# Your Actions Talk: DUET – A Multimodal Dataset for Contextualizable Dyadic Activities

**Cheyu Lin**[1] cheyul@andrew.cmu.edu
**Katherine A. Flanigan, Ph.D.**[1] kflaniga@andrew.cmu.edu
**Sirajum Munir, Ph.D.**[2] sirajum.munir@us.bosch.com

[1]**Department of Civil and Environmental Engineering, Carnegie Mellon University Pittsburgh, PA, USA**

[2]**Bosch Research and Technology Center Pittsburgh, PA, USA**

## Abstract

Human activity recognition (HAR) has advanced significantly with the availability of diverse datasets, yet the field remains constrained by the scarcity of resources focusing on contextualizable two-person, or "dyadic" interactions. Existing datasets primarily aim to help improve the recognition of physical coordination in single-person settings, overlooking the intricate dynamics and kinesics present in interactions between two individuals. To address this gap, we introduce the Dyadic User EngagemenT (DUET), a comprehensive dataset designed to enhance the understanding and recognition of interactions. DUET comprises 12 interactions adopted from a taxonomy rooted in psychology to distill social semantics embedded in bodily movements. The marriage of HAR and social context dependencies sets the stage for refining recognition accuracy, improving the authenticity of telepresence avatar, automating sociological and psychological examinations, and many more. To support applications with different purposes and constraints, every sample spans across four modalities: RGB, depth, infrared, and 3D skeleton joints. Besides, the dataset was collected at three locations, including an open indoor space, a confined indoor space, and an open outdoor space. The variety of data collection locations helps improve the resilience against background variation and investigate the effect ambient environment imposes on HAR algorithms. In total, we collected 14,400 samples utilizing a novel technique that captures interactions from multiple views with a single camera. The technique diversifies how interactions are observed and yields the highest sample-class ratio know to data. We benchmark six state-of-the-art HAR algorithms on DUET, demonstrating the dataset's complexity and current model's limitations in recognizing dyadic interactions. DUET is publicly available at DUET repository, providing a valuable resource for the research community to advance HAR in dyadic settings.

## 1 Introduction

Human activity recognition (HAR), a prominent and rapidly advancing field of artificial intelligence, has achieved significant success across numerous domains [10]. Its ability to analyze and decipher the underlying structure of high-dimensional data and infer patterns in previously unseen samples has achieved many economic goals (e.g., performance, safety). The success of HAR can be attributed to many factors, including publicly available datasets that help refine data-driven deep learning algorithms across contexts. While there are an abundance of datasets already available, the majority

NeurIPS 2024 Workshop on Behavioral Machine Learning.

Table 1: A comparison of existing dyadic datasets. Note: The number of views represents different orientations of the interaction being captured by the sensor, the variation of which adds diversity to how interactions are captured.

| Dataset | | Modalities | #Videos | #Classes | #Videos/ #Classes | #Locations | #Views | Year |
|---|---|---|---|---|---|---|---|---|
| UT Interaction | [1] | RGB | 160 | 6 | 26.7 | 2 | 1 | 2010 |
| SBU Kinect | [2] | RGB+D+J | 300 | 8 | 37.5 | 1 | 1 | 2012 |
| K3HI | [3] | D | 320 | 6 | 53.3 | 1 | 1 | 2013 |
| JPL Interaction | [4] | RGB | 399 | 7 | 57 | 5 | 1 | 2013 |
| G3Di | [5] | RGB+D+J | 168 | 14 | 12 | 1 | 1 | 2015 |
| $M^2I$ | [6] | RGB+D+J | 1,760 | 9 | 195.6 | 1 | 2 | 2015 |
| ShakeFive 2 | [7] | RGB+J | 153 | 8 | 19.1 | 1 | 1 | 2016 |
| NTU RGB+D 120 | [8] | RGB+D+J+IR | 24,828 | 26 | 954.9 | - | 155 | 2019 |
| Air Act2Act | [9] | RGB+D+J | 5,000 | 10 | 500 | 2 | 3 | 2020 |
| **DUET (our dataset)** | | **RGB+D+J+IR** | **14,400** | **12** | **1,200** | **3** | **360** | **2024** |

of datasets pertain to single-person—or monadic—activities, and they focus primarily on refining the recognition of physical movements. This disproportionate effort biased towards monadic HAR overlooks the increased complexity of spatial and temporal coordination between two subjects. The work of Lin *et al.* [11] revealed that monadic algorithms that have outstanding benchmarking records for single-person activities do not perform nearly as well for dyadic interactions, highlighting the disparity between monadic and dyadic activities. Additionally, another understudied aspect of dyadic activities is kinesics. Similar to paralinguistics, kinesics explores beyond the physical movements of body parts—it interprets the nonverbal channel of human communication, along with the expanded variety of expressive and cultural messages in dyadic activities [12]. Not only does the integration of kinesics enhance the performance of HAR [13], it opens the doors to a wide array of downstream applications. For instance, an application inspired by Jupalle *et al.*'s work [14] is to identify emotions by extracting the emotional dependencies embedded in human activities. It can also serve as a pedagogical tool instructors leverage to observe, understand, evaluate, and encourage student collaborations [15]. The contextual comprehension of dyadic activities refines the understanding between stroke survivors and caregivers in healthcare and rehabilitation facilities to improve quality of life [16]. Interactions with humanoid robots are beneficial to the social and cognitive development of autistic children, and the study of social embeddings in human interactions improves the authenticity of the robots [17]. Also, positive interactions between salesmen and customers contribute to the desire of customers to make purchases, and contextualization deepens the understanding of interactions to assess and improve customer services [18]. Despite the identified advantages of kinesics, the few existing dyadic datasets—listed in Table 1—fall short in supporting the extraction of the kinesics. Some of these datasets focus on healthcare activities, while others focus on the mere tracking of bodily movements within their respective specific action categories. A dyadic dataset that is contextualizable is still absent in the research community.

To enhance the performance of HAR for dyadic activities through contextualization, we present the Dyadic User Engagement dataseT (DUET), a dyadic dataset featuring 12 taxonomized interactions, providing a foundation from which to bridge the gap between monadic and dyadic HAR. The selection of interactions is adapted from a classification system rooted in psychology that identifies five fundamental communication functions in human interactions: emblems, illustrators, affect displays, regulators, and adaptors. This methodological choice moves beyond the tracking of bodily movements, contributing to the deeper comprehension of kinesics. In total, 14,400 samples are collected with a novel technique, which captures interactions from multiple views with one camera. The variation in view observes the same interaction from different orientations, which ameliorates the view invariance attribute of HAR algorithms. To support downstream applications with different purposes and constraints, every sample is collected with four modalities, including RGB, depth, infrared (IR), and 3D skeleton joints. The dataset was collected at three locations: an open indoor space, a confined indoor space, and a open outdoor space. Not only does the variety of locations enhance the resilience against background variation, it also helps investigate the effect ambient environment imposes on HAR algorithms. DUET is publicly available at DUET repository [19] and is provided under an MIT License. The repository also contains the **README.md** file that

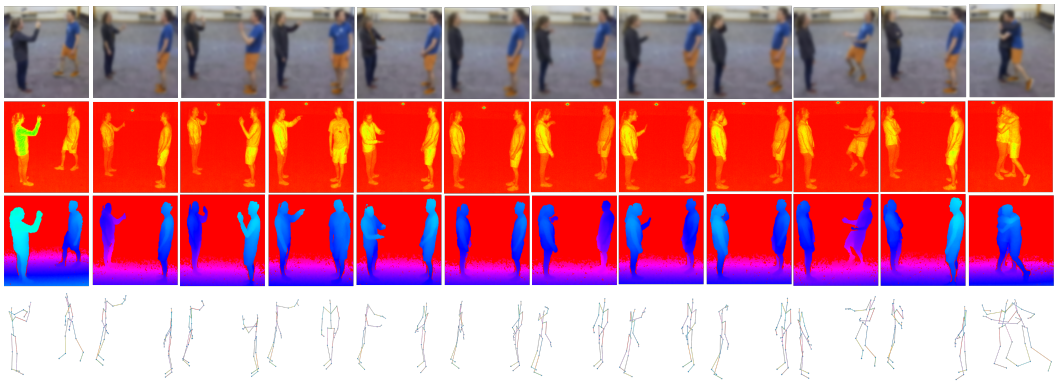


Figure 1: Sample data from 12 interactions. The modalities presented are, from top row to bottom row: RGB, IR, depth, and 3D skeleton joints. The 12 interactions are, from left to right: waving in, thumbs up, waving, pointing, showing measurements, nodding, drawing circles in the air, holding one's palms out, twirling or scratching hair, laughing, arm crossing, and hugging.

details the data acquisition process, including the selection of sensor, data modalities and data format, data collection configurations, biometrics of the subjects, data annotation, and the data division for cross-location and cross-subject evaluations.

The remainder of the paper is organized as follows. Section 2 details the taxonomy adopted to classify human interactions. Following in Section 3, we benchmark six algorithms and outline their corresponding results. Lastly, Section 4 presents key conclusions and takeaways of the work, as well as a discussion of next steps and further development.

## 2 Contextualizing human interactions

Similar to paralinguistics, bodily movements in human interactions carry more than the relocation of body parts—they also deliver social-contextual and cultural embeddings, also known as kinesics. To study kinesics present in dyadic interactions, we propose a dataset, DUET, in this work. A total of 12 dyadic interactions are selected in the dataset, the sample frames of which are displayed in Figure 1. These interactions are adopted from a taxonomy compiled by Ekman *et al.* [20], categorizing human interactions into five groups depending on their fundamental communication functions. This classification system lays the groundwork for efficient extraction of the aforementioned embeddings. The categories encompass emblems, illustrators, affect displays, regulators, and adaptors.

- Emblems: Emblems have direct verbal translation and can be culturally specific. The same gesture might be interpreted differently for different cultures [21]. For instance, a thumbs up indicates "well done" in most Western cultures, but is a derogatory sign in Greece and some Middle Eastern countries. Interactions chosen are *waving in*, *thumbs-up*, and *hand waving*.

- Illustrators: Illustrators are used to clarify the conversation they accompany. Interactions chosen are *pointing* and *showing measurements*.

- Affect displays: Affect displays are gestures that reveal one's affective and emotional state. Interactions chosen are *hugging*, *laughing*, and *arm crossing*.

- Regulators: During interactions, regulators determine the alternation of instigating and receiving. Interactions chosen are *nodding*, *writing circles in the air*, and *holding palms out*.

- Adaptors: Adaptors are habitual movements that satisfy personal needs and can be used to increase or decrease emotional stability [22]. Interactions chosen are *twirling or scratching hair*.

Table 2: Cross-location and cross-subject accuracy comparison for the three modalities considered: RGB, depth, and 3D skeleton joints.

| HAR algorithm | | Modality | Cross-location accuracy ( %) | Cross-subject accuracy ( %) |
|---|---|---|---|---|
| DB-LSTM | [23] | RGB | 9.65 | 17.85 |
| V4D | [24] | RGB | 8.26 | 7.79 |
| DOGV-ST3D | [25] | Depth | 13.15 | 18.77 |
| DB-LSTM | [23] | Depth | 14.94 | 23.18 |
| PAM-STGCN | [26] | 3D skeleton joint | 30.73 | 36.65 |
| DR-GCN | [27] | 3D skeleton joint | 38.17 | 41.57 |

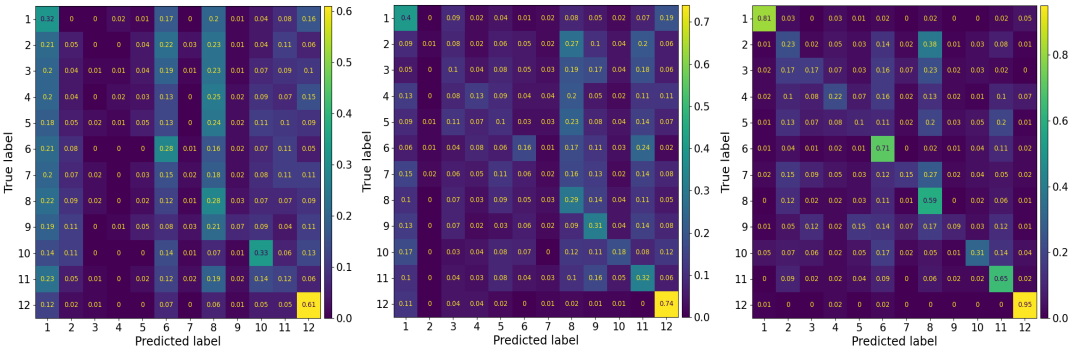
## 3 Benchmarking state-of-the-art HAR algorithms

In this section, we evaluate the performance of six state-of-the-art HAR models with publicly available code. This evaluation features two RGB-based, two depth-based, and two skeleton-based algorithms for different modalities provided in DUET. The results of the evaluation are presented in Table 2.

Overall, cross-subject evaluation outperforms cross-location evaluation in the state-of-the-art algorithms for all modalities, which can be justified for two reasons. First, RGB-based and depth-based algorithms are prone to learning from view-dependent motion patterns. Specifically, they tend to correlate background with motion trajectories during training. In cross-subject evaluation, the training set contains samples collected from three locations, whereas only two locations are taken into account during training in cross-location evaluation. These selected models fail to generalize the unseen background during testing, which results in lower accuracy in cross-location evaluation. Second, the difference in the number of training samples also contributes to the disparity in performance. We use 80% of our dataset for training in cross-subject evaluation, while only two-thirds of our dataset is used as training samples in cross-location evaluation. The performance improves as more samples are dedicated to training. These two phenomena are also present in Liu *et al.*'s work [8].

Another notable observation is the gradual increase in accuracy of state-of-the-art HAR algorithms tested in our work, from RGB to depth, and to 3D skeleton joints, which corresponds to the expansion of the dimension. For RGB-based algorithms, the input is compressed into a 2D plane, which induces low accuracy since human interactions comprise both 3D spatial and temporal coordination [28]. The compression of one dimension sacrifices the spatial comprehension to identify these interactions. The addition of depth to each pixel on an image gives us depth images, providing another degree of information. The refinement of performance given the substitution of colors for depth is evident, especially when we evaluate RGB and depth inputs with the same model (i.e., DB-LSTM) separately. Despite the increase in accuracy from RGB to depth modalities, there still remains room for improvement for both modalities. The substandard accuracy is due to both modalities operating in Euclidean space (i.e., images), where operations are susceptible to varying views. DUET aim to address this issue by including considerable views. Additionally, training in the Euclidean space is easily affected by trivial features. As shown in Figure 2a and Figure 2b, RGB and depth models are confused by common poses shared between these activities—for instance, standing is common in all of these activities. On the other hand, skeleton-based algorithms perform HAR on a non-Euclidean space [29], representing human interactions in a 3D world relative to the camera.

Skeleton-based algorithms outperform other modalities because the activities are captured in a 3D world relative to the camera, which is well-suited for the spatial complexity of human interactions. Regardless of the view, skeleton-based networks are capable of extracting the underlying motion patterns. Additionally, 3D skeletons sparsely represent human skeletons, which prevents the network from learning unrelated features. However, the sparsity is also detrimental to the recognition task for our dataset. In our dataset, it contains dyadic interactions that only differ from each other on a very small scale. For instance, both thumbs up and holding palms out (i.e., label ID 2 and 8) require arm extension, but the former involves raising the thumb, while the latter involves holding the hand vertically. The simplified representation of the human body might not be able to capture minute nuances between these two activities using state-of-the-art HAR algorithms. This is evident in Figure 2c, which shows that these two actions are confused by the algorithm the most. While the nuances




(a) RGB-based DB-LSTM [23] (b) Depth-based DB-LSTM [23] (c) Skeleton-based DR-GCN [27]

Figure 2: Representative confusion matrices for cross-subject evaluations. Note: Class labels represent (1) waving in, (2) thumbs up ,(3) waving, (4) pointing, (5) showing measurements, (6) nodding, (7) drawing circles in the air, (8) holding palms out, (9) twirling or scratching hair, (10) laughing, (11) arm crossing, and (12) hugging, respectively.

are conspicuous in RGB and depth images, which is where the 3D skeletal joints are extracted from, the state-of-the-art skeleton-based algorithms still cannot capture them.

## 4 Discussion and conclusion

In this work, we present DUET, a contextualizable dataset containing 12 interactions, adapted from a psychology classification system that categorizes human interactions based on their communication functions. The extraction of kinesics refines HAR models and sets the stage for considerable downstream applications, such as autonomous vehicles, smart homes, urban infrastructure planning, emotion recognition, and healthcare. 14,400 samples were collected at three locations with a novel technique that collects a considerable amount of data of different modalities (i.e., RGB, depth, IR, and 3D skeleton joints) from various views with one camera. These strategic design decisions improve resilience against background noise and view variations.

To provide a baseline performance for DUET, we evaluate six HAR algorithms—two RGB-based, two depth-based, and two skeleton-based algorithms. The results demonstrate the (1) intricacy of social interactions that has not been captured in literature and (2) the susceptibility of HAR algorithms to changes in view and background, which are newly-found research gaps for future exploration.

Future developments stemming from this work can be broadly categorized in two ways. The first development is the refinement of HAR algorithms of all modalities for contextualizable dyadic interactions, the current limitation of which is demonstrated in Table 2. Additionally, we have laid the groundwork for extracting the kinesics of humna activities by incorporating HAR with the psychological classification system. The next step is to define the framework with more details to numerically map all interactions to their corresponding level of embeddings, benefiting downstream applications, such as automatic sociological and psychological examination, and many more.

## References

[1] M. S. Ryoo, C. C. Chen, J. Aggarwal, and A. Roy Chowdhury, "An overview of contest on semantic description of human activities (sdha) 2010," *Recognizing Patterns in Signals, Speech, Images and Videos: ICPR 2010 Contests, Istanbul, Turkey, August 23-26, 2010, Contest Reports*, pp. 270–285, 2010.

[2] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2012.

[3] T. Hu, X. Zhu, W. Guo, and K. Su, "Efficient interaction recognition through positive action representation," *Mathematical Problems in Engineering*, vol. 2013, no. 1, 2013.

[4] M. S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?" In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[5] V. Bloom, V. Argyriou, and D. Makris, "Hierarchical transfer learning for online recognition of compound actions," *Computer Vision and Image Understanding*, vol. 144, pp. 62–72, 2016.

[6] T. Liu, Z. Chen, H. Liu, Z. Zhang, and Y. Chen, "Multi-modal hand gesture designing in multi-screen touchable teaching system for human-computer interaction," in *Proceedings of the 2nd International Conference on Advances in Image Processing*, 2018.

[7] C. Van Gemeren, R. Poppe, and R. C. Veltkamp, "Spatio-temporal detection of fine-grained dyadic human interactions," in *Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands, October 16, 2016, Proceedings 7*, Springer, 2016.

[8] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "Ntu rgb + d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[9] W. R. Ko, M. Jang, J. Lee, and J. Kim, "Air-act2act: Human–human interaction dataset for teaching non-verbal social behaviors to robots," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 691–697, 2021.

[10] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, 2021.

[11] C. Lin, J. Martins, and K. A. Flanigan, "Read the room: Inferring social context through dyadic interaction recognition in cyber-physical-social infrastructure systems," in *Computing in Civil Engineering 2024: Re-Imagining Civil Engineering—Addressing the World's Most Pressing Problems*, 2024.

[12] M. A. U. Othman, Z. Ismail, C. M. Zaid, *et al.*, "Kinesics as a form of non verbal communication: A textual analysis of the holy quran," *The journal of contemporary issues in business and government*, vol. 27, no. 2, pp. 201–207, 2021.

[13] V. Adeli, E. Adeli, I. Reid, J. C. Niebles, and H. Rezatofighi, "Socially and contextually aware human motion and pose forecasting," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6033–6040, 2020.

[14] H. Jupalle, S. Kouser, A. B. Bhatia, N. Alam, R. R. Nadikattu, and P. Whig, "Automation of human behaviors and its prediction using machine learning," *Microsystem Technologies*, vol. 28, no. 8, pp. 1879–1887, 2022.

[15] S. J. Kühl, A. Schneider, H. A. Kestler, M. Toberer, M. Kühl, and M. R. Fischer, "Investigating the self-study phase of an inverted biochemistry classroom–collaborative dyadic learning makes the difference," *BMC medical education*, vol. 19, pp. 1–14, 2019.

[16] G. Pucciarelli, M. Lommi, G. S. Magwood, *et al.*, "Effectiveness of dyadic interventions to improve stroke patient–caregiver dyads' outcomes after discharge: A systematic review and meta-analysis study," *European Journal of Cardiovascular Nursing*, vol. 20, no. 1, pp. 14–33, 2021.

[17] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "Collaborating with kaspar: Using an autonomous humanoid robot to foster cooperative dyadic play among children with autism," in *2010 10th IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2010.

[18] S. Lee and A. Dubinsky, "Influence of salesperson characteristics and customer emotion on retail dyadic relationships," *The International Review of Retail, Distribution and Consumer Research*, vol. 13, no. 1, 2003.

[19] C. Lin, *DUET Repository*, `https://huggingface.co/datasets/saluslab/DUET`.

[20] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[21] N. A. Hartman, *Nonverbal communication teaching note*, en, [Online]. Available: `https://ocw.mit.edu/courses/15-279-management-communication-for-undergraduates-fall-2012/251fccce2dabe0f6ceafb86218d74c57_MIT15_279F12_nonVerbalComm.pdf`, Accessed on Jun 03 2024.

[22] M. Neff, N. Toothman, R. Bowmani, J. E. Fox Tree, and M. A. Walker, "Don't scratch! self-adaptors reflect emotional stability," in *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*, Springer, 2011.

[23] J. Y. He, X. Wu, Z.-Q. Cheng, Z. Yuan, and Y. G. Jiang, "Db-lstm: Densely-connected bi-directional lstm for human action recognition," *Neurocomputing*, vol. 444, pp. 319–331, 2021.

[24] S. Zhang, S. Guo, W. Huang, M. R. Scott, and L. Wang, "V4d: 4d convolutional neural networks for video-level representation learning," *arXiv preprint arXiv:2002.07442*, 2020.

[25] J. Xiaopeng, Z. Qingsong, C. Jun, and M. Chenfei, "Exploiting spatio-temporal representation for 3d human action recognition from depth map sequences," *KnowledgeBased Systems*, 2021.

[26] C. L. Yang, A. Setyoko, H. Tampubolon, and K.-L. Hua, "Pairwise adjacency matrix on spatial temporal graph convolution network for skeleton-based two-person interaction recognition," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020.

[27] L. Zhu, B. Wan, C. Li, G. Tian, Y. Hou, and K. Yuan, "Dyadic relational graph convolutional networks for skeleton-based human interaction recognition," *Pattern Recognition*, vol. 115, 2021.

[28] G. Lee and J. Kim, "Improving human activity recognition for sparse radar point clouds: A graph neural network model with pre-trained 3d human-joint coordinates," *Applied Sciences*, vol. 12, no. 4, 2022.

[29] W. Peng, J. Shi, T. Varanka, and G. Zhao, "Rethinking the st-gcns for 3d skeleton-based human action recognition," *Neurocomputing*, vol. 454, pp. 45–53, 2021.