VR Driven Unsupervised Classification for Context Aware Human Robot Collaboration

Ali Kamali Mohammadzadeh [1][0000-0003-0346-6570], Carlton Leroy Allen [2][0000-0001-8221-5125], and Sara Masoud [1][0000-0001-7375-3300]

[1] Wayne State University, Detroit MI 48201, USA
[2] Bowling Green State University, Bowling Green, OH 43403, USA
saramasoud@wayne.edu

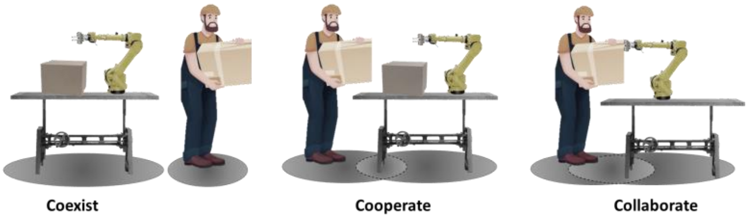
Abstract. Human behavior, despite its complexity, follows structured principles that, if understood, will result in more reliable and effective collaborative automation environments. Characterizing human behavior in collaborative automation systems based on understanding the underlying context allows for novel advances in robotic human behavior sensing, processing, and predicting. Here, virtual reality, through integration of HTC Vive Arena Pro Eye Bundle, Leap Motion, and Unity 3D game engine, is used for safe and secure data collection on humans' movements and body language in human robot collaborative environments. This paper proposes an unsupervised classification framework through integration of dynamic time warping and k-means clustering algorithm to enable robotics agents to understand humans' intentions based on their body movements. Results display that the proposed framework is capable of identifying underlying intentions with an average accuracy, recall, and precision of 85%, 73%, and 75%, respectively.

Keywords: Human Robot Interaction, Virtual Reality, Unsupervised Classification.

1 Introduction

Human-robot interaction (HRI) is a diverse field of research with huge economic impact. The collaborative robots is estimated to reach over $1.43 billion by 2027, having a significant impact on the Gross Domestic Product (GDP) of various economies[1]. HRI is already used in manufacturing environments, space applications, rescue robotics, and more [2-4] Human–robot interaction is defined as "the process of conveying human intentions and interpreting task descriptions into a sequence of robot motions complying with robot capabilities and working requirements" [5]. In manufacturing, industrial robots, equipped with different sensors, can be adapted to do many different industrial tasks [6]. HRI in manufacturing faces challenges such as ensuring safety and efficiency. To address these challenges, it is critical for robots understand humans' intentions and the underlying context [7].

As demonstrated in Fig. 1, HRI can be classified into the following three classes: First, Human–Robot Coexistence, which is defined as the capability of sharing the workspace between humans and robots without requiring mutual contact or coordination of actions and intentions [8]. Second, Human–robot cooperation, in which humans and robots work on the same goal and occupy the same time and space, simultaneously. The cooperation requires advanced technologies and techniques for collision detection and collision avoidance [9]. Third, Human–robot collaboration, performing complex tasks with direct human interactions, where explicit contact between human and robot exists [6].



Aspects	Human Robot Interaction		
Space	Separated	Shared	Shared
Tasks	Unrelated	Linked	Shared
Contact	Impossible	Possible	Desired
Context Awareness	Not necessary	Minimal	Necessary

Fig. 1. The three forms of human-robot interaction [7].

Humans and robots collaborating is on a common objective form a team with complementary skills, in which an agreed-upon strategy is necessary for effective collaboration amongst all parties . Also, humans and robots need to be aware of the intents of the other team members. Based on which, a robot can design its own behaviors that will ultimately result in the achievement of the shared objective through perceiving and understanding the surroundings as well as making decisions and plan ahead [10].

The goal of this study is to create a virtual platform that enables safe yet realistic context-aware human-robot collaboration. To this end, this study develops a Virtual Reality (VR) platform for safe data collection, while proposes an unsupervised classification framework through integration of dynamic time warping (DTW) and k-means algorithm for intent prediction. This enables robotics agents to predict and understand human intents from their physical cues. In the remainder of the paper, Section 2 reviews the literature of the topic. Section 3 provides the methodology of the research. Results and discussion are provided in Section 4 and, Section 5 concludes the paper.

2 Literature review

The literature of human intention recognition (classification) is diverse and has caught the attention of researchers from both industry and academia [11]. Intention recognition is also known as activity recognition, plan recognition, goal recognition, behavior recognition. In this work, we define intention recognition as the process of determining what an observed agent intends to do in the immediate future [12]. Recognizing the goal of the observed agent is an important aspect in human-robot interaction, as this understanding enables effective interaction and also proactive behavior modification in order to prevent accidents or resolve issues that may arise.

Human intention/activity recognition based on images or videos has received plenty of attention recently in the field of computer vision. The occlusion problem, however, reduces the accuracy of visual-based recognitions [13]. Wearable devices, on the other hand, immediately sense human body movements, providing real-time information on the body status. Additionally, a variety of low-cost wearable gadgets are available on the market and are frequently employed in intention recognition [14].

Among the studies that addressed the problem of human intention recognition, Stiefmeire et al. [15] utilized ultrasonic sensors for worker activity recognition using a Hidden Markov Model. In their following study, the authors proposed a string-matching based classification approach using multiple sensors for recognizing worker activities in a manufacturing setting [16]. Koskimaki et al. classified five activities for industrial assembly lines using a wrist-worn Inertial Measurement Unit (IMU) sensor and a K-Nearest Neighbor model [17]. Using inputs from a smartwatch combined with an IMU sensor, Maekawa et al. [18] proposed an unsupervised approach for lead time estimation of manufacturing activities.

Zhu et al. [19] addressed human intention recognition by designing a hidden Markov based recognition algorithm to classify hand gestures using an inertial sensor worn on finger of the subject. Zhu et al. tacked motion intention recognition using an inertial sensor and a deep convolution neural network (CNN) to extract discriminant features from temporal gait period [20]. Sun et al. used muscle electrical signals and joint angle signals as motion data and utilized K-Nearest Neighbor algorithm to identify four gait motion modes including walking naturally, climbing stairs, descending stairs, and crossing obstacles [21].

Wen and Wang [22] proposed an intention recognition algorithm based on multimodal spatiotemporal feature fusion using the data collected by multimodal sensors. Masoud et al. [23] proposed a task recognition framework to identify undergoing tasks in pseudo real-time using a pair of data glove for grafting operations. This study offers a distinct contribution to the existing body of knowledge and provides a fresh and unique approach to the study of human-robot interaction by integrating the advantages of a virtual immersive platform and an unsupervised classification for context aware HRI.

3 Methodology

As displayed in Fig. 2, our proposed framework for intention classification consists of two main phases: First, creating an immersive virtual platform, which is used for safe data collection and second, unsupervised classification using K-means and dynamic time warping. In data acquisition and processing phase, we develop an interactive physics-based model via Unity 3D game engine. The developed model is integrated with HTC-Vive room scale pro eye arena bundle and leap motion controller to create the immersive virtual platform for data collection. The data collected through this immersive platform (Unity) is then cleaned and processed using open source python libraries. In the training and intention classification phase, the processed data is used to train our unsupervised intention classification model.

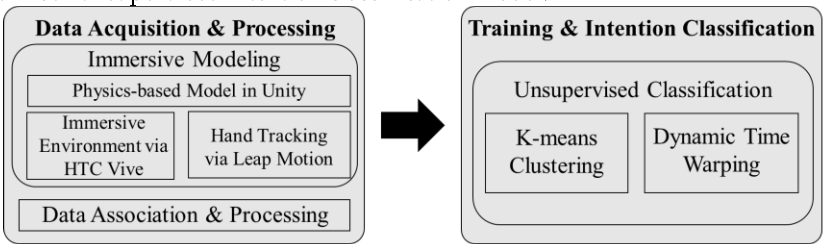

Data Acquisition & Processing
Immersive Modeling
Physics-based Model in Unity
Immersive Environment via HTC Vive
Hand Tracking via Leap Motion
Data Association & Processing
Training & Intention Classification
Unsupervised Classification
K-means Clustering
Dynamic Time Warping

Fig. 2. Our proposed intention classification framework.

3.1 Immersive Virtual Reality Platform

To build the immersive platform, HTC-Vive pro eye arena bundle is used to model the immersive environment while Leap Motion controller replaces the typical controllers to enable users interact with the virtual environment using their hands. Leap motion controller captures the hands' gestures and movements using optical hand tracking sensors with high accuracy. Leap Motion controller can be mounted on the HTC VIVE head mounted display and connects communicate with the system via a USB cable. HTC VIVE also connects to the computation unit (Dell XPS 15 7590) via HDMI cables. HTC VIVE connects to Unity through VIVE port API package, while Leap Motion relies on its own package, UltraLeap plugin. Unity platform allows us to create a safe realistic physics-based manufacturing shop floor, where users can interact with equipment.

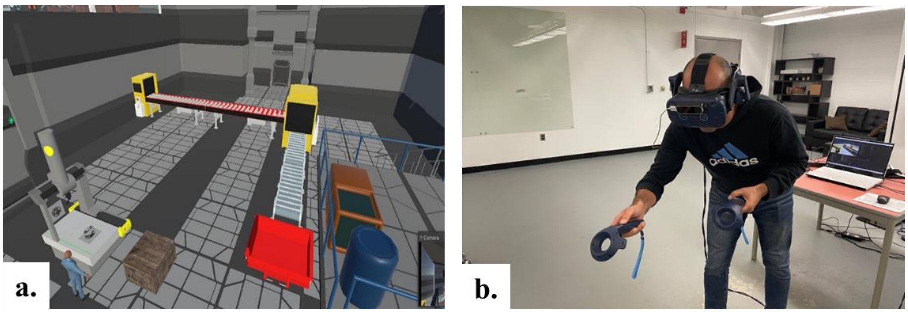


Fig. 3. a. The VR model, developed via Unity 3D game engine, b. The experimental setup for data collection.

Our dataset of activities is established based on the literature of the topic and the available datasets (e.g., UCI, WISDM, GAMI) including activities such as standing idle (ST), walking (WA), bending (BE), and sitting (SI). Then, subjects are recruited and asked to stand in the middle of a designated zone, wear the immersive technology, and perform the assigned activity in a natural way. The data are collected during performing the tasks by the subjects at a frequency of 0.2 seconds. Next, the collected observations are imputed, normalized, and outliers are dropped.

3.2 Unsupervised Classification

The proposed unsupervised classification integrates k-means clustering and dynamic time warping. K-means algorithm is selected due to its scalability, guarantee of convergence, and ease of generalization. To the best of our knowledge, this is the first time this proposed integration is used in intention recognition literature.

K-means Algorithm
K-means algorithm addresses the problem of clustering *m* elements (with *n* features) into *k* groups using the method of the lowest cost function (J), as shown in (1), which is usually the sum of each element's *n-dimensional Euclidean distance* to their closest group centroids. The constituents of the *jth* group are *m(j)*. The *n-dimensional* feature values of each element inside a group determine the centroids. The K-means algorithm recalculates the group centroids, updates the cost function, and assigns all other elements to the closest group. K-means repeatedly performs these tasks until achieving the lowest cost function [24].

$$J(C) = \sum_{j=1}^{k} \sum_{i=1}^{m(j)} \left\| x_i^{(j)} - C_j \right\|^2 \tag{1}$$

where $X = \{x_1, x_2, \dots, x_m\}$ and $C = \{c_1, c_2, \dots, c_k\}$ are the set of elements and centroids. Typically, Euclidean distance is the foundation of classification (or clustering). However, such a simplistic measure is not applicable to the observations that vary in size. As Euclidean distance is susceptible to even the slightest deviations in the size of comparing entities (time axis in case of time series), DTW replaces Euclidean distance in this study.

Dynamic Time Warping
Introduced by Berndt and Clifford [25], DTW provides a one-to-many match rather than being restricted to one-to-one matches. Given two members with *f* and *d* lengths, u_i $(i = 1, 2, 3, \dots, f)$ and v_j $(j = 1, 2, 3, \dots, d)$, a matrix $S_{i,j}$ is calculated as follows [25]:

$$S_{i,0} = S_{0,j} = 0 \tag{2}$$

$$S_{1,1} = (u_1 - v_1)^2 \tag{3}$$

$$S_{i,j} = min\,(S_{i-1,j}, S_{i,j-1}, S_{i-1,j-1}) + (u_i - v_j)^2 \tag{4}$$

where the DTW distance is the minimum value of the sums of $(u_i - v_j)^2$, calculated along several paths. The path that minimizes this sum is typically a warped curve. Dynamic time warping is a well stablished measure of difference between the sequences that vary in length. The observations collected in this study also have different lengths depending on the activities and how they are performed by the subjects.

Here, we start by collecting historical data and labeling the historical time series with the gestures generating them. Then, for any incoming time series read by the sensors, DTW takes place to collect time series of similar shapes. Finally, cluster centroids will be computed (or recomputed) with respect to reported DTW. Here the cluster centroid averages a subset from a set of time series within the measured DTW space. As a result, instead of a point, each centroid is a time series taking the average shape of the time series assigned to the cluster. The proposed framework label new time series by minimizing the DTW distance measures among the time series within clusters and their corresponding centroids.

4 Results and Discussion

In this study, 5 participants performed the assigned activities multiple times and 72216 data points were collected. During data pre-processing, 27966 data points were discarded. The remaining data points formed 125 observations, where each observation contains 16 features and variable number of time stamps (e.g., varying from 39 to 354). The features include headset forward vector (x, y, z), headset position (x, y, z), headset rotation (x, y, z), headset velocity (x, y, z), headset angular velocity (x, y, z), and timestamp. The processed observations are divided into train and test sets according to an 80/20 ratio. Four commonly used metrics, namely accuracy, precision, recall, and F-1 score, (5) to (8), are selected to evaluate the performance of our proposed unsupervised classification algorithm.

$$Accuracy = \frac{True\ Negative\ +True\ Positive}{True\ Nagative\ +True\ Positive+False\ Positive+False\ Nagative} \tag{5}$$

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \tag{6}$$

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \tag{7}$$

$$F - 1\ Score = 2 \times \frac{Percision * Recall}{Percision+Recall} \tag{8}$$

Although accuracy quantifies the ratio of correctly identified classes (intentions here), it cannot handle the imbalanced datasets. As a result, metrics such as precision, recall, and F-1 score are calculated to report the performance of our proposed model given our imbalanced dataset. While recall sheds light on the ratio of the relevant observations correctly classified by our trained models, precision represents the ratio of relevant observations. F-1 score, as the harmonic mean of recall and precision, combines these performance metrics into a single one. Relying on these metrics, we compared the performance of our proposed unsupervised classifier against the traditional k-means as displayed in Fig. 4.

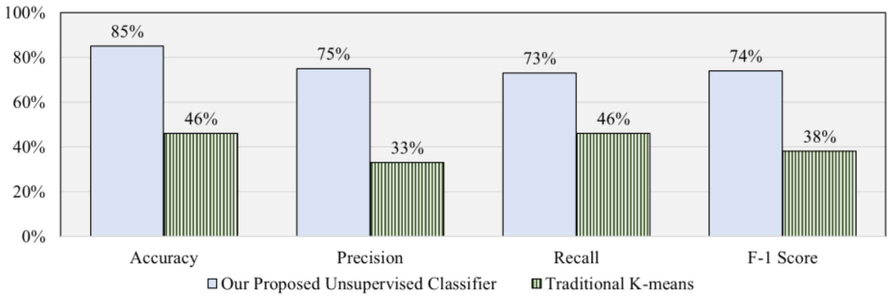


Fig. 4. Accuracy, precision, recall, and F-1 score of our proposed unsupervised classifier and the traditional k-means model.

As displayed in Fig. 4, our proposed model outperforms traditional k-means in accuracy, precision, recall, and F-1 score, by 39% (i.e., 85%-46%), 42% (i.e., 75%-33%), 27% (i.e., 73%-46%), and 36% (i.e., 74%-38%) over the test set, respectively. Justifying the performance gap between the proposed method and traditional k-means can be attributed to superiority of DTW on comparing time series and traditional k-means weakness in handling high dimensional data.

The superiority of DTW on comparing time series is due its on one-to-many matching. While Euclidean distance is vulnerable to even the smallest of distortions in time, DTW optimizes the fit by stretching and/or compressing the time axis over different intervals [25-26]. The second cause contributing to the poor performance of the traditional k-means is its inability to take advantage of a high range of features. While our proposed approach relies on full information available on all 16 available features, the traditional k-means can only process one dimensional time series. For reporting the baseline model’s performance, we trained 16 different traditional k-means models (corresponding to the 16 available features) and reported the best performance (i.e., the model trained on the head rotation time series over the y axis) in Fig. 4.

5 Conclusion

As a fast-growing field, HRI has the potential of revolutionizing manufacturing and production systems through providing safe and efficient human robot teams. To achieve that, the first step is developing platforms that are capable of making robots context aware and enabling them to learn about humans’ intention. In this work, we propose an unsupervised classification framework for intention recognition based humans’ body gestures and motions. The proposed framework is built upon integration of K-means and dynamic time warping algorithms. To train the proposed framework, an immersive VR platform is developed for safe and realistic data collection. Our proposed framework outperformed the traditional k-means and achieved average accuracy, recall, precision, and F-1 score of 85%, 73%, and 75%, and 74%, respectively. Our future research focus on developing more intuitive and seamless interaction methods that enhance the overall user experience and foster a sense of trust and reliability between the human and robot through addressing safety issues.

Acknowledgements

This work has been funded by the National Science Foundation under Grant No. 1950192.

References

1. GlobeNewswire, https://www.globenewswire.com/en/news-release/2020/08/25/2083218/0/en/Collaborative-Robots-Market-Worth-1-43-Billion-by-2027-Growing-at-a-CAGR-of-22-6-from-2020-Pre-and-Post-COVID-19-Market-Opportunity-Analysis-and-Industry-Forecasts-by-Meticulous-Re.html, last accessed 2023/02/02.
2. D. Mourtzis, J. Angelopoulos, and N. Panopoulos, “Closed-Loop Robotic Arm Manipulation Based on Mixed Reality,” *Appl. Sci.*, vol. 12, no. 6, p. 2972 (2022).
3. S. Masoud, M. Zhu, J. Rickli, and A. Djuric, "Challenges and Future Directions for Extended Reality-enabled Robotics Laboratories During COVID-19," *Technology Interface International Journal, vol 23, no.1, pp. 1-22 (2022).*
4. D. Mourtzis, “Simulation in the design and operation of manufacturing systems: state of the art and new trends,” *Int. J. Prod. Res.*, vol. 58, no. 7, pp. 1927–1949 (2020).
5. H. C. Fang, S. K. Ong, and A. Y. C. Nee, “A novel augmented reality-based interface for robot path planning,” *Int. J. Interact. Des. Manuf.*, vol. 8, no. 1, pp. 33–42 (2014).
6. A. Hentout, M. Aouache, A. Maoudj, and I. Akli, “Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017,” *Adv. Robot.*, vol. 33, no. 15–16, pp. 764–799 (2019).
7. R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric, “Human-robot interactions in manufacturing: A survey of human behavior modeling,” *Robot. Comput. Integr. Manuf.*, vol. 78, p. 102404 (2022).
8. A. O. Andrisano, F. Leali, M. Pellicciari, F. Pini, and A. Vergnano, “Hybrid Reconfigurable System design and optimization through virtual prototyping and digital manufacturing tools,” *Int. J. Interact. Des. Manuf.*, vol. 6, no. 1, pp. 17–27 (2012).
9. N. Wang, Y. Zeng, and J. Geng, “A brief review on safety strategies of physical human-robot interaction,” in *ITM Web of Conferences*, vol. 25, p. 1015 (2019).
10. A. Bauer, D. Wollherr, and M. Buss, “Human–robot collaboration: a survey,” *Int. J. Humanoid Robot.*, vol. 5, no. 01, pp. 47–66 (2008).
11. P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, “Machine recognition of human activities: A survey,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488 (2008).
12. K. Vellenga, H. J. Steinhauer, A. Karlsson, G. Falkman, A. Rhodin, and A. C. Koppisetty, “Driver intention recognition: state-of-the-art review,” *IEEE Open J. Intell. Transp. Syst.* (2022).
13. J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308 (2017).

14. W. Tao, M. C. Leu, and Z. Yin, “Multi-modal recognition of worker activity for human-centered intelligent manufacturing,” *Eng. Appl. Artif. Intell.*, vol. 95, p. 103868 (2020).
15. T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz, and G. Troster, “Combining motion sensors and ultrasonic hands tracking for continuous activity recognition in a maintenance scenario,” in *2006 10th IEEE international symposium on wearable computers*, pp. 97–104 (2006).
16. T. Stiefmeier, D. Roggen, G. Ogris, P. Lukowicz, and G. Tröster, “Wearable activity tracking in car manufacturing,” *IEEE Pervasive Comput.*, vol. 7, no. 2, pp. 42–50 (2008).
17. H. Koskimaki, V. Huikari, P. Siirtola, P. Laurinen, and J. Roning, “Activity recognition using a wrist-worn inertial measurement unit: A case study for industrial assembly lines,” in *2009 17th mediterranean conference on control and automation*, pp. 401–405 (2009).
18. T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka, “Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1088–1099 (2016).
19. C. Zhu, W. Sun, and W. Sheng, “Wearable sensors based human intention recognition in smart assisted living systems,” in *2008 International Conference on Information and Automation*, pp. 954–959 (2008).
20. L. Zhu *et al.*, “A novel motion intention recognition approach for soft exoskeleton via IMU,” *Electronics*, vol. 9, no. 12, p. 2176 (2020).
21. B. Sun, G. Cheng, Q. Dai, T. Chen, W. Liu, and X. Xu, “Human motion intention recognition based on EMG signal and angle signal,” *Cogn. Comput. Syst.*, vol. 3, no. 1, pp. 37–47 (2021).
22. M. Wen and Y. Wang, “Multimodal sensor motion intention recognition based on three-dimensional convolutional neural network algorithm,” *Comput. Intell. Neurosci.*, vol. 2021 (2021).
23. S. Masoud, B. Chowdhury, Y.J. Son,C. Kubota, and R. Tronstad, "A dynamic modelling framework for human hand gesture task recognition," arXiv preprint arXiv:1911.03923 (2019).
24. C.H. Chen, W.Y. Lin, and M.Y. Lee, “The Applications of K-means Clustering and Dynamic Time Warping Average in Seismocardiography Template Generation,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1000–1007 (2020).
25. D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series.,” in *KDD workshop*, vol. 10, no. 16, pp. 359–370 (1994).
26. Y. Ida, E. Fujita, and T. Hirose, “Classification of volcano-seismic events using waveforms in the method of k-means clustering and dynamic time warping,” *J. Volcanol. Geotherm. Res.*, vol. 429, p. 107616 (2022).
27. S. Masoud, N. Mariscal, Y. Huang, and M. Zhu, “A Sensor-Based Data Driven Framework to Investigate PM 2.5 in the Greater Detroit Area,” *IEEE Sensors Journal*, Vol 21, no. 14, p.16192-16200 (2021).