Rethinking Ubiquitous Smart Sensing of Social Behaviour In The Wild

Personalized Gesture Range Detection using Transductive Parameter Transfer

Author : Kyungmin Nam | Supervisors : Hayley Hung, Vivian Dsouza, Stephanie Tan | Responsible Professor : Koen Langendoen

Abstract
This research investigate the personalized detection of conversational gestures through torso movements, utilizing accelerometer data with Transductive Parameter Transfer learning. Our work underscores the potential for advanced analysis of social behaviors in-the-wild

INTRODUCTION

- Detecting gestures that naturally occur in conversation can provide valuable insights to human interactions [1]
- We focus on Conflab dataset, collected from a conference social gatherings, people wearing a smart-badge and engaging in spontaneous, unscripted conversations [2]
- From accelerometer data collected from torso-worn badge, we build personalized models that classify different ranges of gesture by employing Transductive Parameter Transfer (TPT)

RELATED WORK

Conversational Gestures [3, 4]

- Goal: Analyze gestures usage in conversation
- Method: Detect gestures from Video recordings in controlled environment
- Research Gap: 1) Lack of studies in natural, unconstrained settings 2) Privacy concerns in relying on video data

Social Behaviors 'In-The-Wild' [5]

- Goal: Analyze social behaviours from data collected in-the-wild, mingling scenario
- Method: Use torso-worn accelerometer sensor or overhead view video data
- Research Gap: 1) Struggled with the variability in individual gesturing styles 2) Overlooking the varying expressiveness of gestures

CONTRIBUTION

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

METHODOLOGY

Automatic Extraction of Gesture Range

- We use Conflab's torso key-points (neck, shoulders, elbows, wrists) for automatic gesture extraction, as they have proven effective in gesture detection with semi-automated annotation [6]
- Gesture is identified by calculating the neck-to-wrist distance, with a threshold determined through empirical testing
- We distinguish 'normal' and 'large' gesture by normalizing wrist-to-wrist distance by shoulder width and elbow-to-elbow distance

Fig 1. Different ranges of gestures identified from Body Key-Points (no gesture, normal gesture - red, wide gesture - blue)

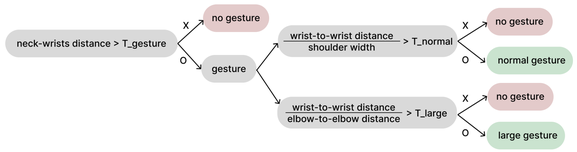


Fig2. Overview of Gesture Range Identification

Data Annotation & Analysis

- 16 participants were labelled using 10 min video
- Using a sliding window of 3 seconds with a 1.5-second shift,
  - if any frame within the window is labeled as a gesture, it is marked as a normal/large gesture; otherwise, it is marked as no gesture.

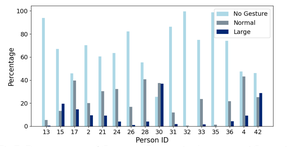


Fig3. Percentage of Gesture Range in Annotated Samples

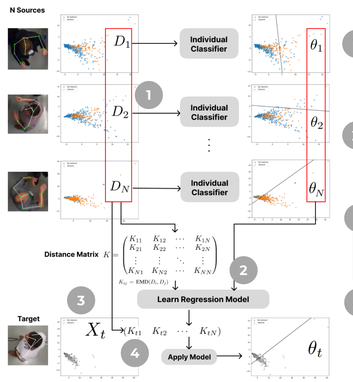
- Person-specific variation in class distribution presents a need for personalized gesture detection.

Feature Extraction

- Using accelerometer data, compute mean, variance, PSD for each x, y, z, |x|, |y|, |z|, magnitude which is proven efficient in analyzing human actions from wearable acceleration [7]
- 70-dimensional feature vector per window

Transductive Parameter Transfer

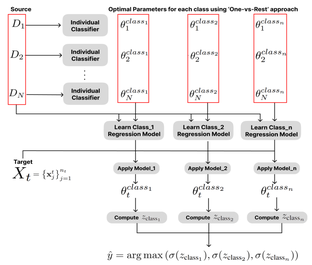
- Person specific nature of gestures and high variability introduce challenge in traditional methods that combine data from different subjects into a single training set
- TPT is an adaptive Transfer Learning that leverages labeled data from multiple sources without needing labeled target data [7, 8]



1. Train individual logistic regression models for each source dataset to learn optimal model parameters.
2. Use Kernel Ridge Regression to map the marginal distributions of sources to their model parameters
3. Compute the similarity between the target and source data distributions
4. Predict the optimal model parameters for the target dataset using the learned mapping

TPT adjusted for Multiclass Classification

- Use logistic regression with a 'one-versus-rest' approach to learn optimal parameters, resulting in distinct sets for each class
- Train n Kernel Ridge Regression models for each class
- Predict n sets of parameters for the target dataset, calculate decision values and classify based on the highest sigmoid value



EVALUATION

TPT in Binary Classification

Experiment setup inspired by [7]

- 'Normal' and 'large' gestures classified as positive
- Personalized models are trained in following three setups, all with Logistic Regressors
  - Person Dependent: Individual model trained/tested on each participant's dataset, using Leave-One-Sample-Out cross-validation
  - Person Independent: Model trained on combined data from all other participants, using Leave-One-Subject-Out cross-validation
  - TPT: Each participant treated as a target set, with models trained using other participants as source sets

Results

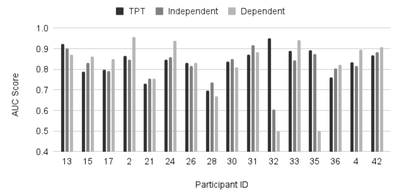


Fig. 4 AUC scores of gesture detection in three setups for each participant

- Average AUC of TPT (0.84) was higher than both person dependent (0.82) and independent (0.82) setups
- 8 out of 16 participants performed better with TPT compared to person-independent
- Participants 13, 32, and 35 performed best with TPT, handling extreme class imbalances effectively

Comparison with other binary classifiers

- TPT was compared with other well-known classification methods in person-independent setup.

Binary 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

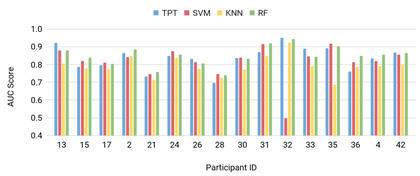


Fig. 5 AUC scores of TPT approach and other binary classification models

TPT in Multiclass Classification

- Classify 'no gesture,' 'normal gesture,' 'large gesture'
- Multiclass TPT compared to person-independent setup with well known multiclass classifiers
- All participants achieved higher performance than random

Multiclass 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

DISCUSSION

Effectiveness of TPT in Personalized Gesture Detection

- Performance significantly exceeded random guessing (0.5 for binary and 0.33 for multiclass)
- Addresses challenges of limited/unlabeled data, especially with highly imbalanced data.
- TPT did not outperform RF in both binary and multiclass possibly due to these factors:
  - Limited number of sources as kernel regression requires at least 20~30 participants for optimal performance to stabilize according to [8]
  - TPT for multiclass classification requires more diverse participant pool due to its increased algorithm complexity
  - Ensemble nature of RF improves reduces overfitting
- RF lacks personalization and computationally inefficient with larger dataset, while TPT offers a balance between computational efficiency and accuracy
- Scalability testing with larger dataset is needed in future research

Speaking and Gesture

- Correlation between speaking and gesticulation was analyzed using speaking status annotations in Conflab dataset which is vital for better understanding of human interactions
- Large variation in speaking and gesturing among participants
- Need for a comprehensive coding scheme for speech-related gestures

CONCLUSION

- The proposed approach using a torso-worn smart badge and automatic annotation based on body key-points effectively reduced manual effort and demonstrated high accuracy in detecting conversational gestures
- TPT approach addressed person-specific patterns in predicting gestures, showing its potential for developing personalized gesture models and handling multiclass classification tasks
- This research highlights the potential of using accelerometer data for practical applications in understanding social behavior in real-life settings, offering insights into individual expressiveness and interactions
- Future applications in healthcare and education could leverage tracking of physical and social behaviors to enrich analysis

REFERENCES

[1] Kendon, A. (2004). Gesture: Visible action as utterance. Cambridge University Press.
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] Naoto Ienaga, Alice Cravotta, Kei Terayama, Bryan WScotney, Hideo Saito, and M Grazia Busa. Semiautomation of gesture annotation by machine learning and human collaboration. Language Resources and Evaluation, 56(3):673–700, 2022.
[7] 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.
[8] 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.