In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, Canada, September 2017

Leveraging Commonsense Reasoning and Multimodal Perception for Robot Spoken Dialog Systems

Dongcai Lu[1], Shiqi Zhang[2], Peter Stone[3], and Xiaoping Chen[1]

***Abstract*— Probabilistic graphical models, such as partially observable Markov decision processes (POMDPs), have been used in stochastic spoken dialog systems to handle the inherent uncertainty in speech recognition and language understanding. Such dialog systems suffer from the fact that only a relatively small number of domain variables are allowed in the model, so as to ensure the generation of good-quality dialog policies. At the same time, the non-language perception modalities on robots, such as vision-based facial expression recognition and Lidar-based distance detection, can hardly be integrated into this process. In this paper, we use a probabilistic commonsense reasoner to "guide" our POMDP-based dialog manager, and present a principled, multimodal dialog management (MDM) framework that allows the robot's dialog belief state to be seamlessly updated by both observations of human spoken language, and exogenous events such as the change of human facial expressions. The MDM approach has been implemented and evaluated both in simulation and on a real mobile robot using guidance tasks.**

I. INTRODUCTION

Language has been the most natural way of interaction among humans. Accordingly, there is the need to equip robots with the same kind of capabilities to achieve more natural human-robot interaction (HRI). Robot spoken dialog systems are designed to enable a robot to converse with a human with voice, and typically include the components for language understanding, dialog management, and language generation. A typical language understanding module includes a speech recognition subsystem that converts spoken language into text, and a parsing subsystem that converts text into symbolic representations. It should be noted that both subsystems are unreliable and introduce noise, placing the main challenge to dialog management, as the next step in the pipeline. This work is aimed at the *dialog manager* component that is used for computing a language policy. This policy takes as input the symbolic, unreliable observations from the language understanding module, and outputs language actions that are synthesized by the language generation component.

Following the *Markov* assumption, i.e., the next state relying on only the current state and being independent of all previous ones (first-order), Markov decision processes (MDPs) have been developed for action selection toward maximizing long-term rewards under the uncertainty in action outcomes. Partially observable MDPs (POMDPs) further assume the partial observability of the current world state, so a *belief state*, in the form of a distribution over all possible states, is maintained for estimating the current state [1]. As a result, POMDPs have been used for dialog management [2]. Such dialog systems share a limitation with many other POMDP-based systems that only a relatively small number of domain variables can be modeled in the state space.

[1]Dongcai Lu and Xiaoping Chen are with the School of Computer Science, University of Science and Technology of China, Hefei, China

[2]Shiqi Zhang is with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland OH, USA

[3]Peter Stone is with the Department of Computer Science, the University of Texas at Austin, Austin TX, USA

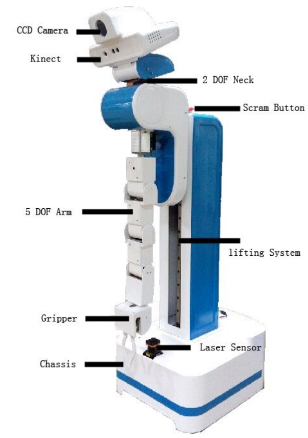


Fig. 1: The mobile robot platform (KeJia robot) that has been used (as a museum guide robot) in the experiments.

In this paper, we decompose standard POMDP problems into a sub-problem of commonsense reasoning in the original state space, and a sub-problem of probabilistic planning in a (much smaller) partial state space, enabling our dialog manager to account for exogenous events *during* human-robot conversations. This work is particularly useful for conversations that last a relatively long time. As the second contribution of this paper, we add the perception modalities of vision and laser range finding into our robot dialog system to augment its sensing capabilities.

The proposed algorithm has been implemented and evaluated both in simulation and on a mobile robot working on museum guide tasks (Fig. 1), where the robot interacts with visitors via spoken language to identify their interests and physically guide visitors to exhibits. In addition to language, this work enables the robot to incorporate vision-based facial expression recognition and Lidar-based localization into spoken dialog systems. For instance, given an "unhappy" face being detected in front of an exhibit of Professor P's research achievement in the $50's$, our robot revises its belief (in probability) toward the visitor being interested in neither Professor P's research nor research in the $50's$. As a result, it is likely that the visitor will be led to exhibits related to other researchers in different years (after confirming with language).[1] Existing dialog management methods cannot produce such robot behaviors.

II. Related Work

Within the context of intelligent robotics, this work lies in the intersection of spoken dialog management, commonsense reasoning, and multimodal perception. We summarize a representative set of algorithms and systems in these areas.

Spoken dialog management has been modeled as a probabilistic planning problem to handle the inherent noise from spoken language understanding in dialog systems. NJFun is one of the early, representative systems that base its dialog manager on an MDP [3]. POMDPs are used in modern dialog systems to model the uncertainty from language understanding in a continuous space [2], and such dialog systems have been applied to real robot systems [4], [5], [6]. Despite the significant advancements achieved in POMDP-based planning algorithms and implementations ([7], [8]) and in speech recognition (e.g., the ones based on Deep Neural Networks [9]), it is still a challenge to apply such dialog systems to real-world problems that frequently involve a large number of domain variables. Also, the perception in such systems is restricted to language.

The term *commonsense* knowledge, in this paper, is used to refer to the knowledge that is normally true but not always, and can be represented using defaults and probabilities. There is a rich history of default reasoning in artificial intelligence [10], and answer set programming (ASP) is one of the popular realizations that has been applied to different reasoning problems [11], [12]. P-log extends ASP to further enable probabilistic commonsense reasoning using causal Bayesian network [13]. As a result, P-log is able to draw conclusions in both quantitative and qualitative forms using commonsense knowledge. We use P-log in this work.

Modern robots are mostly equipped with multiple sensing modalities, such as audio, video, range-finding and haptics [14], [15], [16], [17]. While audio-based language has been the predominant input of spoken dialog systems [2], recent work has shown that other sensing modalities have the potential to significantly improve robots' language capabilities [18], [19]. In this work, we develop an algorithm that integrates multiple sensing modalities (audio, video, and Lidar) for robot spoken dialog systems in a principled way.

The work closest to this research is algorithm CORPP [5] that uses a commonsense reasoner and a probabilistic planner to focus on the "curse of dimensionality" and the "curse of history" (defined in [20]) respectively. The reasoner specifies a partial state space (typically much smaller than the original), on which a POMDP-based planner computes a policy that maximizes its long-term reward. This strategy has enabled dialog management in high-dimensional spaces and long planning horizons that have been impossible before. A recent work further enables dynamically constructing (PO)MDPs using a commonsense reasoner [21]. However, neither of them is capable of incorporating multimodal perception or incorporating exogenous events into state estimation. For instance, an unhappy face of the human in a human-robot conversation can be a negative sign to the robot's belief estimation or current acting policy. In case of such exogenous events (unhappy face), CORPP and its recent extension either discard all information collected so far (to account for the changes) or completely ignore the exogenous events. This paper aims to address these issues.

[1]In this work, a dialog does not end until the visitor leaves the current room.

III. Algorithm

In this section, we first introduce how commonsense knowledge is structured in our multimodal dialog management (MDM) framework, then building on this structure we present the MDM control flow in a general way, and finally we detail the implementation of our MDM-based spoken dialog system on a KeJia robot.

A. Structure of commonsense knowledge

In a factored space, a *world state* can be specified using a set of random variables (or simply variables), denoted by $\mathbf{V}$, and their values, denoted by $\mathbf{v}$. The world states together form a *full world state space* (or simply *full space*). Given a task at hand, we can select a minimal set of random variables for specifying a *partial world state space* (or *partial space*). The partial space includes a set of *partial world states* (or simply *partial states*). The variables modeled in a partial space are called *endogenous* variables and denoted by $\mathbf{V}^{en}$. The elements in $\mathbf{V}^{ex} = \mathbf{V} \setminus \mathbf{V}^{en}$ (set subtraction) are *exogenous* variables.

Our strategy is to use a commonsense reasoner (includes a set of logical and probabilistic rules) to reason within the full space about the variables of $\mathbf{V}^{en} \cup \mathbf{V}^{ex}$ (without considering robot actions); and to use a task-oriented probabilistic planner (corresponds to a probabilistic transition system) to plan within the partial space about only the variables of $\mathbf{V}^{en}$. Intuitively, the reasoner and planner are concerned with "*understanding the current state of the world*" and "*accomplishing the task given the current understanding*" respectively. The task-oriented probabilistic planner is guided by the reasoner – this will be detailed in our dialog system. Standard probabilistic planners (including the POMDP-based spoken dialog systems) do not differentiate these two spaces, directly computing plans in the full world state space.

Fig. 2 illustrates an example world space that includes three endogenous variables and three exogenous variables. The endogenous variables and their conditional dependencies are modeled by a *causal Bayesian network* via a directed acyclic graph (DAG), where the relationships between nodes are required to be causal [22]. This partial state space corresponds to the POMDP state space and is where planning happens. It should be noted that the standard POMDP framework does not allow the modeling of direct dependencies between states (e.g., a POMDP cannot tell which state is the

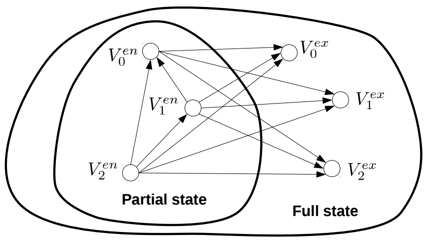


Fig. 2: An illustration of world state space and partial state space, where $\mathbf{V}^{en} = \{V_0^{en}, V_1^{en}, V_2^{en}\}$ is the set of endogenous variables and $\mathbf{V}^{ex} = \{V_0^{ex}, V_1^{ex}, V_2^{ex}\}$ is the set of exogenous variables.

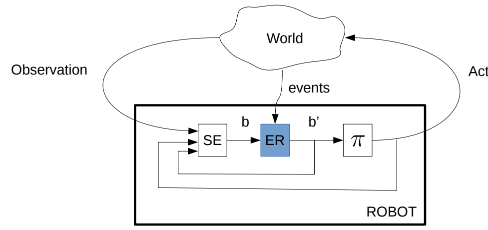


Fig. 3: A flow chart of our multimodal dialog management (MDM) framework. The *event reasoner* (ER) plays a key role in comparison to standard stochastic dialog manager.

most likely *before* taking any actions), whereas MDM does so via the structured knowledge.

The exogenous variables are causally dependent on the endogenous variables. The dependencies are represented by the arrows that connect the partial and non-partial state spaces in Fig. 2. The dependencies between exogenous variables are not modeled because they have default values that can be smoothly replaced by true values when available (accomplished via default reasoning). Modeling the exogenous variables this way allows us to use their dependencies to estimate the values of endogenous variables. This enables POMDP belief state, as a distribution, to be directly updated by the value change of exogenous variables. This belief update mechanism has been absent in the literature.

Consider a museum environment that includes one museum guide robot and two exhibits about science and art respectively. The robot can ask questions to find out (in probability) which exhibit a visitor is most interested in. In this sense, it is necessary to model only one domain variable (either science or art). If the robot is equipped with the facial expression recognition capability (happy/unhappy), another domain variable is needed. In this example, the science/art variable is endogenous, where the robot plans to actively uncover its value; the happy/unhappy variable is exogenous, because the expression recognition algorithm can be running all the time and the robot does not need to consider this variable until its value change is detected.

B. Multimodal Dialog Management

Our control framework is presented in Fig. 3. Comparing our flow chart with that of standard POMDPs [1], our system has an *event reasoner* (ER) that is inserted between *state estimator* (SE) and policy π for action selection. ER is mainly used for two purposes:

1) Initializing beliefs at the very beginning; and
2) Updating beliefs given events being detected.

We say an *exogenous event* (or event) happens when there is a value change of an exogenous domain variable. At the beginning of a conversation, ER reasons about both logical and probabilistic commonsense knowledge (using P-log [13]), including the values of exogenous domain variables and the dependencies of endogenous variables. This reasoning step enables ER to compute a distribution over all possible partial states. This distribution serves as the initial belief for our POMDP. After that, we follow the standard process to use the precomputed POMDP policy for action selection and update the belief based on observations. In each iteration, if there is no event detected, the updated belief is directly used for action selection. Otherwise, we take the current belief as the prior distribution and use the new values of exogenous domain variables to compute the posterior distribution of the POMDP belief (over partial states).

The probability of the current partial world being in partial state s (a POMDP state) is computed based on the distribution over the combinations of values of endogenous domain variables:

$$\Pr(s) = \Pr(V_0^{en} = v_0^{en}, V_1^{en} = v_1^{en}, \cdots, V_{m-1}^{en} = v_{m-1}^{en}) \quad (1)$$

where V_i^{en} is the ith endogenous domain variable, v_i^{en} is the value of that variable, and $\mathbf{v}^{en} = [v_0^{en}, \cdots, v_{m-1}^{en}]$ specifies s. There are totally m endogenous domain variables. It should be noted that these endogenous variables are not necessarily to be independent. Continuing our museum guide example, a visitor who is interested in exhibits about the College of Engineering (*college*) is likely to be interested in researchers working on engineering (*person*). The standard POMDP framework does not support reasoning about dependencies between states (or domain variables), whereas it is achieved via the P-log-based event reasoner in MDM.

Given the values of exogenous domain variables and the belief over partial world (POMDP) states, we can use the standard Bayes' theorem to update the POMDP belief (this process is automatically done in ER):

$$\Pr(s|\mathbf{v}^{ex}) = \eta \cdot \Pr(\mathbf{v}^{ex}|s)\Pr(s) \quad (2)$$

where $\mathbf{v}^{ex}$ is a vector of exogenous domain variable values, and η is a normalizer.

The observation model of exogenous events is $\Pr(\mathbf{v}^{ex}|s)$ and can be computed as follows

$$\Pr(\mathbf{v}^{ex}|s) = \Pr(v_0^{ex}|s) \cdot \Pr(v_1^{ex}|s) \cdots \Pr(v_{n-1}^{ex}|s) \quad (3)$$

where n is the number of exogenous domain variables.

It should be noted that we do not assume that exogenous domain variables are independent. We use *default reasoning* to assign each exogenous variable a default value. When the true value becomes available, this default value can be

```
 1: procedure MDM
 2:     while not in a terminal state do
 3:         Initialize belief b_t by commonsense reasoning in ER
 4:         Select action a using policy π based on belief b
 5:         Make observation o from the world
 6:         Update belief b based on prior belief, a, and o
 7:         if exists(exogenous events) then
 8:             Update b based on prior belief and exogenous events
 9:         end if
10:     end while
11: end procedure
```

Fig. 4: Procedure of our multimodal dialog management (MDM) framework.

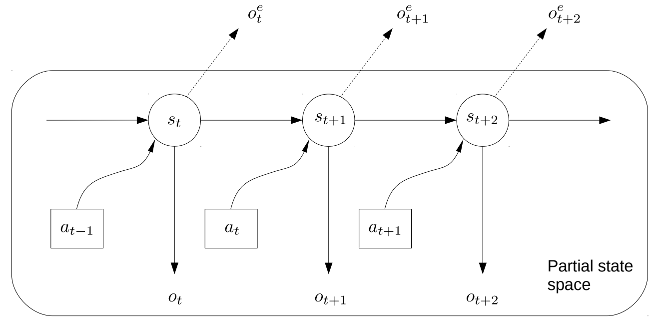


Fig. 5: A visualization of the transitions, actions, and observations in the proposed spoken dialog system.

overwritten without breaking the consistency of the whole reasoning system. In this way, each exogenous domain variable always has a "known" value (default or observed), making it unnecessary to model the dependencies between exogenous domain variables. Modeling the dependencies would improve the value estimation of endogenous variables – this will be investigated in future work.

After updating the belief over partial state space, we use the updated belief for selecting the next action $\pi : b' \rightarrow a$. Intuitively, this new belief includes all information collected from the history and the information from the exogenous domain changes. Finally, the belief can be updated based on observations over the partial world state space:

$$\Pr(s') = \eta \cdot O(o \mid s', a) \sum_{s \in S} T(s' \mid s, a) \Pr(s) \qquad (4)$$

where $O(o \mid s', a)$ is the observation model that defines the probability of receiving an observation in state s' after taking action a (in our case, it describes the speech recognition reliability). T is the state transition model that specifies the probabilities transition system of a POMDP.

Fig. 4 summarizes the procedure of our MDM framework. It is different from standard POMDP control loop in that, when exogenous events are detected (such as a change of human facial expression), belief b is updated based on the exogenous events in Lines 7-9. This enables a robot to update its dialog belief state using multimodal (both language and non-language) perception. Fig. 5 visualizes the belief update process, where the robot actively acquires information via observations of $\langle o_t, o_{t+1}, \cdots \rangle$ and the belief state can be passively updated via observations over exogenous events of $\langle o^e_t, o^e_{t+1}, \cdots \rangle$ at the same time.

IV. IMPLEMENTATION

Our MDM dialog management framework has been implemented on a mobile robot that serves as a tour guide in the Museum of USTC (University of Science and Technology of China). The museum includes gallery rooms, where each gallery room includes a set of exhibits. The exhibits can be characterized using the properties of *identity* (the main person involved in the exhibit, such as Adams, Brown and Clark), *college* (the main college involved in the exhibit, such as Engineering, Education and Business), and *years* (the years that corresponds to the exhibit, such as 50s-70s, 70s-90s, and 90s-present). It should be noted that, given an exhibit, a property might not be applicable. For instance, exhibit $E1$ might be about Professor Brown who served as the Provost of USTC in the 80s, so this exhibit is not corresponding to any colleges. To account for such situations, we add a special value of *na* (standing for "not applicable") into the range of each property.

Since the robot is only interested in which exhibit(s) a visitor is interested in, it makes sense to include only the variables of "identity" , "college", and "years", as endogenous domain variables, in the partial state space.

- V^i: the identity of the main person of an exhibit.
- V^c: the college to which an exhibit is most related.
- V^y: the years of an exhibit, e.g., $50s$, $70s$ and $90s$.

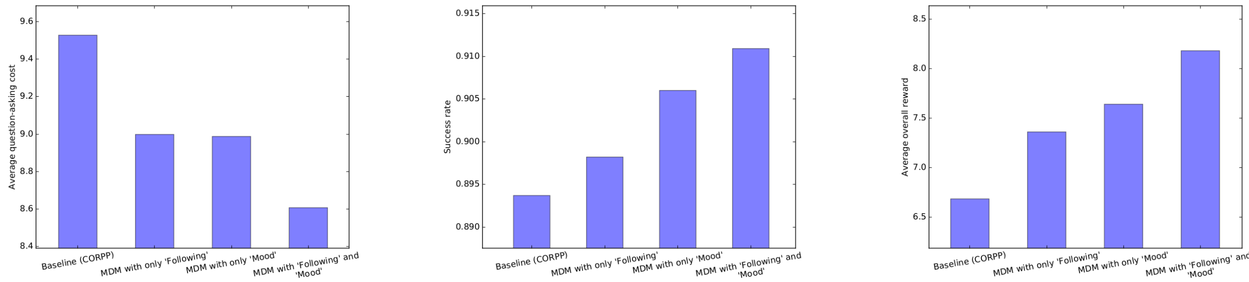
Focusing on the relationships among the endogenous variables, we find a visitor who is interested in college C is likely to be interested in a person (with identity I) affiliated to college C. Accordingly, we model the causal dependency of V^i on V^c. For instance, the probability of a visitor being interested in identity I given the fact this person is affiliated to college C and the visitor is interested in C is set 0.8 arbitrarily in experiments.

Three other variables are not modeled in the partial state space, including "following", "age", and "mood".

- V^f: a boolean variable that represents if a visitor is following the robot.
- V^a: the age of a visitor (<20, $20-40$, and >40).
- V^m: the mood of a visitor (happy and unhappy).

The three endogenous variables depend on variable V^f (following): a visitor staying at the previous exhibit (instead of following the robot to the next) indicates that the visitor is interested in at least one property of the previous exhibit (in probability). The endogenous variables also depend on variable V^m (mood): an unhappy face of a visitor indicates that the visitor is not interested in any property of the current exhibit (in probability).

The goal is to create a spoken dialog system to enable the robot to talk to visitors to find their interests, use vision-based mood recognition and Lidar-based localization to help estimate the visitor's interests, and physically guide visitors to exhibits of the most interest.



(a) Cost. (b) Success. (c) Reward.

Fig. 6: Incorporating sensing modalities of vision and Lidar into a robot spoken dialog system. The x-axis represents the four dialog management strategies that use: language only, language and Lidar, language and vision, and all three. Our MDM framework generally reduces the question-asking cost, as shown in (a), and increases the interest identification accuracy, as shown in (b), in comparison to the baseline.

V. MULTIMODAL PERCEPTION

In this section, we first give an overview of our robot's hardware, and then describe the perception modules on which we build our spoken dialog system.

Our robot is based on a two-wheel driving chassis of size $62 \times 53 \times 32$ in centimeter. A lifting system is mounted on the chassis and is attached to the robot's upper body. Assembled with the upper body is a six degrees-of-freedom (DOF) arm. It is able to reach objects over 83 centimeters far from the mounting point and the maximum payload is about 500 grams when fully stretched. The robot is powered by a 20Ah battery that guarantees the robot a continuous run of at least one hour (as a tour guide robot, the time is significantly longer as the robot spends most of the time talking instead of moving). As for real-time perception needs, our robot is equipped with a Kinect sensor, a high-resolution CCD camera, two Lidar sensors and a microphone. A workstation laptop is used to meet the computational need.

a) Speech recognition: The speech recognition module converts speech signals (voice) into natural language text. For speech recognition, we use directional microphone hardware to reduce background noise and we use the Speech Application Programming Interface (SAPI) package for speech recognition.[2] On this basis, we have encapsulated it into a ROS [23] package in our code base.

b) Emotion recognition: The emotion recognition module is realized via a facial expression recognition package. We first track the speaker's face using OpenFace, an open source face detection and recognition software building on deep neural networks [24]. Then, the saved face image is sent to our emotion recognition module that is built on Emotion Recognition Application Programming Interface (ERAPI) from Microsoft Cognitive Services.[3] Finally, ERAPI recognizes the emotion(s) of one or more people in the image, based on which a happy/unhappy emotion is returned.

c) Age recognition: In order to recognize human age, we save the face image and send it to the Face API of (again) Microsoft Cognitive Services. The Face API returns face locations and face attributes in comma-separated strings like "returnFaceAttributes=age,gender", from which human age is extracted. Other attributes supported by the API include age, gender, headPose, smile, facialHair, and glasses.

[2]http://www.iflytek.com/en/
[3]https://www.microsoft.com/cognitive-services/en-us/

VI. EXPERIMENTS

The MDM framework has been implemented both in simulation and on a real robot. In a museum environment, the robot (simulated or physical) needs to estimate the interest of a visitor via multimodal perception, and decide where (which exhibit) to guide the visitor to. After a brief self-introduction, the robot actively asks questions to acquire information about the visitor's interests (such as "*Are you interested in the history of the College of Education?*"), while at the same time passively using vision and Lidar to estimate the visitor's interests. For instance, an unhappy face indicates the visitor might not be interested in the current exhibit, and not following the robot to the next exhibit indicates the visitor is interested in at least some aspects of the last exhibit. *The goal is to evaluate if (and how much) the non-language perceptual capabilities can contribute to the dialog management, using the MDM framework.*

a) Simulation experiments

In simulation experiments, the robot interacts with visitors in a gallery room that has a set of nine exhibits. For instance, the first exhibit is about Dr. Brown's achievement in the $70s$ when Dr. Brown was with the College of Engineering. Imperfect sensing capabilities are simulated. For instance, when the visitor answers "yes/no" to the robot's question "*Are you interested in the history of the College of Education?*", the robot correctly recognizes the answer in probability 0.8 unless specified otherwise.

We simulate human emotion (reflected by facial expression) in such a way that the more properties of the current exhibit the human is not interested in, the higher probability that the visitor has an unhappy face: the probability ranges from 0.2 (zero uninterested property) to 0.8 (three uninterested properties). At the same time, we simulate that a "not following" behavior indicates the visitor is interested in some properties of the previous exhibit (more properties result in a high probability). We cannot evaluate the statistical

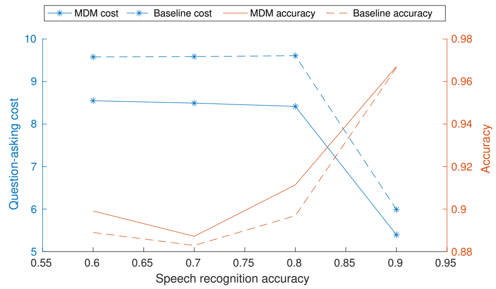


Fig. 7: Question-asking cost (left y-axis) and accuracy of visitor interest identification (right y-axis). Our MDM framework is especially useful when the speech recognizer is unreliable (the left end of the curves).

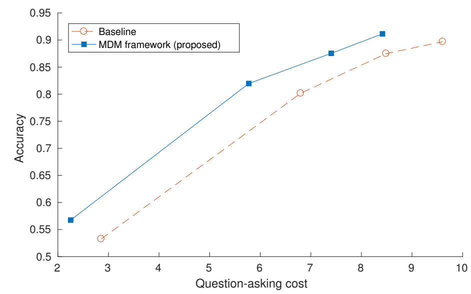


Fig. 8: Adjusting the trade-off between cost and accuracy by changing the success bonus and failure penalty ($[5, 10, 15, 20]$ from left to right for each curve). Our MDM framework consistently performs better than the baseline given different trade-offs between cost and accuracy.

significance in following experiments, because we sample human interests in each trial, making direct comparisons over different trials infeasible. Each data point corresponds to 1000 trials in all figures.

Fig. 6 reports the results of the first set of experiments in simulation. The evaluation is based on question-asking cost (time needed before the robot starts to physically guide the visitor), success rate (correctly identifying the exhibit that the visitor is most interested in is a success), and overall reward (that combines question-asking cost and success bonus/failure penalty). We can see our MDM framework generally performs better than the baseline (CORPP [5]) in all three criteria (question-asking cost, overall reward, and success rate). Especially, when all three sensing modalities are used, the robot produces the best accuracy in success rate, while requiring the least question-asking cost. The results support our hypothesis that MDM improves spoken dialog management (in both accuracy and efficiency) by leveraging vision-based and Lidar-based perception.

We further evaluate the performance of MDM given different speech recognition accuracies. Our hypothesis is that multimodal perception plays a more significant role when speech recognition is more unreliable. For the sake of easy analysis of the results, we only change the recognition accuracy of visitors' "yes/no" answers. Fig. 7 shows the question-asking cost and success rate (accuracy) given different speech recognition accuracies. First, we can see given a better-quality speech recognizer, both the baseline and our MDM perform better (in both cost and accuracy). Second, the performance difference between the baseline and MDM decreases as the speech recognizer becomes more reliable. The results support our hypothesis that our MDM framework is especially useful in environments where good-quality speech recognition is challenging.

In the last set of simulation experiments, we evaluate the performance of our MDM-based dialog manager given different success bonus, R^+ (a positive real number), and failure penalty, R^- (a negative real number). Their values are adjusted in a way that $|R^+| = |R^-|$. Intuitively, these values can be used to adjust the trade-off between visitor interest identification accuracy and question-asking cost. Fig. 8 reports the results, where the value of R^+ varies from 5 to 20. We can see MDM generally performs better than the baseline (CORPP) given different values of R^+ and R^-. The results can serve as a reference in deciding how long a conversation needs to be. For instance, it makes sense to give a longer time to a conversation with senior people (thanks to our age recognition module) who might be physically challenged in walking, and hence the visitor's interest can be more accurately identified (minimizing walking distance).

Fig. 9: Our KeJia robot interacting with a "visitor" in an artificial museum.

b) Trials on a real robot

Our MDM framework has been implemented on a KeJia robot. Figure 9 shows a picture of the robot interacting with a visitor in artificial museum environment.[4] Section IV detailed the implementation of MDM on our KeJia robot, including the components for speech recognition, emotion recognition, and age recognition, in addition to standard robotic capabilities such as localization and navigation.

We present the spoken language in a human-robot conversation, where the conversation starts near an exhibit about $\langle Adam, Engineering, Seventies\rangle$, whereas the visitor's real interest is $\langle Brown, Education, Nineties\rangle$.

- Robot: *Hi! I'm a tour guide robot. Do you need my assistance to find your favorite exhibit?*

* Visitor: *Sure!*

- Robot: *Sounds good. May I ask which college of our university you are the most interested in?*

[4] Even though the project is on a museum guide robot, the experiments were all conducted in our research lab so far, where we try our best to capture the challenges from the real museum environment.

* Visitor: *I am interested in the College of Education.*
- Robot: *Did you say that you are interested in the College of Education?*
* Visitor: *Yes!*
- Robot: *I see. Are you interested in its history in the nineties?*
* Visitor: *Yes!*
- Robot: *Did you say you are interested in its history in the nineties?*
* Visitor: *Yes!*
- Robot: *Got it! Please follow me.*

It should be noted that the robot detected an unhappy facial expression, right before it received the underlined human feedback. As a result, the robot has a prior belief that the visitor might not be interested in exhibits about *Adam*, *Engineering*, or *Seventies* (the properties of its current exhibit), and only four questions (not considering the self-introduction one) are needed before the robot is confident about the visitor's interest. In comparison, using the baseline approach, the robot would have asked five questions on *college*, confirming college, person *Brown*, *years*, *confirming years*. An illustrative trial can be viewed online.[5]

VII. CONCLUSIONS

In this paper, we develop a multimodal dialog management (MDM) framework that leverages multimodal perception and commonsense reasoning for robot spoken dialog systems. In addition to language, sensing modalities of vision (for facial expression recognition) and Lidar (for localization) have been integrated into our dialog manager. We build our commonsense reasoner on P-log, a declarative language that is good at both logical and probabilistic commonsense reasoning. As a result, MDM enables the robot to actively acquire information via spoken language and passively update its belief based on observations over exogenous events at the same time. Experiments conducted in simulation and on a real robot show that MDM reduces the communication cost while increasing accuracy of acquired information.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China under grant No. U1613216. A portion of this work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), AFOSR (FA9550-14-1-0087), Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1, pp. 99–134, 1998.
[2] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.
[3] S. Singh, D. Litman, M. Kearns, and M. Walker, "Optimizing dialogue management with reinforcement learning: Experiments with the njfun system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 105–133, 2002.
[4] N. Roy, J. Pineau, and S. Thrun, "Spoken dialogue management using probabilistic reasoning," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 93–100.
[5] S. Zhang and P. Stone, "Corpp: Commonsense reasoning and probabilistic planning, as applied to dialog with a mobile robot," in *Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
[6] N. Gopalan and S. Tellex, "Modeling and solving human-robot collaborative tasks using pomdps," in *Robotics: Science and Systems 2015: Workshop on Model Learning for Human-Robot Communication*, 2015.
[7] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: an anytime algorithm for pomdps," in *Proceedings of the 18th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc., 2003, pp. 1025–1030.
[8] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based pomdp solvers," *Autonomous Agents and Multi-Agent Systems*, pp. 1–51, 2013.
[9] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.
[10] R. Reiter, "A logic for default reasoning," *Artificial intelligence*, vol. 13, no. 1-2, pp. 81–132, 1980.
[11] M. Gelfond and Y. Kahl, *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge University Press, 2014.
[12] E. Erdem, M. Gelfond, and N. Leone, "Applications of answer set programming," *AI Magazine*, vol. 37, no. 3, pp. 53–68, 2016.
[13] C. Baral, M. Gelfond, and N. Rushton, "Probabilistic reasoning with answer sets," *Theory and Practice of Logic Programming*, vol. 9, no. 01, pp. 57–144, 2009.
[14] X. Chen, J. Xie, J. Ji, and Z. Sui, "Toward open knowledge enabling for human-robot interaction," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 100–117, 2012.
[15] M. Veloso, J. Biswas, B. Coltin, and S. Rosenthal, "Cobots: Robust symbiotic autonomous mobile service robots," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
[16] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrová, J. Young, J. Wyatt, D. Hebesberger, T. Körtner, *et al.*, "The strands project: Long-term autonomy in everyday environments," *IEEE Robotics and Automation Magazine*, 2016.
[17] P. Khandelwal, S. Zhang, J. Sinapov, M. Leonetti, J. Thomason, F. Yang, I. Gori, M. Svetlik, P. Khante, V. Lifschitz, *et al.*, "Bwibots: A platform for bridging the gap between ai and human–robot interaction research," *The International Journal of Robotics Research*, 2017.
[18] S. Tellex, R. A. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics." in *Robotics: Science and systems*, vol. 2, no. 3, 2014.
[19] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning multi-modal grounded linguistic semantics by playing "i spy"," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, New York City, 2016, pp. 3477–3483.
[20] H. Kurniawati, Y. Du, D. Hsu, and W. S. Lee, "Motion planning under uncertainty for robotic tasks with long time horizons," *The International Journal of Robotics Research*, vol. 30, no. 3, pp. 308–323, 2011.
[21] S. Zhang, P. Khandelwal, and P. Stone, "Dynamically constructed (po) mdps for adaptive robot planning," in *Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.
[22] J. Pearl, *Causality*. Cambridge university press, 2009.
[23] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, 2009, p. 5.
[24] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[5] https://youtu.be/8fBR6D_HLgo