Knowledge-based Visual Context-Aware Framework for Applications in Robotic Services

Doosoo Chang
Korea Telecom R&D Center, Korea
doosoo.chang@kt.com

Bohyung Han
Seoul National University, Korea
bhhan@snu.ac.kr

Abstract

Recently, context awareness in vision technologies has become essential with the increasing demand for real-world applications, such as surveillance systems and service robots. However, implementing context awareness with an end-to-end learning-based system limits its extensibility and performance because the context varies in scope and type, but related data are mostly rare. To mitigate these limitations, we propose a visual context-aware framework composed of independent processes of visual perception and context inference. The framework performs logical inferences using the abstracted visual information of recognized objects and relationships based on our knowledge representation. We demonstrate the scalability and utility of the proposed framework through experimental cases that present stepwise context inferences applied to robotic services in different domains.

1. Introduction

The need for context awareness has emerged not only in the Internet of Things (IoT) or in recommender systems [1, 23] but also in vision applications, owing to the increase in the use of real-world services, such as surveillance systems and service robots [10, 11]. Context awareness in vision implies a high-level understanding of complete visual scenes that are beyond simple recognition, similar to human cognition. Its scope of recognition varies from general situations, such as children having a picnic (rather than just the actions of sitting or running around), to anomalies, such as traffic accidents and intrusions.

Therefore, scalability and low data dependency are important factors for context awareness, because its domains are extremely diverse to enumerate, but the related dataset is mostly rare or non-existent [8]. From this point of view, when implementing context awareness in an end-to-end learning-based system, the feasibility as a real-world application is significantly reduced. Although existing methods [9, 28] exhibit quality performance in specific tasks, additional learning on the entire system is continuously required as an extension of those tasks, leading to unavoidable additional costs and a reduction in effectiveness. This constraint can be maximized especially, particularly when the network-based methods are applied to service robots that need to adapt to various environmental changes.

To mitigate this limitation, we propose a visual context-aware framework (Vis-CAF) structured to maintain independence in the processes of visual perception and context inference for a flexible real-world application of context-aware tasks. The perception process visually recognizes objects/relationships, and the inference process infers their contexts based on a high-level knowledge representation using the abstraction of perceived results. This Vis-CAF structure ensures quality recognition performance in various categories using relatively rich object/relationship data. In addition, the Vis-CAF structure enables the flexible application of additional context-aware tasks through a simple extension of knowledge concepts and reasoning rules. Although the recognition models may need to be re-learned for added object/relationship categories as context-aware tasks increase, they can be fine-tuned at lower costs and higher performances compared with those of end-to-end systems.

The main contributions of the Vis-CAF are as follows:

Scalability and flexibility. Vis-CAF can maintain its scalability and flexibility even with variations in context-aware tasks and service domains because of the independent perception and inference processes. The structure of Vis-CAF facilitates an easy application of variations through simple customization at the knowledge level.

Modularity. Because the processes of Vis-CAF are performed by data flows between the individual modules, the existing detection models can be applied with minimal restrictions. Further, modularity enables parallel management with regard to independent development and advancement.

Availability in multi-modality. Vis-CAF can

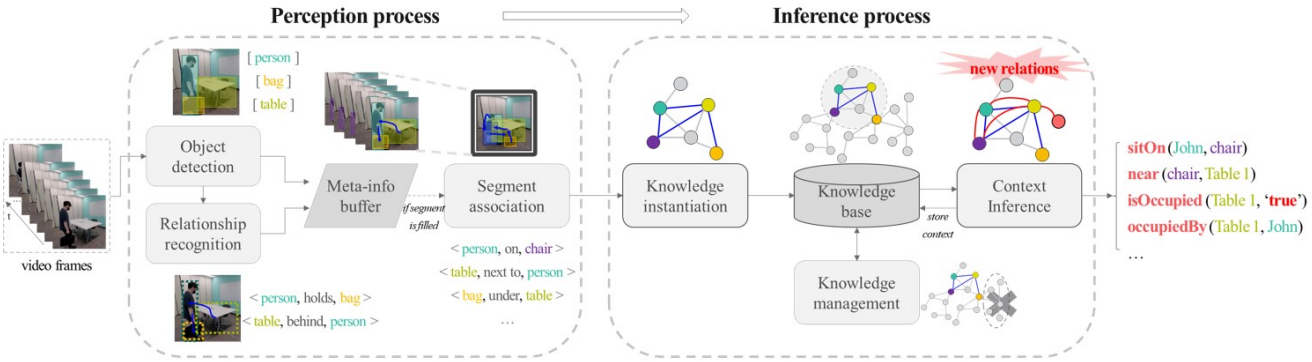

Figure 1. Architecture of Vis-CAF. The colors in boxes, words and vertices represents the same objects in the example.

leverage sensed data from modalities, other than vision, such as audio, text, and signal, by their abstractions into the common domain knowledge and definitions of corresponding reasoning rules. If the knowledge is shared in real time, multi agent service is also feasible.

2. Proposed Framework

2.1. Overall Structure of Vis-CAF

Vis-CAF aims to infer appropriate contexts by leveraging the information perceived from an input video. As shown in Figure 1, it comprises two main processes: the perception and inference processes.

Perception Process The perception process focuses on the visual analysis of each image frame in a video based on deep neural networks, and the integration of the analyzed results on consecutive frames (indicated as a *segment*). The perception process comprises three modules: object detection, relationship recognition, and segment association. The purpose of the first perception module is to detect and classify objects of interest using localization from a raw two-dimensional (2D) image. The classified categories and the corresponding bounding boxes are stored in a frame buffer and used as inputs for relationship recognition. Any model that outputs a similar format of detection can be applied to the first module because modeling object detection is not considered in the primary scope of this study.

Based on the output from object detection, the second perception module performs relationship recognition that recognizes the inter relationships of detected object pairs using neural network-based prediction. The detailed model architecture and procedures in the relationship recognition module are presented in section 2.2.

The analyzed results per image frame obtained from the perception modules are stacked in a meta-information buffer with a fixed size, implying that the old results are removed. As the meta-information buffer stacks a new segment, consecutive image frames of the same temporal length using a sliding window, the segment association module performs two functions: tracking and association.

First, the segment association module successively identifies a tracking ID of each object in the new segment by applying a tracking algorithm to all meta-information of the tracked objects in the buffer. Note that the meta-information buffer contains not only objects in the new segment, but also those in previous segments that are recently analyzed because the temporal length of the buffer is set to be longer than that of the segment. After tracking, the module associates each pair of tracked objects by the statistical deduction for its relationship among the entire frames of the new segment, and excludes the pairs that are below a frequency threshold. Consequently, the associated meta-information of the objects and relationships for a segment is used to infer potential contexts through the inference process.

Inference process The inference process comprises three modules: knowledge instantiation, context inference, and knowledge manager, as shown in Figure 1. The knowledge instantiation module aims to abstract low-level meta-information into high-level domain knowledge represented in the web ontology language OWL [2] format. The module maps the detected objects to individuals in a knowledge namespace by matching their locations or identities based on SPARQL [22], and generates temporary individuals denoting as unidentified for mismatched objects. Further, relationships are anchored in the corresponding ontological properties or concepts and linked to the mapped individuals of related objects by a *subject/object* relation. Because our knowledge is represented based on an event-oriented ontology, cognitions of objects and relationships are instantiated as event individuals. We present the details of the knowledge representation in section 2.3.

In our knowledge base, we define reasoning rules in the semantic web rule language (SWRL) [12], a language used to express rules for the semantic web. Vis-CAF performs context inference on individuals in the knowledge base based on reasoning rules and fundamental ontological characteristics of hierarchy and property. Vis-CAF au-

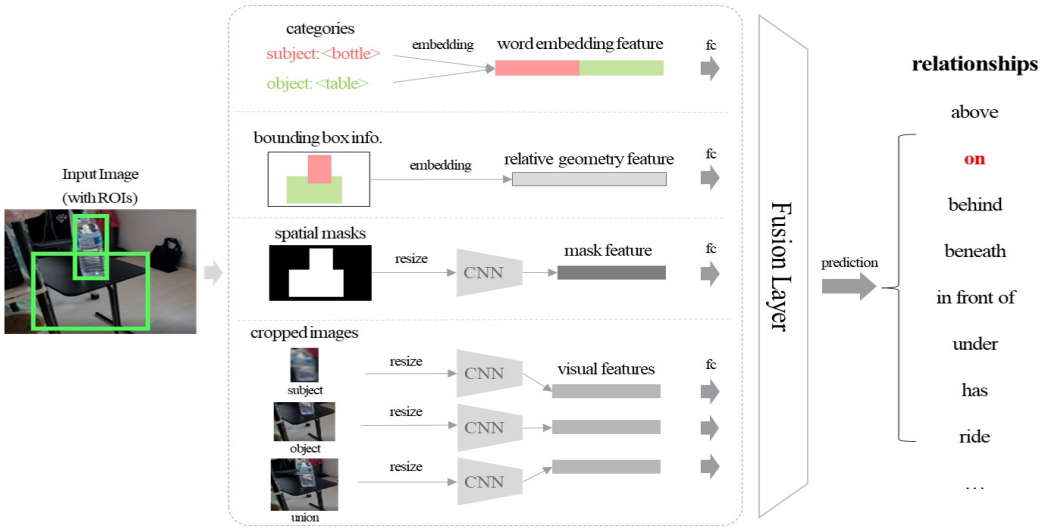

Figure 2. The modeling of visual relationship recognition in Vis-CAF with an example.

tomatically stores the newly inferred axioms and outputs some of those corresponding to the context of interest in a triple format. Because the point of context inference is the inclusion of ontological expressions and logics appropriate for particular situations, we present exemplarily defined rules through experiments.

The scale of knowledge base gradually increases as the number of input videos increases, causing high spatial and temporal costs for knowledge queries and reasoning. To reduce the spatial and temporal complexities, the knowledge manager automatically discards volatile individuals, such as unidentified people/objects and event individuals that have not been detected over a period of time. In addition, from a conceptual perspective, the knowledge manager also performs manual updates to the ontology schema and reasoning rules, and integration of external knowledge, such as environment maps for a service robot. These conceptual functions are not included in the automated context-aware process and are performed at ordinary times. Consequently, a series of inference processes for an input video can be performed continuously and without errors while maintaining the scale, consistency, and rules of knowledge.

2.2. Visual Relationship Recognition

Vis-CAF leverages the relationships between objects that are visually analyzed on a video for context inference as aforementioned. We model visual relationship recognition (VRR) as a deep neural network-based method that predicts the relationship between a pair of detected objects. Note that VRR is performed on each image in a segment. The model architecture of VRR is illustrated in Figure 2. VRR takes an image with the meta-information of the detected objects, including bounding boxes and categories, as input and outputs a predicted relationship category. We use four multi-modal features to maximally capture the information potential of an image: word embedding, relative geometry, mask, and visual features.

Word Embedding Feature To capture the language potential, we use GloVe [20] pre-trained on Wikipedia 2014 and Gigaword 5. The concatenated vector of the subject and object category embedding vectors is fed into a fully connected (FC) layer to learn the correlation. The use of the language-contextual information in the cognition of visual relationships was demonstrated as effective in previous studies [18, 16].

Relative Geometry Feature The relative geometry of objects involves a significant cue in VRR, particularly for categories of spatial relationships. Given bounding boxes for the subject and object as (x, y, w, h) and (x', y', w', h'), respectively, where (x, y) denotes a center point, the relative location is first calculated as $[\log \frac{|x-x'|}{w}, \log \frac{|y-y'|}{h}, \log \frac{w}{w'}, \log \frac{h}{h'}]$ based on the widely used form of spatial features [13]. Subsequently, we embed this 4 dimensional vector into a high-dimensional space with reference to [19]. The embedding process is based on the computation of cosine and sine functions of different wavelengths, which is a method of positional encoding in a transformer [26]. We set the wavelength factor to 1000 and the encoding dimension to 16 in the experiments.

Mask Feature Coordinate values have a limitation to fully represent the relativity of two objects. In addition, we use a binary spatial mask of the bounding boxes in an image, based on [16], which demonstrates its effectiveness. The spatial mask comprises two channels for the subject and object, with non-zero pixels in the bounding box area. Subsequently, the spatial mask that is down-sampled to a size of 32×32 is compressed using two convolutional layers, each of 5×5 and 3×3 filters and one FC layer.

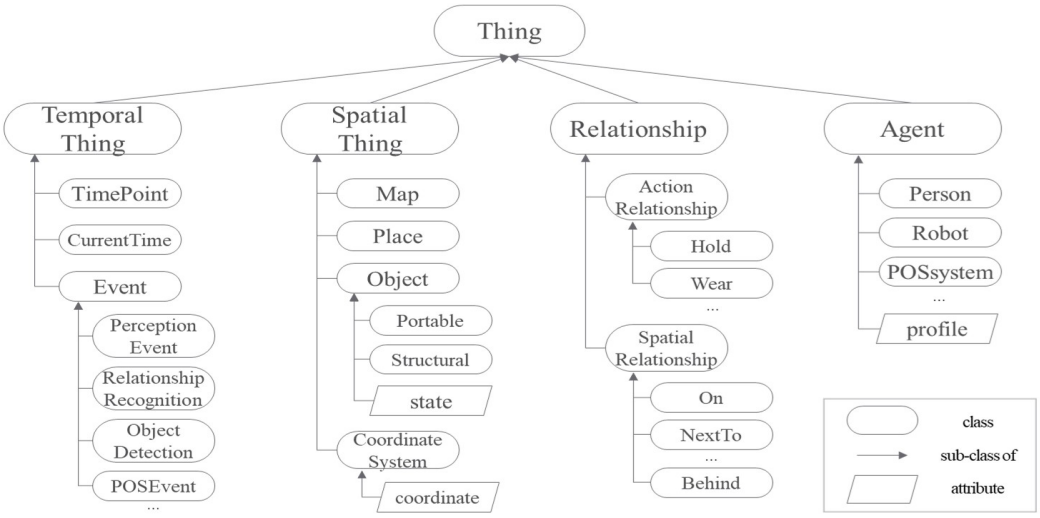

Figure 3. The main schematic structure of the ontology-based knowledge.

Visual Features We leverage visual appearances because the VRR model needs to distinguish a pose of an object and a topological state of an overlapped region. Each cropped image for the subject and object bounding boxes and their union is independently input to a convolutional neural network (CNN)-based feature extractor and further to an FC layer to generate a visual feature. We adopted VGG-16 [24] as the backbone network in our experiments.

The four generated features are concatenated and fed into a fusion layer to learn the joint representations. The fusion layer comprises two FC layers with a classification output layer. Section 3.1.1 specifies the dimensions of the VRR networks used for experimental evaluation.

2.3. Ontology-based Knowledge Model

For the scalability and flexibility of our proposed framework, the main focus is on the use of knowledge. We define upper-level concepts through hierarchical clustering, which comprises four main sub-models: temporal, spatial, relationship, and agent. Figure 3 depicts the schematic structure of the main parts of the upper-level concepts. Further, we define bidirectional semantic relations internally and externally among the sub-models to represent their associations. We introduce them schematically in this section, the details of which are shown with an example graph instance in the experiment section 3.2. For the construction of the knowledge schema, we refer to some of the event-oriented manners and concepts of previous studies [6, 3] that emphasized extensibility and utility in robotic service applications.

Temporal Model The proposed knowledge is an event-oriented ontology; therefore, a *PerceptionEvent* individual is first generated to include all the perceptions of a video segment. Each perception of objects or relationships is instantiated in *ObjectDetection* or *RelationshipRecognition* event individuals, associated with the corresponding object or relationship individual in the knowledge base and linked to the *PerceptionEvent* individual in inclusion form by the *subEvent* relation. We define several temporal properties—*startTime* and *occursAt*—to indicate the occurrence of time points and places of events. *CurrentTime* is a purposeful concept to use when reasoning only on the present event.

Spatial Model The spatial model includes general concepts for objects and places and domain concepts for coordinate systems and robot maps. We categorize object concepts into *structural* objects, which represent immovable objects (for example, tables, chairs, and beds) that can be described in a specific map, and *portable* objects, which are unidentifiable objects (for example, bags, and bottles) that can move irregularly. We define the attributes of object concepts as movable, loss, or occupancy states. The coordinate systems comprises two classes—*BoundingBox* and *RotationMatrix*. Vis-CAF uses rotation matrix-based coordinates when applied to real-world services with 3D spatial robotic maps and uses bounding boxes when applied to simple 2D videos. In real-world applications, the localization of objects and places in a service environment, and their linkages to the robot map are available using the object properties *location* and *describedInMap*, respectively.

Relationship Model Relationships classified from VRR are defined in this model. Relationship concepts are primarily categorized into action and spatial relationships. Each relationship is represented as having a subject and an object in a triple format via the inter-relations of the aforementioned *ObjectDetection* and *RelationshipRecognition* individuals. Furthermore, the measured distance from the perception process of each relationship is annotated in the knowledge base and is used when reasoning the adjacent degree of the corresponding subject and object.

Table 1. Category-wise composition of a VRR training set. The numbers in the brackets indicate augmented instances.

Category	# of instances (aug.)
behind	1,890 (1,050)
in	3,836 (-)
in front of	1,050 (1,890)
next to	3,004 (3,004)
on	19,157 (895)
under	895 (19,157)

Agent Model The agent model represents an actor who is the subject of an applied service or activity, and it can also be an object to be detected. In this model, we define data properties as profile attributes (such as name, face-id, and gender) for *person*, and properties (such as specifications) for *robot* or other agents. The agent model is closely associated with spatial and temporal sub-models. Using the ontological relations, *worksOnMap* and *actorOf*, we conceptualize the current map of the robot, which is used in localization, and the subject of perception events, respectively. Furthermore, concepts in the agent model inherit properties of the spatial model through subclass hierarchical relations.

3. Experimental Results

3.1. Evaluation of Visual Relationship Recognition

3.1.1 Experimental configuration

Category The VRR training involves two types of categories—the input object and output relationship. We selected 18 generic object categories (for example, person, table, chair, bag, bottle, and so on) and 6 spatial relationships (for example, behind, in, in front of, next to, on, and under). Even in the limited circumstance where only using a few spatial relationships, we experimentally verified the availability and versatility of Vis-CAF.

Dataset We used images in the visual relationship dataset (VRD) [18], visual genome [14], and open images [15] for training, resulting in our training set, as shown in Table 1. While extracting relation-annotated images for the above categories from these datasets, sub-categories were merged into the corresponding super-categories (for example, 'stand behind' into 'behind', and 'handbag' into 'bag'). We augmented the training set by switching subject and object and/or changing their spatial relationship to the opposite. For example, from instances, 'A in front of B' and 'C next to D', new instances, 'B behind A' and 'D next to C' were generated. Consequently, the total number of image instances was 55,828, which are randomly divided category-wisely into a 9:1 ratio for training and validation, respectively. For the test set, we randomly sampled 30 instances for each category from the test images in VRD.

Feature Dimensions For word embedding vectors of object categories, we used 100-d GloVe. Before propagation to the fusion layer, the final dimensions of the word embedding, mask, and each visual feature were 64, 64 and 128, respectively, each of which was the output dimension of the FC layer. The final relative geometry feature was a 64D embedded vector without an FC layer. For visual features, a cropped input image for the subject, object, or union was resized to $128 \times 128 \times 3$ prior to CNN-based feature extraction. These multi-modal features were associated with a 64D hidden layer and a 6D output layer in the fusion layer.

Configuration in Training We adopted an ADAM optimizer with a learning rate initialized to 0.001 and reduced to 0.00001 at a minimum and set the batch size to 32, resulting in approximately 50 epochs by early-stopping. For the ablation study, we varied the loss functions, balancing mini-batch, and fine-tuning the CNN-based backbone network. To alleviate the data imbalance between categories, we used a focal loss function [17] to compare of categorical cross-entropy loss and stochastically balancing mini-batches in the data ratio by category. The top-4 layers of CNN-based backbone networks for visual features were unfreezed during fine-tuning.

Table 2. Ablation study of VRR by varying its configurations.

Model configuration	Precision	Recall	F1 score
CE	0.66	0.6	0.629
CE + BLC	0.68	0.64	0.659
FL	0.7	0.64	0.669
FL + BLC	0.69	0.66	0.675
FL + BLC + FT	**0.74**	**0.68**	**0.704**

Notations: CE (cross-entropy loss), BLC (balancing), FL (focal loss), FT (fine tuning)

Table 3. Category-wise performance with FL + BLC + FT.

Category	Precision	Recall	F1 score
behind	0.76	0.43	0.549
in	1.0	0.37	0.54
in front of	0.59	0.67	0.627
next to	0.58	0.7	0.634
on	0.58	1.0	0.734
under	0.9	0.9	0.9

3.1.2 Recognition Performance

Table 2 presents the results of the ablation study using our VRR model. We used the mean values of precision, recall, and f1 score for measurements of category-wise performances. The results indicate that VRR with focal loss achieves improved performance compared to VRR with cross-entropy loss. Balancing mini-batches during training alleviates the data unbalance problem by comparable performance improvements. However, the use of both focal

loss and balancing performs not relatively effective because of the duplication of the alleviation. Fine-tuning the CNN-based backbone networks that extract visual features results in reasonable effectiveness in both precision and recall. This implies that fine-tuning enables the model to improve its learning ability to capture the interactions between objects. For a concrete validation, we excluded misannotated test instances and re-evaluated the performance of our VRR model with the configuration with best performance, FL + BLC + FT (focal loss + balancing + fine-tuning). Consequently, the actual performance with a precision of 0.82 and a recall of 0.8 on 132 refined test instances was achieved.

The category-wise performance of the best model is shown in Table 3. Categories with a relative lack of data were predicted by high precision and low recall performance, perticularly with respect to the *in* relationship, implying that the data imbalance problem was not fully relaxed. However, we verified the significant effect of our training methods and data augmentation on the *under* relationship, outperforming the other relationships, although its actual number of instances was only 895. The aforementioned evaluations are further discussed in section 3.3.

3.2. Case Studies of Context-Awareness

We conducted several case studies that were assumed to have been applied to robotic services in different domains. The case studies show that Vis-CAF can implement various context-awareness by defining additional knowledge-based reasoning rules without changing the processes, demonstrating the scalability and utility of Vis-CAF. In this experiment, based on the objects and relationships in 2D images of a video obtained from the perception process, we assume that the robot's localization can be performed based on a map in 3D rotation matrix format using the video depth information. For the perception process, we applied YOLOv2 trained on the COCO dataset and a simple algorithm based on intersection-of-union (IOU) as the object detector and tracker, respectively. Only the relationships between objects within a certain distance, the measurement of which was obtained with reference to [7], were recognized to reduce the time complexity for real-time services. We used a pellet reasoner [25] to apply our rules for context inference.

General reasoning cases Service robots must have the ability to provide personalized services to their customers, independent of a specific domain. Vis-CAF can identify personal attributes or relationships without applying additional detectors trained for specific purposes. For instance, the recognition of users with a disability using the following SWRL rule enables personalized reception.

```
Person(?p), WheelChair(?wc), on(?p,?wc)
                        -> disabled(?p, true)
```

The relationship information is significant in context-awareness because, even in this case, the simultaneous detection of a person and a wheelchair should not be regarded as disabled. Similarly, a robot acquires the ability to actively approach and cope with a new-comer who is inferred as a customer, using relationships, such as 'person-on-seat (in a waiting area)' or 'person-in front of-kiosk'.

When trained to recognize action relationships beyond spatial relationships, VRR can naturally make more diverse general reasoning. For instance, the training enables more various personalized services, such as setting a seat in advance for an infant and suggesting storage of carry-on luggage, which can be inferred using 'person-hold-person' and 'person-carry-luggage' relationships, respectively.

Anomaly inference in hospital The implementation of anomaly detection based on deep learning has a severe problem because of the scarcity of a dataset [5], implying that structured reasoning can be more efficient in terms of cost and scalability. Vis-CAF can detect an anomaly by defining reasoning rules for abnormal situations and further leveraging the localization capability of the robot.

Assuming a circuit service scenario in a hospital, a robot detects an emergency when a patient falls based on its map information, according to the following rule:

```
Person(?p), ?Bed(?b), under(?p,?b),
?HospitalRoom(?hr), locatedIn(?b,?hr)
                  -> isInEmergency(?hr, true)
```

A situation in which a person is lying *under* (rather than *on*) a bed may be a rare emergency case, especially in a hospital room. Similarly, a robot performing a crime prevention scenario at night identifies a suspicious person close to objects (for example, a cash register or an access-restricted door) located in a secured area. Subsequently, the robot can execute appropriate service tasks, such as immediately calling the medical staff and raising alarm on the suspicious.

Complex context-awareness in restaurant We present a more complex context-awareness of Vis-CAF in an actual video, assuming a service scenario for a serving robot in a restaurant environment. The purpose of the case study of the Occupancy and Loss scenario is for the robot to infer the occupancy state of a restaurant table and the occurrence of a lost object while locomoting.

The example knowledge individuals for the scenario are shown in Figure 4. Based on the event-oriented approach of our knowledge, we represent each object in a *has* relation linked to the corresponding event. In addition, we associate the perceived place and related map information of objects with corresponding object properties and specify their attributes, such as user profile, coordinate, distance, and occupancy/loss state, with data properties. These associations enable the robot to identify and localize a visible person and a table. Using such knowledge graph representations for segment scenes, Vis-CAF performs three levels of stepwise context inference in the Occupancy and Loss scenario.

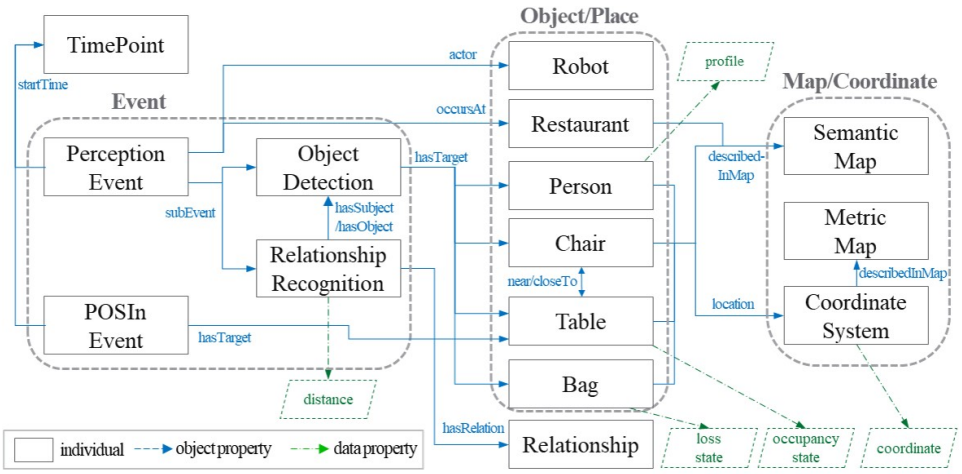

Figure 4. Example of knowledge instantiation in the Occupancy and Loss scenario.

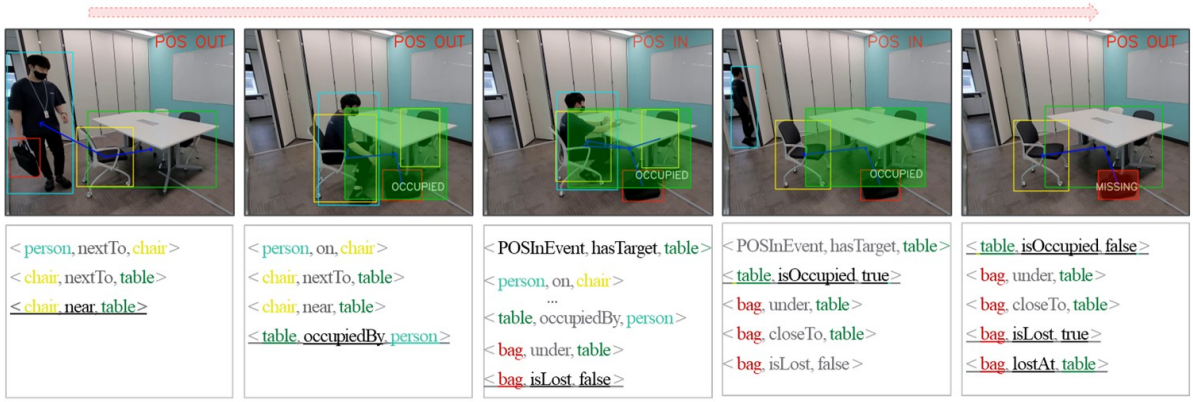

Figure 5. Qualitative results in Occupancy and Loss scenario assuming a restaurant service environment. Underlined triples represent new inferred relations obtrained using the inference process in each scene of video segments. The red words in the upper-right region indicate the assumed situation of whether the recognized table is registered in POS system. Only primary relationships are illustrated in this figure.

Rules

```
targetObject(?s_od, ?sbj), PerceptionEvent(?pe), subEvent(?pe, ?rd), targetObject(?o_od, ?obj),
hasSubjectEvent(?rd, ?s_od), distance(?rd, ?dis), RelationshipRecognition(?rd), CurrentTime(?ct),
startTime(?pe, ?tp), startTime(?ct, ?tp), lessThanOrEqual(?dis, 0.8), hasObjectEvent(?rd, ?o_od) ->
near(?sbj, ?obj), near(?obj, ?sbj)
```

Figure 6. Actual SWRL rule on Protégé for *near* relation.

Step 1. Reasoning the adjacency between objects: Figure 5 shows the key scenes of the scenario and the qualitative results. The first scene is a situation in which a customer approaches an unoccupied table. The robot continuously recognizes objects and spatial relationships and uses the distances to infer their adjacency relations following the rule in Figure 6. This actual rule is defined as fairly complicated because it is expressed based on event individuals for temporal expression; however, the representation of rules in this paper is conceptually simplified for readability as follows:

```
Relationship(?r), Subject(?s), Object(?o),
hasSbj(?r,?s), hasObj(?r,?o),
distance(?r,?d), swrl:lessThan(?d,0.8)
-> near(?s, ?o)
```

The criterion value of 0.8 is based on the numerical range of the distance measurement. Similarly, we also defined a reasoning rule for *closeTo* with a criterion under a distance of 0.4. *near* and *closeTo* properties have a symmetric characteristic for the opposite relations. Consequently, the robot perceived one of the contexts, 'the chair is near the table'.

Step 2. Context awareness of occupancy: In this step, we present a higher-level contextual inference that leverages the inferred adjacency relations. A serving robot requires the ability to recognize the occupancy contexts of tables in sight, despite not receiving a menu order. If a person sits on a chair for a certain duration, the table to which the chair belongs can be reasonably considered as occupied by the person. Assuming that a chair belongs to its nearby table, we represent the above cognition by the following rule:

```
Person(?p), Chair(?c), on(?p,?C),
Table(?t), near(?c,?t),
```

```
-> isOccupied(?t,true), occupiedBy(?t,?p)
```

This enables awareness of the entire context of table occupancy beyond the simple perception of 'sitting-on' behavior. If a robot stores the knowledge for the inferred occupancy context, other reception robots sharing the same knowledge can avoid the confusion of guiding new customers to a pre-occupied table. This ability implies that Vis-CAF has directivity and availability for multi-robot applications.

Using another nonvision modality such as a POS system, the robot can distinguish the occupancy of tables. The third scene shows an assumed situation in which a menu order is registered in the POS system. The registered information of the POS system is abstracted to an *POSInEvent* individual and the related properties in the knowledge representation. We define an additional rule based on POS system events.

```
POSInEvent(?pi), Table(?t),
hasTarget(?pi,?t)   -> isOccupied(?t,true)
```

Essentially, our proposed framework is capable of using multi-modalities by the abstraction of their sensed data or stored information to the shared knowledge base.

Step 3. Discovering a lost object: Vis-CAF performs a more complex context-awareness for the cognition of a loss event based on the previous inference logics, as follows:

```
Table(?t), isOccupied(?t,false),
Bag(?b), under(?b,?t), closeTo(?b,?t)
        -> isLost(?b,true), lostAt(?b,?t)
```

This rule integrally includes the occupancy and adjacency inferences and spatial relationship recognition. Additionally, defining other objects, spatial relationships, and adjacency relations, rather than *bag*, *under*, and *closeTo*, in the form of logical disjunction may enable the awareness of other types of lost objects in more generic situations.

In the third scene, the robot found a bag under the table but inferred that it was not a lost object because the robot regarded the bag as the possession of a customer occupying the table. The fourth scene assumes that the customer leaves the seat before checking out. Therefore the table was still perceived to be occupied although visually unknown by whom due to the remaining *POS-In* status. Therefore, the bag remained in the default loss state, *false*, similar to the third scene. Finally, after the check out (that is, the unregistered status in the POS system), the table was re-expressed as unoccupied; subsequently, the bag close to the table was inferred as lost with its location. If a counter is notified immediately upon discovery, the robot can perform appropriate services, such as delivering the lost item to the customer before leaving or requesting storage in 'Lost and Found'.

We verified that the stepwise context inference could be designed by merging or hierarchically constructing reasoning rules. The results of the experimental cases demonstrate the extensive and effective use of the proposed framework for various context-awareness.

3.3. Discussions

On analyzing the test results of VRR, we observed a tendency for predictions to be biased according to language modal features rather than other modal features. For instance, the predicate, 'person-under-chair', was not recognized even in an intended scene. Overall performance can be further improved using detailed techniques, such as calibrating the learning ratio of modal features in VRR and subdividing a balancing data manner using not only the relationship category but also the object category.

We demonstrate that fine-tuning the VGG-16 backbone networks in our VRR model was effective in improving the performance, indicating the potential for better enhancement when the backbone is replaced with a more suitable model. A scene graph generation model can better capture the associations between objects and extract more elaborate visual features for relationship prediction [27].

In rule-based context inference, Vis-CAF tends to depend on the consistency of rules. Therefore, in terms of knowledge management, the application of an additional method is necessary wherein the method periodically inspects rule conflicts, monitors conflicted rules, and temporarily excludes them during contextual inference.

The reasoning rule is quite specifically represented and defined in a duplicate form because of the limitations in the variety of SWRL expressions. Our framework can be made aware of more abundant contexts by applying the Prolog [4] or Drool rule engine [21], which have more diverse logical expressions, such as negation and individual generation, and fully leverage ontological characteristics. For instance, the anomaly inference for a patient's fall can be more accurate by adding a rule that a person should not be detected on a bed based on a negation expression.

4. Conclusions

In this study, we proposed Vis-CAF, a context-aware framework that performs knowledge-based inference using visually perceived information. Vis-CAF is structured to maintain process independence in perception and inference to ensure scalability and flexibility for variations in context-aware tasks. Further, we present the VRR model for perception and ontology-based domain knowledge with rules. We demonstrated the effectiveness of Vis-CAF through experimental case studies on different robotic service domains.

In our future work, we intend to apply scene graph generation methods to fully capture object associations for VRR. In robotic applications, we plan to use video depth information to clarify the distinction between relationships, such as 'behind' and 'in front of'. For context-awareness, we intend to extend the ability to express rules and adopt a Bayesian network and graph neural networks to increase inference diversity and apply pattern recognition for surveillance.

References

[1] Zahrah A Almusaylim and Noor Zaman. A review on smart home present state and challenges: linked to context-awareness internet of things (iot). *Wireless networks*, 25(6):3193–3204, 2019.

[2] Sean Bechhofer, Frank Van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, Lynn Andrea Stein, et al. Owl web ontology language reference. *W3C recommendation*, 10(2):1–53, 2004.

[3] Michael Beetz, Daniel Beßler, Andrei Haidu, Mihai Pomarlan, Asil Kaan Bozcuoğlu, and Georg Bartels. Know rob 2.0—a 2nd generation knowledge processing framework for cognition-enabled robotic agents. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519. IEEE, 2018.

[4] Bruno Blanchet et al. An efficient cryptographic protocol verifier based on prolog rules. In *csfw*, volume 1, pages 82–96. Citeseer, 2001.

[5] Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

[6] Doo Soo Chang, Gun Hee Cho, and Yong Suk Choi. Ontology-based knowledge model for human-robot interactive services. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2029–2038, 2020.

[7] Doo Soo Chang, Gun Hee Cho, and Yong Suk Choi. Zero-shot recognition enhancement by distance-weighted contextual inference. *Applied Sciences*, 10(20):7234, 2020.

[8] Xilun Chen and Claire Cardie. Multinomial adversarial networks for multi-domain text classification. *arXiv preprint arXiv:1802.05694*, 2018.

[9] Dan-Bi Cho, Hyun-Young Lee, and Seung-Shik Kang. Multi-channel long short-term memory with domain knowledge for context awareness and user intention. *Journal of Information Processing Systems*, 17(5):867–878, 2021.

[10] Mohamed Elhoseny. Multi-object detection and tracking (modt) machine learning model for real-time video surveillance systems. *Circuits, Systems, and Signal Processing*, 39(2):611–630, 2020.

[11] Shuqiang Guo, Qianlong Bai, Song Gao, Yaoyao Zhang, and Aiquan Li. An analysis method of crowd abnormal behavior for video service robot. *IEEE Access*, 7:169577–169585, 2019.

[12] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, Mike Dean, et al. Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, 21(79):1–31, 2004.

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3588–3597, 2018.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[15] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[16] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[18] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[19] Ruotian Luo, Ning Zhang, Bohyung Han, and Linjie Yang. Context-aware zero-shot recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11709–11716, 2020.

[20] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[21] Mark Proctor. Drools: a rule engine for complex event processing. In *International symposium on applications of graph transformations with industrial relevance*, pages 2–2. Springer, 2011.

[22] E. Prud'hommeaux and A. Seaborne. Sparql query language for rdf - w3c candidate rec, 2008. http://www.w3.org/TR/rdf-sparql-query/.

[23] Shaina Raza and Chen Ding. Progress in context-aware recommender systems—an overview. *Computer Science Review*, 31:84–97, 2019.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 5(2):51–53, 2007.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Wentao Xie, Guanghui Ren, and Si Liu. Video relation detection with trajectory-aware multi-modal features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4590–4594, 2020.

[28] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *2018 IEEE International conference on data mining (ICDM)*, pages 727–736. IEEE, 2018.