Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation

Michal Balazia*
michal.balazia@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

Philipp Müller*
philipp.mueller@dfki.de
DFKI Saarbrücken
Saarbrücken, Germany

Ákos Levente Tánczos
akos.tanczos@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

August von Liechtenstein
august.liechtenstein@dfki.de
DFKI Saarbrücken
Saarbrücken, Germany

François Brémond
francois.bremond@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

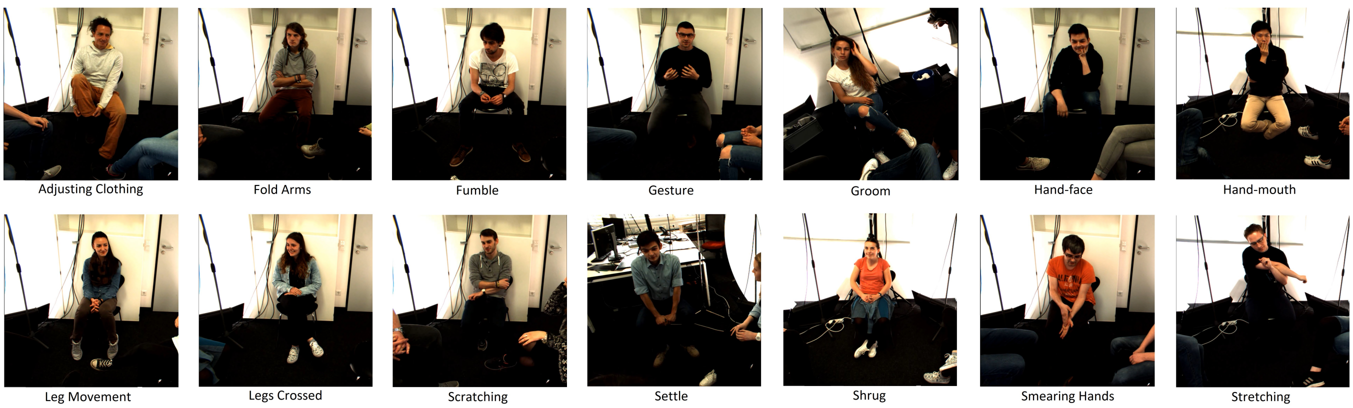


Figure 1: Examples of annotated bodily behaviors.

ABSTRACT

Body language is an eye-catching social signal and its automatic analysis can significantly advance artificial intelligence systems to understand and actively participate in social interactions. While computer vision has made impressive progress in low-level tasks like head and body pose estimation, the detection of more subtle behaviors such as gesturing, grooming, or fumbling is not well explored. In this paper we present BBSI, the first set of annotations of complex Bodily Behaviors embedded in continuous Social Interactions in a group setting. Based on previous work in psychology, we manually annotated 26 hours of spontaneous human behavior in the MPIIGroupInteraction dataset with 15 distinct body language classes. We present comprehensive descriptive statistics on the resulting dataset as well as results of annotation quality evaluations. For automatic detection of these behaviors, we adapt the Pyramid Dilated Attention Network (PDAN), a state-of-the-art approach for human action detection. We perform experiments using four variants of spatial-temporal features as input to PDAN: Two-Stream Inflated 3D CNN, Temporal Segment Networks, Temporal Shift Module and Swin Transformer. Results are promising and indicate a great room for improvement in this difficult task. Representing a key piece in the puzzle towards automatic understanding of social behavior, BBSI is fully available to the research community.

CCS CONCEPTS

• **Computing methodologies** → **Image and video acquisition**; **Activity recognition and understanding**; • **Human-centered computing** → *Collaborative and social computing*; • **Applied computing** → *Psychology*.

KEYWORDS

dataset, body pose, gesture, social signals, behavior detection

*Authors contributed equally.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548363

ACM Reference Format:
Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548363

1 INTRODUCTION

Bodily movements and poses are a key aspect of human behavior in social interaction [67] and are indicative of a large variety of personal and interpersonal information [49]. For example, leaning of the torso was found to be related to liking of the addressee [38], and behaviors like fumbling, grooming or face touching are related to the regulation of stress [5]. Furthermore, body language was shown to have distinct effects both on its perceivers as well as on its producers. Dominant, open bodily displays can be perceived as attractive [66] and gesturing was shown to lighten the cognitive load [25, 52] and improve memory [14]. As a result, machines that are supposed to understand and participate in social interactions need to be able to accurately sense and interpret body language.

Over recent decades, huge advances were made in human body- and hand pose estimation [3, 9, 28, 47]. At the same time, a large number of works investigated the prediction of high-level attributes based on bodily behavior [1, 7, 42]. For example, body movements were utilized for detection of emergent leadership [7] and recognition of emotions [42] or personality types [1]. These approaches typically use generic feature sets extracted from pose estimates or rely on CNN-based visual representations. While such approaches have the advantage of being relatively task-agnostic, they run the danger of missing subtle differences in behavior, such as between scratching and fumbling, that can only be exploited with fine-grained annotation. They further suffer from subjective or ambiguous annotation and from the lack of interpretability associated with a psychologically-motivated mid-level representation of behavior [24, 48, 56, 65], which is especially important if a behavior analysis is supposed to be accepted by practitioners like clinical or organizational psychologists.

Despite the advantages of a mid-level representation of bodily behavior in human interactions, automatic approaches for the detection of such behaviors are scarce [6, 35]. The main reason for this is the lack of suitable datasets for training and evaluation. The few existing datasets either only cover a single behavior like touching the face with the hands [6], or focus on single people only and are at present not publicly available [35]. To overcome this limitation, we present the first publicly available annotations of a comprehensive set of body language classes embedded in continuous group conversations. Our choice of behavior classes is motivated by previous work in psychology [65]. As a basis for annotation, we make use of a naturalistic multi-view group interaction dataset [44, 45] which will enable future research to study body language in the context of high-level social phenomena such as leadership, rapport, or liking.

Our specific contributions are threefold: First, we introduce Bodily Behaviors in Social Interaction (BBSI), a set of novel annotations for 15 bodily behavior classes on the MPIIGroupInteraction dataset [45]. BBSI comprises 2.87 million frames of annotated behavior classes from 26 hours of human behavior embedded in continuous group interactions. Second, we provide detailed descriptive analyses on the collected annotations as well as the results of a dedicated experiment quantifying annotator agreement. Third, we evaluate several state-of-the-art action detection approaches on BBSI, reaching 61.3% True Positive Rate with the Pyramid Dilated Attention Network [15] and Swin Transformer [37] features.[1]

[1]Data and code are available at https://git.opendfki.de/body_language/acm_mm22.

2 RELATED WORK

Our work is related to the function of body language in social interactions, to approaches for the recognition of actions and body language, as well as to existing human body language datasets.

2.1 Body Language in Social Interaction

Body language has been actively researched by psychologists for decades [25, 38, 66]. Early work by Mehrabian [38, 39] found that, among other signals, backward leaning of the torso is indicative of liking. Dominant and open nonverbal displays, as opposed to folded arms and crossed legs, are perceived as attractive when meeting with strangers [66]. In a meta-analysis, [26] found significant correlations between perceived social verticality and for example self-touching and gesturing. A further study by [10] indicated that people believe power is expressed with nonverbal cues like open posture (i.e. no arms crossed or legs crossed), more gesturing, and less self-touching (both hands and face). Furthermore, leaning towards the interlocutor was shown to be associated with rapport [58], and crossed arms were shown to be associated with emotion expressions [68]. Displacement behaviors such as grooming, face touching or fumbling are related to anxiety and stress regulation [5, 40, 41].

As a consequence of these manifold connections of body language with important personal and social attributes, body language analysis has been a focus of automatic approaches attempting to infer high-level attributes such as emotion [23, 42, 53], leadership role [7, 43], or personality type [1, 54]. In contrast to the human science studies discussed above, these automatic approaches commonly lack an explicit intermediate representation of functional bodily behavior categories. Instead, they rely on a generic feature representation encoding body postures and movements [7, 42, 43] or on deep learning approaches [53, 54] without easily interpretable internal structure. While such representations can be effective in prediction scenarios, they often lack interpretability and may miss subtle but meaningful differences, e.g. between fumbling and scratching. In this work, we draw upon the ethological rating scheme of functional body language categories described in [65] to derive a set of bodily behaviors that are intuitively interpretable and allow to train models for fine-grained behavior distinctions.

2.2 Recognition of Actions and Body Language

RGB-based human action recognition has often been addressed by three main approaches. Two-stream 2D Convolutional Neural Networks [29, 60, 73] generally contain two 2D CNN branches taking different input features extracted from the RGB videos for action recognition. Recurrent Neural Networks (RNN) [17, 36, 72] usually employ 2D CNNs as feature extractors for an LSTM model. 3D CNN-based methods [20, 63, 64] extend 2D CNNs to 3D structures, to simultaneously model the spatial and temporal context information in videos that is crucial for action recognition.

Among the many available human action recognition methods we choose the following three for our evaluations: A well-cited two-stream 2D CNN architecture by Wang et al. [69] which divides each video into three segments and processes each segment with a two-stream network, fusing the individual classification scores by an average pooling method to produce the video-level prediction. A revolutionary method by Carreira and Zisserman [11] which

Table 1: Datasets annotated with body language classes described in the literature. *Behaviors* indicates the number of annotated body language classes, *Participants* the number of human individuals, *Length* the length of annotated behavior, *Views* the number of synchronized camera views on each participant, *Group Size* the number of participants that were synchronously annotated, *Spontaneous* whether behavior was shown spontaneously, and *Public* whether the dataset is publicly available. NTU is in italic, as only a subset of its classes are body language.

Name	Behaviors	Participants	Length	Views	Group Size	Spontaneous	Public
iMiGUE [35]	32	72	35h	1	1	✓	✗
PAVIS Face-Touching [6]	1	64	22h	1	4	✓	✓
EMILYA [21]	7	11	6h	1	1	✗	✓
NTU RGB+D 60/120 [34, 57]	*60/120*	*40/106*	*133h/266h*	*80/155*	*1–2*	✗	✓
BBSI (ours)	15	78	26h	3	3–4	✓	✓

introduces the two-stream Inflated 3D CNN inflating the convolutional and pooling kernels of a 2D CNN with an additional temporal dimension. And the best performance method tested by [35] on body language recognition by Lin et al. [33] of a parameter-free Temporal Shift Module, which shifts a part of the channels along the temporal dimension to perform temporal interaction between the features from adjacent frames. We also experiment with the transformer method by Liu et al. [37] that was designed for natural language processing but its application has been recently extended to computer vision tasks [18, 31].

In contrast to action recognition, which typically considers freely moving people [16, 30, 59], the much thinner work on body language recognition addresses more constrained social interaction scenarios. For example, Yang et al. [70] generate sequences of body language predictions from estimated human poses and feed them to an RNN for emotion interpretation and psychiatric symptom prediction. Kratimenos et al. [32] extract a holistic 3D body shape, including hands and face, from a single image and feed them also to an RNN for sign language recognition. Singh et al. [61] use handcrafted features to analyze body language for estimating a person's emotions and state of mind. Santhoshkumar et al. [55] use Feedforward Deep CNNs for detecting emotions from full body motions. We observe that the common denominator of body language analysis methods are the employment of a general action recognition method and the lack of a benchmark body language dataset.

2.3 Human Body Language Datasets

In contrast to datasets with annotations of high-level attributes like emotions [42, 53], leadership [7, 45], or personality [50], datasets annotated with concrete classes of bodily behavior are sparse. Table 1 summarizes four relevant datasets with manual body language annotations. Research that extracted body language automatically but did not provide human annotations is not included [19, 23, 53].

In EMILYA [21], actors were asked to express different emotions while performing daily actions such as walking, sitting down, or moving objects. Two datasets NTU RGB+D 60/120 [34, 57] contain a large number of general action classes that also include a number of body language classes. However, these datasets do not consist of spontaneous human behavior. The two most relevant datasets for our work are the PAVIS Face-Touching dataset [6] and iMiGUE [35]. PAVIS Face-Touching is similar to BBSI as it also consists of recordings of group discussions. In contrast to the 15 behavior classes annotated in BBSI, PAVIS Face-Touching only has binary annotations of whether a participant touches her face or not. Furthermore it only has a single frontal view on each participant. The recently introduced iMiGUE dataset [35] consists of annotations of 32 behaviors classes of speakers at sports press conferences. Annotations are only provided for a single person (i.e. no annotations of discussion partners), and only a single view on the target person is provided. At the time of submission, the iMiGUE videos are not publicly accessible due to privacy issues[2]. We hereby present the first publicly available annotations of body language on a multi-view dataset of three to four people engaged in spontaneous group discussions.

3 DATASET

BBSI builds upon the MPIIGroupInteraction dataset [45]. This dataset comprises of 22 three- to four-person group discussion on controversial topics, each lasting for 20 minutes. In total, it consists of 78 participants and 26 hours of behavior recordings. Every interaction was recorded by 8 frame-synchronized cameras as well as with 4 microphones. After the discussions, participants rated their perceived leadership, competence, dominance and liking of all other members, as well as their feelings of rapport towards each other. In addition to rapport and emergent leadership prediction [43, 45], the dataset was further annotated and used for eye contact detection [22, 46] and for next speaker prediction [8, 44]. This wealth of already existing annotations makes the MPIIGroupInteraction dataset a perfect choice for the collection of body language labels as it will allow future research on the connections and the utility of body language information with key group phenomena.

3.1 Body Language Annotation

We densely annotated the full MPIIGroupInteraction dataset with 15 body language classes (see Figure 1 and Table 2). Our set of behavior classes is based on the the Ethological Coding System for Interviews (ECSI) [65]. This coding system includes many bodily behaviors that were shown to be connected to different social phenomena, as described in Section 2.1. We selected all ECSI behaviors involving the limbs and torso and excluded behavior classes based on facial behavior, gaze, and head pose as these are not the focus of this work and highly accurate methods to analyze such behaviors

[2] According to a note dating from September 2021 on the official github page of iMiGUE (https://github.com/linuxsino/iMiGUE), the file containing the links to the videos used in the dataset has been removed for privacy protection.

Table 2: Behavior classes in the dataset, including descriptions, number of annotated frames, annotation instances, and annotator agreement.

Behavior	Description	# Frames	# Instances	Agreement
Adjusting Clothing	Clothing is adjusted	23k	250	0.77
Fold Arms	Arms are folded across the chest	251k	200	0.82
Fumble	Twisting and fiddling finger movements	422k	1374	0.54
Gesture	Variable hand and arm movements during speech	373k	2607	0.85
Groom	Fingers are passed through the hair in a combing movement	17k	282	0.71
Hand-face	Hand(s) in contact with the face	79k	535	0.79
Hand-mouth	Hand(s) in contact with the mouth	55k	318	0.74
Lean Towards	Leaning forward from the hips towards the interlocutor	5k	72	0.13
Leg Movement	Repetitive movement of legs	14k	860	0.51
Legs Crossed	Legs are crossed	1397k	77	0.87
Scratch	Fingernails are used to scratch parts of the body	72k	519	0.61
Settle	Adjusting movement into a more comfortable posture in the chair	40k	290	0.54
Shrug	Shoulders are raised and dropped again	8k	192	0.57
Smearing Hands	Smearing hands on clothing	21k	298	0.54
Stretching	Stretching of body parts	4k	31	0.61

already exist [4, 62]. We also excluded the two classes *Crouch* and *Relax*, as they were only very rarely annotated (Crouch: 411 frames, Relax: 2k frames), rendering estimation of classification performance meaningless. In addition to the bodily behaviors included in ECSI, we scanned the MPIIGroupInteraction dataset for additional behaviors that occur frequently and carry potential meaning in a social situation. As a result, we included the five additional classes: *Adjusting Clothing*, *Leg Movement*, *Legs Crossed*, *Smearing Hands*, *Stretching*.

To achieve high-quality annotations while keeping costs manageable, we designed the following annotation procedure. First, we trained three annotators on the task by providing examples and discussing edge cases jointly. In this way, we made sure that the annotators arrived at a common understanding of the body language classes. Each of the 78 participants of the MPIIGroupInteraction dataset was fully annotated by one of the annotators. Subsequently, each of the resulting annotations was checked by another annotator to further improve quality. This procedure of annotation followed by checking proved to be much more economical than collecting several separate annotations of the same video. We used a separate experiment to quantify annotation quality (see Section 3.2).

3.2 Analysis of Annotations

3.2.1 Descriptive Statistics. In total, 2.87 million frames of body language were annotated across all classes for the full 26 hours of video. Each annotation instance is defined by a specific behavior label and a start time and an end time between which the behavior appears continuously on all frames. Table 2 shows that the annotation across the 15 behavior classes has highly uneven number of annotated frames and instances. The most frequently annotated class, *Legs Crossed* was annotated for 1397k frames, while *Stretching* was only annotated for 4k frames. As a complementary view on the quantity of annotations, *Legs Crossed* has the highest number of annotated frames, but only over 77 annotation instances, meaning that participants remained for a crossed-leg position for extended periods of time. On the other hand, *Gesture* is annotated on less frames (373k), but consists of many more distinct instances (2607).

Another important aspect of BBSI is its multi-label characteristic, that is, several body language classes can occur simultaneously. Figure 2 shows the co-occurrence patterns of body language classes. Strong co-occurrences can be observed between the lower body classes (*Legs Crossed*, *Leg Movement*) and upper-body behaviors. Co-occurrences between upper-body behaviors do exist, but are more sparse. As a result, BBSI creates a challenging multi-label classification problem.

3.2.2 Annotation Quality. To obtain a numerical estimate of annotation quality, we performed a dedicated experiment based on the collected annotations. We sampled 800 4-second clips from the full dataset that were classified into body language classes separately by all three annotators. These samples were drawn randomly from the whole dataset with the following constraints: First, we considered a 4-second window to be a sample of a body language class if either the class is annotated for at least 2 seconds of this window, or if the 4-second window completely encompasses the corresponding annotation instance. Second, we drew 50 samples of each behavior class. Estimated with the rate of class co-occurrences, the precise number of instances for each class in the 800 samples may be larger than 50. For comparability, we used the same metric as [35] which computes the agreement of two annotators by dividing twice the number of annotated behaviors for which they agree by the total number of behaviors annotated by both.

Table 2 shows the resulting agreements for each class separately. Very high agreements above 0.8 are reached for frequent classes such as *Legs Crossed*, *Gesture*, or *Fold Arms*. All other classes are in the range of 0.5 to 0.8 with the only exception of *Lean Towards* which was proven very challenging to annotate with only 0.13 agreement. Liu et al. [35] do not provide class-specific annotator agreement but only a global measure in which frequent classes contribute more than less frequent classes, i.e. micro average. To get an estimate how our annotator agreement relates to the agreement of 0.81 reported in [35], we weight our class-specific agreements by the frame-wise label distribution on BBSI, reaching an agreement of 0.78. Note however, that these numbers are not directly comparable due to different behavior classes and annotation protocols.

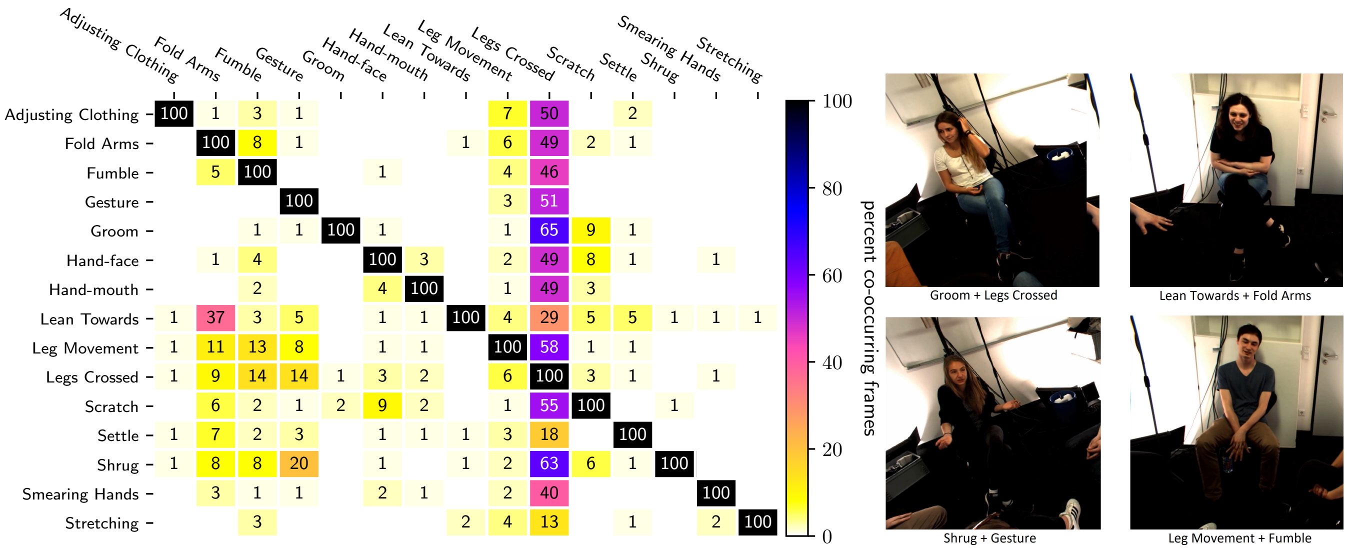


Figure 2: Co-occurrences of body language classes. Each row shows the percentage with which other classes are annotated at the same time as the class designated on the y-axis.

4 METHOD

For detecting the behaviors in the long input videos, we propose a baseline method based on the Pyramid Dilated Attention Network [15] for action detection. The model is fed with features extracted by four types of action recognition architectures.

4.1 Feature Extraction Networks

We examine the following four established algorithms that are designed for general action recognition tasks.

4.1.1 Two-Stream Inflated 3D CNN. Extent of a pre-training boost depends on the ability of a model architecture to adapt to a given pre-training dataset. As the 3D image classification backbone, the Two-Stream Inflated 3D CNN (I3D) [11] uses the ImageNet-pretrained Inception V1 with batch normalization. Filters and pooling kernels of very deep 2D image classification CNNs are inflated into 3D to learn spatio-temporal feature extractors from video.

4.1.2 Temporal Segment Networks. An obvious problem of the two-stream CNNs is their inability to model long-range temporal structure due to their access to only a limited stack of frames. Temporal Segment Networks (TSN) [69] operate on a sequence of short video clips sparsely sampled from the entire video. Each clip in this sequence will produce its own preliminary prediction of the action classes. Prediction over the full video is then derived from a consensus among the partial clip predictions. In the learning process, the loss values of video-level predictions, other than those of clip-level predictions which were used in two-stream CNNs, are optimized by iteratively updating the model parameters.

4.1.3 Temporal Shift Module. Traditional 3D convolution uses a 3D convolution kernel to perform convolution operations between adjacent multiple frames at the same time, which can extract the spatio-temporal feature information in the video at the cost of an increase in calculation. Temporal Shift Module (TSM) [33] uses a simple data preprocessing method to convert the invisible temporal information in a single frame into extractable spatial feature information. Several consecutive frames are stacked to form the original tensor and the channels are moved forward and backward in the temporal dimension to perform a simple feature fusion between the consecutive frames. The fusion makes an independent single frame contain certain temporal information, and simple 2D convolution can be used to achieve spatiotemporal feature extraction.

4.1.4 Swin Transformers. Adapting the network architectures in natural language processing to the domain of computer vision suffers from large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text. Building upon the Transformer designed for sequence modeling and translation tasks, the Swin Transformer (Swin) [37] is a hierarchical Transformer whose representation is computed with Shifted windows. The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.

4.2 Training Feature Extractors

As human body language ground truth contains temporally overlapping labeled segments as well as unlabeled sections, we investigate two training settings for the feature extraction networks: *Single-Label* training and *Multi-Label* training. In Multi-Label training, a label is assigned to a clip if it overlaps an annotated segment in at least half of its duration. On BBSI, 25% of the samples have no label assigned, 49% samples have one label and the remaining 26% on an overlap have between two and five labels. For the Single-Label setting, we selected the samples from the Multi-Label setting with at least one label. Each sample was assigned precisely one label and those originally with multiple labels are copied multiple times, each time with a single and unique label. The loss function we used was the cross entropy loss followed by the softmax activation function. For each sample, each feature extractor returns a confidence score for each behavior class. Prior to the output layer, each feature extractor produces a 2048-dimensional feature vector that is used as input to the behavior detection method.

4.3 Behavior Detection

Detection of behaviors from a long video is done by feeding the extracted feature vectors into an action detection architecture. The Pyramid Dilated Attention Network (PDAN) [15] uses a self-attention mechanism to capture temporal relations. The layers that make up the network are called Dilated Attention Layers (DAL). A DAL takes each segment as a center segment and concatenates its feature representation with the feature representation of segments being D-far in both directions from the center segment, where D is the dilation rate. At this point, it applies self-attention on the extracted segment-feature representations. PDAN is based on a pyramid of DALs with same kernel sizes and dilation rates that exponentially increase their temporal receptive field. The output of PDAN consists of a list of predicted behaviors with their beginnings and endings tied to the segmentation cuts, and their confidences.

4.4 Implementation Details

The feature extraction methods operate on fixed inputs of length 16 frames and size 224×224 pixels. Consequently, we resize the videos appropriately and cut the long dataset videos into 16-frame video clips. These clips are assigned with the corresponding behavior class label and treated as independent samples for training and evaluation. This splitting of videos does not disconnect the flow of the actions as the annotated behaviors are mostly non-transitional, that is, the actions described by these behaviors do not change people's body poses from one to another. For instance, a 64-frame-long behavior *Hand-mouth* can be split into four 16-frame-long clips in which a person keeps touching their mouth. Advantages are that the number of samples increases significantly and that the fixed-length clips can be input to all methods with an equal FPS.

All action recognition models are pre-trained on ImageNet and Kinetics-400 and the action detection model is used without any pre-training. Fine-tuning on BBSI is performed on both levels, action recognition and action detection. For comparability, all models were trained for 15 epochs. Learning rates are set to: I3D 10^{-2}, TSN 10^{-3}, TSM $7.5 \cdot 10^{-4}$, Swin 10^{-3} with AdamW optimizer, and PDAN 10^{-1}. Our implementation uses the open-source toolbox MMaction2 [13] built on top of PyCharm.

5 EVALUATION

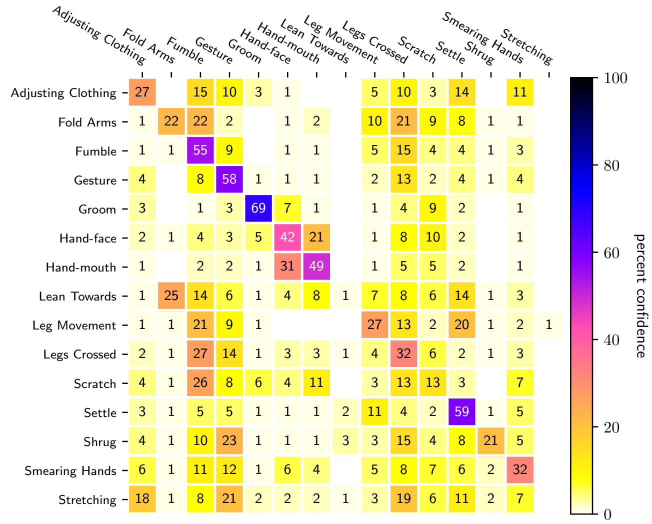
We provide evaluation results of our baseline method at two levels: quality of the extracted features and quality of the final detection. As feature extractors are trained as classifiers, they are evaluated with standard classification metrics, and the final detector is evaluated on standard detection metrics. In all experiments, we use the training/validation split of MPIIGroupInteraction reported in [44]: training recordings 07, 10–25; validation recordings 08, 09, 26–28.

5.1 Classification Evaluation

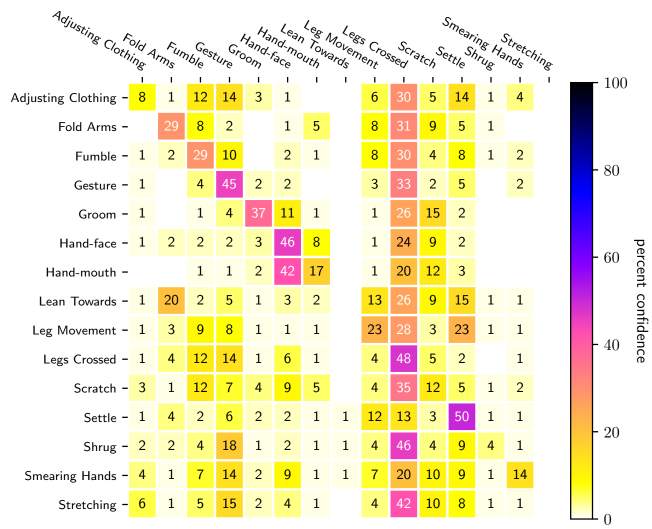
Table 3 shows classification performance of the feature extraction networks, both for the Multi-Label as well as for the Single-Label training setting. We also include the random classifier as a baseline. Results are reported in terms of Mean Average Precision (MAP) using micro averaging (same weight for each sample) and macro averaging (same weight for each class). The best result is achieved by Swin Transformers in the multi-label training scenario, reaching 0.75 micro averaging MAP and 0.37 macro averaging MAP. The

Table 3: Five action recognition architectures, four proposed methods and one random, in both labeling settings. Each value is MAP score computed using micro/macro averaging.

Method	Single-Label	Multi-Label
random	0.258 / 0.067	0.377 / 0.106
I3D [11]	0.445 / 0.212	0.624 / 0.284
TSN [69]	0.520 / 0.232	0.661 / 0.308
TSM [33]	0.508 / 0.228	0.721 / 0.313
Swin [37]	0.601 / 0.305	0.745 / 0.374



(a) Single-Label Swin [37]



(b) Multi-Label Swin [37]

Figure 3: Confidence matrices of behavior recognition by the Swin Transformer trained in both labeling settings.

Table 4: Effects of exclusion of static classes and class balancing on behavior recognition by TSM [33]. Each value is a MAP using micro/macro averaging.

static included	balancing	Single-Label	Multi-Label
✓	✓	0.508 / 0.228	0.721 / 0.319
✓	✗	0.595 / 0.294	0.746 / 0.384
✗	✓	0.601 / 0.279	0.618 / 0.305
✗	✗	0.658 / 0.333	0.639 / 0.336

second best method in the Multi-Label setting was TSM with 0.72 and 0.31 micro- and macro averaging MAP. In the Single-Label scenario, Swin Transformers also reached the best performance. The second best method in this case is TSN. All feature extraction networks clearly outperformed the random baseline.

Evaluation of all methods can be visualized by aggregating all confidence vectors into a confidence matrix. Rows of this 15×15 square matrix are ground truth classes and columns are prediction confidences. The matrix is constructed by adding all confidence vectors into the corresponding ground truth rows and then dividing each row by the number of its summands. In the ideal case, this matrix would coincide with the co-occurrence matrix presented in Figure 2. See Figure 3 for the confidence matrices of behavior recognition by the Swin Transformer for both Single-Label and Multi-Label training. Compared to Single-Label training, the Multi-Label network is able to more accurately model class co-occurrences, especially with *Legs Crossed*. We report further confidence matrices in the supplementary material.

We performed additional ablation experiments with the TSM model. First, as the behaviors *Legs Crossed* and *Fold Arms* make forms of static positional body pose rather than dynamic motion actions, they can be recognized on a frame level and with an eventual aid of a generally non-temporal skeleton estimation technique. We evaluated the training and evaluation scenario with only 13 classes, excluding these two static classes. And second, as the dataset has considerably imbalanced class frequencies (ranging from 4k annotated frames to more than a million, see Table 2), overrepresented behaviors have a too high impact on training compared to underrepresented ones. Therefore, we evaluated the influence of class balancing by randomly selecting 20k samples from each class to counteract weight of overrepresented classes while keeping all samples of the underrepresented classes. Table 4 shows the effects of static class exclusion and class balancing. We observe systematic advantage of excluding static classes in the Single-Label setting and of including static classes in the Multi-Label setting, and of no class balancing overall.

5.2 Detection Evaluation

In addition to MAP, the evaluation metrics are calculated from the detection confidence vector on the frame level: True Positives (TP) as the sum of confidences in true classes, False Positives (FP) as the sum of confidences in false classes, and False Negatives (FN) as the sum of 1 minus confidences in true classes. From TP, TN and FN on the frame level, we calculate the F1 score globally using both micro and macro averaging. As expected from behavior recognition, Swin achieves the best results with F1 0.728/0.544 and MAP 0.742/0.415

Figure 4: Illustrative examples of true positive, false positive and false negative predictions.

on the Single-Label setting, and F1 0.726/0.511 and MAP 0.691/0.367 on Multi-Label, using micro/macro averaging respectively. Figure 4 illustrates examples of true positive, false positive and false negative predictions. As in behavior detection without class balancing, there is a high inter-class performance variance. The most frequent class *Legs Crossed* reaches the highest performance among all classes.

6 DISCUSSION

6.1 Annotations

We presented the first publicly available annotations of 15 body language classes on a multi-view group discussion dataset. Annotating human bodily behavior is challenging due to the subtle and often subjective nature of body language. To evaluate the agreement of our annotators, we conducted a dedicated experiment. Our class-based analysis of annotator agreement revealed clear differences between the agreement for different behavior classes, which should be taken into account by potential users of the dataset. Restricting to a subset of the annotated classes to those with a proper relevance to the particular application and a high inter-annotator agreement can be a good practice for any body language analysis system. On the other hand, if body language annotations are used to train a feature representation that is used in a downstream task, even low agreement classes can still be useful. On the other hand, if body language annotations are used to train a feature representation that is used in a downstream task, even classes with low annotator agreement can still be useful. We include classes with low agreement scores for full transparency and leave it to the users to decide which classes to use depending on their preferences. Supplementary material provides class-specific evaluation results to facilitate comparison with researchers who choose a subset of classes.

6.2 Achieved Performance

In line with the recent trend on computer vision tasks [18, 31], the effectiveness of transformers is also reflected on BBSI. Even the Tiny version of Swin Transformers has outperformed all other

Table 5: Evaluation of PDAN [15] with four types of features trained in both labeling settings in terms of F1 and MAP using micro/macro averaging.

Features	Single-Label F1	Single-Label MAP	Multi-Label F1	Multi-Label MAP
random	0.348 / 0.114	0.312 / 0.092	0.348 / 0.115	0.312 / 0.092
I3D [11]	0.542 / 0.340	0.502 / 0.276	0.581 / 0.339	0.557 / 0.271
TSN [69]	0.550 / 0.309	0.624 / 0.291	0.531 / 0.343	0.658 / 0.311
TSM [33]	0.660 / 0.419	0.600 / 0.311	0.627 / 0.347	0.626 / 0.293
Swin [37]	0.729 / 0.545	0.742 / 0.415	0.726 / 0.511	0.691 / 0.367

CNN-based architectures in every setup where it was applied. This is usually followed by TSM and TSN, although I3D has a higher potential due to its 10-times larger number of parameters.

Class balancing degrades the performance in any setup. Although it was introduced to counteract the dominance of static classes, the MAP drop is the highest in those setups where the static classes are included. Our assumption is that equally balancing the dataset is not adequate in this case as the distribution of instance numbers per classes are exponential. Despite giving equal weight to the classes of very few instances increases their performance on the training set, they are not possible to achieve good recognition on unseen data. Not only it does not improve testing inference, the metrics of other classes fall as well.

Applying the Single-Label setting on a detection task inherently produces incorrect predictions. As in most of the cases there is at least one static class involved in concurrent actions, excluding static classes results in a classification problem of a significantly reduced rate of multiple labels. Thus, the difference between the Single-Label and Multi-Label experiments when the static classes are excluded is almost negligible compared to the case when all the classes are included, which is in the range between 0.005–0.088 MAP if there is no class balancing applied.

6.3 Applications

The primary intended application for BBSI annotations is to train and evaluate algorithms that predict body language classes. However, our annotations can also be useful in a pre-training step or for auxiliary training of approaches that address high-level behavior interpretation tasks such as leadership detection [7, 43] or personality prediction [1, 50] for which only limited amount of training data is available. Furthermore, it can be of interest for behavioral scientists to use our annotations for research on the expression of nonverbal behavior in group interactions and how it relates to aspects like leadership, rapport, or interpersonal synchrony.

As BBSI is based on a rating scheme developed in the context of psychiatric interactions [65], we expect our body language predictions to be highly useful in clinical tasks, e.g. for depression detection [71] or to estimate the quality of the therapist-patient relationship [27]. Using a set of psychologically motivated behaviors as an intermediate representation instead of generic pose-based features or deep learning representations will allow for better interpretability and build trust with clinicians and patients alike. Our presented prediction methods can also be integrated into existing conversation analysis tools [51], which at present do not have the ability to detect fine-grained body language.

6.4 Limitations and Future Work

Our novel annotations and state-of-the-art evaluations represent an important step towards automatic analysis of body language in social interaction. At the same time, several challenges remain that need to be addressed in future work. While the BBSI set of behavior classes is motivated by previous work linking those classes to social attributes like leadership, rapport, or emotions, this link needs to be solidified by investigating the predictive power of bodily behaviors for such downstream tasks. Furthermore, as the MPIIGroupInteraction dataset consists of participants recruited at a German university, future work should collect comparable datasets with more diverse cultural backgrounds. A key challenge on BBSI is the large class imbalance that makes it difficult to train accurate models for classes that occur seldomly in natural behavior. Future work could investigate generation of synthesized training examples or advanced data augmentation techniques. The detection and classification approaches presented in this paper learn a single model that is applied to all participants. While this is a meaningful first step to approach the task, the expression of body language is highly individual. Future work should investigate personalization and test-time adaptation [2, 12] to model personal idiosyncrasies adequately. Another possible improvement is to use multi-channel inputs, e.g. by exploiting all three views on a person, or adding pose information [9].

7 CONCLUSION

In this work, we presented BBSI, the first publicly available set of annotations of subtle bodily behaviors in group interactions. The novel annotations consist of 15 body language classes that were densely annotated for 26 hours of human behavior recorded from 78 participants on the publicly available MPIIGroupInteraction dataset. We provided results of descriptive analyses of the annotations as well as a dedicated experiment on annotation quality, as they were done manually by our human annotators. Furthermore, we presented the results of state-of-the-art action recognition approaches evaluated on the MPIIGroupInteraction dataset with the BBSI annotations. As such, our work is a key contribution to advance in-depth analyses of subtle body language cues in human interactions.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency under the UCA^{JEDI} Investments into the Future, project number ANR-15-IDEX-01, and by the German Ministry for Education and Research, grant number 01IS20075.

REFERENCES

[1] Tanay Agrawal, Dhruv Agarwal, Michal Balazia, Neelabh Sinha, and François Bremond. 2022. Multimodal Personality Recognition using Cross-attention Transformer and Behaviour Encoding. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*. INSTICC, SciTePress, 501–508. https://doi.org/10.5220/0010841400003124

[2] Sikandar Amin, Philipp Müller, Andreas Bulling, and Mykhaylo Andriluka. 2014. Test-time adaptation for 3d human pose estimation. In *German conference on pattern recognition*. Springer, 253–264.

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. 2008. People-tracking-by-detection and people-detection-by-tracking. In *2008 IEEE Conference on computer vision and pattern recognition*. IEEE, 1–8.

[4] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[5] M Bardi, T Koone, S Mewaldt, and K O'Connor. 2011. Behavioral and physiological correlates of stress related to examination performance in college chemistry students. *Stress* 14, 5 (2011), 557–566.

[6] Cigdem Beyan, Matteo Bustreo, Muhammad Shahid, Gian Luca Bailo, Nicolo Carissimi, and Alessio Del Bue. 2020. Analysis of face-touching behavior in large scale social interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 24–32.

[7] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a leader: Detecting emergent leadership in small groups using body pose. In *Proceedings of the 25th ACM international conference on Multimedia*. 1425–1433.

[8] Chris Birmingham, Kalin Stefanov, and Maja J Mataric. 2021. Group-Level Focus of Visual Attention for Improved Next Speaker Prediction. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4838–4842. https://doi.org/10.1145/3474085.3479213

[9] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 43, 01 (jan 2021), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

[10] Dana R. Carney, Dana. 2005-06-01. Beliefs about the nonverbal expression of social power. *Journal of nonverbal behavior*. 29, 2 (2005-06-01).

[11] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[12] James Charles, Tomas Pfister, Derek Magee, David Hogg, and Andrew Zisserman. 2016. Personalizing human video pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3063–3072.

[13] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[14] Susan Wagner Cook, Terina KuangYi Yip, and Susan Goldin-Meadow. 2010. Gesturing makes memories that last. *Journal of memory and language* 63, 4 (2010), 465–475.

[15] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. 2021. PDAN: Pyramid Dilated Attention Network for Action Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/WACV48630.2021.00301

[16] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. 2019. Toyota Smarthome: Real-World Activities of Daily Living. In *The IEEE International Conference on Computer Vision (ICCV)*. 833–842.

[17] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 677–691. https://doi.org/10.1109/TPAMI.2016.2599174

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

[19] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*. Springer, 488–500.

[20] M. Fayyaz, E. Bahrami, A. Diba, M. Noroozi, E. Adeli, L. Van Gool, and J. Gall. 2021. 3D CNNs with Adaptive Temporal Feature Resolutions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 4729–4738. https://doi.org/10.1109/CVPR46437.2021.00470

[21] Nesrine Fourati and Catherine Pelachaud. 2014. Emilya: Emotional body expression in daily actions database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 3486–3493.

[22] Eugene Yujun Fu and Michael W Ngai. 2021. Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4873–4877. https://doi.org/10.1145/3474085.3479230

[23] Mihai Gavrilescu. 2015. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *2015 23rd Telecommunications Forum Telfor (TELFOR)*. IEEE, 720–723.

[24] Jeffrey M. Girard, Jeffrey F. Cohn, Mohammad H. Mahoor, S. Mohammad Mavadati, Zakia Hammal, and Dean P. Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing* 32, 10 (2014), 641–647. https://doi.org/10.1016/j.imavis.2013.12.007 Best of Automatic Face and Gesture Recognition 2013.

[25] Susan Goldin-Meadow, Howard Nusbaum, Spencer D Kelly, and Susan Wagner. 2001. Explaining math: Gesturing lightens the load. *Psychological science* 12, 6 (2001), 516–522.

[26] Judith Hall, Erik Coats, and Lavonia LeBeau. 2005. Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis. *Psychological bulletin* 131 (12 2005), 898–924. https://doi.org/10.1037/0033-2909.131.6.898

[27] Adam O Horvath and Lester Luborsky. 1993. The role of the therapeutic alliance in psychotherapy. *Journal of consulting and clinical psychology* 61, 4 (1993), 561.

[28] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*. Springer, 34–50.

[29] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732. https://doi.org/10.1109/CVPR.2014.223

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. https://doi.org/10.48550/ARXIV.1705.06950

[31] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in Vision: A Survey. *ACM Comput. Surv.* (dec 2021). https://doi.org/10.1145/3505244

[32] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. 2021. Independent Sign Language Recognition with 3d Body, Hands, and Face Reconstruction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4270–4274.

[33] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7082–7092. https://doi.org/10.1109/ICCV.2019.00718

[34] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701. https://doi.org/10.1109/TPAMI.2019.2916873

[35] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10631–10642.

[36] Zhenbing Liu, Zeya Li, Ruili Wang, Ming Zong, and Wanting Ji. 2020. Spatiotemporal Saliency-Based Multi-Stream Networks with Attention-Aware LSTM for Action Recognition. *Neural Comput. Appl.* 32, 18 (sep 2020), 14593–14602. https://doi.org/10.1007/s00521-020-05144-7

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10012–10022.

[38] Albert Mehrabian. 1968. Relationship of attitude to seated posture, orientation, and distance. *Journal of personality and social psychology* 10, 1 (1968), 26.

[39] Albert Mehrabian and John T Friar. 1969. Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology* 33, 3 (1969), 330.

[40] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. 2013. Displacement behaviour is associated with reduced stress levels among men but not women. *PloS one* 8, 2 (2013), e56355.

[41] Changiz Mohiyeddini, Stephanie Bauer, and Stuart Semple. 2015. Neuroticism and stress: The role of displacement behavior. *Anxiety, stress, & coping* 28, 4 (2015), 391–407.

[42] Philipp Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 663–669.

[43] Philipp Müller and Andreas Bulling. 2019. Emergent leadership detection across datasets. In *2019 International Conference on Multimodal Interaction*. 274–278. https://doi.org/10.1145/3340555.3353721

[44] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4878–4882. https://doi.org/10.1145/3474085.3479219
[45] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*. 153–164. https://doi.org/10.1145/3172944.3172969
[46] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–10.
[47] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. 2021. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 11865–11874.
[48] Lya K. Paas Oliveros, Javier Villanueva Valle, Susana I. González Arredondo, Ana Fresán, Iván Arango de Montis, Martin Brüne, and Jairo Muñoz Delgado. 2015. Study of application and validation of the Ethological Coding System for Interviews (ECSI). *Salud Mental* 38, 1 (2015), 41–46. https://doi.org/10.17711/SM.0185-3325.2015.005
[49] Gizem Oneri Uzun. 2021. A Review of Communication, Body Language and Communication Conflict. *International Journal of Psychosocial Rehabilitation* 24, 9 (04 2021), 2833–2844.
[50] Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. 2021. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1–12.
[51] Anna Penzkofer, Philipp Müller, Felix Bühler, Sven Mayer, and Andreas Bulling. 2021. ConAn: A Usable Tool for Multimodal Conversation Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 341–351. https://doi.org/10.1145/3462244.3479886
[52] Raedy Ping and Susan Goldin-Meadow. 2010. Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science* 34, 4 (2010), 602–619.
[53] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.
[54] Marta Romeo, Daniel Hernández García, Ting Han, Angelo Cangelosi, and Kristiina Jokinen. 2021. Predicting apparent personality from body language: benchmarking deep learning architectures for adaptive social human–robot interaction. *Advanced Robotics* 35, 19 (2021), 1167–1179.
[55] R. Santhoshkumar and M. Kalaiselvi Geetha. 2019. Deep Learning Approach for Emotion Recognition from Human Body Movements with Feedforward Deep Convolution Neural Networks. *Procedia Computer Science* 152 (2019), 158–165. https://doi.org/10.1016/j.procs.2019.05.038 International Conference on Pervasive Computing Advances and Applications- PerCAA 2019.
[56] Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 32, 10 (2014), 648–658. https://doi.org/10.1016/j.imavis.2014.06.001 Best of Automatic Face and Gesture Recognition 2013.
[57] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1010–1019. https://doi.org/10.1109/CVPR.2016.115
[58] Christopher F. Sharpley and Anastasia Sagris. 1995. When does counsellor forward lean influence client-perceived rapport? *British Journal of Guidance & Counselling* 23, 3 (1995), 387–394. https://doi.org/10.1080/03069889508253696 arXiv:https://doi.org/10.1080/03069889508253696
[59] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. *ArXiv* abs/1604.01753 (2016).
[60] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 568–576.
[61] Saurabh Singh, Vishal Sharma, Kriti Jain, and Ravi Bhall. 2015. EDBL - algorithm for detection and analysis of emotion using body language. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*. 820–823. https://doi.org/10.1109/NGCT.2015.7375234
[62] Neelabh Sinha, Michal Balazia, and François Bremond. 2021. FLAME: Facial Landmark Heatmap Activated Multimodal Gaze Estimation. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–8.
[63] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. 2020. D3D: Distilled 3D Networks for Video Action Recognition. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 614–623. https://doi.org/10.1109/WACV45572.2020.9093274
[64] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6450–6459.
[65] Alfonso Troisi. 1999. Ethological research in clinical psychiatry: the study of nonverbal behavior during interviews. *Neuroscience & Biobehavioral Reviews* 23, 7 (1999), 905–913.
[66] Tanya Vacharkulksemsuk, Emily Reit, Poruz Khambatta, Paul W Eastwick, Eli J Finkel, and Dana R Carney. 2016. Dominant, open nonverbal displays are attractive at zero-acquaintance. *Proceedings of the National Academy of Sciences* 113, 15 (2016), 4009–4014.
[67] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.
[68] Harald G Wallbott. 1998. Bodily expression of emotion. *European journal of social psychology.* 28, 6 (1998).
[69] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*.
[70] Zhengyuan Yang, Amanda Kay, Yuncheng Li, Wendi Cross, and Jiebo Luo. 2020. Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation. *CoRR* abs/2011.00043 (2020). arXiv:2011.00043 https://arxiv.org/abs/2011.00043
[71] Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. 2019. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 65–71.
[72] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond Short Snippets: Deep Networks for Video Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[73] Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. 2021. Motion saliency based multi-stream multiplier ResNets for action recognition. *Image and Vision Computing* 107 (2021), 104108. https://doi.org/10.1016/j.imavis.2021.104108

Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation

Supplementary Material

Michal Balazia*
michal.balazia@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

Philipp Müller*
philipp.mueller@dfki.de
DFKI Saarbrücken
Saarbrücken, Germany

Ákos Levente Tánczos
akos.tanczos@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

August von Liechtenstein
august.liechtenstein@dfki.de
DFKI Saarbrücken
Saarbrücken, Germany

François Brémond
francois.bremond@inria.fr
INRIA Sophia Antipolis
Sophia Antipolis, France

CCS CONCEPTS

• **Computing methodologies** → **Image and video acquisition**; **Activity recognition and understanding**; • **Human-centered computing** → *Collaborative and social computing*; • **Applied computing** → *Psychology*.

KEYWORDS

dataset, body pose, gesture, social signals, behavior detection

ACM Reference Format:
Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily Behaviors in Social Interaction: Novel Annotations and State-of-the-Art Evaluation: Supplementary Material. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3503161.3548363

1 CONFIDENCE MATRICES OF BEHAVIOR RECOGNITION METHODS

Evaluation of all methods can be visualized by aggregating all confidence vectors into a confidence matrix. See Figure 1 and Figure 2 for the confidence matrices of all four behavior recognition algorithms.

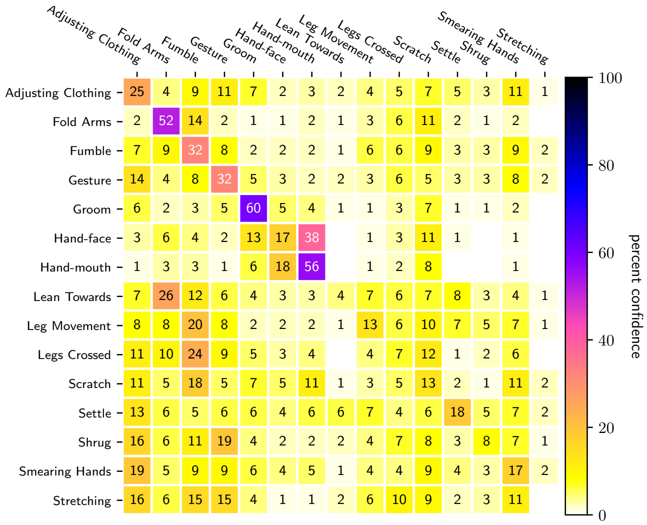
2 ILLUSTRATIVE EXAMPLES OF PDAN PREDICTIONS

We analyzed four major sources of impact on performance: additional person in the scene, occlusion, low class agreement, distinction between visually similar classes. Figure 3 illustrates these examples in the context of ground truth and prediction.

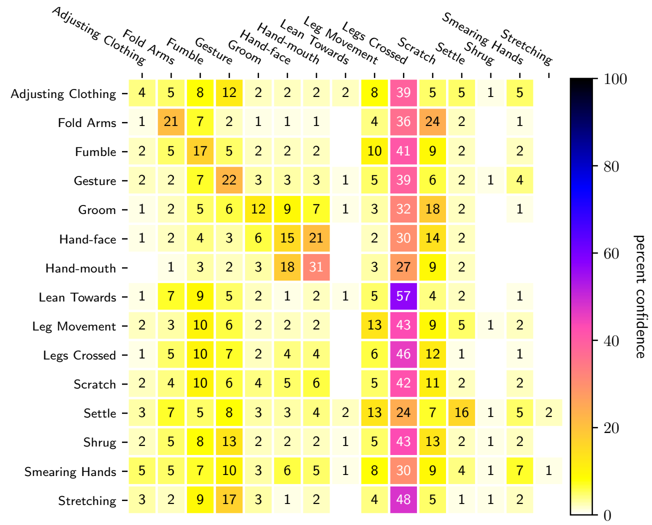
*Authors contributed equally.

MM '22, October 10–14, 2022, Lisboa, Portugal

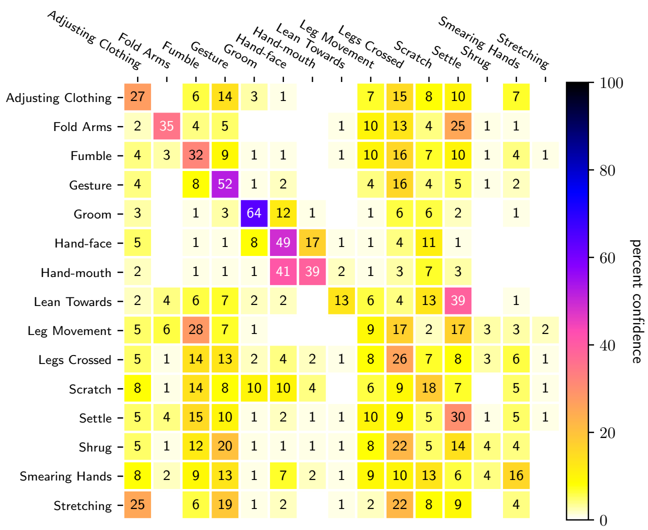
ACM ISBN 978-1-4503-9203-7/22/10...$15.00
https://doi.org/10.1145/3503161.3548363



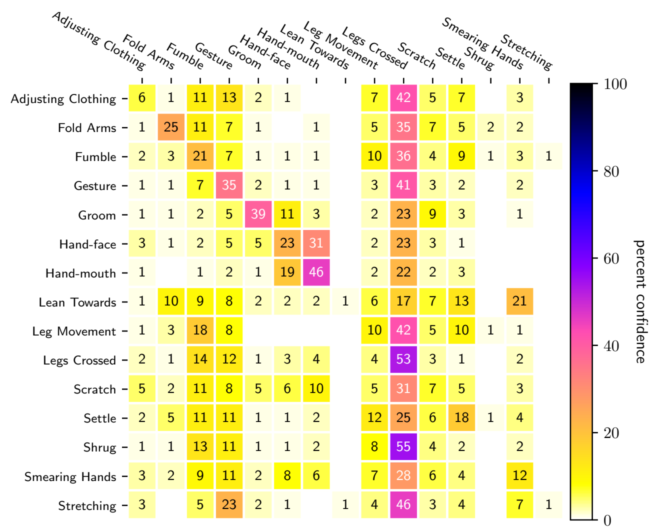
(a) Single-Label I3D [?]



(b) Multi-Label I3D [?]

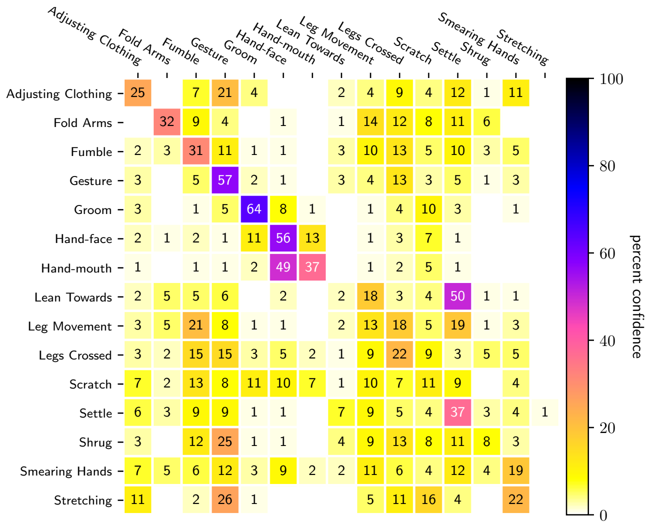


(c) Single-Label TSN [?]

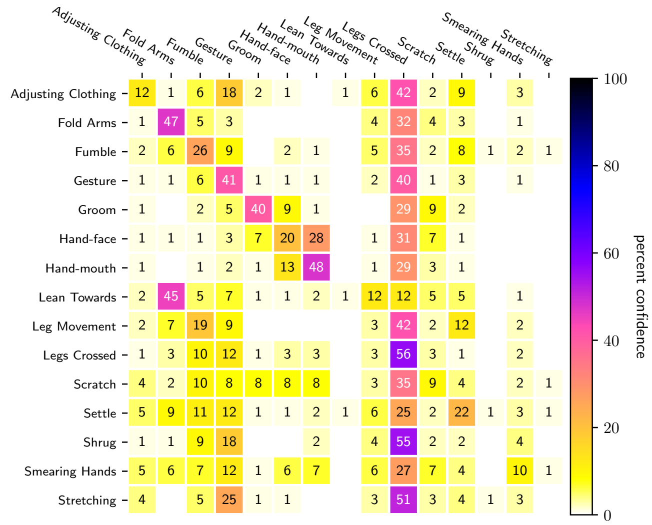


(d) Multi-Label TSN [?]

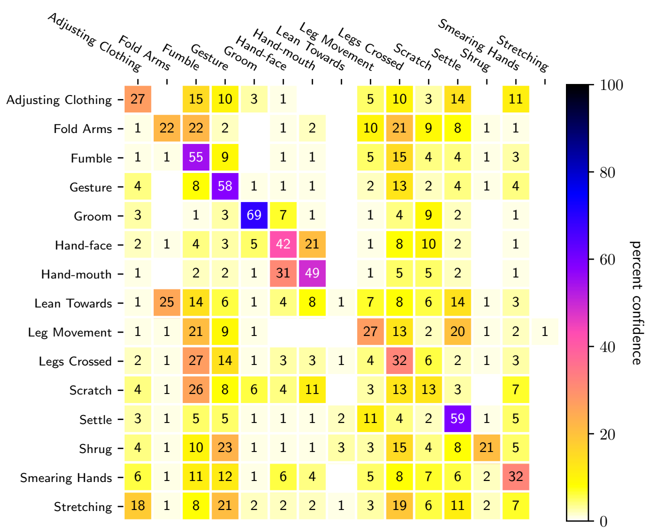
Figure 1: Confidence matrices of behavior recognition by I3D and TSN, trained in both labeling settings.



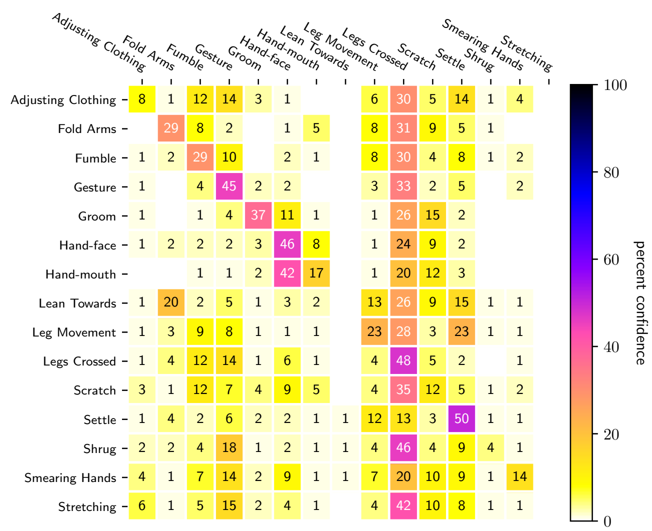
(a) Single-Label TSM [?]



(b) Multi-Label TSM [?]



(c) Single-Label Swin [?]



(d) Multi-Label Swin [?]

Figure 2: Confidence matrices of behavior recognition by TSM and Swin, trained in both labeling settings.

Figure 3: Illustrative examples of PDAN predictions on the frame level. Ground truth and acceptance threshold of 0.3 determine one of the three statuses: True Positive (TP), False Positive (FN) and False Negative (FN). First column shows a decrease in performance caused by an additional person present in the scene. Second column shows the effect of occlusion. Third column indicates that classes with lower level of agreement have a generally worse performance. Fourth column highlights the great challenge of distinguishing between classes of very subtle differences.