ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Stable Worker Intention Recognition via Transformer and CRF-Ontology Decoding for Human–Robot Collaboration

Hwijin Park[1], Hyun-Bin Kwon[1], Hyundo Lee[1], Cheolwoo Park[1], and Hak Yi[1]

***Abstract*— This paper proposes a transformer-based single-stream model with CRF–ontology decoding for stable worker intention recognition in human–robot collaboration(HRC). Although existing intention recognition methods achieve high accuracy, they often suffer from temporal prediction instability and logically inconsistent combinations among actions, tools, parts, and intentions. To address these issues, the proposed approach employs a transformer encoder to integrate worker actions and part-related information, thereby capturing the task context and jointly predicting actions, tools, parts, and intentions. For intention prediction, a conditional random field (CRF) is applied to enforce temporal consistency and improve prediction stability. In addition, an ontology-based post-processing step removes infeasible combinations under a given task intention and reselects predictions that satisfy structural constraints. Experimental results show that the CRF reduces the intention change rate from 7.9% to 3.0%, improving temporal stability, while ontology-based decoding decreases the violation rate from 26.5% to 6.9% by eliminating inconsistent predictions. When combined, the proposed method achieves both a low change rate (3.0%) and a low violation rate (3.7%), demonstrating its effectiveness for reliable intention recognition in HRC.**

I. INTRODUCTION

Recent advances in manufacturing have positioned human–robot collaboration (HRC) as a key component of Industry 5.0, increasing the importance of intelligent systems that can understand worker actions and intentions and respond accordingly [1]. In assembly scenarios, the same action can have different meanings depending on the task context, making it difficult to reliably infer worker intentions based solely on action recognition [2]. Existing approaches primarily focus on short-term action recognition or rely on separately recognizing actions and contextual information; however, such approaches fail to adequately capture the relationship between actions and intentions and do not explicitly consider temporal continuity or structural constraints, leading to unstable predictions and inconsistencies among actions, tools, parts, and intentions [3], [4], [5].

To address these limitations, this paper proposes a Transformer-based single-stream framework with CRF–ontology decoding for worker intention recognition. The proposed method integrates 3D skeleton, tool, and part information to represent the task context and jointly predicts actions, tools, parts, and intentions. In particular, a conditional random field (CRF) is applied to intention prediction to enforce temporal continuity and reduce label fluctuations, while ontology-based constraints eliminate infeasible combinations and reselect predictions that satisfy structural consistency.

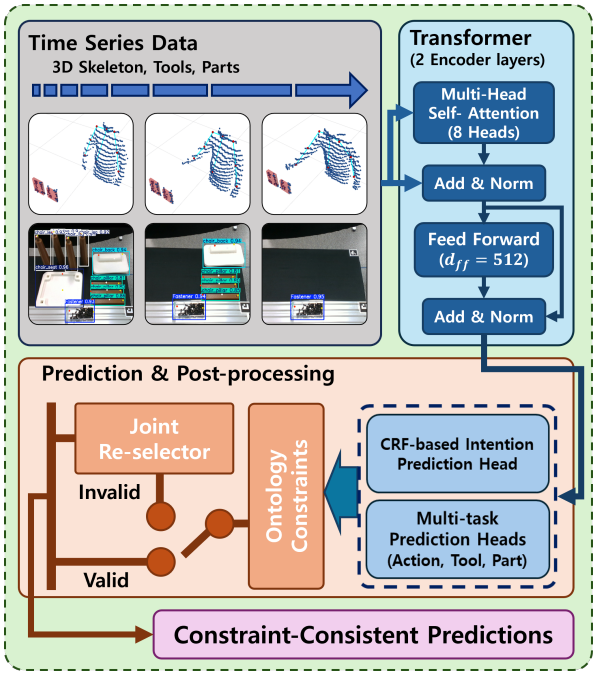

Fig. 1. Overview of the proposed framework, which integrates context-aware Transformer encoding with CRF–ontology-based decoding for stable and consistent intention recognition.

II. METHOD

A. Context-Aware Transformer Encoding

As shown in Fig. 1, the proposed framework takes time-series data composed of 3D skeleton-based joint positions, tools, and parts as input. The input sequence is represented as 60 frames (4s), where each frame consists of a 150-dimensional feature vector. These multimodal features are unified into a single sequence representation, capturing both worker motion and task context. The sequence is then processed by a transformer encoder composed of self-attention and feed-forward layers, enabling the model to learn temporal dependencies and contextual relationships. Based on the

This research was supported by the Ministry of Trade, Industry, and Energy (MOTIE) and the Korea Evaluation Institute of Industrial Technology (KEIT) in 2024 through the "Development of Innovative Smart Painting Technology for the Shipbuilding Process" project.(No.RS-2024-00431769)

This research, undertaken at Kyungpook National University, was supported by the Regional Innovation System & Education (RISE) program through the Daegu RISE Center, funded by the Ministry of Education (MOE) and the Daegu Metropolitan City, Republic of Korea.(2026-RISE-03-001)

[1] Hwijin Park, Hyun-Bin Kwon, Hyundo Lee, Cheolwoo Park, and Hak Yi are with the School of Mechanical Engineering, Kyungpook National University, Daegu, Republic of Korea. yihak@knu.ac.kr

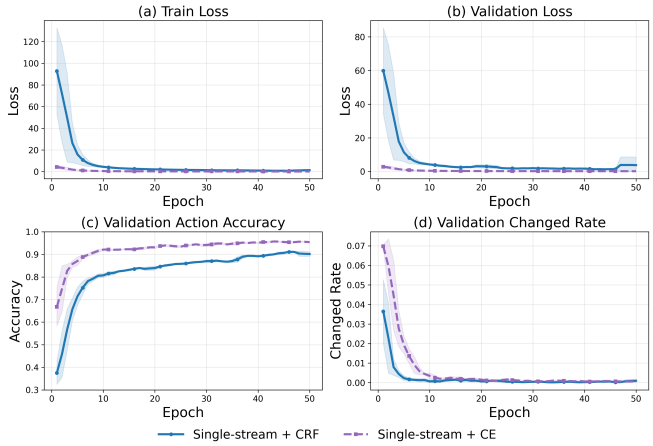

Fig. 2. Training curves of the single-stream model with and without CRF, including train/validation loss, validation action accuracy, and intention changed rate.

TABLE I

ACCURACY COMPARISON OF DECODING VARIANTS

Method	Act. (%)	Tool (%)	Part (%)	Intent (%)
Base	89.809	98.841	95.161	99.174
+ CRF	89.809 (+0.000)	98.841 (+0.000)	95.161 (+0.000)	99.151 (-0.023)
+ Ont.	89.968 (+0.158)	98.921 (+0.080)	95.273 (+0.113)	99.174 (+0.000)
+ CRF+Ont.	90.000 (+0.191)	98.921 (+0.080)	95.231 (+0.070)	99.151 (-0.023)

encoded representation, multi-task prediction heads jointly estimate action, tool, part, and intention labels.

B. CRF–Ontology Decoding

As illustrated in the *Prediction & Post-processing* stage in Fig. 1, a two-stage decoding strategy is employed to improve prediction reliability. First, action, tool, and part are trained using cross-entropy loss at each time step. For intention prediction, CRF is applied to model temporal dependencies, and the sequence-level log-likelihood is maximized as:

$$\mathcal{L}_{\text{CRF}} = -\log P(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

where $\mathbf{x}$ denotes the input sequence and $\mathbf{y}$ represents the intention label sequence. This enables modeling of transition probabilities between adjacent time steps, reducing abrupt label changes.

Then, ontology constraints are applied to validate consistency among action, tool, part, and intention. Invalid combinations are discarded, and the final prediction is selected from the feasible set $\mathcal{V}(t)$ as:

$$\hat{y} = \arg\max_{y \in \mathcal{V}(t)} \left(\log P(a \mid t) + \log P(u \mid t) + \log P(p \mid t)\right) \tag{2}$$

where a, u, and p denote action, tool, and part, and t is the intention. The Joint Re-selector selects the most consistent prediction under ontology constraints.

III. EXPERIMENTS

As shown in Fig. 2, both models converge stably, while the model with CRF exhibits superior temporal stability. In particular, the intention changed rate decreases from 7.9% to 3.0%, indicating that CRF effectively suppresses prediction fluctuations.

TABLE II

STABILITY AND CONSISTENCY COMPARISON

Metric	Baseline (%)	+CRF (%)	+Ont. (%)	+CRF+Ont. (%)
Changed Rate	7.9	3.0 (-4.9)	7.9 (+0.0)	3.0 (-4.9)
Violation (GT)	26.5	26.5 (+0.0)	6.9 (-19.6)	3.7 (-22.8)

Table I shows that all variants achieve comparable performance across action, tool, part, and intention accuracy, with only marginal differences. This suggests that accuracy alone is insufficient to highlight the effectiveness of the proposed method.

In contrast, Table II demonstrates the impact of the proposed decoding strategy on stability and consistency. With CRF, the intention changed rate is reduced from 7.9% to 3.0%, confirming its effectiveness in enforcing temporal continuity. Meanwhile, ontology-based decoding reduces the violation rate from 26.5% to 6.9%, eliminating structurally infeasible combinations. The violation rate is defined as the percentage of frame-wise predictions that violate ontology constraints among action, tool, part, and intention combinations. When combined, the proposed method achieves both a low changed rate (3.0%) and a low violation rate (3.7%), improving both temporal stability and structural consistency.

IV. CONCLUSION

This paper proposed a Transformer-based single-stream framework with CRF–ontology decoding for stable worker intention recognition in HRC. The proposed method integrates worker actions and task context to jointly infer actions, tools, parts, and intentions, while CRF enforces temporal stability and ontology-based constraints ensure structural consistency. Experimental results showed that, although accuracy remains comparable to baseline methods, the proposed approach significantly improves both temporal stability and structural consistency by reducing intention fluctuations and eliminating infeasible combinations. These results demonstrate the effectiveness of the proposed framework for reliable intention recognition in HRC.

REFERENCES

[1] A. Adel, "Future of Industry 5.0 in society: Human-centric solutions, challenges and prospective research areas," *Journal of Cloud Computing*, vol. 11, no. 1, p. 40, 2022.

[2] S. Aberbach-Goodman and R. Mukamel, "Temporal hierarchy of observed goal-directed actions," *Scientific Reports*, vol. 13, no. 1, p. 19701, 2023.

[3] G. Hoffman, T. Bhattacharjee, and S. Nikolaidis, "Inferring human intent and predicting human action in human–robot collaboration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2024.

[4] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001.

[5] W. Fang, T. Zhang, Z. Wang, and J. Ding, "A multi-modal context-aware sequence stage validation for human-centric AR assembly," *Computers & Industrial Engineering*, vol. 194, p. 110355, 2024.