Personalized Gesture Range Detection Using Transductive Parameter Transfer

Rethinking Ubiquitous Smart Sensing of Social Behaviour In The Wild

Kyungmin Nam[1]

Supervisor(s): Hayley Hung[1]**, Stephanie Tan**[1]**, Vivian DSouza**[1]
Responsible Professor: Koen Langendoen[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 23, 2024

Name of the student: Kyungmin Nam
Final project course: CSE3000 Research Project
Thesis committee: Koen Langendoen, Hayley Hung, Stephanie Tan, Vivian DSouza, Qun Song

Abstract

This research investigates the detection of gesticulation using a torso-worn accelerometer sensor. Using the Conflab dataset, we focus on gestures during conversations in mingling scenarios. Due to significant variability in gesture styles among individuals, traditional methods face challenges in building personalized models. Our experiments demonstrate that Transductive Parameter Transfer (TPT), an adaptive transfer learning method, can more effectively model these individual differences in gesturing. To gain insights into individual expressiveness, we classify gestures into three classes: 'no gesture,' 'normal,' and 'large' gestures. TPT performed an average AUC score of 0.84 in binary classification and 0.77 in multiclass classification. These findings highlight the potential of using a single torso-worn accelerometer to understand social behavior in naturalistic settings.

1 Introduction

Gestures are integral to human communication, serving as non-verbal cues and tools that complement speech to emphasize and structure it [9]. These gestures provide insights into the speaker's emotions, personality traits, and intentions [13]. Analyzing gestures can be useful in particular settings, such as inferring the overall mood of gatherings, monitoring kids' behaviors for education, or observing patients in healthcare settings. This makes the analysis of hand gestures a valuable method for providing insights into human interactions [16].

Gesture recognition in human-computer interaction (HCI) has been studied to enable control and interaction with machines using hand and body movements, without the need for physical contact. While such studies often focus on symbolic or predefined gestures, natural hand gestures are more commonly encountered in real-life conversations. For example, [24, 33] analyzed the frequency of gestures during conversations using a side-view video dataset while [19,20] have proposed methods for automatically capturing upper body communicative cues in seated conversation. [2] delved into gestures during free-standing conversations in social settings, yet the field remains relatively understudied.

Studying social interactions in natural, unconstrained environments where conversations occur organically offers a wealth of meaningful data. In these settings, known as "in the wild" scenarios, individuals interact freely, providing a rich context for analyzing conversational gestures. Our goal is to address the challenges in real-life social interactions to develop more practical gesture detection systems.

This research focuses on studying social interactions in mingling scenarios, a specific type of "in the wild" setting where people engage in spontaneous, unscripted conversations. The Conflab dataset [26,27], which captures such interactions provides an excellent basis for our study. This dataset was recorded from social gatherings of a conference event, including overhead view video and sensor data collected from all participants wearing a smart badge hung on their necks. The dataset also includes manually annotated speaking status and full body key-points.

Previous studies on conversational gestures have predominantly relied on vision-based techniques or wearable devices positioned on the hands and arms [10,14,18,22,30,32]. However, our research focuses on detecting gestures using a torso-worn smart badge equipped with an accelerometer.

Gestures and their ranges vary by individual, which offers insight into the expressiveness and communicative intent [4,23]. Our methodology involves using body key-points to identify different ranges of gestures and classifying these gestures using accelerometer data. Unlike the traditional approach of concatenating various subjects into one training set, we investigate whether personalized models can improve the detection of individual gestures. We employed Transductive Parameter Transfer (TPT) [8], a machine-learning technique that adapts models to individuals' unique gestural patterns. TPT has shown promise in personalized model adaptation, which can address the variability in individuals' gesture styles.

The contributions of this research are threefold:

1. Demonstrate the effectiveness of combining body key-points and accelerometer data to classify gesture ranges.
2. Propose a personalized gesture detection model by applying Transductive Parameter Transfer (TPT) approach and validate the method.
3. Extend the TPT approach to handle multiclass classification tasks to differentiate between multiple ranges of gestures.

Through the evaluation of the Conflab dataset, we hope to contribute to developing effective gesture recognition systems that can operate in natural social settings.

2 Related Work

This section reviews existing literature on gesture recognition methodologies, with a focus on vision-based and sensor-based systems, conversational gesture detection, and social behavior analysis. By examining these studies, we highlight our unique challenges within the field of gesture recognition.

2.1 Gesture Recognition for Human-Computer Interaction

Gesture recognition has been extensively studied for human-machine interaction (HMI) and human-robot interaction (HRI), particularly in the context of interpreting hand sign languages. There are two primary implementations of gesture recognition systems: sensor-based and vision-based [30]. Vision-based recognition systems focus on interpreting gestures through video data, where individuals perform predefined gestures in front of a camera. However, these systems typically address only a subset of the wide variety of possible gestures. There remains an issue related to complex trajectories and occlusions, which can impact performance in real-time applications.

In contrast, sensor-based solutions utilize accelerometers or gyroscopes to detect gestures [10]. These systems often involve special gloves or handheld devices to capture hand

movements [22]. For instance, [32] introduced a method using a three-dimensional accelerometer embedded in a Wiimote controller. Their approach extracts temporal and spectral features from the accelerometer data and employs a multiclass SVM for classification. Similarly, [18] investigated accelerometer-based gesture recognition using a wearable watch, utilizing feature-weighted Naïve Bayesian classifiers and dynamic time warping. [14] explored the feasibility of using smartbands for social gesture recognition. The study involved 32 participants performing 12 predefined social gestures while wearing a smartband equipped with a tri-axial accelerometer and classified with logistic regression.

While these studies have specific objectives like interacting with a computer, the reliance on wearable devices can make these systems cumbersome and less convenient for users. In contrast, we aim to analyze social interactions based on a person's gestures detected by a sensor positioned on the torso. The context of using a smart badge sensor to monitor and interpret social gestures presents unique challenges and requirements that differ from traditional HCI applications.

2.2 Detecting Conversational Gestures

Researchers have investigated gestures within conversational contexts which is a crucial aspect of understanding nonverbal communication in social interactions. A feasibility study by [29] focused on hand gesture recognition during natural conversations. They involved an experimental assistant and two subjects, capturing hand gestures naturally occurring during dialogues. These gestures were manually annotated to analyze their frequency and duration, revealing that hand gestures significantly impact communication, often reinforcing verbal messages.

[33] presented work on analyzing frequency properties of hand gestures during conversations. They applied a windowed Fourier transform and wavelet transform to detect and extract gesticulatory oscillations. Similarly, [20] and [19] addressed the detection of gestures during seated encounters within the context of job interviews. Their first work focused on detecting upper body movements using monocular video to approximate a 3D upper body pose, including hand positions. Their subsequent study used these features to identify adaptors—unintentional gestures typically performed when a person is fidgety—and beat gestures, which emphasize speech rhythms.

While these previous works were oriented toward conversational gestures, the primary focus has been on controlled environments with limited variability. Moving towards detecting social behaviors in more dynamic and natural settings provides a broader understanding of human interactions and the challenges involved.

2.3 Detecting Social Behaviors In-The-Wild

Detecting social behaviors in crowded and dynamic environments, where multiple individuals interact closely, has been approached using various methodologies. Some studies highlight the feasibility of using only accelerometer data, while others benefit from multimodal approaches by combining visual and acceleration data.

Previous work [11] using a single body-worn accelerometer has demonstrated the estimation of actions such as speaking, laughing, gesturing, drinking, and stepping. This emphasized the advantages of accelerometers in noisy, crowded environments where traditional sensors like cameras and microphones struggle due to occlusions and auditory interference. The results for gesture classification showed the lowest performance, with an F1 score of 0.34. This can be attributed to the variability in individual gesturing styles, making it challenging for classifiers to recognize person-specific movements.

Another study [8] focused on detecting speech in crowded environments using a single body-worn triaxial accelerometer. They used transductive parameter transfer learning to address the high variability in body movements during speech. This method allows the adaptation of learned models to new, unseen subjects using only unlabeled data, improving detection performance over state-of-the-art methods. These findings highlight the importance of adaptive models in recognizing speech-related body movements.

The most similar work to ours [2], introduced a method for detecting conversational hand gestures in crowded mingle scenarios. The authors propose a multimodal approach with a fusion of video data and wearable acceleration data collected from smart badges. The acceleration data was effectively used at the decision level, enhancing the detection performance.

Although the fusion approach for gesture detection [2] has demonstrated effectiveness, using video data still poses several challenges, such as privacy concerns. While relying solely on accelerometer data may seem simple and robust, [11] has shown limitations in gesture detection using accelerometer data alone. Therefore, a better approach is needed to handle the variability in how individuals gesture.

Many gesture detection studies [2,11] focus on a simple binary classification task, overlooking the varying expressiveness of gestures. The size of gestures is crucial for understanding the saliency of social interactions as it often correlates with the intensity of the emotion or meaning being conveyed [4]. For example, larger, more expansive gestures tend to express stronger intentions. Therefore, we need a system detecting these varying ranges that can help better understand the expressiveness behind the gestures.

3 Methodology

We propose a novel approach to automatically detect and classify conversational gestures using accelerometer data from a torso-worn smart badge. This section includes the annotation of gesticulation from video data, feature extraction from accelerometer data, and the application of the Transductive Parameter Transfer (TPT) learning method.

3.1 Automatic Gesture Annotation

As defined by the Cambridge Dictionary, gesticulation refers to "movements with your hands or arms intended to express something or to emphasize what you are saying" [3]. It is the most common type of gesture, occurring spontaneously with speech. We see gesticulations as movements between

two consecutive 'non-gestures' where the hands and arms remain still in a resting position. This allows us to differentiate between active gesticulation and periods of inactivity.

Manual annotation of gestures through video inspection is a valid method but has significant limitations. It is time-consuming and labor-intensive, requiring substantial effort to mark the gesture start and end points in video data [12]. This process becomes even more arduous when multiple annotators are needed to ensure objectivity [15]. Ensuring consistency over time is challenging, and important details may be missed. These limitations have motivated research into automated approaches to address these challenges.

Our approach focuses on torso body key-points, such as the positions of shoulders, elbows, and wrists. Semi-automated gesture annotation approaches have identified such key-points (shoulders, elbows, and wrists) as key features contributing to accurate gesture detection [12]. Additionally, we calculate the size of the gestures based on these key-points. From the analysis of spatial information, we classify gestures into normal' and 'large' categories. This classification captures how extensively individuals use their hands and arms away from their body.

For each participant in a video frame, we perform two main processes: normalizing the positions of torso key-points and calculating the distances between necessary key-points. If these key-points meet specific conditions for being a 'normal' or 'large' gesture, the frame is labeled accordingly. The detailed process for gesture identification is discussed in 4.1

We annotated 16 participants who engaged in conversational groups with video intervals of 10 minutes at 60 fps. A sliding window size of 3 seconds was used with a 1.5-second shift. If at least one of the 180 frames within a window is marked as a gesture, the window is labeled as a gesture. Conversely, if all frames in the window are labeled as non-gesture, the window is labeled as non-gesture.

The class distributions for each participant are depicted in Fig.1. On average, 32% of the samples across all participants were positive (indicating gesturing), with a standard deviation of 20%. Among these, 23% were normal range gestures, and 9% were large gestures. Participant 30 had the highest percentage of positive samples (74%), including the most large gestures (37%). Participant 32 had the fewest positive samples (0.5%) with no 'large' gestures. This person-specific variation in class distribution presents a need for personalized gesture detection.

3.2 Feature Extraction

We extracted features using a sliding window approach, similar to the annotation process. We use the same approach that has been proven efficient in analyzing human actions from wearable acceleration [8]. For each participant, a torso-worn smart-badge recorded acceleration data at 50 Hz. The data is processed to produce time series of triaxial acceleration, the absolute value of each axis, and the magnitude of the acceleration ($|\text{accel}| = \sqrt{x^2 + y^2 + z^2}$), resulting in seven different time series: x, y, z, $|x|$, $|y|$, $|z|$, and $|\text{accel}|$. The triaxial time series can address movements where the direction is important, while the magnitude and absolute values allow focusing on direction-invariant movements and overall intensity. Then,

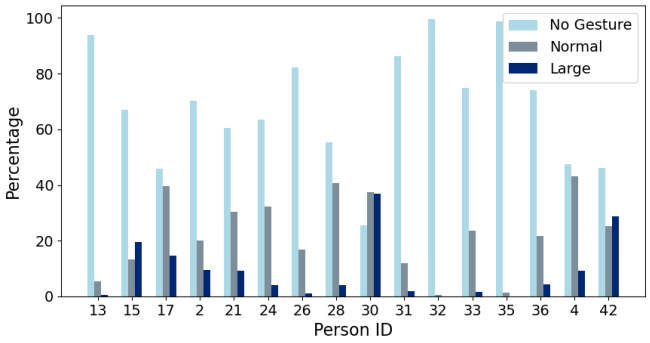

Figure 1: Percentage of gesture range for each participant

the following 10 features are extracted from each time series: mean, variance, and power spectral density (using eight bins).

This process resulted in 398 samples (windows) with 70 dimensions (7 time series $\times$ 10 features each). These feature vectors were then used for classifiers in multiple experiments, which are detailed in Section 5

3.3 Transductive Parameter Transfer

Gesticulation varies significantly from person to person due to factors like cultural background, personal habits, and physical characteristics [6]. Instead of using a traditional approach that combines data from different subjects into a single training set, we explore whether personalized models can better detect individual gestures.

To achieve this, we employ an adaptive transfer learning method called Transductive Parameter Transfer (TPT). TPT is a machine learning method that personalizes models by leveraging labeled data from multiple source domains to adapt to a new target domain without requiring labeled data from the target (Fig.3). The parameters from the source models are adapted to fit the target individual's data distribution. This approach has been successfully used in personalized speech detection [8] and in facial expression analysis [28]. Given the variability in gestural patterns among individuals, TPT is suited for our study as it enables personalized models that can adapt to each individual's data distribution. The detailed steps of our TPT process are described in 4.2 where we outline the procedure along with the algorithm pseudocode.

4 Implementation Details

This section provides an in-depth look at the implementation aspects of our research. We discuss the criteria for identifying gesture ranges based on body key-points and the detailed steps of applying the TPT approach.

4.1 Automatic Extraction of Gesture Range

Individual differences exist in how speakers use gesture space during communication [23]. Mcneill [21] categorized spatial form of gestures into center-center, center, periphery, and extreme periphery range, as visualized in Appendix A.1. These gesture ranges vary based on factors such as a person's speech content, emotion, and cultural background [31], which indicate a person's expressiveness and gestural habits [17].

We identify these different gesture ranges by analyzing annotated body key-points from the Conflab dataset, which provide a consistent viewing angle relative to the gestures being performed. Our focus is on gestures occurring in the periphery and extreme periphery spaces, which we refer to as 'normal' and 'large' gestures. We extract these gestures by calculating the distance between the neck and wrist points. If this distance exceeds T_{gesture}, the person is using their hands outside their body's boundary, signifying gesticulation. This threshold was determined through empirical testing on diverse participants to identify when hands are used for expressive movements rather than being at rest.

[1] suggests that we can determine the 'size' of the gesture by using the distance between the wrists normalized by shoulder width. We identify a 'normal' gesture by checking if the wrist-to-wrist distance/shoulder width ratio exceeds T_{normal}, indicating that the arm span extends beyond the torso into the periphery range. For a 'large' gesture, the hand-to-hand distance/elbow-to-elbow distance ratio should exceed T_{large}, indicating that the arms are extended widely without bending the elbows. Algorithm 1 outlines the pseudocode for determining gesture size.

Algorithm 1 Gesture Range Identification

Input: Annotated body key points from Conflab dataset
Output: Classified gestures as 'normal' or 'large'
for each frame **do**
  Extract and normalize p_{neck}, $p_{\text{left\_wrist}}$, $p_{\text{right\_wrist}}$, $p_{\text{left\_shoulder}}$, $p_{\text{right\_shoulder}}$, $p_{\text{left\_elbow}}$, $p_{\text{right\_elbow}}$
  if $D(p_{\text{neck}}, p_{\text{left\_wrist}}) > T_{\text{gesture}}$ **and** $D(p_{\text{neck}}, p_{\text{right\_wrist}}) > T_{\text{gesture}}$ **then**
    $W_{\text{shoulder}} = D(p_{\text{left\_shoulder}}, p_{\text{right\_shoulder}})$
    $W_{\text{elbow}} = D(p_{\text{left\_elbow}}, p_{\text{right\_elbow}})$
    $W_{\text{wrist}} = D(p_{\text{left\_wrist}}, p_{\text{right\_wrist}})$
    $R_{\text{wrist\_shoulder}} = \frac{W_{\text{wrist}}}{W_{\text{shoulder}}}$
    $R_{\text{wrist\_elbow}} = \frac{W_{\text{wrist}}}{W_{\text{elbow}}}$
    if $R_{\text{wrist\_shoulder}} > T_{\text{normal}}$ **then**
      Mark the frame as 'normal' gesture
    else if $R_{\text{wrist\_elbow}} > T_{\text{large}}$ **then**
      Mark the frame as 'large' gesture
    end if
  else
    Mark the frame as 'no gesture'
  end if
end for
Return: List of frames classified with gestures

Fig. 2 shows a snippet of a frame with body key-points that indicate 'no gesture', 'normal gesture' (red), and 'wide gesture' (blue). After automation of marking all the frames' body key-points, we went through several rounds of manual inspection. This process involved playing the video continuously frame by frame, with the colored key-points indicating the gestures. This confirmed that there were no false positives or misinterpretations of gestures.

Figure 2: Different gesture ranges of body key-points (no gesture, normal gesture - red, wide gesture - blue)

4.2 Algorithm of TPT

The TPT approach accounts for the variance in participants' gesture movements when personalizing models. This involves computing optimal classifier parameters for a target dataset based on source datasets and their optimal classifiers (Fig. 3). Our algorithm is based on previous works [8, 28] and their codebase [7]. First, we train individual logistic regression models for each participant in the source dataset to learn optimal model parameters. Then, we use Kernel Ridge Regression to map the data distributions of sources to their respective model parameters. Next, we compute the similarity between the target and source data distributions. Using this learned mapping, we predict the optimal model parameters for the target dataset based on these similarities.

The detailed steps are in Algorithm 2. From N source datasets with label information and the unlabeled target dataset, the goal is to compute the optimal parameters (w_t, c_t) for X_t (where w and c correspond to regression coefficients and the intercept, respectively:

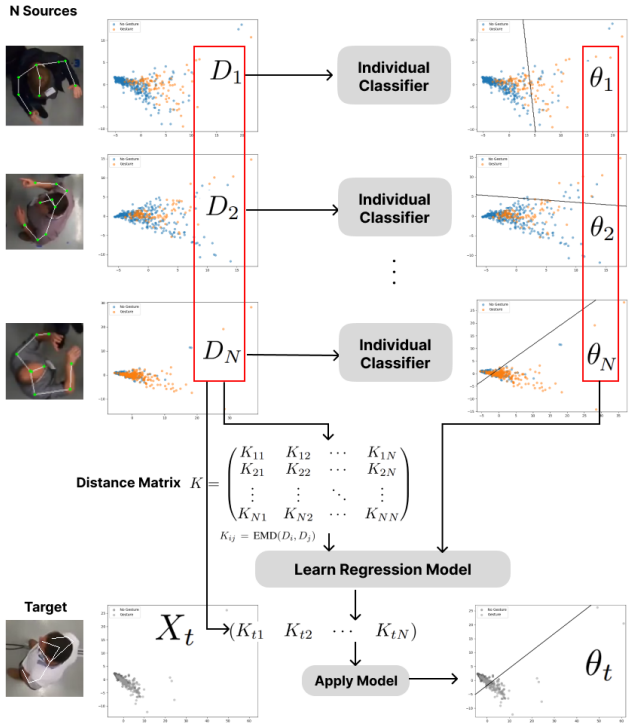

Figure 3: Overview of Transductive Parameter Transfer (TPT) approach for gesture detection

Algorithm 2 Transductive Parameter Transfer (TPT) Algorithm

Input: Source datasets $\mathcal{D}_i^s = \{(\mathbf{x}_j^s, y_j^s)\}_{j=1}^{n_i^s}$, target dataset $X^t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$

Output: Personalized model parameters (w_t, b_t) for target data

1. Train a classifier on each source dataset $D_{i\,i=1}^{s\,N}$ to obtain the classifier parameters $\theta_i = (w_i, c_i)$ using logistic regression.
2. Construct a training set $\mathcal{T} = (X_i^s, \theta_i)_{i=1}^N$, where X_i^s represents the feature vectors of the source dataset $\mathcal{D}_i^s$ without the labels.
3. Compute the kernel matrix K where $K_{ij} = \kappa(X_i^s, X_j^s)$ represents Earth Mover's distance between the data distributions of the source datasets X_i^s and X_j^s.
4. Using K and $\mathcal{T}$, compute the mapping function $\hat{f}$ between marginal distributions of the datasets and their optimal parameters, with Kernel Ridge Regression
5. Apply the mapping function $\hat{f}$ to the target dataset X^t to predict the parameters (w_t, b_t).

Return: (w_t, b_t)

Using the Algorithm 2, we can compute the parameter vector θ_t, for any new target dataset by plugging X^t into the mapping function $\hat{f}$. Then, the classification of the samples is obtained by $y = \text{sign}(w_t x + c_t)$.

4.3 Algorithm of TPT for Multiclass Classification

The TPT approach needs to be adjusted for multiclass classification to classify the gesture range into three classes (no gesture, normal gesture, large gesture) as identified in 4.1. Previous studies have only applied the TPT approach to binary classification, so we have evolved the existing method to address this multiclass problem.

We use logistic regression with a ' one-versus-rest ' approach when training classifiers for each source set to obtain the optimal parameters. This method provides distinct parameters for each class by comparing each class against the others, resulting in three parameter sets. Similar to the binary TPT case, kernel ridge regression maps the distribution of the source datasets to their parameters. Consequently, separate regressors are trained for each class's parameters, resulting in three individual regressors.

Using the learned mapping functions between the distribution and the parameters, we put in the target distribution to obtain three sets of parameters. Decision values are then calculated from each class's parameters and passed through a sigmoid function. The sample is classified based on the class with the highest sigmoid value. Fig. 4 shows the overview of the multiclass TPT procedure.

5 Evaluation

A series of experiments was conducted using the Conflab dataset. This section describes the experimental setup, the results obtained from both binary and multiclass classification tasks, and a comparison of of our TPT models with other classification techniques.

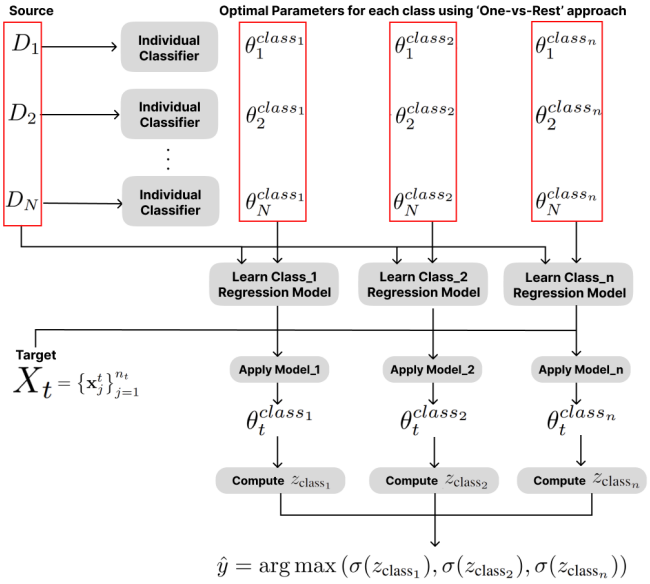

Figure 4: Overview of TPT approach for muticlass classification

To systematically evaluate classification performance, we selected the Area Under the Curve (AUC) as our metric. For personalized models, where data distribution and class balance can vary by participant, AUC is particularly effective in handling data imbalances.

5.1 TPT in Binary Classification

5.1.1 Experimental Setups

For each participant, a personalized model was trained in three setups: person-dependent, person-independent, and TPT, with the first two serving as baselines for evaluating TPT, inspired by the experiments described in [8]. The task is binary classification, where both 'normal' and 'large' gestures are classified as the positive class. Logistic regressors were used, and optimal regularization was determined using k-fold cross-validation on the training set. To address potential bias due to class imbalance, we adjusted the class weights by setting them inversely proportional to class frequencies in the training data. The details for each experimental setup are the following:

- **Person Dependent Setup** Individual model was trained and tested on each participant's dataset. Due to the limited number of samples in each dataset, we use a Leave-One-Sample-Out cross-validation approach.
- **Person Independent Setup** Each model is trained using the concatenated data from all other participants, following a Leave-One-Subject-Out cross-validation setup. The model was tested and evaluated on the participant's data that was left out.
- **TPT** Each participant is treated as a target set while a model is trained using the other participants as source

sets, similar to the Leave-One-Subject-Out setup in the person-independent setup.

5.1.2 Results

Fig. 5 shows the scores for each participant across all three setups.

Person-dependent setup resulted in an average AUC score of 0.82 with a standard deviation of 0.14, ranging from 0.5 to 0.97. This high variation is attributed to the class distribution being highly skewed towards the negative class in some participants (32 and 35) with an AUC score of 0.5. However, we cannot guarantee that balanced datasets always lead to higher performance, as seen with Participant 28 (0.67).

Person-independent setup obtained an average AUC score of 0.82 with a standard deviation of 0.075, ranging from 0.6 to 0.91. Most participants (10 out of 16) had higher scores in the person-dependent setup than in the person-independent setup. Typically, training with more samples improves model performance in machine learning, but this was not the case here. This suggests that training an individual model using combined datasets from different participants may create a decision boundary that does not accurately reflect the unique probability distributions inherent to each participant's data.

TPT approach resulted in an average AUC of 0.84 with a standard deviation of 0.068, slightly higher than both person-independent and person-dependent setups (0.82). A one-tailed paired t-test showed no significant difference between the TPT setup and either the person-independent or person-dependent setups.

Half of the participants (8 out of 16) performed better with TPT compared to the person-independent setup. In the person-dependent setup, 10 out of 16 participants performed better than with TPT, although the overall average for the dependent setup was lower. We consider the person-dependent setup as an upper bound on performance because it benefits from the personalized nature of the setting.

Interestingly, Participants 13, 32, and 35 performed best in the TPT setup compared to both the person-independent and person-dependent setups. A common factor for these participants was their highly skewed data distribution, with over 90% of their data belonging to the negative class. Notably, participant 32 had the most negative class and showed the largest difference between TPT and the other two setups, highlighting the effectiveness of TPT in handling extreme class imbalances.

5.1.3 Comparison with Other Binary Classifiers

The performance of TPT was compared with other well-known classification methods: Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF). For each participant, these models were trained in person-independent setups. The AUC scores for all models for each participant are shown in Fig. 6. As illustrated in Table 1 including average AUC score over all models, TPT outperformed SVM and KNN, and the paired t-test revealed a statistically significant difference between TPT and KNN. On the other hand, RF showed slightly better performance than TPT, achieving the highest average AUC among the classifiers.

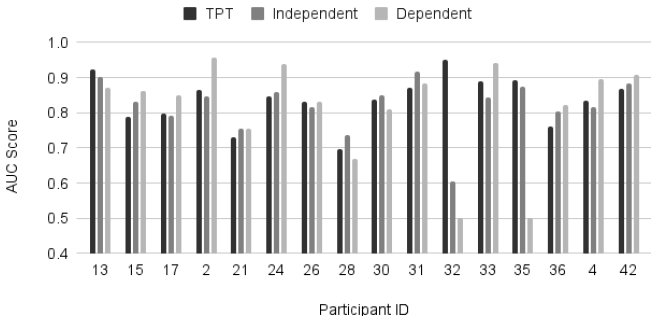

Figure 5: AUC scores of gesture detection in three setups for each participant (TPT, person-independent, person-dependent)

Figure 6: AUC scores of TPT approach and other binary classification models for each participant

5.2 TPT in Multiclass Classificatiom

5.2.1 Experimental Setups

The multiclass classification task aims to classify three classes: 'no gesture,' 'normal gesture,' and 'large gesture.' We used data from 14 participants who exhibited all three classes in their datasets. Participants 32 and 35, who did not perform any large gestures, were discarded since the library models used for multiclass classification were infeasible to handle datasets with only two classes.

All experiments were conducted in a Leave-One-Subject-Out manner. The multiclass TPT experiment followed the same approach as the binary classification TPT. It was then compared with the person-independent setup using four other multiclass classification models: Logistic Regression (LR), Support Vector Machine (SVM) with a one-vs-rest approach, k-Nearest Neighbors (KNN), and Random Forest (RF).

5.2.2 Results

The multiclass TPT performed an average AUC of 0.77, ranging from 0.63 to 0.89. All participants achieved higher performance than random (0.33)

As shown in Table 2, the average AUC of TPT outperformed SVM, with 12 out of 14 participants performing better. In contrast, TPT underperformed compared to LR and RF, with paired t-test showing no significant differences. The AUC scores for each participant across all models are shown in Appendix A.2.

6 Discussion

This section discusses the results and the evaluation of our proposed methods. We also consider the potential improvements and future research directions.

Setup	TPT	Person-Independent			
		LR	SVM	KNN	RF
AVG AUC	0.84	0.82	0.81	0.79	0.85
STDEV	0.068	0.07	0.097	0.057	0.054

Table 1: Average and Stdev AUC of different binary classification setups

Setup	TPT	Person-Independent			
		LR	SVM	KNN	RF
AVG AUC	0.77	0.79	0.71	0.71	0.78
STDEV	0.07	0.07	0.056	0.056	0.071

Table 2: Average and Stdev AUC of different multiclass classification setups

6.1 Effectiveness of Body Key-Points and Accelerometer Data in Gesture Detection

We demonstrated the automatic extraction of gestures and analyzed the size of gestures, deriving spatial information from body key-points. This method is more effective than manual annotation, which is tedious and requires inter-annotator agreement to eliminate subjectivity. In contrast, automatic extraction provides an initial guideline for annotation, with manual inspection adding a layer of verification. This validation process is straightforward, as the gestures are visualized by coloring the key-points in the frame when detected.

However, the presented algorithm is specifically designed for body key-points from an overhead view, as seen in the Conflab dataset. Additionally, it requires the body key-points to be annotated initially, which introduces a dependency. Misidentification of gestures might occur due to errors in the algorithm or inaccuracies in the annotated body key-points. For example, natural body movements or accidental gestures could potentially lead to false positives. Nonetheless, such issues can still be detected through quick manual inspection, and the method can significantly streamline the overall annotation process.

We demonstrated the effectiveness of gesture classification using accelerometer data collected from a smart badge worn around the neck. Various classification models were tested, yielding average AUC scores between 0.79 and 0.85. Additionally, these models successfully identified different gesture ranges in a multiclass classification task, with accuracies ranging from 0.71 to 0.79. In both binary and multiclass scenarios, the performance significantly exceeded random guessing benchmarks (0.5 for binary and 0.33 for multiclass). These results confirm that gestures and a person's expressiveness are predictive from torso movements.

6.2 Evaluation of TPT Approach

We focused on applying TPT to overcome the challenge of developing individual models with limited or unlabeled data. TPT was particularly effective for participants with highly imbalanced data, such as participants 13, 32, and 35, who are expected to have very subtle movements during conversations. TPT's kernel regression model finds it easier to identify similarities between these participants due to the consistency in their movement patterns, leading to better performance. This shows that TPT effectively addressed the difficulty of detection introduced by the imbalance in class distribution, which is very person-specific.

However, the overall performance of TPT in both binary and multiclass setups did not outperform the independent setup with RF classifier. This could be due to several factors. First, the data had a relatively low number of participants. TPT relies on the diversity of participants' data, and for multiclass classification, the model requires even more diversity due to increased algorithm complexity. The previous TPT research showed that TPT performance stabilizes to high performance when there are at least 20-30 sources [28]. With a limited number of participants (16), the kernel regression learning used in TPT might struggle to find optimal parameters for a new target. Therefore, future studies should especially test our newly proposed multiclass TPT with a larger participant pool across different contexts.

Second, the independent setup, which leverages the combined data from all participants, was inherently suitable for RF to handle complex patterns. Specifically, RF becomes more effective than LR as the dataset size increases due to its ensemble nature, which combines multiple decision trees to improve generalization and reduce overfitting. This ability to manage larger datasets more effectively could explain why our study's independent setup with RF was better than TPT.

Despite these advantages, the independent setup also has downsides. As more participants are included, these models become computationally expensive due to the need to process large amounts of data. Moreover, they may fail to capture the unique patterns and behaviors specific to each individual as the models are generalized across the combined data from all participants. This lack of personalization can lead to suboptimal performance for individual participants, especially in tasks where personal differences are critical.

On the other hand, TPT offers a balance between computational efficiency and accuracy as the dataset grows, becoming more effective in creating personalized models by learning from diverse gesture styles. This is because adding new source participants involves adapting parameters rather than retraining on the entire concatenated dataset. This approach makes it more scalable and efficient compared to traditional methods. However, the scalability of TPT still needs to be tested with larger datasets to validate its efficiency in such scenarios.

6.3 Speaking and Gesture

Speaking and gesticulation are correlated, with gestures often complementing speech to emphasize and structure it [9]. This correlation can be analyzed using the speaking status annotations from the Conflab dataset. Fig. 7 shows the sample percentages of speaking status, gesture status, and their co-occurrences.

The general trend indicates that people tend to gesture while speaking, although the extent varies significantly among individuals. Participants who speak less, such as Participants 13, 32, and 35, also gesture less. However, high speaking percentages do not necessarily correspond to high

gesturing percentages, as seen in Participants 28 and 36. Some participants show more gesturing than speaking, which could be due to several factors, such as stretching, which might be incorrectly identified as gestures, leading to misinterpretation.

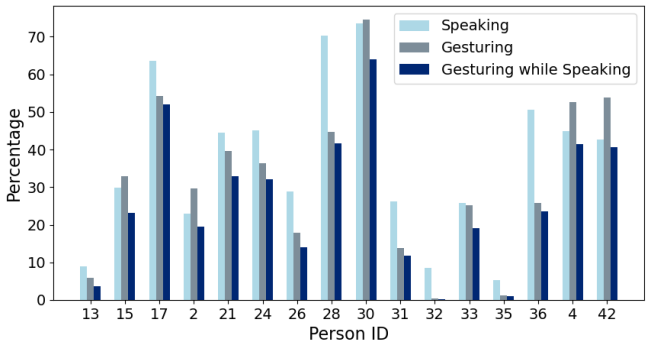

Figure 7: Percentage of Speaking, Gesturing, and their concurrency for each participant

The mutual influence between speech and gesture and their shared cognitive and communicative functions make analyzing their correlation highly meaningful for understanding human interactions. This insight is vital for developing privacy-sensitive methods for social behavior analysis that emphasize body language over spoken language, thereby avoiding the intrusive of recording private conversations.

A more comprehensive and realistic coding scheme for gesture annotations is needed to analyze speech-related gestures. The gesture research community widely uses four categories of gestures (beat, deictic, iconic, metaphoric) proposed by [21], which are complex and subjective in the real world. Our gesture range annotation method based on body key-points can serve as a foundational point for future schemes that better understand speech and gesture patterns.

For future work, our research can be used to infer personal speaking patterns and analyze social behavior. This can be applied in various practical scenarios, such as healthcare settings where tracking patients' physical and social activities can enhance diagnosis and treatment plans, and educational environments where monitoring students' attentiveness and interactions can improve learning outcomes. By exploring these applications, future research can expand the utility of torso movements in detecting and analyzing human behaviors across diverse contexts.

7 Responsible Research

This section reflects on the ethical use of the dataset, addressing potential biases, and the reproducibility of the research.

7.1 Ethical Considerations in Dataset Usage

The Conflab dataset, which underpins our study, was collected following ethical guidelines. Participants were fully informed about the data collection and consented to participate. The video recordings adhere to ethical standards by using an overhead view camera setup, which minimizes the capture of identifiable facial features. Our research complies with data-sharing agreements and ethical guidelines. We ensure any sharing or subsequent use of the dataset strictly follows the agreed terms.

7.2 Addressing Bias

Among the 16 participants in our study, there is a gender imbalance, with only 2 women. This discrepancy could introduce bias in the gesture detection models, as gestural habits may vary by gender. However, our use of TPT aims to mitigate these biases by personalizing models to individual users. Further research should strive to improve the fairness and inclusivity of the models.

7.3 Ensuring Reproducibility

We have documented our methodologies in detail to allow other researchers to validate our findings and build upon our work. Our TPT algorithm is based on the codebase provided by [7]. The modified code for the TPT and multiclass TPT algorithms is available in a public repository [5], complete with documentation and usage instructions. Full reproducibility requires access to the Conflab dataset from the Socially Perceptive Computing lab at Delft University of Technology. For those using their dataset, body key-points must be annotated as described in [25]. Once annotated, one can follow the gesture extraction described in 4.1 and our other methodologies.

8 Conclusions

This research explored the detection of conversational gestures using a torso-worn smart badge in real-world mingling scenarios. Utilizing the Conflab dataset, we focused on video with annotated body key-points and accelerometer data from 16 participants. By computing the distance between torso body key-points, we automatically annotated gestures, which significantly reduced the effort required for manual annotation. Gestures were classified into 'normal' and 'large' based on spatial information. For developing personalized gesture models, we experimented with the Transductive Parameter Transfer (TPT) approach to address person-specific patterns in predicting gestures. The TPT approach was extended to handle multiclass classification tasks, enabling the differentiation of various ranges of gestures. As a result, we achieved an average AUC score of 0.84 in binary classification and 0.77 in multiclass classification. This research demonstrated the effectiveness of the proposed approach for detecting gestures in real-life interactions and highlighted its potential for practical applications in understanding social behavior.

References

[1] Kipp M Neff M Albrecht. I an annotation scheme for conversational gestures: how to economically capture timing and form lang. *Resour. Eval*, 41(3):325, 2007.

[2] Laura Cabrera-Quiros, David MJ Tax, and Hayley Hung. Gestures in-the-wild: Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. *IEEE Transactions on Multimedia*, 22(1):138–147, 2019.

[3] Cambridge Dictionary. Gesticulation, n.d. Accessed: 2024-06-02.

[4] Mingyuan Chu, Antje Meyer, Lucy Foulkes, and Sotaro Kita. Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, 143(2):694, 2014.

[5] TU Delft CSE. Research project. https://github.com/TU-Delft-CSE/Research-Project?tab=readme-ov-file, 2024.

[6] Pierre Feyereisen and Jacques-Dominique De Lannoy. *Gestures and speech: Psychological investigations*. Cambridge University Press, 1991.

[7] Ekin Gedik. Tpt: Transductive parameter transfer. https://github.com/ekingedik/TPT/blob/master/TPT.py, 2018.

[8] Ekin Gedik and Hayley Hung. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing*, 21:723–737, 2017.

[9] Susan Goldin-Meadow and Martha Wagner Alibali. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64:257–283, 2013.

[10] Raghav Gupta, Shashank Chaudhary, Akshat Vedant, Niladri Paul Choudhury, and Vandana Ladwani. Gesture detection using accelerometer and gyroscope. In *Emerging Research in Computing, Information, Communication and Applications: Proceedings of ERCICA 2022*, pages 99–116. Springer, 2022.

[11] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 207–210, 2013.

[12] Naoto Ienaga, Alice Cravotta, Kei Terayama, Bryan W Scotney, Hideo Saito, and M Grazia Busa. Semi-automation of gesture annotation by machine learning and human collaboration. *Language Resources and Evaluation*, 56(3):673–700, 2022.

[13] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[14] Jonathan Knighten, Stephen McMillan, Tori Chambers, and Jamie Payton. Recognizing social gestures with a wrist-worn smartband. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 544–549. IEEE, 2015.

[15] Anthony Pak-Hin Kong, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam. A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (dosage). *Journal of nonverbal behavior*, 39:93–111, 2015.

[16] Robert M Krauss, Yihsiu Chen, and Purnima Chawla. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In *Advances in experimental social psychology*, volume 28, pages 389–450. Elsevier, 1996.

[17] Margaux Lhommet and Stacy C Marsella. 19 expressing emotion through posture and gesture. *The Oxford handbook of affective computing*, page 273, 2014.

[18] David Mace, Wei Gao, and Ayse Coskun. Accelerometer-based hand gesture recognition using feature weighted naïve bayesian classifiers and dynamic time warping. In *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent user interfaces companion*, pages 83–84, 2013.

[19] Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, and Daniel Gatica-Perez. Capturing upper body motion in conversation: An appearance quasi-invariant approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 327–334, 2014.

[20] Alvaro Marcos-Ramiro, Daniel Pizarro-Perez, Marta Marron-Romera, Laurent Nguyen, and Daniel Gatica-Perez. Body communicative cue extraction for conversational analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[21] David McNeill. Hand and mind1. *Advances in Visual Semiotics*, 351, 1992.

[22] M Popa. Hand gesture recognition based on accelerometer sensors. In *The 7th International Conference on Networked Computing and Advanced Information Management*, pages 115–120. IEEE, 2011.

[23] Matthias A Priesters and Irene Mittelberg. Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study. In *Proceedings of the Tilburg Gesture Research Meeting (TiGeR)*, pages 19–21, 2013.

[24] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193, 2002.

[25] Jose Vargas Quiros, Stephanie Tan, Chirag Raman, Laura Cabrera-Quiros, and Hayley Hung. Covfee: an

extensible web framework for continuous-time annotation of human behavior. In *Understanding social behavior in dyadic and small group interactions*, pages 265–293. PMLR, 2022.

[26] 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, 2022.

[27] 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. *Advances in Neural Information Processing Systems*, 35:23701–23715, 2022.

[28] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 357–366, 2014.

[29] Dian Christy Silpani, Keishi Suematsu, and Kaori Yoshida. A feasibility study on hand gesture recognition in natural conversation. In *2021 5th IEEE International Conference on Cybernetics (CYBCONF)*, pages 085–090. IEEE, 2021.

[30] Xiaolong Teng, Bian Wu, Weiwei Yu, and Chongqing Liu. A hand gesture recognition system based on local linear embedding. *Journal of Visual Languages & Computing*, 16(5):442–454, 2005.

[31] Frank R Wilson. *The hand: How its use shapes the brain, language, and human culture*. Vintage, 1999.

[32] Jiahui Wu, Gang Pan, Daqing Zhang, Guande Qi, and Shijian Li. Gesture recognition with a 3-d accelerometer. In *Ubiquitous Intelligence and Computing: 6th International Conference, UIC 2009, Brisbane, Australia, July 7-9, 2009. Proceedings 6*, pages 25–38. Springer, 2009.

[33] Yingen Xiong and Francis Quek. Hand motion gesture frequency properties and multimodal discourse analysis. *International Journal of Computer Vision*, 69:353–371, 2006.

A Appendix

A.1 Mcneil's Gesture Space

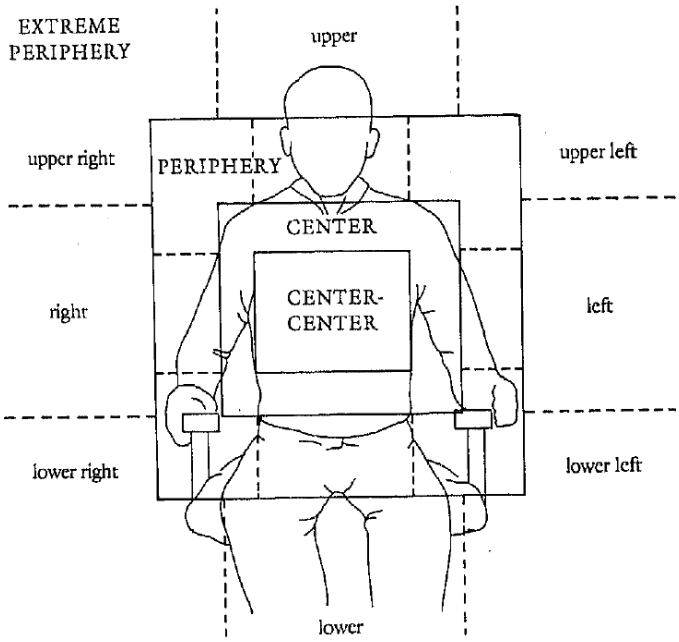

Figure 8: Gesture Space visualized by Mcneil [21]

A.2 Multiclass Classification Results

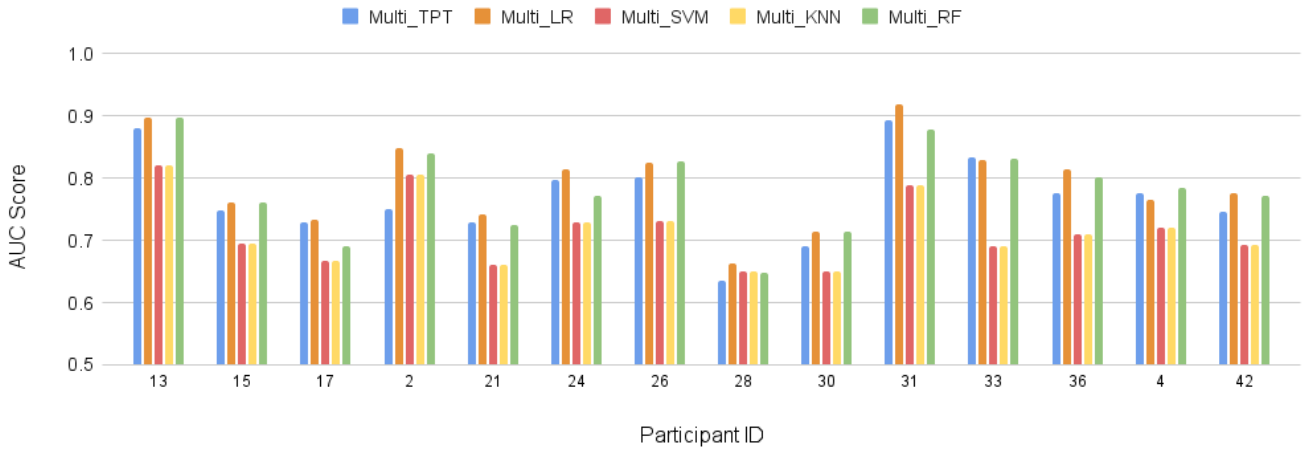

Figure 9: AUC scores of TPT approach and other multiclass classification models for each participant