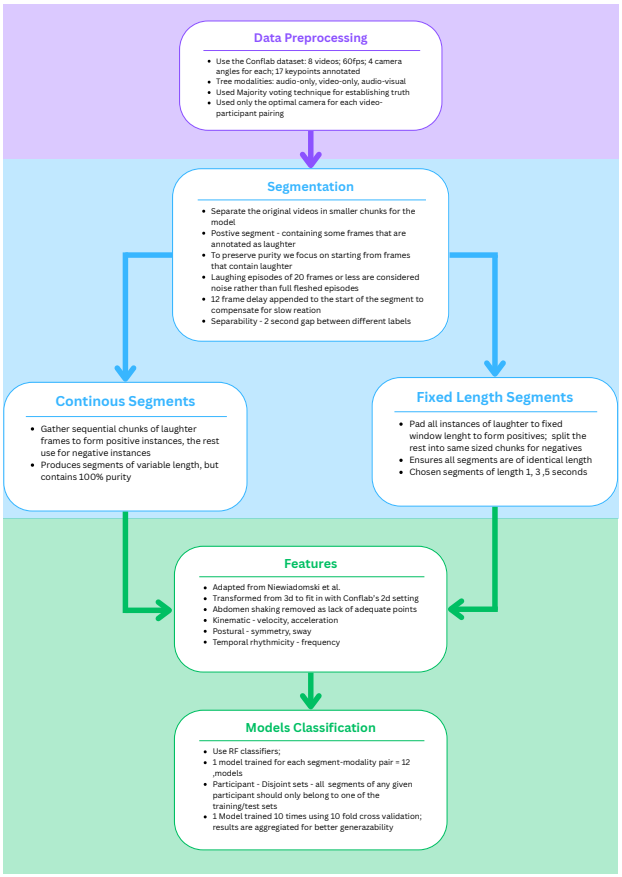

Figure 11: A detailed diagram explaining our methodology

B Generative AI Prompts

The following lists presents example prompts we have used in the project:

- Table generation: Hello, I have a python, pandas dataframe containing metrics for the performance of a Random Forest classifier. There are twelve models in total paired with the columns "segmentation" and "modality" serving as a group key. Please generate me a box plot in python for each of the metric in the data frame, so I can compare their performance.
- Data transfer: The attached csv file contains the results of my research. Please convert it to a LaTeX table. DO NOT ALTER ANY VALUES AND PRESENT THEM EXACTLY AS THEY ARE IN THE FILE!
- Docs and code reformatting: Refactor the provided code to make it neater and less repetitive. Generate documentation for the explicit methods and write in-line comments for important points. Code: *pasted code*

[3]https://github.com/vguenov/RP-laughing-pose-estimation

- Text paraphrasing: The following sentence is too long and hard to understand for a reader who do not know the field of my research in depth. Please paraphrase it in a more digestible manner. Sentence: *pasted sentence*
- LaTeX help: To following diagram is behaving not as intended. It goes all the way down after reference. I want it placed exactly in the place I have placed in; can you help? *paste figure code*

References

[1] W. Curran, G. J. McKeown, M. Rychlowska, E. André, J. Wagner, and F. Lingenfelser. Social context disambiguates the interpretation of laughter. *Frontiers in Psychology*, 8:1–12, 2018.

[2] C. Darwin. Expression of the emotions in man and animals. *Nature*, 36(926):294–295, 1887.

[3] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th International Conference on Pervasive Computing Technologies for Healthcare*, pages 1–10. ACM, 2019.

[4] Stéphane Dupont, Hüseyin Çakmak, Will Curran, Thierry Dutoit, Jennifer Hofmann, Gary McKeown, Olivier Pietquin, Tracey Platt, Willibald Ruch, and Jérôme Urbain. Laughter research: A review of the ilhaire project. In *Toward Robotic Socially Believable Behaving Systems*, pages 147–181. Springer, 2016.

[5] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. Robust laughter detection in noisy environments. In *Proceedings of Interspeech 2021*, pages 2481–2485, 2021.

[6] Harry J Griffin, Min SH Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives. *IEEE Transactions on Affective Computing*, 6(2):165–178, 2015.

[7] Josip Tomo Licardo, Mihael Domjan, and Tihomir Orehovački. Intelligent robotics—a systematic review of emerging technologies and trends. *Electronics*, 13(3):542, 2024.

[8] R. Niewiadomski, J. Urbain, C. Pelachaud, and T. Dutoit. Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases. In *Proc. 4th Int. Workshop Corpora Res. Emotion*, pages 25–32, 2012.

[9] Radoslaw Niewiadomski, Maurizio Mancini, Giovanna Varni, Gualtiero Volpe, and Antonio Camurri. Automated laughter detection from full-body movements. *IEEE Transactions on Human-Machine Systems*, 46(1):113–123, 2016.

[10] S. Petridis, B. Martinez, and M. Pantic. The mahnob laughter database. *Image and Vision Computing*, 31(2):186–202, 2013.

[11] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. Conflab: A data collection concept, dataset, and benchmark for machine analysis of free-standing social interactions in the wild. In *NeurIPS Datasets and Benchmarks*. NeurIPS, 2022.

[12] Gordon Rennie. *Automatic Detection of Laughter in Spontaneous Conversations*. Phd thesis, University of Glasgow, 2024.

[13] Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183, 2012.

[14] Jose David Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild. *IEEE Transactions on Affective Computing*, 15(2):519–534, 2024.

[15] Adrienne Wood and Paula Niedenthal. Developing a social functional account of laughter. *Social and Personality Psychology Compass*, 12(4):e12383, 2018.

[16] David L. Woods, John M. Wyma, E. William Yund, Timothy J. Herron, and Bruce Reed. Factors influencing the latency of simple reaction time. *Frontiers in Human Neuroscience*, 9:131, 2015.