Joint Inference of States, Robot Knowledge, and Human (False-)Beliefs

Tao Yuan¹  Hangxin Liu¹  Lifeng Fan¹  Zilong Zheng¹  Tao Gao¹,³  Yixin Zhu¹,²  Song-Chun Zhu¹,²

Abstract— Aiming to understand how human (false-)belief—a core socio-cognitive ability—would affect human interactions with robots, this paper proposes to adopt a graphical model to unify the representation of object states, robot knowledge, and human (false-)beliefs. Specifically, a parse graph (pg) is learned from a single-view spatiotemporal parsing by aggregating various object states along the time; such a learned representation is accumulated as the robot’s knowledge. An inference algorithm is derived to fuse individual pg from all robots across multi-views into a joint pg, which affords more effective reasoning and inference capability to overcome the errors originated from a single view. In the experiments, through the joint inference over pg, the system correctly recognizes human (false-)belief in various settings and achieves better cross-view accuracy on a challenging small object tracking dataset.

I. INTRODUCTION

The seminal Sally-Anne [1] study has spawned a vast research literature in developmental psychology regarding Theory of Mind (ToM); in particular, human’s socio-cognition in understanding false-belief—the ability to understand other’s belief about the world may contrast with the true reality. A cartoon version of the Sally-Anne test is shown in the left of Fig. 1. Sally puts her marble in the box and left. While Sally is out, Anne moves the marble from the box to a basket. The test would ask a human participant where Sally would look for her marble when she is back. In this experiment, the marble would still be inside the box according to Sally’s false-belief, even though the marble is actually inside the basket. To answer this question correctly, an agent should understand and disentangle the object state (observation from the current frame), the (accumulated) knowledge, the belief of other agents, the ground-truth/reality of the world, and importantly, the concept of false-belief.

The prior study suggests that at the age of 4 years old, children begin to develop the capability to understand false-belief [2]. Such abilities to ascribe the mental belief to the human mind, to differentiate belief from the physical reality, and even to recognize false-belief and perform psychological reasoning, is a significant milestone in the acquisition of ToM [3], [4]. Such evidence emerged from developmental psychology in the past few decades call for integrating such socio-cognitive aspects into a modern social robot [5].

In fact, false-belief is not rare in our daily life. Two examples are depicted in the middle and the right of Fig. 1. (i) Where does Bob think his cup1 is after Charlie put cup2 (visually identical to cup1) on the table while Dave took cup1 away? (ii) Which milk box should Alice give to Bob if she wants to help? The one closer to Bob but empty, or the one further to Bob but full? Although such false-belief tasks are primal examples for social and cognitive intelligence, current state-of-the-art intelligent systems are still facing challenges in acquiring such a capability in the wild with noisy visual input (see Related Work for discussion).

One fundamental challenge is the lack of proper representation for modeling the false-belief from visual input; it has to be able to handle the heterogeneous information of a system’s current states, its accumulated knowledge, agent’s belief, and the reality/ground-truth of the world. Without a unified representation, the information across all these domains cannot be easily interpreted, and the cross-domain reasoning of the events is infeasible.

Largely due to this difficulty, prior work that takes noisy sensory input can only solve a sub-problem in understanding false-belief. For instance, sensor fusion techniques are mainly used to obtain better state estimation by filtering the measurements from multiple sensors [6]. Similarly, the Multiple View Tracking (MVT) in computer vision is designed to combine the observations across camera views to better track an object. Visual cognitive reasoning (e.g., human intention/attention predictions [7], [8], [9], [10]) only
targets to model human mental states. These three lines of work are all crucial ingredients but developed independently; a unified cross-domain representation is still largely missing.

In order to endow such an ability to understand the concept of false-belief to a robot system from noisy visual inputs, this paper proposes to use a graphical model represented by a parse graph $(pg)$ [11] to serve as the unified representation of a robot’s knowledge structure, fused knowledge across all robots, and the (false-)beliefs of human agents. A $pg$ is learned from the spatiotemporal transition of humans and objects in the scene perceived by a robot. A joint $pg$ can be induced by merging and fusing the individual $pg$s from each robot to overcome the errors originated from a single view. In particular, our system enables the following three capabilities with increasing depth in cognition:

1) **Tracking small objects with occlusions across different views.** Human-made objects in an indoor environment (e.g., cups) are oftentimes small with a similar appearance. Tracking such objects could be challenging due to occlusions with frequent human interactions. The proposed method can address the challenging multi-view multi-object tracking problem by properly maintaining cross-view object states using the unified representation.  

2) **Inferring human beliefs.** The state of an object normally does not change unless a human interacts with it; this observation shares a similar spirit in human cognition known as object permanence [12]. By identifying the interactions between humans and objects, our system also supports the high-level cognitive capability; e.g., knowing which object is interacted with which person, whether a person knows the state of the object has been changed.  

3) **Assisting agents by recognizing false-belief.** Giving the above object tracking and cognitive reasoning of human beliefs, the proposed algorithm can further infer whether and why the person has false-belief, thereby to better assist the person given a specific context.  

A. Related Work

**Robot ToM**, aiming at understanding human beliefs and intents, receives increasing research attentions in human-robot interaction and collaboration [13], [14]. Several false-belief tasks akin to the classic Sally-Anne test were designed. For instance, Warnier et al. [15] introduced a belief management algorithm, and the reasoning capability is subsequently endowed to a robot to pass the Sally-Anne test [16] successfully. More sophisticated human-robot collaboration is achieved by maintaining a human partner’s mental state [17]. More formally, Dynamic Epistemic Logic is introduced to represent and reason about belief and false-belief [18], [19]. These successes are, however, limited to the symbolic-based belief representations, requiring handcrafted variables and structures, making the logic-based reasoning approaches brittle in practice to handle noises and errors. To address this deficiency, this paper utilizes a unified representation by $pg$, a probabilistic graphical model that has been successfully applied to various robotics tasks, e.g., [20], [21], [22]; it accumulates the observations over time to form a knowledge graph and robustly handles noisy visual input.

**Multi-view Visual Analysis** is widely applied to 3D reconstruction [23], object detection [24], [25], cross-view tracking [26], [27], and joint parsing [28]. Built on top of these modules, Multiple Object Tracking (MOT) usually utilizes tracking-by-detection techniques [29], [30], [31]. This line of work primarily focuses on combining different camera views to obtain a more comprehensive tracking, lacking the understanding of human (false-)belief.

**Visual Cognitive Reasoning** is an emerging field in computer vision. Related work includes recovering incomplete trajectories [32], learning utility and affordance [33], inferring human intention and attention [9], [10], etc. As to understanding (false-)belief, despite many psychological experiments and theoretical analysis [34], [35], [36], [37], very few attempts have been made to solve (false-)belief with visual input; handcrafted constraints are usually required for specific problems in prior work. In contrast, this paper utilizes a unified representation across different domains with heterogeneous information to model human mental states.

B. Contribution

This paper makes three contributions:

1) We adopt a unified graphical model $pg$ to represent and maintain heterogeneous knowledge about object states, robot knowledge, and human beliefs.  

2) On top of the unified representation, we propose an inference algorithm to merge individual $pg$s from different domains across time and views into a joint $pg$, supporting human belief inference from multi-view to overcome the noises and errors originated from a single view.  

3) With the inferred $pg$s, our system can keep track of the state and location of each object, infer human beliefs, and further discover false-belief to better assist human.

C. Overview

The remainder of the paper is organized as follows. Sections II and III describe the representation and the detailed probabilistic formulation, respectively. We demonstrate the efficacy of the proposed method in Section IV and conclude the paper with discussions in Section V.

II. REPRESENTATION

In this work, we use the parse graph $(pg)$—a unified graphical model [11]—to represent (i) the location of each agent and object, (ii) the interactions between agents and objects, (iii) the beliefs of agents, and (iv) the attributes and states of objects; see Fig. 2 for an example. Specifically, three different types of $pg$s are utilized:

- **Robot pg**, shown as blue circles, maintains the knowledge structure of an individual robot, which is extracted from its visual observation—an image. It also contains attributes that are grounded to the observed agents and objects.  

- **Belief pg**, shown as red diamonds, represents the inferred human knowledge by each robot. Each robot maintains the parse graph for each agent it observed.  

- **Joint pg** fuses all the information and views across a set of distributed robots.

**Notations and Definitions:** The input of our system can be represented by $M$ synchronized video sequences $I = \{I^k_{t=1:T}\}$ with length $T$ captured from $M$ robots. Formally, a scene $R$ is expressed as

$$R = \{(O_t, H_t) : t = 1, 2, \ldots, T\},$$

$$O_t = \{o^i_t : i = 1, 2, \ldots, N_o\},$$

$$H_t = \{h^j_t : j = 1, 2, \ldots, N_h\},$$

(1)
where $O_t$ and $H_t$ denote the set of all the tracked objects ($N_o$ objects in total) and the set of all the tracked agents ($N_h$ agents in total) at time $t$, respectively.

Object $o_i$ is represented by a tuple: bounding box location $b_i^t$, appearance feature $\phi_i^t$, states $s_i^t$, and attributes $a_i^t$,

$$o_i^t = (b_i^t, \phi_i^t, s_i^t, a_i^t),$$  

where $s_i^t$ is an index function: $s_i^t = j, j \neq 0$ indicates the object $o_i$ is held by the agent $h_j$ at time $t$, and $s_i^t = 0$ means it is not held by any agent at time $t$.

The agent $h_i^t$ is represented by its body key-point position $h_i^t = (k_i^t, \phi_i^t)$.

Robot Parse Graph is formally expressed as

$$\tilde{pg}_{\mathcal{R}}^k = \{(o_i^t, h_i^t): o_i^t, h_i^t \in I^k_t\},$$

where $I^k_t$ is the area where $k$th robot can observe at time $t$.

Belief Parse Graph is formally expressed as

$$\tilde{pg}_{\mathcal{B}}^{k,j} = \{(o_i^{t'}, h_i^{t'}): o_i^{t'}, h_i^{t'} \in I^k_t\},$$

where $\tilde{pg}_{\mathcal{B}}^{k,j}$ represents the inferred belief of agent $h_j$ under robot $k$'s view; $t'$ is the last time that the robot $k$ observes the human $h_j$. We assume that the agent $h_j$ only keeps the objects $s$he observed last time in this area, which satisfies the Principle of Inertia: an agent’s belief is preserved over time unless the agent gets information to the contrary.

Joint Parse Graph keeps track of all the information across a set of distributed robots, formally expressed as

$$pg_t = \{(o_i^t, h_i^t): i = 1, 2, ..., N_o; j = 1, 2, ..., N_h\}. $$

Objective: The objective of the system is to jointly infer all the parse graphs $PG = \{ pg, \tilde{pg}_{\mathcal{R}}, \tilde{pg}_{\mathcal{B}} \}$ so that it can (i) track all the agents and objects across scenes at any time by fusing the information collected by a distributed system, and (ii) infer human (false-)beliefs to provide assistance.

III. PROBABILISTIC FORMULATION

We formulate the joint parsing problem as a maximizing a posterior (MAP) inference problem

$$PG^* = \arg \max_{PG} p(PG | I) = \arg \max_{PG} p(I | PG) \cdot p(PG),$$

where $p(PG)$ is the prior, and $p(I | PG)$ is the likelihood.

A. Prior

The prior term $p(PG)$ models the compatibility of the robot $pg$s and the joint $pg$, and the compatibility of the joint $pg$ over time. Formally, we can decompose the prior as

$$p(PG) = p(pg_1)\prod_{t=1}^{T-1} p(pg_{t+1} | pg_t) \prod_{k=1}^{M} \prod_{t=1}^{T} p(\tilde{pg}_{\mathcal{R}}^k | pg_t),$$

where the first term $p(pg_{t+1} | pg_t)$ is the transition probability of the joint $pg$ over time, further decomposed as

$$p(pg_{t+1} | pg_t) = \frac{1}{Z} \exp\{-E(pg_{t+1} | pg_t)\}.$$  

The second term $p(\tilde{pg}_{\mathcal{R}}^k | pg_t)$ is the probability which models the compatibility of individual $pg$s and the joint $pg$. Its energy can be decomposed into three energy terms

$$p(\tilde{pg}_{\mathcal{R}}^k | pg_t) = \frac{1}{Z} \exp\{-E(pg_t, \tilde{pg}_{\mathcal{R}}^k)\}.$$

$$= \frac{1}{Z} \exp\{-E_A(pg_t, \tilde{pg}_{\mathcal{R}}^k) - E_S(pg_t, \tilde{pg}_{\mathcal{R}}^k) - E_{Attr}(\tilde{pg}_{\mathcal{R}}^k)\}.$$  

Below, we detail the above six energy terms $E(.)$ in Eqs. (10) and (11).

Motion Consistency: The term $E_M$ measures the motion consistency of objects and agents in time, defined as

$$E_M(b_i^{t+1}, b_i^t, s_i^t) = \begin{cases} \delta(D(b_i^{t+1}, b_i^t) > \tau) & \text{if } s_i^t = 0 \\ \delta(D(\kappa_i^{t+1}, \kappa_i^t) > \tau) & \text{if } s_i^t = j \end{cases}$$

where $D$ is the distance between two bounding boxes or human poses, $\tau$ is the speed threshold, and $\delta$ is the indicator function. If an object $i$ is held by an agent $j$, we use the agent’s location to calculate $E_M$ of the object.

State Transition Consistency: The term $E_ST$ is the state transition energy, defined as

$$E_ST(s_i^{t+1}, s_i^t) = -\log p(\delta(s_i^{t+1} = 0) | \delta(s_i^t = 0)),$$

where the state transition probability $p(\delta(s_i^{t+1} = 0) | \delta(s_i^t = 0))$ is learned from the training data.

Appearance Consistency: $E_A$ measures appearance consistency. In robot $pg$s, the appearance feature vector $\phi$
Spatial Consistency: Each object and agent in the robot’s 2D view should also have a corresponding 3D location in the real-world coordinate system, and such correspondence should remain consistent when projected any location in the real-world coordinate system, and such correspondence should remain consistent when projected any location in the real-world coordinate system, and such correspondence should remain consistent when projected any location in the real-world coordinate system. Thus, spatial consistency is defined as

\[ \mathcal{E}_{S}(p_{g_k}, p_{\tilde{g}_k}) = \sum_{c \in O_k \cup H_k} ||\Lambda^c - f(\Lambda^c)||_2, \]  

(15)

where \( \Lambda \) is the 3D positions in the real-world coordinate, and \( f \) is the transformation function that projects the points from the robot’s 2D view to the 3D real-world coordinate.

Attribute Consistency: Attributes of each entity should remain the same across time and viewpoints. Such an attribute consistency is defined by the term \( \mathcal{E}_{Attr} \)

\[ \mathcal{E}_{Attr}(p_{g_k}, p_{\tilde{g}_k}) = \sum_{i=1}^{N_c} \delta(a_i^x \neq \tilde{a}_i^x). \]  

(16)

B. Likelihood

The likelihood term \( p(I|PG) \) models how well each robot can ground the knowledge in its \( pg \) to the visual data it captures. Formally, the likelihood is defined as

\[ p(I|PG) = \prod_{k=1}^{M} \prod_{t=1}^{T} p(I^k_t | \tilde{p}_{g_k}^t). \]  

(17)

The energy of term \( p(I^k_t | \tilde{p}_{g_k}^t) \) can be further decomposed as

\[ p(I^k_t | \tilde{p}_{g_k}^t) = \frac{1}{Z} \exp\{-\mathcal{E}(I^k_t | \tilde{p}_{g_k}^t)\}, \]  

(18)

\[ \mathcal{E}(I^k_t | \tilde{p}_{g_k}^t) = \sum_{i=1}^{N_c} \mathcal{E}_D(b_i^t, \phi_i^t) + \mathcal{E}_C(b_i^t, \phi_i^t, a_i^t) + \sum_{j=1}^{N_c} \mathcal{E}_D(p_j^t, \phi_j^t), \]  

(19)

where \( \mathcal{E}_D \) can be calculated by the score of object detection or human pose estimation, and \( \mathcal{E}_C \) can be obtained by the object attributes classification scores.

C. Inference

Given the above probabilistic formulation, we can infer the best \( \{pg^*, \tilde{p}_{g_k}^t\} \) by an MAP estimate. It can be solved by two steps: (i) Each robot individually processes the visual input; the output (e.g., object detection, and human pose estimation) can be aggregated as the proposals for the second step. (ii) The MAP estimate can be transformed to an assignment problem given the proposals, solvable using the Kuhn-Munkres algorithm [39], [40] in polynomial time.

Based on Eq. (5), robot \( k \) can generate belief parse graphs \( \tilde{p}_{g_k}^t \) for agent \( j \) after obtaining the robot graphs \( pg_k \).

IV. EXPERIMENT

We evaluate the proposed method in two setups: cross-view object tracking and human (false-)belief understanding. The first experiment evaluates the accuracy of object localization using the proposed inference algorithms, focusing on the robot parse graphs \( pg \) and the joint parse graph \( p_{\tilde{g}_k} \). The second experiment evaluates the inference of the belief parse graphs \( pg, i.e., human beliefs regarding the object states (e.g., locations) in both single-view and multi-view settings.

A. Dataset

The dataset includes two subsets, a multi-view subset and a single-view subset. Ground-truth tracking results of objects and agents, and states and attributes of objects are all annotated for evaluation purpose.

- The single-view subset includes 5 different false-belief scenarios with 12532 frames. Each scenario contains at least one kind of false-belief test or helping test. In this subset, objects are not limited to the cups.
- The multi-view subset consists of 8 scenes, each shot with 4 robot camera views, making a total number of 72720 frames. Each scenario contains at least one kind of false-belief test. The objects in each scene are, however, limited to the cups: 12-16 different cups made with 3 different materials (plastic, paper, and ceramic) and 4 colors (red, blue, white, and black). In each scene, three agents interact with cups by performing actions depicted in Fig. 3.

B. Implementation Details

Below, we detail the implementation of the system.

- Object detection: we use the RetinaNet model [41] pre-trained on the MS COCO dataset [42]. We keep all the bounding boxes with a score higher than the threshold 0.2, which serve as the proposals for object detection.
- Human pose estimation: we apply the AlphaPose [43],
- Object attribute classification: A VGG16 network [44] was trained to classify the color and the material of the objects.
- Appearance feature: A deep person re-id model [38] was fine-tuned on the training set.

Due to the lack of multi-view in the single-view setting, we locate the object that an agent plan to interact by simply finding the object closest to the direction the agent points at according to the key points on the arm.
C. Experiment 1: Cross-view Object Localization

To test the overall cross-view tracking performance, 2000 queries are randomly sampled from the ground-truth tracks. Each query $q$ can be formally described as

$$q = (k, t, b, t_q),$$

where the tuple $(k, t, b)$ indicates the object shown in robot $k$'s view located in bounding box $b$ at time $t$. Such a form of the query can be very flexible. For instance, if we ask about the location of that object at time $t_q$, the system should return an answer in the form of $(k_o, b_o)$, meaning that the system predicts the object is shown in robot $k_o$'s view at $b_o$.

The system generates the answer in two steps. It firstly locates the query of the object by searching the object $i$ in $pg^k_i$ with the smallest distance to the bounding box $b$. Then it returns the location $b_{t_q}^i$ from $pg^k_i$. The accuracy of model $M$ can be calculated as

$$\text{acc}(M) = \frac{1}{N_q} \sum_{i=1}^{N_q} \delta(\text{IoU}(b_{t_q}^i, b_o^i) > 0.5) \cdot \delta(k_{gt} = k_o),$$

where $N_q$ is the number of queries, $b_{gt}$ is the ground-truth bounding box, and $b_{t_q}$ is the inferred bounding box returned by model $M$. We calculate the Intersection over Union (IoU) between the answer and the ground-truth bounding boxes; the answer is correct if and only if the answer predicts the right view and the IoU is larger than $\xi = 0.5$.

Table I shows the ablative study by turning on and off the joint parsing component that models human interactions, i.e., whether the model parses and tracks objects by reasoning about the interaction with agents. "# interactions" means how many times the object was interacted by agents. The result shows that our system achieves an overall 88% accuracy. Even without parsing humans, our system still possesses the ability to reason about object location by maintaining other consistencies, such as spatial consistency and appearance consistency. However, its performance drops significantly if the object was moved to different rooms. Figure 4 shows some qualitative results.

<table>
<thead>
<tr>
<th># interactions</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsing w/o humans acc.</td>
<td>0.98</td>
<td>0.82</td>
<td>0.78</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td>Joint parsing acc.</td>
<td>0.98</td>
<td>0.86</td>
<td>0.85</td>
<td>0.82</td>
<td>0.88</td>
</tr>
</tbody>
</table>

D. Experiment 2: (False-)Belief Inference

In this experiment, we evaluate the performance of belief and false-belief inference, i.e., whether an agent’s belief $pg$ is the same as the true object states. The evaluations were conducted on both single-view and multi-view scenarios.

Multi-view: We collected 200 queries with ground-truth annotations that focus on the Sally-Anne false belief task. The query is defined as

$$q = (k_o, t_o, b_o, k_h, t_h, b_h, t_q),$$

where first three terms $(k_o, t_o, b_o)$ define the objects in robot $k_o$'s view located at $b_o$ at time $t_o$. Similarly, another three terms $(k_h, t_h, b_h)$ define an agent in robot $k_h$'s view located at $b_h$ at time $t_h$. The question is: where does the agent $(k_h, t_h, b_h)$ think the object $(k_o, t_o, b_o)$ is at time $t_q$?

Our system generates the answer in three steps: (i) search for the object $i$ and the agent $j$ in robot parse graphs $pg^k_{t_o}$ and $pg^k_{t_h}$, (ii) retrieve all the belief parse graphs $pg^k_{t_q}$ at time $t_q$ to find the object $i$’s location $b_{t_q}^k$ in human $j$’s belief, and (iii) find an object $i'$ in robot parse graph, which has the same attributes as $i$’s and has smallest distance to $b_{t_q}^k$. The system finally returns $i'$’s location $b_{t_q}^k$ as the answer.

Since there is no publicly available code on this task, we compare our inference algorithm with a random baseline model as the reference for future benchmark; it simply returns an object with the same attributes as the query object at $t_q$. The result shows that our system achieves 81% accuracy, while the baseline model only has 30% accuracy.

Single-view: We collected a total of 100 queries, including two types of belief inference tasks: the Sally-Anne false-belief task and the helping task, as shown in Fig. 1. The queries have two forms

$$q = (t_o, b_o, b_h, t_q), \quad \text{and} \quad q = (b_h, t_q),$$

indicating two different types of questions: (i) where does the agent $b_h$ think the object $(t_o, b_o)$ is at time $t_q$? (ii)
V. CONCLUSION AND DISCUSSIONS

In this paper, we describe the idea of using $pg$ as a unified representation for tracking object states, accumulating robot knowledge, and reasoning about human (false-)beliefs. Based on the spatiotemporal information observed from multiple camera views of one or more robots, robot $pg$ and belief $pg$ are induced and merged to a joint $pg$ to overcome the possible errors originated from a single view. With such a representation, a joint inference algorithm is proposed, which possesses the capabilities of tracking small occluded objects across different views and inferring human (false-)beliefs. In experiments, we first demonstrate that the joint inference over the merged $pg$ produced better tracking accuracy. We further evaluate the inference on human true- and false-belief regarding objects’ locations by jointly parsing the $pg$s. The high accuracy demonstrates that our system is capable of modeling and understanding human (false-)beliefs, with the potential of helping capability as demonstrated in developmental psychology.

ToM and Sally-Anne test are interesting and difficult problems in the area of social robotics. For a service robot to interact with humans in an intuitive manner, it must be able to maintain a model of the belief states of the agents it interacts with. We hope the proposed method using a graphical model has demonstrated a different perspective compared to prior methods in terms of flexibility and generalization. In the future, a more interactive and active set up would be more practical and compelling. For instance, by integrating activity recognition modules, our system should be able to perceive, recognize, and extract richer semantic information from the observed visual input, thereby providing more subtle (false-)belief applications. Communication, gazes, and gestures are also crucial in intention expression and perception in collaborative interactions. By incorporating these essential ingredients and taking advantage of the flexibility and generalization of the model, our system should be able to go from the current passive query to active response to assist agents in real-time.