From Actions to Kinesics: Extracting Human Psychological States Through Bodily Movements

CHEYU LIN and KATHERINE A. FLANIGAN

ABSTRACT

Understanding the dynamic relationship between humans and the built environment is a key challenge in disciplines ranging from environmental psychology to reinforcement learning (RL). A central obstacle in modeling these interactions is the inability to capture human psychological states in a way that is both generalizable and privacy preserving. Traditional methods rely on theoretical models or questionnaires, which are limited in scope, static, and labor intensive. We present a kinesics recognition framework that infers the communicative functions of human activity—known as kinesics—directly from 3D skeleton joint data. Combining a spatial-temporal graph convolutional network (ST-GCN) with a convolutional neural network (CNN), the framework leverages transfer learning to bypass the need for manually defined mappings between physical actions and psychological categories. The approach preserves user anonymity while uncovering latent structures in bodily movements that reflect cognitive and emotional states. Our results on the Dyadic User EngagemenT (DUET) dataset demonstrate that this method enables scalable, accurate, and human-centered modeling of behavior, offering a new pathway for enhancing RL-driven simulations of human-environment interaction.

INTRODUCTION

In *Space is the Machine*, Hiller argues that buildings are not merely shelters but also serve physical, spatial, and social functions that are deeply interconnected [1]. The physical structure defines and supports spatial configurations, while architectural style and details reflect both preferences and broader cultural meanings. These spatial layouts, in turn, shape accessibility, patterns of interaction, and overall social dynamics. The complex and reciprocal relationship between the built environment and human behavior is evident in various real-world contexts. For example, in nursing homes, the design and quality of physical spaces can influence residents' psychological well-being and sense of home [2]. Similarly, the physical condition and organization of school buildings affect students' health, cognitive development, and academic outcomes [3]. Such examples not

Cheyu Lin[1], Katherine A. Flanigan[2], Ph.D (Corresponding author). Email: {cheyul[1], kflaniga[2]}@andrew.cmu.edu. Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

only illustrate the mutual influence between humans and the urban fabric but also suggest that we can intentionally design physical, spatial, and social elements to promote social benefits [4, 5]. In practice, this has motivated both post-occupancy evaluation (POE) studies that assess how spaces are experienced and used after construction, and modeling approaches that aim to simulate such dynamics in advance. One prominent approach is agent-based modeling (ABM), which allows researchers to represent individual behavior and interaction within specific spatial layouts.

Agent-based modeling (ABM) captures the dynamics between interacting agents—such as humans and the built environment—and is particularly valued for its ability to encode decision-making processes within human agents [6]. This enables the simulation of human behaviors in hypothetical scenarios. However, agent rationale is often based on theoretical frameworks of planned behavior [7] or from manually collected data via questionnaires and expert opinions [8], both of which limit ABM's potential in key ways. First, theoretical models tend to oversimplify human decision making, neglecting the inherent heterogeneity of human behavior. Second, while questionnaires may better reflect cognitive complexity [9], they are time consuming, costly, and constrained by the limited number of processes humans can report or interpret. Moreover, such data only captures a snapshot of a person's psychological state, lacking the adaptability of continuously sensed information. These limitations hinder ABM's ability to represent human reasoning and preferences in a faithful and timely way. *This underscores the need for sensing technologies that can infer psychological states while remaining human-centered.* Crucially, such technologies must prioritize user privacy—not all sensing approaches are appropriate for this task [10, 11]. We argue that effective solutions must extract rich, socially relevant signals that reflect mental reasoning, while explicitly excluding identifiable features to maintain trust between stakeholders and users.

To address the limitations of manually sourced cognitive data, we turn to an underutilized yet powerful modality for understanding human reasoning: body language. Humans naturally externalize thought processes through a range of channels, and among them, body language plays a critical role by conveying unspoken cues about an individual's mental and emotional state through movement [12]. While these movements may initially seem unorganized, they are in fact structured by a psychological taxonomy developed by Ekman and Friesen [13]. This taxonomy, known as kinesics, classifies body language into five functional categories: illustrators, regulators, affect displays, adaptors, and emblems. Each category provides a well-defined link between specific bodily actions and the meanings or intentions they express. For instance, a hug is classified as an affect display, signaling warmth and emotional closeness. This taxonomy enables a principled mapping between physical activity and underlying psychological state, setting the foundation for extracting cognitive insight from bodily behavior.

By pairing this taxonomy with human activity recognition (HAR) techniques—which use sensor data such as RGB video, depth maps, or 3D skeletal keypoints to identify actions—it becomes possible to infer the kinesic category of a given movement. However, this approach faces a major limitation: the sheer variety of human actions makes it infeasible to manually define mappings for every possible movement. Consequently, the generalizability of existing frameworks remains limited. To truly decode human reasoning from bodily movements, we must move beyond dictionary-based mappings and toward methods capable of learning a generalized translation between physical actions

and their cognitive or affective significance.

To extract the kinesic function of human activities without relying on a predefined dictionary, we propose a kinesics recognition framework based on transfer learning. This framework leverages patterns inherently embedded in human activity data to infer the corresponding kinesic categories [14]. Specifically, our approach combines a frozen HAR model with a trainable convolutional neural network (CNN). The HAR component is implemented using a Spatial-Temporal Graph Convolutional Network (ST-GCN) [15], a skeleton-based model originally developed to classify activity types. Rather than using ST-GCN for activity recognition, we extract the latent features from its final hidden layer and use these as input to a CNN that classifies the activity's kinesic function. This architecture not only eliminates the need for a manually defined mapping between actions and kinesic categories, but also preserves user privacy by relying solely on 3D skeleton keypoints, which contain no identifiable features. The framework is applied to the Dyadic User EngagemenT (DUET) dataset [16], a HAR dataset explicitly inspired by the psychological taxonomy of kinesics.

The remainder of this paper is structured as follows: Section 2 introduces the DUET dataset and the supporting kinesic taxonomy. Section 3 details the data preprocessing steps, the structure of the machine learning models, and the experimental results. The full implementation is publicly available on Hugging Face [17]. We conclude in Section 4 with a discussion of key findings and directions for future work.

DUET – DYADIC USER ENGAGEMENT DATASET

Integrating psychological theory with HAR is essential for interpreting behavioral coherence as an expression of human thought processes. In this section, we present the psychological taxonomy of kinesics and introduce the DUET dataset, a HAR dataset explicitly derived from this taxonomy.

DUET is a two-person (or "dyadic"), multimodal HAR dataset. It contains 12 human activities spanning four sensing modalities, including RGB, depth, infrared, and 3D skeleton joints. The work in this paper only adopts 3D skeleton joints as the sensing modality, as shown in Figure 1, to align with the privacy-preserving requirement of human-centered applications [18]. The dataset was collected at an open indoor space, a confined indoor space, and an open outdoor space, allowing users to investigate the effects ambient environments impose on HAR algorithms. The 12 activities are adopted from the taxonomy of kinesics developed by Ekman and Friesen [13], which, to the best of our knowledge, is the only dataset that integrates HAR with a scientifically grounded psychology study. As described in Table I, there are five categories in the taxonomy: emblems, illustrators, regulators, adaptors, and affect displays.

KINESICS RECOGNITION FRAMEWORK

Building on DUET's foundation for integrating HAR and psychological theory, we develop a framework for recognizing kinesic functions without relying on a predefined kinesic dictionary. As shown in Figure 2, the framework consists of three components: (1) skeleton joint data preparation, (2) ST-GCN, and (3) CNN for kinesics recognition.

TABLE I. Overview of Ekman and Friesen's kinesic taxonomy [13] with representative gestures from the DUET dataset.

Taxonomy	Interaction Description	Example(s)
Emblems	Emblems are culturally specific gestures that have direct verbal equivalents. Their meaning is widely understood within a given culture but may differ significantly across cultural contexts. For example, a "thumbs up" signals approval in many Western cultures but is offensive in some Middle Eastern regions [19].	*waving in (0)*, *thumbs-up (1)*, *hand waving (2)*
Illustrators	Illustrators are gestures that complement and clarify spoken language by visually reinforcing verbal messages. They provide additional context to the verbal exchange between speakers.	*pointing (3)*, *showing measurements (4)*
Regulators	Regulators control the flow and pacing of conversation, helping to signal turn-taking or conversational shifts. For instance, "nodding" indicates agreement, while "drawing circles in the air" suggests speeding up, and "holding palms out" signals a desire to pause.	*nodding (5)*, *drawing circles in the air (6)*, *holding palms out (7)*
Adaptors	Adaptors are unconscious, self-directed movements that help individuals manage internal emotional states or satisfy personal needs. These gestures often arise in moments of stress or discomfort [20].	*twirling or scratching hair (8)*
Affect displays	Affect displays are gestures that reveal a person's emotional state, often without accompanying speech. These include both spontaneous expressions and socially conditioned emotional cues.	*laughing (9)*, *arm crossing (10)*, *hugging (11)*

Notes: The numbers in parentheses represent each interaction's activity label.

The full codebase and model architecture are publicly available in the DUET kinesics recognition repository [17]. In this section, we describe each component of the framework in detail and present the results of its implementation. All notations introduced are used consistently throughout the section.

Skeleton Data Preparation

The first step of the kinesics recognition framework is to process the data so that it is aligned with the ST-GCN data format. In DUET, one skeleton joint data sample is captured as a `csv` file, which stores an array of shape $T \times (M \times V \times C)$. T, M, V, and C stand for the number of frames, the number of subjects (i.e., 2), the number of keypoints (i.e., 32), and dimensions (i.e., x, y, and z coordinates), respectively. This format is not compatible with that of ST-GCN. In fact, ST-GCN calls for an assembly of skeleton keypoints and their metadata to be condensed in a nested Python dictionary, which is serialized as a pickle (`pkl`) file as displayed in "Skeleton data" in Figure 2. The `DUET.pkl` file contains two nested dictionaries: `split` and `annotation`. The `split` dictionary specifies the training and validation partitions for ST-GCN and stores

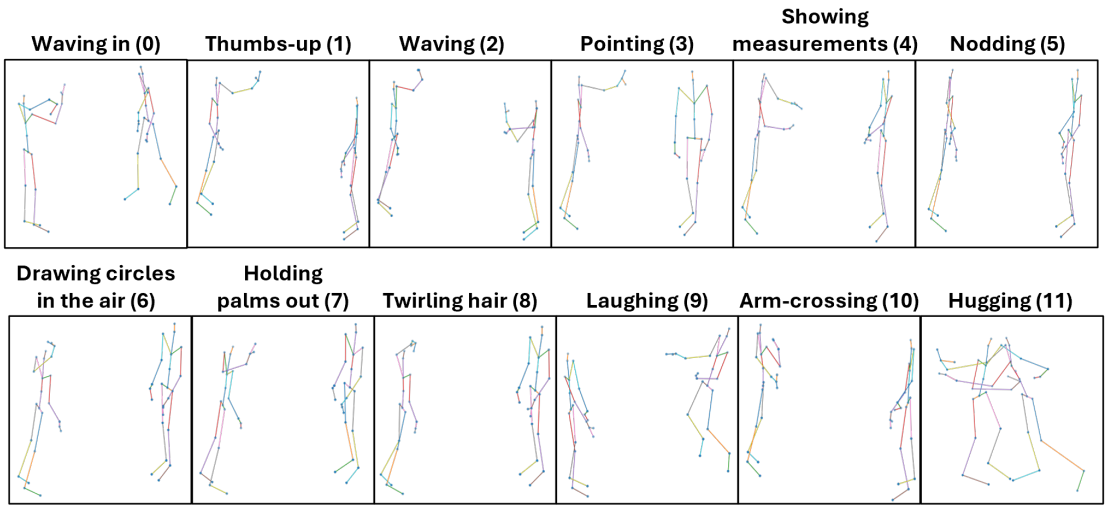

Figure 1. Sample frames for each interaction (class label is denoted in parentheses).

them in the lists `xsub_train` and `xsub_value`, respectively. Every element in the lists is the sample name in the form `LLIISS_t1_t2`. Here, `LL` stands for the location, which can be `CC` (a confined indoor space), `CM` (an open indoor space), or `CL` (an open outdoor space). `II` denotes numbers ranging from 0-11, which are the enumeration of interactions listed in Section 2. `SS` identifies the subject pairs ranging from 1-10. Last, `t1` and `t2` are the start and end timestamps of the interaction. The annotation dictionary stores a list of samples, each containing skeleton data along with associated metadata. Each sample includes the sample name (`frame_dir`), activity class label (`label`), total number of frames (`total_frames`), and skeleton keypoints (`keypoint`). A notable detail is that we extract only 25 of the 32 skeleton joints provided in DUET, as ST-GCN operates on a reduced set of keypoints. For our experiments, we perform cross-subject evaluation by designating participants `CCII01` and `CMII10` as the test set, with the remaining data used for training. This split is applied to both the ST-GCN and CNN components. Once the data are compiled in the required format, they are fed into the ST-GCN model to capture the structural patterns encoded in the skeleton keypoints.

Kinesics Recognition

Human movements can appear unpredictable—let alone when represented as skeleton keypoint data—making it challenging to identify patterns that convey their underlying kinesic functions. To recognize the kinesic functions of human activities, we construct a transfer learning model that extracts intrinsic patterns from skeleton data to infer the communicative purpose of each activity. As illustrated as the "ST-GCN" and "CNN" layers in Figure 2, the model consists of ST-GCN as fixed layers and CNN as modifiable layers. In this model, ST-GCN is not used in its conventional HAR role for classifying activity types. Instead, it serves to compress the high-dimensional skeleton data into a more compact representation that preserves essential activity features. We extract the output from ST-GCN's final hidden layer and use it as the input to the CNN. Leveraging the CNN's pattern recognition capabilities and the condensed keypoint representation, the model learns to distinguish the structural features that define each kinesic function.

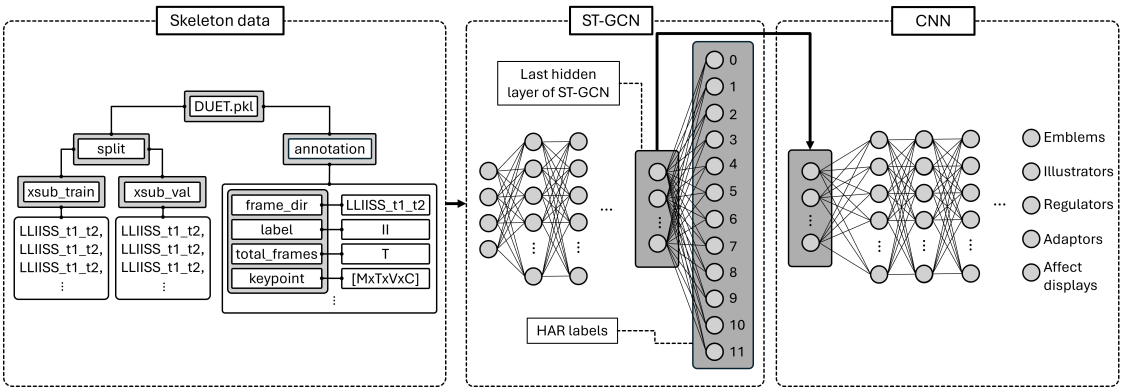

Figure 2. The kinesics recognition framework comprises skeleton data preparation, ST-GCN, and CNN. (Note: In "Skeleton Data,' double-bounded boxes represent dictionary keys, and each connecting line points to the corresponding value in the layer below.)

Experiment on DUET

To evaluate the performance of the kinesics recognition framework, we implemented the framework on five different subsets of DUET, as shown in Table II. This design choice was motivated by the initial underperformance of ST-GCN. When applied to the full set of 12 activities, ST-GCN struggled to accurately classify the activities, which in turn limited its ability to generate meaningful low-dimensional representations. These weak representations negatively impacted the downstream CNN, which failed to reliably recognize the underlying kinesic patterns in the skeleton data.

To address this issue, we constructed four additional subsets of DUET by selecting activities with more distinct movement patterns—thus maximizing the likelihood of improved ST-GCN accuracy. As shown in Table II, ST-GCN performance improves as the number of activity types decreases. With more accurate feature compression in ST-GCN, the final hidden layer can then be more effectively used as input to the CNN for kinesics classification. In general, the trends of ST-GCN and CNN performances align with each other—CNN improves when ST-GCN improves.

The results in Table II demonstrate a clear parallel between the performance of ST-GCN and CNN: the CNN is able to more accurately extract the kinesics of human activities when ST-GCN effectively encodes the distinguishing characteristics of each activity into a lower-dimensions representation. This finding has two key implications. First, improving the performance of ST-GCN alone can lead to better results for both HAR and kinesics classification. However, current limitations of ST-GCN hinder its ability to accurately classify all activities in DUET, particularly those involving subtle hand gestures. These fine-grained movements are difficult to capture with skeletal data, as each hand is represented by only three joints—fingertip, thumb, and wrist [15]. If future developments enable ST-GCN to reliably classify these nuanced interactions, we anticipate that the corresponding kinesic functions will also become more accurately identifiable. Second, the observed alignment between ST-GCN and CNN performance suggests that there is an underlying structure within the skeleton data that governs the kinesic function of movements. To the best of our knowledge, this latent structure has not yet been systematically explored and presents a promising direction for future research.

TABLE II. Results of five subsets of DUET tested on the kinesics recognition framework.

Number of interactions	Activity labels	ST-GCN Accuracy (%)	CNN Accuracy (%)
4	2, 4, 8, 11	77	85
6	2, 4, 5, 7, 8, 11	75	81
8	0, 2, 4, 6, 7, 8, 9, 11	70	70
10	0, 1, 2, 3, 5, 6, 7, 8, 10, 11	67	58
12	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	55	48

CONCLUSIONS AND FUTURE WORK

We introduce a kinesics recognition framework that uses transfer learning to infer the communicative functions of human activities based on a psychologically grounded taxonomy. By integrating ST-GCN with a CNN, the framework extracts latent structures from 3D skeleton joint data that reflect the kinesic categories—emblems, illustrators, regulators, adaptors, and affect displays. Traditionally, recognizing kinesic functions has relied on a one-to-one mapping between activities and categories, a method that lacks generalizability and incurs significant time and cost. Our approach eliminates the need for this manual mapping by uncovering intrinsic patterns within the data that govern kinesic expression. This not only improves accuracy and scalability, but also provides a richer, real-time input source for reinforcement learning models simulating human-environment interactions. Importantly, the use of anonymous 3D skeleton data preserves user privacy, aligning with ethical considerations central to human-centered research.

Future work will focus on establishing the functional relationship between the performances of ST-GCN and CNN. While our experimental results suggest a consistent trend between the two, a rigorous statistical analysis is needed to quantify this relationship—specifically, to assess its linearity and correlation. Establishing this link will enable the joint refinement of both components, advancing the framework's ability to translate human activities into deeper representations of psychological and cognitive states.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under Grant #2425121.

REFERENCES

1. Hillier, B. 2007. *Space is the Machine: A Configurational Theory of Architecture*, Space Syntax.
2. Eijkelenboom, A., H. Verbeek, E. Felix, and J. Van Hoof. 2017. "Architectural factors influencing the sense of home in nursing homes: An operationalization for practice," *Frontiers of architectural research*, 6(2):111–122.
3. Schneider, M. 2002. "Do school facilities affect academic outcomes?" Tech. rep., National Clearinghouse for Educational Facilities.
4. Doctorarastoo, M., K. Flanigan, M. Bergés, and C. McComb. 2023. "Exploring the potentials and challenges of cyber-physical-social infrastructure systems for achieving human-

centered objectives," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Istanbul, Turkey, pp. 385–389.
5. Doctorarastoo, M., K. A. Flanigan, and M. Bergés. 2024. "Preference-aware human spatial behavior modeling in cyber-physical-human systems," *IFAC-PapersOnLine*, 58(30):115–120.
6. Crooks, A. T. and A. J. Heppenstall. 2011. "Introduction to agent-based modelling," *Agent-based models of geographical systems*.
7. Ajzen, I. 1991. "The theory of planned behavior," *Organizational behavior and human decision processes*, 50(2):179–211.
8. Liang, X., T. Lu, and G. Yishake. 2022. "How to promote residents' use of green space: An empirically grounded agent-based modeling approach," *Urban Forestry & Urban Greening*, 67:127435.
9. Lin, C., M. Doctorarastoo, and K. Flanigan. 2024. "Your actions talk: Automated sociometric analysis using kinesics in human activities," in *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Hangzhou, China, pp. 271–278.
10. Doctorarastoo, M., K. Flanigan, M. Bergés, and C. McComb. 2023. "Modeling human behavior in cyber-physical-social infrastructure systems," in *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, Istanbul, Turkey, pp. 370–376.
11. Lin, C., J. Martins, and K. A. Flanigan. 2024. "Read the room: Inferring social context through dyadic interaction recognition in cyber-physicalsocial infrastructure systems," in *ASCE International Conference on Computing in Civil Engineering*.
12. Sharan, N. N., A. Toet, T. Mioch, O. Niamut, and J. B. van Erp. 2022. "The relative importance of social cues in immersive mediated communication," in *Human Interaction, Emerging Technologies and Future Systems V: Proceedings of the 5th International Virtual Conference on Human Interaction and Emerging Technologies*, France, pp. 491–498.
13. Ekman, P. and W. V. Friesen. 1969. "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, 1(1):49–98.
14. Torrey, L. and J. Shavlik. 2010. "Transfer learning," *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*.
15. Yan, S., Y. Xiong, and D. Lin. 2018. "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32.
16. Lin, C., K. A. Flanigan, and M. Sirajum. 2024, "DUET Repository," Available online at: `https://huggingface.co/datasets/saluslab/DUET`.
17. Lin, C. 2025, "Kinesics Recognition Hugging Face repository," Available online at: `https://huggingface.co/saluslab/DUET_kinesics_recognition`.
18. Martins, J., C. Lin, K. A. Flanigan, and C. McComb. 2025. "HM-SYNC: A multimodal dataset of human interactions with advanced manufacturing machinery," *Journal of Mechanical Design*, 147(4):044504.
19. Hartman, N. A. 2024, "Nonverbal communication teaching note," https://dspace.mit.edu/bitstream/handle/1721.1/106670/15-281-spring-2009/contents/readings/MIT15$_2$81s09$_r$$ead$02.$pdf$.
20. Neff, M., N. Toothman, R. Bowmani, J. E. Fox Tree, and M. A. Walker. 2011. "Don't scratch! Self-adaptors reflect emotional stability," in *Intelligent Virtual Agents: 10th International Conference, IVA 2011*, Reykjavik, Iceland, pp. 398–411.