ASME Journal of Computing and Information Science in Engineering
Online journal at:
https://asmedigitalcollection.asme.org/computingengineering

Xinyao Zhang
Department of Environmental Engineering Sciences,
University of Florida,
Gainesville, FL 32611
e-mail: xinyaozhang@ufl.edu

Sibo Tian
Department of Mechanical and Aerospace Engineering,
University at Buffalo,
Buffalo, NY 14260
e-mail: sibotian@buffalo.edu

Xiao Liang
Assistant Professor
Department of Civil, Structural, and Environmental Engineering,
University at Buffalo,
Buffalo, NY 14260
e-mail: liangx@buffalo.edu

Minghui Zheng
Associate Professor
Department of Mechanical and Aerospace Engineering,
University at Buffalo,
Buffalo, NY 14260
e-mail: mhzheng@buffalo.edu

Sara Behdad[1]
Associate Professor
Department of Environmental Engineering Sciences,
University of Florida,
Gainesville, FL 32611
e-mail: sarabehdad@ufl.edu

Early Prediction of Human Intention for Human–Robot Collaboration Using Transformer Network

Human intention prediction plays a critical role in human–robot collaboration, as it helps robots improve efficiency and safety by accurately anticipating human intentions and proactively assisting with tasks. While current applications often focus on predicting intent once human action is completed, recognizing human intent in advance has received less attention. This study aims to equip robots with the capability to forecast human intent before completing an action, i.e., early intent prediction. To achieve this objective, we first extract features from human motion trajectories by analyzing changes in human joint distances. These features are then utilized in a Hidden Markov Model (HMM) to determine the state transition times from uncertain intent to certain intent. Second, we propose two models including a Transformer and a Bi-LSTM for classifying motion intentions. Then, we design a human–robot collaboration experiment in which the operator reaches multiple targets while the robot moves continuously following a predetermined path. The data collected through the experiment were divided into two groups: full-length data and partial data before state transitions detected by the HMM. Finally, the effectiveness of the suggested framework for predicting intentions is assessed using two different datasets, particularly in a scenario when motion trajectories are similar but underlying intentions vary. The results indicate that using partial data prior to the motion completion yields better accuracy compared to using full-length data. Specifically, the transformer model exhibits a 2% improvement in accuracy, while the Bi-LSTM model demonstrates a 6% increase in accuracy. [DOI: 10.1115/1.4064258]

Keywords: human intent recognition, early prediction, transformer, hidden Markov model, human–robot collaboration, manufacturing, artificial intelligence, manufacturing automation

1 Introduction

In recent years, human–robot collaboration (HRC) has gained increasing popularity for common co-assembly tasks in manufacturing settings. One widely used application involves humans retrieving components and placing them, followed by robots picking up the placed components and assembling them into a final product [1]. Moreover, in the quest for efficient mechanical assembly, robots play an essential role in the manufacturing planning process [2]. However, although humans and robots work together, they often are treated as independent agents. This is because humans can exhibit more flexibility in their actions, while robots are typically programmed for fixed automation modes. In addition, humans possess the ability to perceive the actions of others and infer their intentions which makes them capable of initiating relevant complementary actions. Equipping robots with such capabilities proves to be challenging. Therefore, there is a need for a higher level of understanding of human intent and enabling robots to rapidly adapt accordingly.

Unlike other physical features, such as location coordinates or distance traveled, human intent is implicitly contextual and not directly observable. However, it is encoded and expressed through human actions [3]. Specifically, the movement and orientation of workers have a significant impact on the recognition of intent in a warehouse [4]. Observing and interpreting abundant information embedded in human actions can be beneficial for understanding human intent. Recent research has proposed new approaches to cooperation between humans and robots by using the recognition of human intent in robotic control and process planning [5]. For instance, the prediction of the sequence of assembly activities relies on modeling human motion to recognize intent [6]. Another application in the assembly process involves measuring quality assurance and detecting human failure through the recognition of intent [7].

[1]Corresponding author.

Manuscript received July 11, 2023; final manuscript received December 6, 2023; published online January 8, 2024. Assoc. Editor: Caterina Rizzi.

Inspired by the necessity of intent recognition and the legibility of actions, our research is driven by the goal of achieving explicit human intent recognition. Leveraging advancements in deep learning, state-of-the-art algorithms show great promise in providing intelligent solutions [8]. To name a few methods, convolutional recurrent neural networks have been effectively used to learn the temporal and spatial relationships embedded in human body actions [9]. Also, recursive Bayesian filtering methods have been used to explore the correlation between intent and non-verbal behavior [10].

However, despite existing case studies that assess human intention recognition, the importance of early prediction has been overlooked. To overcome this gap, we aim to design an intention recognition framework by using motion data, as shown in Fig. 1. To predict human intentions, human motion data are fed into a deep learning model. Moreover, to achieve early prediction, a Hidden Markov Model (HMM) is used to find state transitions. The data before the state transitions are used to accomplish early prediction.

We propose a framework for motion-based human intention recognition. In terms of model selection, we employ two types of architectures: Transformer and Bidirectional Long short-term memory (Bi-LSTM). The Transformer architecture is selected due to its capability to capture important information by calculating attention values and assigning weights to sequences. Bi-LSTM architecture is chosen as it is capable of learning long-range dependencies in the inputs in both forward and backward directions.

Besides intent recognition, we suggest the concept of earlier prediction, which involves predicting intent before the movement is complete. We extract features from joint distances by leveraging human motion trajectories. Recognizing that HMM models possess the capability of continuous action division and unsupervised learning, we incorporate joint distance data into an HMM to identify the point in time when intention transition occurs, i.e., uncertainty to certainty. Data from the onset of motion to the time of state transition can be used for early prediction of human intentions.

We design two experimental cases: one where the intentions of reaching two adjacent targets are grouped into one class, and another where the intentions of reaching two adjacent targets are grouped into separate classes. The latter case poses increased difficulty in intention recognition as the motion trajectories are similar, but the intentions are completely different.

In this study, human intent is defined as the judgment of the operator's goal based on the observed trajectory of reaching movements. The operator's arm motion, captured by the Vicon system, serves as the input to the model, which subsequently predicts the intent based on the motion. We provide a detailed comparison of the performance of the Transformer and Bi-LSTM models in both cases and offer recommendations for model selection based on task specificity. In addition, we use the HMM to compute state transitions for each reaching trajectory and evaluate the performance of early predictions compared to predictions using full-length data.

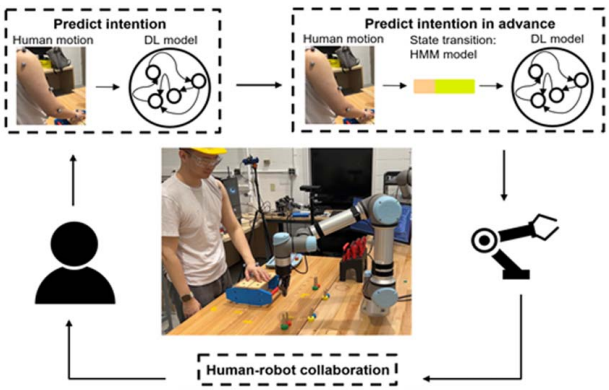

Fig. 1 Intention recognition and early prediction framework

The remainder of the paper is structured as follows: Sec. 2 compares related studies on intent recognition. Section 3 describes the proposed Transformer and Bi-LSTM architectures, and Sec. 4 presents the experimental design, the dataset, and practical results. Finally, Sec. 5 concludes the paper and outlines potential future work.

2 Related Work

In this section, we summarize related work and introduce the importance of intentional learning, and its perceptual, predictive, and state transition methods.

2.1 Importance of Intention Learning. Intention prediction from object trajectory has been an active area of research in different domains such as vehicle driving [11], pedestrian intention prediction [12], aerial targets [13], and human–robot collaborations [14]. As the concept of autonomous vehicles and robotics is emerging, the need for accurate prediction of intention and specifically early intention prediction becomes essential.

Intention learning is essential for human–robot teaming. In team environments, coordination among team members relies on the ability to predict each other's intentions. While humans possess this knowledge, it remains a challenge to enable robots to accurately predict and adjust their actions accordingly. For example, in manufacturing environments, if collaborative robots are programmed in a fixed offline manner, it is labor-intensive to recode the corresponding unexpected collaborations that are likely to occur with a change in human intent [14,15]. On the other hand, considering situational needs, people have been shown to unconsciously adjust their behavior, such as movement speed and execution paths [16,17]. This situation has a high probability of happening in a manufacturing workplace where an operator has multiple trajectories of motion to pick up and place many tools or parts during assembly. Both the speed of movement and the path of movement are not stable, so predicting human intent is informative, and understanding it becomes crucial. In our study, intent learning mainly refers to understanding the intent behind human actions, especially in human–robot collaboration in manufacturing where operators' action patterns naturally follow task-specific intent.

2.2 Perception Methods. There are various approaches to allow robots to perceive human intentions in specific task scenarios. For example, by collecting electroencephalography (EEG) signals on a person's scalp, it is possible to understand the person's intentions, as the EEG signal fluctuates in different patterns when a person wants to move different parts of the body [9]. Similarly, surface electromyography signals can be used to estimate associated biomechanics motion by measuring the velocity or acceleration of muscles [18]. However, there are some limitations to collecting bioelectrical signals, as the collected signals contain too much noise, and the sensor equipment affects the flexibility of experimenters.

Using visual data to collect workers' movements in the complex manufacturing environment offers insight into the operations they perform. For example, spatial-temporal information from disassembling hard disks is captured using video recordings and processed with an unsupervised learning framework to recognize human activities [19]. Likewise, the actions performed by the operator when assembling a car engine are identified through the development of a neural network architecture [20]. Furthermore, the analysis of optical flow images serves as a supplementary source of temporal information to enhance the ability to predict movements in static images [21]. Nonetheless, processing visual data that contain motion information to extract intention is computationally intensive and requires a lot of manual labeling.

As an alternative motion detection technology, inertial sensors are capable of quantitatively characterizing human motion with less labeling effort. Mounting inertial sensors on various body frames makes it possible to collect complementary information such as angular velocity and magnetic fields [22]. The application of tracking systems like radio-frequency identification (RFID) tags is considered a powerful means of recording task-level human motion in manufacturing operations [23]. Additionally, researchers have used the Vicon system due to its portability freedom from experimental constraints and its ability to perform 3D trajectory capture [24,25]. Therefore, in this study, we take advantage of the motion-tracking sensor system, the Vicon system, to infer intent directly from the motion trajectory data.

2.3 Intention Prediction Methods. The prediction of operator's intent is a key factor for achieving safety in the HRC system as it serves as a prerequisite for adjusting the robot's behavior accordingly. A wide range of machine learning and deep learning models have been developed for intentional learning. To name several examples, support vector machines, and random forest algorithms have been used for daily motion intent classification [26]. Implementing a nonlinear support vector machine model enables the projection of inertial tracking signals to 12 physical activities [27]. The Random Forest algorithm has been employed in which muscle signals have been used as inputs and three motor parameters are extracted to classify motions [28]. Also, Bayesian estimation has been utilized to estimate human stiffness parameters from force data to infer the intent behind human–robot collaborative actions [29].

Besides machine learning, deep learning models such as recurrent neural networks (RNNs), Gated Recurrent Units (GRU), and LSTM have been widely used in intent estimation. To name several examples, Nicolis et al. fed reaching trajectories into the RNNs to estimate the user's intent [30]. Mavsar et al. proposed an RNN consisting of an LSTM layer for inferring intentions from hand positions [31]. Maceira et al. inferred task intent by processing force signals using a fully connected RNN [32]. However, despite their specialization in modeling time-series data, RNNs often struggle to capture temporal dependencies. Another stream of literature used GRU networks, which can capture long-term dependencies by mitigating the vanishing gradient problem commonly observed in RNNs [33]. The ability of GRU to capture temporal dependencies has been validated in time-series classification tasks [33,34]. For instance, Liu et al. found that the addition of GRU enhances pedestrian intention prediction results compared to the same structure without GRU [35]. Moreover, LSTM-based structures have been used to learn linear and nonlinear features of motion sequences and overcome the weakness of time dependence [36,37]. To name several studies, Xin et al. proposed a lane intent recognizer based on an LSTM network to facilitate vehicle trajectory prediction [38]. In a similar application, Shi and Zhang combined a hierarchical over-sampling bagging method with LSTM to overcome the challenge of imbalanced datasets [39]. By designing a stacked LSTM network, Saleh et al. solved the time-dependent problem and achieved predictions 4 s in advance [40]. While LSTM networks have demonstrated performance in capturing temporal dependencies, they exhibit certain limitations. Specifically, LSTM networks suffer from a shallow structure that limits their ability to handle long-range sequential data. In addition, in the case of closely located items where motion sequences for reaching neighboring targets are highly similar, LSTM networks show difficulties as they emphasize only the dependence of inputs in the forward network stream in explicitly separating intentions.

To address these issues, we propose two corresponding models: a Transformer and a Bi-LSTM. Our motivation for applying a Transformer for intention classification in manufacturing stemmed from its successful application in predicting pedestrian intentions and trajectories [41]. Furthermore, Patterson and Falkman compared the results of Transformers with LSTM and showed that Transformer achieved higher accuracy in gaze-based intent recognition [42]. The attention mechanism within the Transformer provides

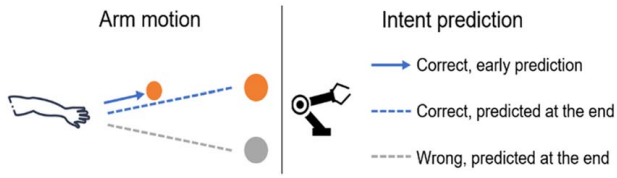

Fig. 2 Illustration of early prediction

an advantage in discerning unique information from similar datasets [43]. Besides Transformers, Bi-LSTM models have shown promising outcomes for intention prediction in HRC as shown by Gao et al. [44]. The use of Bi-LSTM architecture helps learn data context in both the forward and reverse network streams and facilitates the representation of long-range data dependencies [45]. In this study, we extend the Transformer and Bi-LSTM models for intent recognition in an HRC environment. Specifically, we evaluate the performance of these models in predicting pickup action intentions based on arm motion trajectories. In addition, we demonstrate the dynamic prediction capability of the models by gradually increasing the length of the motion trajectory.

2.4 Intention State Transition. In addition to the task of intent recognition, it has become important to predict the intent in advance and test whether it is necessary to use full-length sequences or partial motion sequences and is adequate to achieve accurate predictions. As depicted in Fig. 2, early prediction aids robots in recognizing human intent before the human reaches the target object without necessitating the use of full motion data.

Considering state transition is a critical information factor for early prediction of intent. State space models are a class of models that contain observable and hidden states to describe state transitions. Available methods include Kalman filters and Markov models. A Kalman filter is a mathematical tool used for estimating the state of a moving object, particularly when the object follows a linear system of motion [46]. However, arm movements change dynamically over time, and observations are not necessarily ordered in time. For nonlinear motion, a Markov chain was applied to discover the transitions that perform goal changes [47]. The use of Markov chains requires that the state be directly observable, but it is not practical in an HRC setting as a robotic agent can only observe motion, not directly infer states.

To handle the division of hidden states, HMM-based algorithms are considered unsupervised model-driven methods for learning the correlation of the states of time-series data [48]. HMMs are widely used in the domain of estimating driving behavior. They take the input motion features and then determine the behavior with the highest probability [49,50]. In the field of human–robot interaction, Peddi et al. used an HMM model to calculate the probability of a human crossing a robot's path [51]. Zhang et al. applied HMM to realize the state division of disassembly activities [19]. In a study to realize robots' understanding of human intentions, Kelley et al. not only included hidden states in the HMM but also incorporated visible states that encode changes in position or angular motion [52]. In addition to the basic HMM model, extensions of the HMM are also well known, one example being the autoregressive HMM (ARHMM) [53]. Unlike HMMs where the current observation depends only on the current hidden state, ARHMM assumes that the current observation depends on the current hidden state as well as on previous observations. Using an autoregressive process, ARHMM can model interdependencies in a sequence [54]. While both HMM and ARHMM are tools for modeling sequences with hidden states, the choice between them should be based on the characteristics of the dataset. Therefore, in this study, we first perform feature extraction on the motion data and then use both HMM and ARHMM to model observations and evaluate their performance in state transition division.

3 Methodology

The proposed method consists of utilizing Transformer and Bi-LSTM models to learn the relationship between human intentions and motion trajectories. Furthermore, HMM is used to identify the transition from uncertain intent to certain intent and is employed to determine the appropriate length of sequence data that are fed as input to the models.

3.1 Transformer Model. The Transformer was initially introduced for its innovative utilization of the attention mechanism [55]. The attention mechanism allows modeling sequential dependencies regardless of their position in the input or output. Referring to this capability for trajectory modeling, the attention mechanism can simultaneously observe all inputs and assign weights to these observations, rather than processing the data sequentially. By employing an attention mechanism, the architecture learns features in long time-series and computes correlations.

The Transformer architecture consists of two modules: the encoder and the decoder. Each module includes three blocks: a feedforward fully connected block, a multi-head attention block, and two residual connections following each of the aforementioned blocks [56]. In contrast to tasks like language translation that necessitate a decoder module to generate output sequences, human intent recognition can be achieved by employing only the encoder module and replacing the decoder with a probabilistic classifier [57].

The proposed Transformer encoder model is displayed in Fig. 3. First, the sequence of trajectories of the arm movements collectively forms a matrix $\chi \in \mathbb{R}^{l\times d}$, where l is the length of the sequence and d is the feature dimension. The matrix is then normalized before being input into the attention mechanism. Within the attention mechanism, an element of the input matrix is represented as a query (Q) vector, while the remaining elements are referred to as key (K) vectors. The output of assigning weights to the sequence elements is termed value (V) vectors. The scale dot-product attention computes the attention value of each input element as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where $\sqrt{d_k}$ is the scale factor.

Second, the multi-head attention block is formed by combining multiple dot-product attention at various scales, represented as follows:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$MultiHead(Q, K, V) = Concat_i(head_i)W^O \tag{3}$$

where W_i and W^O are the learnable weight matrices, and *concat* denotes concatenating the results. The multi-head attention will parallelly compute and join the complex information of more representations at different positions of input data. Since no recurrence or convolution calculation is required, each input element is provided to the feedforward network along with the associated positional information. Finally, all embedded elements are passed through a normalization layer to speed up the learning, and then, a classifier with a SoftMax activation function is used to determine the intent class.

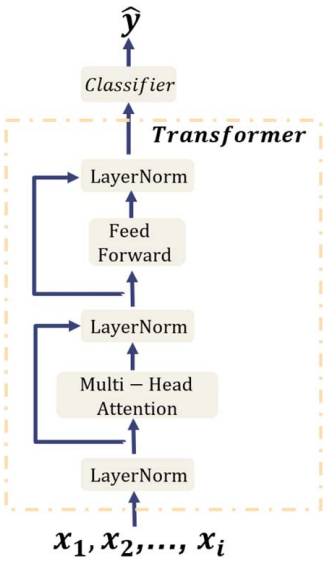

Fig. 3 The transformer model's structure

3.2 Bi-LSTM Model. Before the emergence of Transformers, LSTM architectures were commonly used to learn dependencies in sequential data. The results of the previous study demonstrate that an LSTM network is capable of learning temporal features and accurately recognizing human activities [58]. Nevertheless, a unidirectional LSTM only learns data structures in a fixed direction, i.e., after starting from the motion, but this is not sufficient to distinguish between highly similar data.

In contrast to the unidirectional LSTM, our approach consists of a Bi-LSTM architecture for the processing of motion steps. This architecture operates in both the forward and backward contexts to provide a more comprehensive analysis of the data. This bidirectional approach further enhances performance in capturing the correlation between motion and human intent.

As illustrated in Fig. 4(*a*), the LSTM cells are stacked bidirectionally, meaning there are two directions for processing the input motion sequences, each with its time-steps and features. The forward layer processes the sequence in the standard order (past to future), whereas the backward layer processes it in the reverse order (future to past). After processing the input through the Bi-LSTM network, the output is then passed through a SoftMax activation layer, which computes the probability distribution over each intent class for the input sequence. In detail, Fig. 4(*b*)

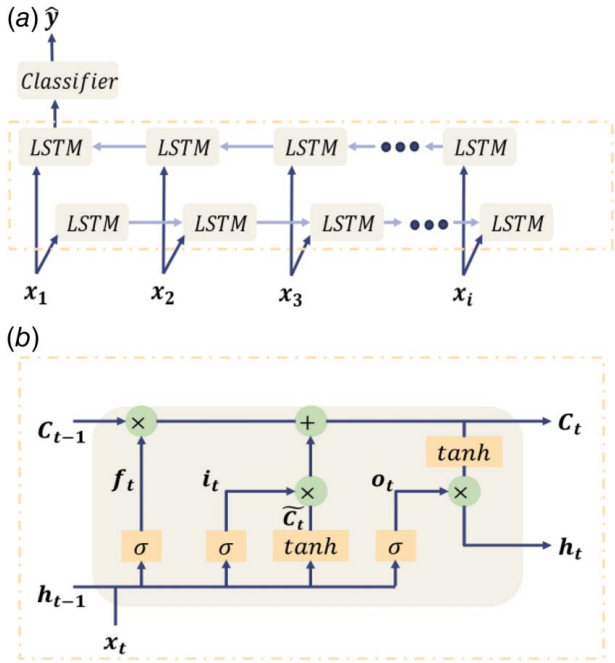

Fig. 4 (*a*) The BI-LSTM model's structure and (*b*) The LSTM cell workflow

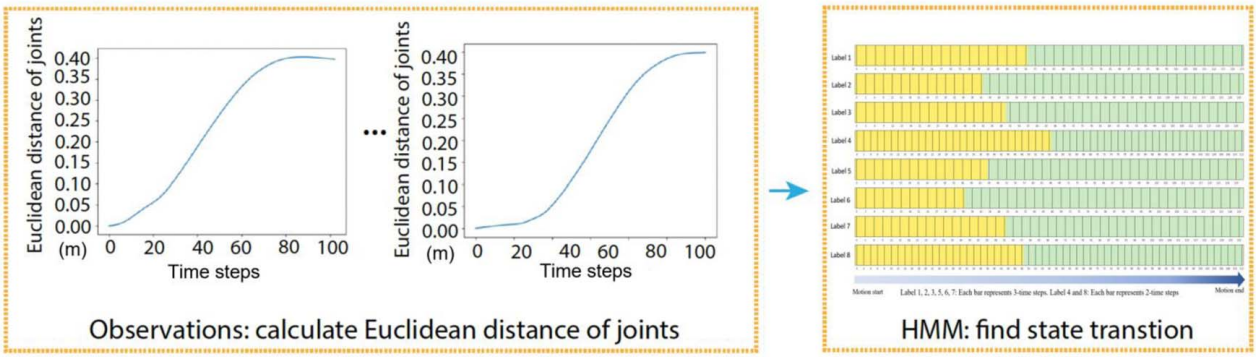

Fig. 5 The process of identifying state transition using the hidden Markov model: Calculating Euclidean distances of joints and putting it into the HMM. The HMM results plot is the same as Fig. 12.

illustrates how each LSTM cell performs operations. Equations (4)–(8) are mathematically interpreted as follows:

$$f_t = \sigma(w_{fx}x_t + w_{fh}h_{t-1} + b_f) \tag{4}$$

$$i_t = \sigma(w_{ix}x_t + w_{ih}h_{t-1} + b_i) \tag{5}$$

$$o_t = \sigma(w_{ox}x_t + w_{oh}h_{t-1} + b_o) \tag{6}$$

$$c_t = c_{t-1} \odot f_t + i_t \odot \tanh(w_{ct}x_t + w_{ch}h_{t-1} + b_c) \tag{7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{8}$$

where f_t, i_t, and o_t are namely the forget gate, input gate, and output gate. x_t and h_t are input elements and hidden states. $\odot$ represents element-wise vector multiplication.

3.3 Select a Hidden Markov Model for State Transition. In addition to deep learning-based intention learning methods, we have applied an HMM to perform state transitions. Our objective is to show the effectiveness of HMM in partitioning states without the need for explicit supervision. We extend the utility of HMM to motion trajectory analysis, where they are employed to separate trajectories into two discernible states that emphasize the identification of intention states. The application of HMM for the early prediction of intention remains relatively unexplored within the existing literature. Moreover, we carefully finetune the parameters of both basic HMM and ARHMM models to make them well-suited for this problem domain.

For a given input sequence, an HMM-based algorithm can model the data as different states by measuring the likelihood of observations and hidden states. In practice, an HMM-based algorithm can separate motion trajectories into states where the intention is uncertain and those where the intention is certain. Uncertainty of intention means that the experimenter has no clear intention at the beginning of the action, or the motion changes to a small extent. And toward the end of the motion, intentions will gradually become clear. Thus, the implied state of intention regarding the behavior shifts from uncertainty to clarity. To achieve it, we first calculate the Euclidean distance variation of the joint with motion, i.e., the joint position at each time point minus the joint position at the beginning. As shown in Fig. 5, the motion starts at a slow speed but gradually moves away from the original position. Second, these Euclidean distances about human joints are the observation variables of an HMM model. Further, we set the number of hidden states in an HMM to two. An HMM will classify the sequences into two continuous states based on the distance. We fit HMM parameters using an expectation–maximization (EM) algorithm to ensure that it accurately captures state transitions [59]. The basic HMM and ARHMM have the same inputs and learning process, and we will show the results of their division in Sec. 4.4. Last, we ultimately care about the time of the state transition since we extract the length of the data from the beginning of the motion to the state transition as input to the intention classification model for early prediction.

4 Experimental Studies

To show the application of the proposed methods, we design an experiment within a collaborative human–robot environment in manufacturing. In this experiment, eight screws have been positioned in four distinct locations in pairs as shown in Fig. 6. We aim to access the accuracy of two models in predicting the operator's intentions for four locations (four intent labels) as well as their accuracy in predicting the operator's intentions for individual screws (eight intent labels). We also validate the concept of early prediction of human intentions, which implies the ability to recognize intentions before the completion of the corresponding movement. Early intention prediction helps robots understand human intent faster to provide timely and proactive assistance.

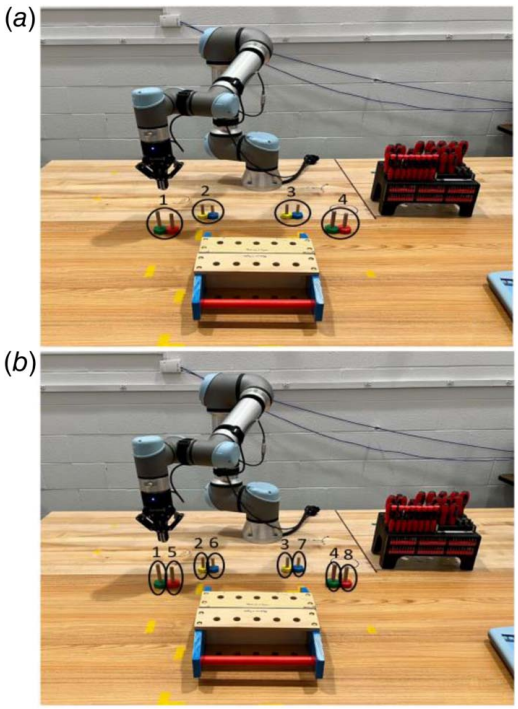

Fig. 6 The experiment with (*a*) four-label intentions and (*b*) eight-label intentions

4.1 Experimental Design and Dataset. The human–robot collaboration setup includes a robotic manipulator that shares a common workspace with an operator and executes a predetermined path. The human operator stands facing the robot and moves to four different locations to pick up targets and place them in a collection box. Simultaneously, the manipulator moves back and forth between these four target locations and the collection box.

Each location contains two adjacent screws. As a result, the reaching motions for two screws at the same location are similar, but the human operator's intent is different. Therefore, we have two objectives. One is to predict the target location that the human operator is reaching for among four distinct areas displayed in Fig. 6(*a*), while the other objective is to predict which screw the human wants to retrieve among the eight screws displayed in Fig. 6(*b*).

To collect the data, a Vicon motion capture system is used to track the movement of the human operator's right arm. Two markers are attached to each side of the wrist, elbow, and shoulder of the participant. The data are recorded as a sequence of Cartesian coordinates for each marker, at a frequency of 50 Hz, resulting in a trajectory time interval of 0.02 s. The center of each rotation joint can be easily estimated by taking the mean of the two markers' positions.

To show the similarity of trajectories for neighboring targets, two sets of trajectories from the beginning to the end, starting from left to right, have been visualized in Fig. 7, as examples. We can observe that the motion trajectories are separate at the beginning and progressively become similar over time, especially at the end of the motion, when the trajectories almost overlap. This increases the complexity of predicting intentions on very similar trajectories.

4.2 Results of Intent Classification With Four Labels. We have 232 motion data in total with each eight labels having an equal amount of data. The length of each single motion data is approximately 2–3 s. The proposed models were built using Keras and TensorFlow.

Adam optimizer with a fixed learning rate of 0.001 was applied. To ensure the reproducibility of results, we fixed the values of random seeds to a total of five seeds. The training epoch was set to 500 epochs. The experiment was carried out using a single Nvidia 3080 GPU, with 70% of the data for training and 30% for testing.

To compare the classification performance of different models, boxplots are used to show outliers as well as the distribution of the results. In addition, heatmaps are used to visualize the classification output for different intentions. We trained and predicted models by consequently increasing the data length from 20% to 100% and using 20% as an interval.

As seen in Fig. 8, a slight decline in classification accuracy was observed for the Bi-LSTM model when the data length was increased 100%, while the accuracy of the Transformer model continued to improve with increasing time-steps. The Bi-LSTM's decline in performance could be primarily attributed to overfitting, which occurs when a model begins to memorize the training data rather than generalize it. The Bi-LSTM model can capture long-term dependencies and nuances in the data. However, when all the time-steps are fed into the model, especially as the input dimensions increase, the model starts to overfit due to fewer categories.

The Transformer architecture, due to its self-attention mechanism, alternatively has the ability to assign varying degrees of importance as the number of time-steps increases within a sequence. Consequently, the Transformer can dynamically adjust the attention weights and perform optimally when processing data of full temporal length. As a result, the Transformer outperforms Bi-LSTM when using 100% length trajectory data to make predictions.

In addition, when using full-length trajectories, we used a heat map and compared the classification results, as illustrated in Fig. 9. The Transformer is 100% accurate, while Bi-LSTM incorrectly predicts label 2 as label 1 due to the close location of the two objects. Depending on the results of our experiments, we recommend prioritizing the use of the Transformer model if neighboring targets are labeled as a group. Hence, Transformers yield better results when classifying intentions with significant action differences (e.g., four distinct labels).

4.3 Results of Intent Classification With Eight Labels. As previously mentioned, it is crucial to analyze the intentions when the targets are near each other. Especially in manufacturing sites, many tools or parts needed during operation are often placed together. Apart from that, training models with a dataset of eight labels increases the computational time and complexity. This drives us to evaluate the robustness of the models by dealing with intention recognition with similar trajectories.

The performance of the Bi-LSTM model exhibits dynamic variations in accuracy when processing data of different temporal lengths of data, as depicted in Fig. 10. When the temporal length approaches 80%, a slight decline in accuracy is observed, accompanied by a broader distribution in results. Referring to

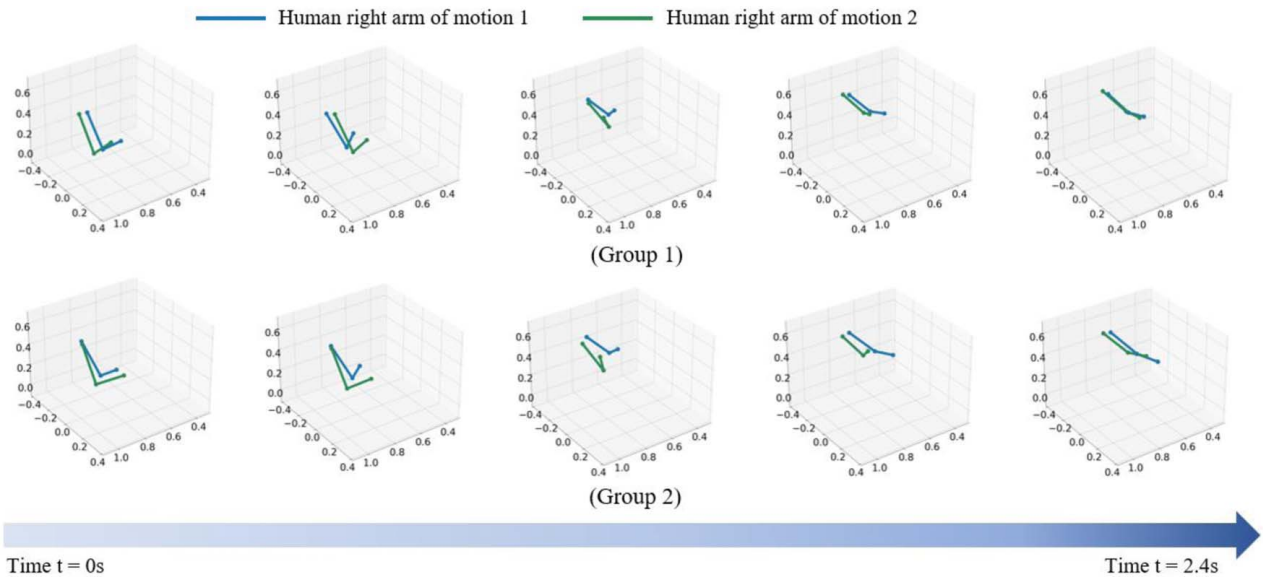

Fig. 7 Visualization of the trajectory of two approaching targets

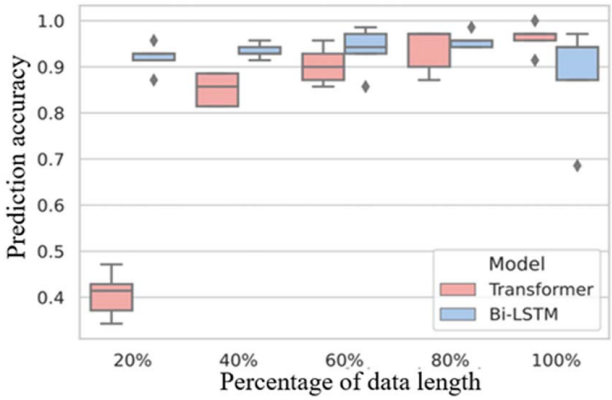

Fig. 8 Four-label classification results of different models on test data

Fig. 7, motions 1 and 2 exhibit distinct trajectories in the earlier parts of their sequences. The Bi-LSTM model competently leverages this distinction when processing shorter temporal lengths. However, as the sequence progresses and these initial discriminative features account for a smaller proportion of the entire sequence, the model might struggle to maintain the same level of classification accuracy.

Unlike recurrent models, the Transformer processes all time-steps in parallel, without including sequential bias. This means that for sequences with longer temporal lengths, the Transformer can continue to extract meaningful information without being overwhelmed. As seen in Fig. 10, the Transformer model achieves the best performance when full-length data are used. On the other hand, the Transformer's self-attention mechanism demands more data for optimal learning due to its complexity. This could explain the Transformer's lower performance compared to Bi-LSTM, especially since eight-label categorization results in fewer motion sequences per category.

To conclude, Bi-LSTM architecture proves to be a superior choice for intention prediction when dealing with similar trajectories, such as with eight labels. Essentially, the dynamic performance of these two models emphasizes the significance of selecting the appropriate temporal length for motion data and suggests the potential for implementing early prediction.

The intentions of two targets nearby are easily misclassified in both models, e.g., intentions labeled 7 are confused with label 3, as shown in Fig. 11. Nonetheless, Bi-LSTM is more robust in intention prediction for closely located object;, e.g., label 2 is correctly recognized by Bi-LSTM but is partially misidentified as label 6 by Transformer.

4.4 Trajectory State Transition Results From Hidden Markov Model. After conducting predictions using data of different lengths, we found that training with longer data lengths, i.e., complete data lengths, to achieve higher accuracy is not always effective. Therefore, the use of an HMM is necessary to help us find the best length series to achieve better accuracy as well as earlier predictions.

We aim to evaluate the performance of the basic HMM and ARHMM and select the most appropriate one for the early prediction framework. Since dividing eight labels is a more complex case study, our focus is on the eight-label case. Figure 12 shows the average state transitions of the basic HMM for each label, and Fig. 13 shows the results of the ARHMM transitions. In both plots, each status bar represents three consequent time-steps except for label 4 and label 8 where each bar represents two time-steps. Each bar is displayed in yellow to indicate an uncertain intent, and in green to indicate a certain intent.

In our comparative analysis of the segmentation results, the HMM consistently segmented the observations and produced clear and coherent segmentation results. In contrast, the ARHMM produced segmentation results that showed irregular patterns which make them less intuitive and more challenging to interpret. The main reason for the superior performance of the HMM in our dataset is that the data do not exhibit a strong time dependence in the observations outside the hidden state. Introducing an autoregressive component to the ARHMM may add unnecessary complexity and lead to overfitting or misinterpretation. Subsequently, we tend to use HMM in early prediction due to the consistency and interpretability of its segmentation results.

In addition to presenting the HMM results in the time domain, we also transfer them to the distance domain. First, the Euclidean distance between each pair of neighboring time points about the wrist is calculated. Then, based on the principle of using straight-line distances as an approximate representation of curved distances, we use the Euclidean distance as an approximation of the physical distance between these two neighboring points. Adding up the distances obtained for each two points along the trajectory is the complete physical distance of the entire trajectory. Finally, we divide the distances according to the HMM transformation results, as shown in Fig. 14. In this figure, the blue part indicates the distance of uncertain intent, and the orange color indicates the distance of wrist movement when the intent is determined. We should clarify that in this study we just used the HMM results in the time domain and did not use the distance domain as inputs to the prediction

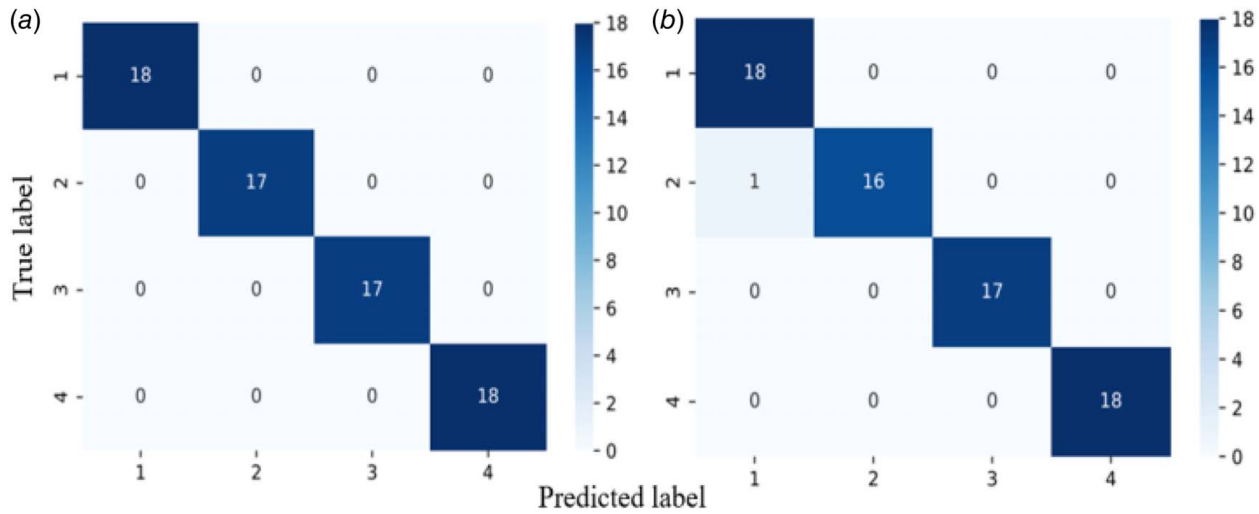

Fig. 9 Heatmap results for four-label case: (*a*) transformer and (*b*) Bi-LSTM

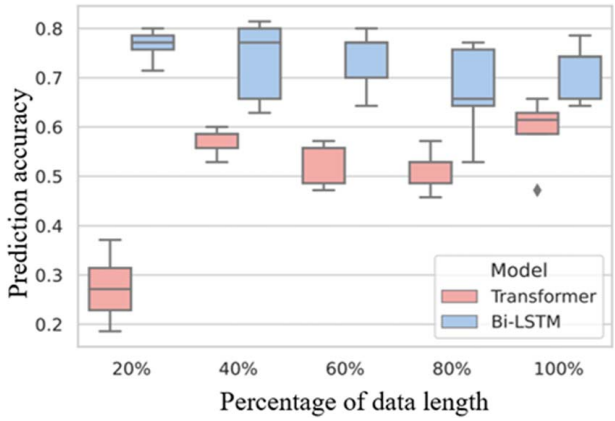

Fig. 10 Eight-label classification results of different models on test data

models, However, we believe that such a distance interpretation can be used in practical human–robot collaboration in the design of working positions or task assignment for both humans and robots.

However, it is worth noting that the state transition results associated with the HMM affect the performance of the intent classification models. The choice of state transition method depends on the specific dataset characteristics.

4.5 Intent Early Prediction With Eight Labels. To implement the concept of early intent prediction, we plan to validate the two models through experiments with eight labels. First, we prepare the data whose sequence length is the length from the start point to the time transition point determined by the HMM. Second, we train the prepared dataset with both models and compare the results with the model's performance on the full-length sequences dataset.

A summary of the comparison is illustrated in Fig. 15. For both models, the length of the data elements calculated using the HMM achieved better prediction accuracy compared to using all data elements. Looking at the results of individual models and comparing the values of the highest accuracy, the accuracy of the Transformer model was improved by 2% and the accuracy of the Bi-LSTM model was improved by 6%. While the Transformer's optimal accuracy only saw a slight enhancement, its early predictive model exhibits greater stability and a narrower range of accuracy when compared to using full-length data. In addition, the overall performance of the Bi-LSTM model surpasses that of the Transformer model because it is better suited for analyzing data sequences with high similarities, as discussed in Sec. 4.3.

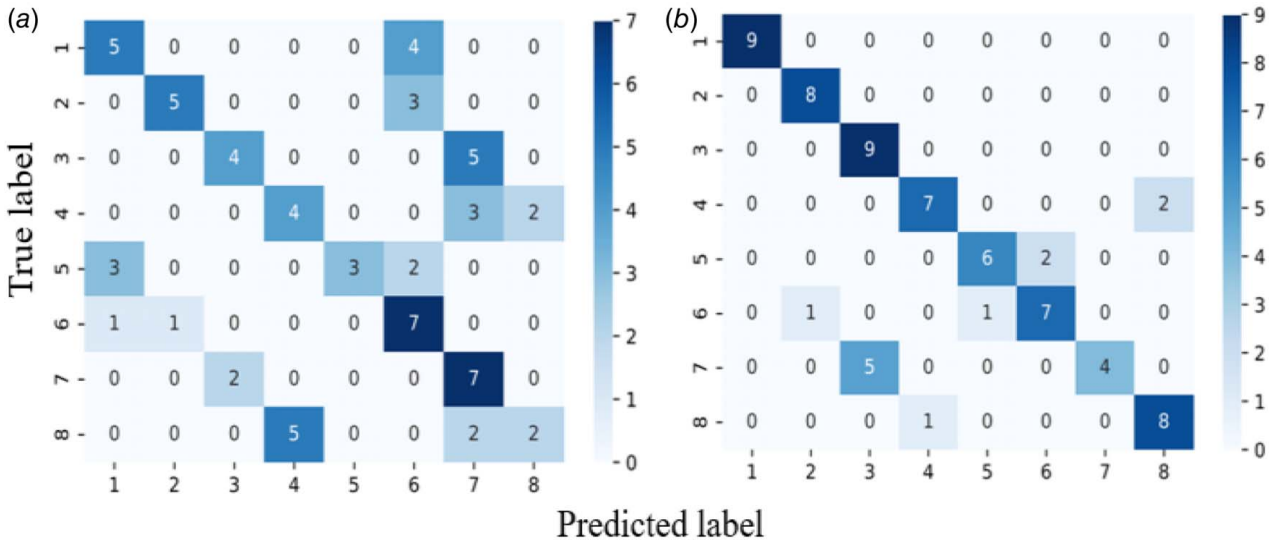

Fig. 11 Heatmap results for eight-label case: (*a*) transformer and (*b*) Bi-LSTM

Fig. 12 Average state transition results of the HMM in a time domain

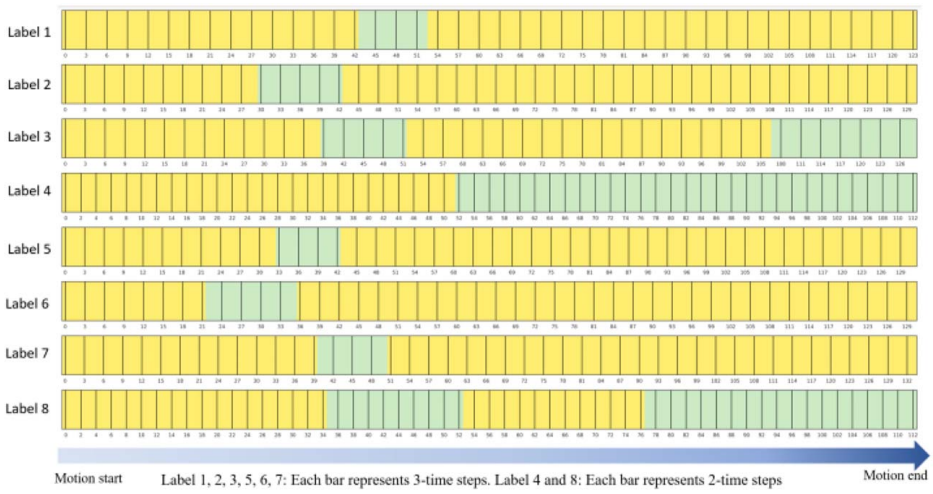

Fig. 13 Average state transition results of the ARHMM in a time domain

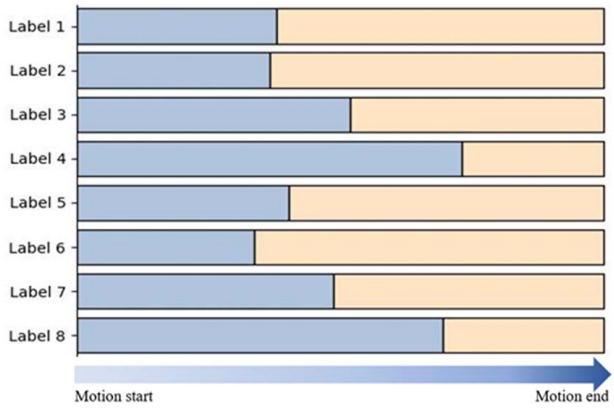

Fig. 14 Average state transition results in a distance domain for different intents

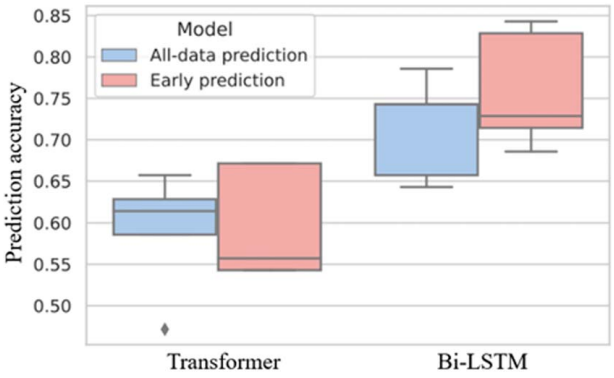

Fig. 15 Comparison of predictions using early data with all-data predictions in eight-label classification

5 Conclusion and Future Work

In this study, we propose a framework for intent prediction based on human movement data. The framework includes the use of Transformer and Bi-LSTM models to learn motion data and HMM to determine the intention shifts. Our experimental study reveals that the Transformer architecture yields better results in classifying intentions with significant action differences, whereas the Bi-LSTM architecture demonstrates greater robustness in identifying similar actions. Furthermore, leveraging the state transition information from HMM leads to early prediction and higher accuracy compared to using full-length data. Combining an HMM with classification models allows for early prediction of intent before completing the action. We assess the suggested framework within a human–robot collaboration context, with a focus on identifying intent when picking up targets in a manufacturing environment.

The proposed work enhances human–robot collaboration in multiple ways. First, by accurately predicting human intent, robots can anticipate future actions and provide timely assistance, thereby reducing time and improving overall efficiency. Second, the recognition of human intent enables robots to identify hazardous situations, fostering the creation of safe work environments.

The proposed framework holds potential for several extensions. First, for data with similar motion trajectories, we can utilize deep feature extraction techniques to achieve a prediction accuracy exceeding 90%. This becomes especially important when trajectories are similar, yet the underlying intention differs substantially. Second, in scenarios involving non-sequential or coordinated tasks, it is essential to explore how dynamic interactions between humans and robots affect the recognition of human intent. Finally, while the current frameworks primarily rely on joint movements to predict human intent, it is worthwhile to investigate how small-scale movements at the wrist and finger levels can be utilized for learning and predicting human intent.

To provide further guidance for adapting the framework to complex applications, we recommend conducting a comprehensive hyperparameter search. Hyperparameters, particularly in models utilizing attention mechanisms and LSTM cells, play a pivotal role in determining the performance of the model in diverse industrial scenarios. After a global optimization, domain-specific fine-tuning can be conducted, where the model is further refined based on specific industrial scenarios' data. In addition, to enrich the generalizability of the proposed framework, data augmentation techniques can be tailored for motion trajectories, which will artificially expand the current dataset. Meanwhile, a transfer learning approach can be leveraged, bridging the gap between proprietary and public datasets and addressing potential over-specialization.

Acknowledgment

This material is based upon work supported by the National Science Foundation–USA under Grant Nos. 2026276 and 2026533. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

References

[1] Kaipa, K. N., Morato, C. W., and Gupta, S. K., 2018, "Design of Hybrid Cells to Facilitate Safe and Efficient Human–Robot Collaboration During Assembly Operations," ASME J. Comput. Inf. Sci. Eng., **18**(3), p. 031004.

[2] Bhatt, P. M., Kulkarni, A., Malhan, R. K., Shah, B. C., Yoon, Y. J., and Gupta, S. K., 2021, "Automated Planning for Robotic Multi-Resolution Additive Manufacturing," ASME J. Comput. Inf. Sci. Eng., **22**(2), p. 021006.

[3] Stulp, F., Grizou, J., Busch, B., and Lopes, M., 2015, "Facilitating Intention Prediction for Humans by Optimizing Robot Motions," Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, Sept. 28–Oct. 2, pp. 1249–1255.

[4] Petković, T., Puljiz, D., Marković, I., and Hein, B., 2019, "Human Intention Estimation Based on Hidden Markov Model Motion Validation for Safe Flexible Robotized Warehouses," Rob. Comput. Integr. Manuf., **57**, pp. 182–196.

[5] Losey, D. P., McDonald, C. G., Battaglia, E., and O'Malley, M. K., 2018, "A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human–Robot Interaction," ASME Appl. Mech. Rev., **70**(1).

[6] Manns, M., Tuli, T. B., and Schreiber, F., 2021, "Identifying Human Intention During Assembly Operations Using Wearable Motion Capturing Systems Including Eye Focus," Proc. CIRP, **104**, pp. 924–929.

[7] Gajjar, N. K., Rekik, K., Kanso, A., and Müller, R., 2022, "Human Intention and Workspace Recognition for Collaborative Assembly," IFAC-PapersOnLine, **55**(10), pp. 365–370.

[8] Nahavandi, S., 2019, "Industry 5.0-A Human-Centric Solution," Sustainability, **11**(16), p. 4371.

[9] Zhang, D., Yao, L., Chen, K., Wang, S., Chang, X., and Liu, Y., 2019, "Making Sense of Spatio-Temporal Preserving Representations for EEG-Based Human Intention Recognition," IEEE Trans. Cybern., **50**(7), pp. 3033–3044.

[10] Jain, S., Argall, B., Abilitylab, S. R., Jain, S., and Argall, B., 2019, "Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics," ACM Trans. Human-Rob. Interact. (THRI), **9**(1), pp. 1–23.

[11] Huang, H., Zeng, Z., Yao, D., Pei, X., and Zhang, Y., 2021, "Spatial-Temporal ConvLSTM for Vehicle Driving Intention Prediction," Tsinghua Sci. Technol., **27**(3), pp. 599–609.

[12] Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., and Özgüner, Ü., 2022, "Predicting Pedestrian Crossing Intention With Feature Fusion and Spatio-Temporal Attention," IEEE Trans. Intell. Vehicles, **7**(2), pp. 221–230.

[13] Zhou, T., Chen, M., Wang, Y., He, J., and Yang, C., 2020, "Information Entropy-Based Intention Prediction of Aerial Targets Under Uncertain and Incomplete Information," Entropy, **22**(3), p. 279.

[14] Wang, W., Li, R., Chen, Y., and Jia, Y., 2018, "Human Intention Prediction in Human-Robot Collaborative Tasks," Proceedings of the Companion of the 2018 ACM/IEEE International Conference on Human–Robot Interaction, Chicago, IL, Mar. 5–8, pp. 279–280.

[15] Wang, W., Li, R., Chen, Y., Sun, Y., and Jia, Y., 2022, "Predicting Human Intentions in Human–Robot Hand-Over Tasks Through Multimodal Learning," IEEE Trans. Autom. Sci. Eng., **19**(3), pp. 2339–2353.

[16] Koppenborg, M., Nickel, P., Naber, B., Lungfiel, A., and Huelke, M., 2017, "Effects of Movement Speed and Predictability in Human–Robot Collaboration," Human Factors Ergon. Manuf. Service Ind., **27**(4), pp. 197–209.

[17] Tabar, R. S., Lindkvist, L., Wärmefjord, K., and Söderberg, R., 2022, "Efficient Joining Sequence Variation Analysis of Stochastic Batch Assemblies," ASME J. Comput. Inf. Sci. Eng., **22**(4), p. 040905.

[18] Tahmid, S., Font-Llagunes, J. M., and Yang, J., 2023, "Upper Extremity Joint Torque Estimation Through an Electromyography-Driven Model," ASME J. Comput. Inf. Sci. Eng., **23**(3), p. 030901.

[19] Zhang, X., Yi, D., Behdad, S., and Saxena, S., 2023, "Unsupervised Human Activity Recognition Learning for Disassembly Tasks," IEEE Trans. Ind. Inform.,

[20] Wang, P., Liu, H., Wang, L., and Gao, R. X., 2018, "Deep Learning-Based Human Motion Recognition for Predictive Context-Aware Human-Robot Collaboration," CIRP Ann., **67**(1), pp. 17–20.

[21] Xiong, Q., Zhang, J., Wang, P., Liu, D., and Gao, R. X., 2020, "Transferable Two-Stream Convolutional Neural Network for Human Action Recognition," J. Manuf. Syst., **56**, pp. 605–614.

[22] Digo, E., Pastorelli, S., and Gastaldi, L., 2022, "A Narrative Review on Wearable Inertial Sensors for Human Motion Tracking in Industrial Scenarios," Robotics, **11**(6), p. 138.

[23] Liu, H., and Wang, L., 2017, "Human Motion Prediction for Human-Robot Collaboration," J. Manuf. Syst., **44**, pp. 287–294.

[24] Schlagenhauf, F., Sreeram, S., and Singhose, W., 2018, "Comparison of Kinect and Vicon Motion Capture of Upper-Body Joint Angle Tracking," Proceedings of the 2018 IEEE 14th International Conference on Control and Automation, Anchorage, AK, June 12–15, pp. 674–679.

[25] Tian, S., Liang, X., and Zheng, M., 2023, "An Optimization-Based Human Behavior Modeling and Prediction for Human-Robot Collaborative Disassembly," Proceedings of the American Control Conference (ACC), San Diego, CA, May 31–June 2, pp. 3356–3361.

[26] Vrigkas, M., Nikou, C., and Kakadiaris, I. A., 2015, "A Review of Human Activity Recognition Methods," Front. Rob. AI, **2**, p. 28.

[27] Attal, F., Mohammed, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., and Amirat, Y., 2015, "Physical Human Activity Recognition Using Wearable Sensors," Sensors, **15**(12), pp. 31314–31338.

[28] Vu, C. C., and Kim, J., 2018, "Human Motion Recognition by Textile Sensors Based on Machine Learning Algorithms," Sensors, **18**(9), p. 3109.

[29] Yu, X., He, W., Li, Y., Xue, C., Li, J., Zou, J., and Yang, C., 2019, "Bayesian Estimation of Human Impedance and Motion Intention for Human–Robot Collaboration," IEEE Trans. Cybern., **51**(4), pp. 1822–1834.

[30] Nicolis, D., Zanchettin, A. M., and Rocco, P., 2018, "Human Intention Estimation Based on Neural Networks for Enhanced Collaboration With Robots," Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, Oct. 1–5, pp. 1326–1333.

[31] Mavsar, M., Deniša, M., Nemec, B., and Ude, A., 2021, "Intention Recognition With Recurrent Neural Networks for Dynamic Human-Robot Collaboration," Proceedings of the 2021 20th International Conference on Advanced Robotics (ICAR), Ljubljana, Slovenia, Dec. 6–10, pp. 208–215.

[32] Maceira, M., Olivares-Alarcos, A., and Alenya, G., 2020, "Recurrent Neural Networks for Inferring Intentions in Shared Tasks for Industrial Collaborative Robots," Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, Aug. 31–Sept. 4, pp. 665–670.

[33] Dua, N., Singh, S. N., and Semwal, V. B., 2021, "Multi-Input CNN-GRU Based Human Activity Recognition Using Wearable Sensors," Computing, **103**(7), pp. 1461–1478.

[34] Zhu, X., Li, L., Zhang, W., Rao, T., Xu, M., Huang, Q., and Xu, D., 2017, "Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition," Proceedings of the 26th International Joint Conference on Artificial Intelligence IJCAI, Melbourne, Australia, Aug. 19–25, pp. 3595–3601.

[35] Liu, B., Adeli, E., Cao, Z., Lee, K. H., Shenoi, A., Gaidon, A., and Niebles, J. C., 2020, "Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction," IEEE Rob. Autom. Lett., **5**(2), pp. 3485–3492.

[36] Yan, L., Gao, X., Zhang, X., and Chang, S., 2019, "Human-Robot Collaboration by Intention Recognition Using Deep LSTM Neural Network," Proceedings of the 2019 IEEE 8th International Conference on Fluid Power and Mechatronics (FPM), pp. 1390–1396.

[37] Steven Eyobu, O., and Han, D. S., 2018, "Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable IMU Sensor Data Using a Deep LSTM Neural Network," Sensors, **18**(9), p. 2892.

[38] Xin, L., Wang, P., Chan, C. Y., Chen, J., Li, S. E., and Cheng, B., 2018, "Intention-Aware Long Horizon Trajectory Prediction of Surrounding Vehicles Using Dual LSTM Networks," Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, Nov. 4–7, pp. 1441–1446.

[39] Shi, Q., and Zhang, H., 2021, "An Improved Learning-Based LSTM Approach for Lane Change Intention Prediction Subject to Imbalanced Data," Transp. Res. Part C: Emerg. Technol., **133**, p. 103414.

[40] Saleh, K., Hossny, M., and Nahavandi, S., 2017, "Intent Prediction of Vulnerable Road Users From Motion Trajectories Using Stacked LSTM Network," Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, Oct. 16–19, pp. 327–332.

[41] Sui, Z., Zhou, Y., Zhao, X., Chen, A., and Ni, Y., 2021, "Joint Intention and Trajectory Prediction Based on Transformer," Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, Sept. 27–Oct. 1, pp. 7082–7088.

[42] Pettersson, J., and Falkman, P., 2023, "Comparison of LSTM, Transformers, and MLP-Mixer Neural Networks for Gaze Based Human Intention Prediction," Front. Neurorob., **17**, p. 1157957.

[43] Henderson, M., Casanueva, I., Mrkšić, N., Su, P. H., Wen, T. H., and Vulić, I., 2019, "ConveRT: Efficient and Accurate Conversational Representations From Transformers," arXiv preprint arXiv:1911.03688.

[44] Gao, X., Yan, L., Wang, G., and Gerada, C., 2023, "Hybrid Recurrent Neural Network Architecture-Based Intention Recognition for Human–Robot Collaboration," IEEE Trans. Cybern., **53**(3), pp. 1578–1586.

[45] Zhou, H., Yang, G., Wang, B., Li, X., Wang, R., Huang, X., Wu, H., and Wang, X. V., 2023, "An Attention-Based Deep Learning Approach for Inertial Motion Recognition and Estimation in Human-Robot Collaboration," J. Manuf. Syst., **67**, pp. 97–110.

[46] Prevost, C. G., Desbiens, A., and Gagnon, E., 2007, "Extended Kalman Filter for State Estimation and Trajectory Prediction of a Moving Object Detected by an Unmanned Aerial Vehicle," Proceedings of the 2007 American Control Conference, New York, NY, July 9–13, pp. 1805–1810.

[47] Jin, Z., and Pagilla, P. R., 2020, "Operator Intent Prediction With Subgoal Transition Probability Learning for Shared Control Applications," Proceedings of the 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, Sept. 7–9, pp. 1–6.

[48] Linderman, S., Antin, B., Zoltowski, D., and Glaser, J., 2020, SSM: Bayesian Learning and Inference for State Space Models.

[49] Deng, Q., and Söffker, D., 2021, "A Review of HMM-Based Approaches of Driving Behaviors Recognition and Prediction," IEEE Trans. Intell. Vehicles, **7**(1), pp. 21–31.

[50] Liu, S., Zheng, K., Zhao, L., and Fan, P., 2020, "A Driving Intention Prediction Method Based on Hidden Markov Model for Autonomous Driving," Comput. Commun., **157**, pp. 143–149.

[51] Peddi, R., Di Franco, C., Gao, S., and Bezzo, N., 2020, "A Data-Driven Framework for Proactive Intention-Aware Motion Planning of a Robot in a Human Environment," Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, Oct. 24, 2020–Jan. 24, 2021, pp. 5738–5744.
[52] Kelley, R., Tavakkoli, A., King, C., Nicolescu, M., Nicolescu, M., and Bebis, G., 2008, "Understanding Human Intentions via Hidden Markov Models in Autonomous Mobile Robots," Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, Amsterdam, The Netherlands, Mar. 12–15, pp. 367–374.
[53] Mor, B., Garhwal, S., and Kumar, A., 2021, "A Systematic Review of Hidden Markov Models and Their Applications," Archiv. Comput. Methods Eng., **28**(3), pp. 1429–1448.
[54] Ramezani, S. B., Killen, B., Cummins, L., Rahimi, S., Amirlatifi, A., and Seale, M., 2021, "A Survey of HMM-Based Algorithms in Machinery Fault Prediction," Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–9.
[55] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., and Polosukhin, I., 2017, "Attention Is All You Need," Adv. Neural Inf. Process. Syst., p. 30.
[56] Giuliari, F., Hasan, I., Cristani, M., and Galasso, F., 2021, "Transformer Networks for Trajectory Forecasting," Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), pp. 10335–10342.
[57] Yan, H., Deng, B., Li, X., and Qiu, X., 2019, "TENER: Adapting Transformer Encoder for Named Entity Recognition," arXiv preprint arXiv:1911.04474.
[58] Chen, Z., Zhang, L., Cao, Z., and Guo, J., 2018, "Distilling the Knowledge From Handcrafted Features for Human Activity Recognition," IEEE Trans. Ind. Inform., **14**(10), pp. 4334–4342.
[59] Rabiner, L., and Juang, B., 1986, "An Introduction to Hidden Markov Models," IEEE ASSP Mag., **3**(1), pp. 4–16.