Semantic-Spatial Reasoning for Unique Role Recognition in Dynamic Environment

Chule Yang[1], Yufeng Yue[1], Jun Zhang[1], Mingxing Wen[1] and Danwei Wang[1]

***Abstract*—Autonomous reasoning is a crucial problem for intelligent system when dealing with unspecified targets in dynamic environment. In this paper, a probabilistic reasoning method is proposed that associates semantic labels with spatial observations to recognize the person who is distinctive in attachments and movements in the scene. Roles are categorized as a binary representation of the unique target and others. Two types of observation models, namely, *Object Existence Model (OEM)* and *Human Action Model (HAM)*, have been established by analyzing the corresponding semantic-interaction and spatio-temporal features. Then, OEM and HAM results of each person are compared with the overall distribution, respectively. Finally, the role can be comprehensively inferred through the fusion of two observation models. Experiment results show that the proposed method is feasible in a variety of scenarios.**

***Index Terms*—Semantic Perception, Hierarchical Probabilistic Reasoning.**

I. Introduction

Autonomous systems are highly anticipated operating in the increasingly complex environment and understand human behavior [1]. However, in highly uncertain and dynamic environments, potential targets may only have an abstract description or an ambiguous appearance. Thus, the ability to infer implicit information from a limited perception is beneficial for robots to perform high-level task planning. Moreover, due to the abstractness, dynamic changes and potential occlusions, recognition methods that rely on low-level features or single attribute are inefficient. To solve these problems, it is necessary to integrate multimodal information from both spatial and semantic perspectives to tolerate the uncertainty, achieve comprehensively contextual understanding, and reasoning from it.

Therefore, this research aims to develop a scalable and flexible reasoning strategy that can autonomously exploit semantic labels and spatial observations to recognize the person who is distinctive in attachments and movements in the scene. As an extension of previous work [2], this research categorizes roles as a binary representation of the unique target and others by dynamically analyzing semantic-interaction features and spatio-temporal features in the 3D space. After comparing the observation result of an individual with the overall distribution, a hierarchical graphical model is developed to integrate multimodal information into a comprehensive decision scheme by fusing of multiple observation models. Roles are categorized as a binary representation; namely, the unique target ($\mathcal{I}_1$) and others ($\mathcal{I}_2$), $I \in \{\mathcal{I}_1, \mathcal{I}_2\}$.

*This research was partially supported by the ST Engineering - NTU Corporate Lab through the NRF corporate lab@university scheme.

[1]The authors are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (e-mail: {yang0438, yyue001, jzhang061, mingxing001, edwwang}@ntu.edu.sg).

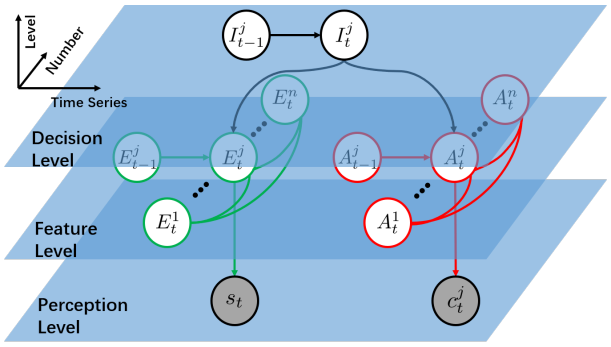


Fig. 1. The hierarchical graphical model for recognizing the role of person j at the time t.

II. Role Inference from Semantic-Spatial Fusion

This problem is formulated as a Bayesian filter, where the hidden variable $I \in \mathcal{I}$, is the role of a person that currently inferred. The robot observes the person's features including semantic-interaction and spatio-temporal features, G, and at each time step estimates a distribution over the current role. To estimate this distribution, we alternately perform a measurement update and a time update.

$$P_j(I_t|G_{1:t}) \propto \underbrace{P_j(G_t|I_t)}_{\text{Decision Model}} \cdot \underbrace{P_j(I_t|G_{1:t-1})}_{\text{Previous Information}} \tag{1}$$

The measurement update contains all the channels of information by fusing all available observations from semantic-interaction features (Θ^{SI}) and spatio-temporal features (Θ^{ST}).

$$P_j(G_t|I_t) \propto \prod_{i=1}^{m} \underbrace{P_j(\Theta^{SI}_{i,t}|I_t)}_{\text{Semantic-Interaction}} \cdot \prod_{k=1}^{n} \underbrace{P_j(\Theta^{ST}_{k,t}|I_t)}_{\text{Spatio-Temporal}} \tag{2}$$

The time update contains the transition probability from the previous role to the current role and the previous belief, where the current role is dependent either on the previous role or the previous measurement.

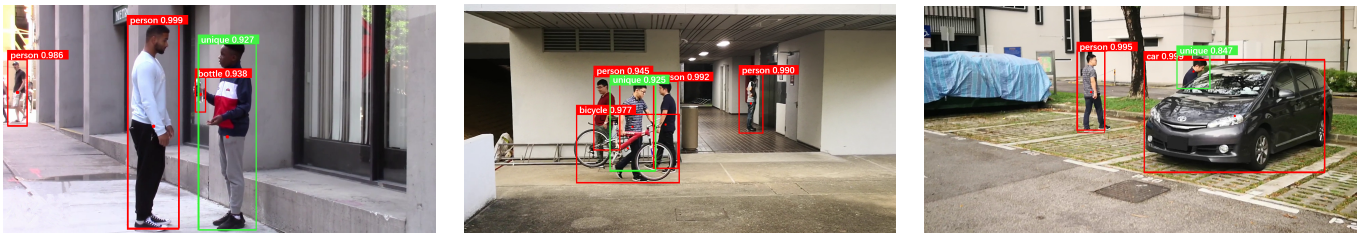
$$P_j(I_t|G_{1:t-1}) = \sum_{I_{t-1}\in\mathcal{I}} \underbrace{P_j(I_t|I_{t-1})}_{\text{Decision Transition}} \cdot \underbrace{P_j(I_{t-1}|G_{1:t-1})}_{\text{Previous Belief}} \tag{3}$$

A. Decision Level Inference

The decision model calculates the probability of the respective feature given the role. Each feature is a set of performed action (A) and the existence of objects (E). By taking the logarithm, decision model for a single person j is formed as:

$$P_j(G_t|I_t) \propto \log(P_j(G_t|I_t)) = L_j(A_t|I_t) + L_j(E_t|I_t) \tag{4}$$

where L denotes the function after taking the logarithm.



(a) People drinking in public. (b) People carrying (stealing) the bicycle. (c) People behind (damaging) the car.

Fig. 2. Three different outdoor scenarios for unique role recognition. The green bounding box is the recognized unique target in the scene.

B. Observation Level Inference

The observation model calculates the probability of the observation given corresponding features. Each observation is a set of object's bounding box (s) and person's position (c). Two observation models are established as follows.

1) Object Existence Model (OEM): The existence of an object is modeled as the intersection area between a person j ($S^h_{j,t}$) and the specified object (S^o_t) at each time step. Let $P_j(s_t|E_t)$ define the probability of the person j carrying the specified object at time t, which is set to be proportional to the area of the intersection: $P_j(s_t|E_t) \propto \frac{S^h_{j,t} \cap S^o_t}{S^o_t}$.

2) Human Action Model (HAM): Human action is modeled as a positional difference ($c^h_{j,t} - c^h_{j,t-1}$) between two consecutive time steps. Let $P_j(h_t|A_t)$ define the probability of the person j taking a specific action, which is set to be proportional to a function f of the movement vector: $P_j(c_t|A_t) \propto \frac{1}{1+e^{f(c^h_{j,t} - c^h_{j,t-1})}}$.

3) Comparison with Overall Distribution: Final inference results of OEM and HAM are calculated separately by measuring the degree of deviation of a single inferred result from the overall probability distribution. Let $x_{j,t} \in \{P_j(s_t|E_t), P_j(c_t|A_t)\}$ define the random variable (inferred result from either OEM or HAM) associated with the node (person) j at the time step t. ϕ_j is the degree of derivation of $x_{j,t}$ from the overall distribution, which can be formed as:

$$\phi_j(x_{j,t}) = \frac{1}{Z} \cdot \prod_{c \in \mathcal{C}} x_{c,t}, \quad Z = \sum_{x_{j,t}} \prod_{c \in \mathcal{C}} x_{c,t} \tag{5}$$

C. Final Role Inference Result

Similar to (1), the final role of each person can be obtained from OEM and HAM, as shown in (6) and (7), respectively.

$$L_j(E_t|I_t) \propto \phi_j(P_j(s_t|E_t)) \cdot P_j(E_t|s_{1:t-1}) \tag{6}$$

$$L_j(A_t|I_t) \propto \phi_j(P_j(c_t|A_t)) \cdot P_j(A_t|c_{1:t-1}) \tag{7}$$

1) Models Fusion: As illustrated in Fig. 1, by acquiring the information from all channels, the entire recognition process forms a hierarchical graphical model. Thus, the role of each person can be finally inferred as follows:

$$P_j(I_t|G_{1:t}) \propto \underbrace{(\underbrace{L_j(E_t|I_t)}_{\text{OEM}} + \underbrace{L_j(A_t|I_t)}_{\text{HAM}})}_{\text{Decision Model}} \cdot \sum_{I_{t-1} \in \mathcal{I}} \underbrace{P_j(I_t|I_{t-1})}_{\text{Decision Transition}} \underbrace{P_j(I_{t-1}|A_{1:t-1}, E_{1:t-1})}_{\text{Previous Belief}} \tag{8}$$

TABLE I

EVALUATION OF UNIQUE ROLE RECOGNITION IN THREE DIFFERENT SCENARIOS. P AND R DENOTE PRECISION AND RECALL, RESPECTIVELY.

Method	Scenario 1		Scenario 2		Scenario 3	
	P	R	P	R	P	R
OEM	74.63	27.03	87.22	88.90	97.05	91.86
HAM	63.10	53.00	78.96	76.12	27.40	14.91
OEM+HAM	**92.17**	**81.82**	**88.85**	**89.14**	**98.21**	**99.79**

III. EXPERIMENTS

In this experiment, the algorithm is tested under different sizes of specified objects to assess the performance. Besides, the parameters and observation features of the probabilistic model can be easily modified according to different cases.

Three different scenarios were collected from the street, college campus and parking lots, respectively. As shown in Fig. 2, each scenario has an specified object, that is, the bottle used to recognize people drinking in public in scenario 1; the bicycle used to recognize people carrying (stealing) the bicycle in scenario 2; the car used to recognize people behind (damaging) the car in scenario 3. The experiment first evaluates two observation models separately and then combines them together for another evaluation.

As shown in Table I, the result validated that the proposed method could tolerate single model failure and fuse them into a comprehensive decision scheme to achieve higher accuracy. The proposed method OEM+HAM relied more on HAM when the size of the specified object was small since OEM had difficulty detecting small objects.

IV. CONCLUSIONS

In this paper, a hierarchical probabilistic reasoning approach was proposed that enables the intelligent system to associate semantic labels with spatial observations to recognize the role of people and to distinguish the unique target from others in unknown environments. This semantic-spatial reasoning framework can be used in many other application, such as human intention estimation and goal-oriented navigation.

REFERENCES

[1] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE transactions on circuits and systems for video technology*, vol. 23, no. 11, pp. 1993–2008, 2013.

[2] C. Yang, Y. Zeng, Y. Yue, P. Siritanawan, Z. Jun, and D. Wang, "Knowledge-based role recognition by using human-object interaction and spatio-temporal analysis," in *Robotics and Biomimetics (ROBIO), 2017 IEEE International Conference on*. IEEE, 2017.