TASK-ORIENTED MULTI-MODAL QUESTION ANSWERING FOR COLLABORATIVE APPLICATIONS

Hui Li Tan, Mei Chee Leong, Qianli Xu, Liyuan Li,
Fen Fang, Yi Cheng, Nicolas Gauthier, Ying Sun, Joo Hwee Lim

{hltan, leong_mei_chee, qxu, lyli,
fang_fen, cheng_yi, nicolas_gauthier, suny, joohwee}@i2r.a-star.edu.sg
Institute for Infocomm Research (I^2R), A*STAR

ABSTRACT

Cobots that can work in human workspaces and adapt to human need to understand and respond to human's inquiry and instruction. In this paper, we propose new question answering (QA) task and dataset for human-robot collaboration on task-oriented operation, *i.e.,* task-oriented collaborative QA (TC-QA). Differing from conventional video QA for answering questions about what happened in video clips constrained by scripts and subtitles, TC-QA aims to share common ground for task-oriented operation through question answering. We propose an open-end (OE) format of answer with text reply, image with annotated related objects, and video with operation duration to guide operation execution. Designed for grounding, the TC-QA dataset comprises query videos and questions to seek acknowledgement, correction, attention to task-related objects, and information on objects or operation. Due to the flexibility of real-world task with limited training sample, we propose and evaluate a baseline method based on a hybrid approach. The hybrid approach employs deep learning methods for object detection, hand detection and gesture recognition, and symbolic reasoning to ground question on observation for providing the answer. Our experiments show that the hybrid method is effective for the TC-QA task.

Index Terms— question answering, multi-modal grounding, human-robot collaboration, hybrid system, corpora

1. INTRODUCTION

Technology is progressing from traditional industrial robot designed to work autonomously in isolation from human, to cobots designed to interact with humans in shared spaces. Effective human-robot collaboration on operational task requires effective grounding [1], the construction and maintenance of a shared conception of the operational task. QA in collaboration provides a mechanism to provide acknowledgement of existing understanding, correct wrong facts or concepts, draw attention to task-related objects, and provide novel information about task-related objects and operation.

Despite the advancements in video QA, existing video QA tasks, datasets, and solutions are not designed for such scenario. Existing video QA systems are mostly designed for understanding what happened in the video clip with the constraints of scripts or subtitles [2] [3]. The multiple-choice (MC) QA systems [2] are also heavily constrained for grounding, requiring visual-text association encoding for producing the answers. Although open-end (OE) QA systems [4] have been explored, the short words or phrases text answers are insufficient for clarification nor providing task-related information. We propose a new QA task for human-robot collaboration on task-oriented operation. Referring to Fig. 1, in the task-oriented collaborative QA (TC-QA) task, given the query query text (speech) and video, the aim is to provide a text reply of answer or instruction, an image reply indicating objects involved in the operation, and a video reply of inquired operation to guide operation execution.

We further create a novel TC-QA dataset, with QAs that are designed to provide acknowledgement, correction, attention indication, and novel information presentation to support natural and flexible collaborations in real-world industrial tasks. Since pointing gestures with pronouns in questions are frequently used in inquiry in task collaboration, especially for solving referential ambiguities, TC-QA is designed to comprise multi-modal grounding. As the probe video, question and ground truth answer are not constrained by pre-defined scripts and subtitles, it poses new challenges to learning approaches for QA tasks. We propose and evaluate a hybrid approach which employs deep learning methods for object detection, hand detection and gesture recognition, and symbolic reasoning for grounding question on observation to provide the answer. Our experiments show the effectiveness of the hybrid method for the TC-QA task and its performance on the TC-QA dataset can be used as baseline for future investigations.

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Award A18A2b0046) and the National Research Foundation, Singapore under its NRF-ISF Joint Call (Award NRF2015-NRF-ISF001-2541).

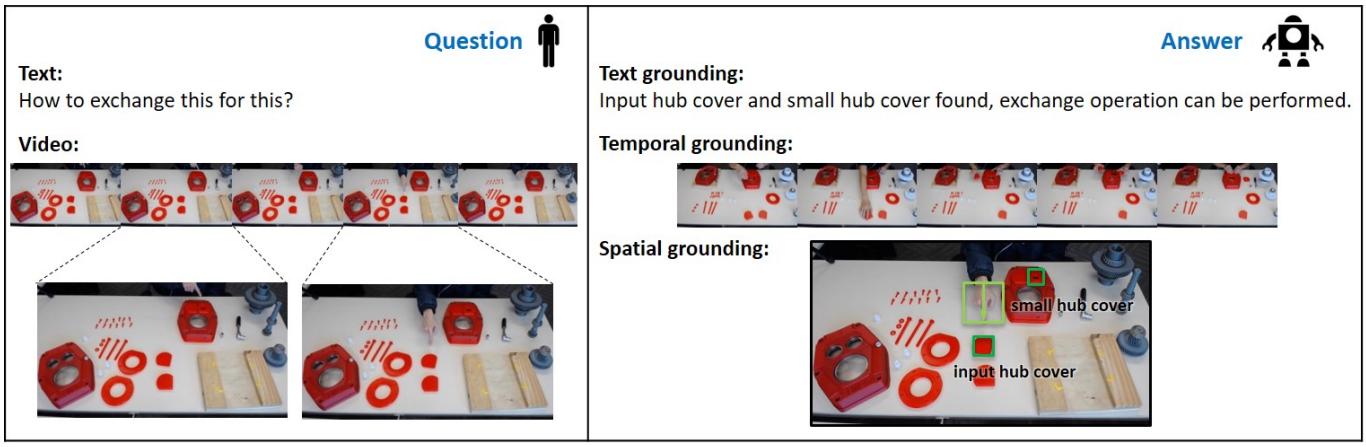

Fig. 1. Task-orientated collaborative question answering (TC-QA).

The contributions of our paper are three-fold: 1) We propose a novel TC-QA task for supporting natural collaborations in tasks without constraints of scripts and subtitles. We propose an extended answer format which includes a text reply about answer and instruction, an image reply with related object indicated, and a video reply for operation guidance. 2) We propose a new TC-QA dataset which extends the capacity of existing QA systems beyond visual and text encoding to multi-modal grounding which requires association between speech, gesture and visual detection to understand the inquiry and instruction. 3) We propose a hybrid method that combines deep learning and symbolic reasoning as our baseline for evaluation on the proposed TC-QA benchmark dataset.

2. DESIGN OF TC-QA TASK AND DATASET

QA is a challenging problem that has received increasing attention from both the natural language processing and computer vision communities. In particular, in visual QA [5], given an image or video and a text question about it, the answer is to be inferred from the visual features, other constraints, and general knowledge. Various image QA datasets such as COCO-QA [6], Visual Madlibs [7], Visual7W [8], CLEVR [9], and KB-IQA [10], as well as solutions have been proposed to investigate the task. Video QA is a natural but challenging extension to image QA, with additional challenges of spatio-temporal understanding. Similarly, various video QA datasets such as MovieQA [2], TGIF-QA [11], TVQA+ [3], and ActivityNet-QA [4] have been proposed. Extensions of image QA solutions for video QA [4] have also been attempted. Despite the advancements in visual QA, existing datasets, tasks, and solutions designed for post-understanding of text and visual content only partially address the needs for grounding in human-robot collaboration on operational task.

Tables 1 and 2 highlight the differences of TC-QA with some state-of-the-art video QA solutions. TC-QA seeks to establish common ground for operational task understanding and execution by providing acknowledgement, correction, attention and information. As shown in Table 1, existing QA solutions do not fully address these intentions for common ground sharing, especially on providing correction when wrong concepts are identified, and draw attention to task-related objects. Achieving these intentions require more fine-grained grounding beyond short MC or OE text which are the focus of existing video QA systems. As shown in Table 2, TC-QA provides OE text with spatial and temporal localization necessary for more fine-grained grounding of task-related objects and tasks.

Table 1. Comparison with state-of-the-art video QA, where TC-QA aims at establishing common ground for operational task understanding and execution by providing acknowledgement (Ack.), correction of wrong concept (Cor.), attention to most related object (Att.) and information of operation (Info).

	Ack.	Cor.	Att.	Info
VideoQA [12]	✓			✓
ActivityNet [4]	✓			✓
TVQA+ [3]	✓		✓	✓
TutorialVQA [13]				✓
TC-QA (Ours)	✓	✓	✓	✓

The TC-QA dataset is designed to address the TC-QA task. The TC-QA dataset involves real-world tasks with combinations of related objects, gestures, operations, questions, and answers to address the interaction issues in collaborations to perform an industrial task. It includes 991 QA pairs with 495 video question segments relating to gearbox assembly. The videos are recorded in a workspace with a static frontal cameras, where 20 subjects are involved in the performance of 25 querying actions. The subjects could be pointing to sin-

Table 2. Comparison of answer format with state-of-the-art video QA, where TC-QA provides OE text reply, image reply with marked objects, and video reply of related operation.

	Text MC	Text OE	Image	Video
VideoQA [12]		✓		
ActivityNet [4]		✓		
TVQA+ [3]	✓		✓	✓
TutorialVQA [13]				✓
TC-QA (Ours)	✓	✓	✓	✓

gle or multiple objects to seek acknowledgement, correction, attention to task related object, or information on object or operation. The video segments are on average five seconds, and range from one second to 16 seconds. Some examples of QA are given in Table 3. For instance, in the case of "Is this O_A?", with the subject pointing to certain object, the system should be able to either provide acknowledgement that the object is correct or an answer of correction if the object is incorrect. The questions are multi-modal, requiring both text and video for visual-linguistic referencing. Similarly, the answers are also in multi-modal format, grounded in text, image and video. The breakdown of QA quantities for each question type are also provided in Table 3. For the QA pairs for inquired operation video retrieval and duration localization, there are 495 unique task operation videos which are on average seven seconds and range from one second to 23 seconds.

3. PROPOSED BASELINE METHOD

Existing methods for image QA or video QA are mainly deep learning approaches based on end-to-end learning on ground truth QA pairs. For image QA, approaches such as joint embedding networks [6, 14], attention mechanisms [8, 15], compositional/modular [16, 17], memory models [18], and knowledge base enhanced networks [10] have shown their effectiveness on image QA datasets. For video QA, extensions of image QA approaches as well as new approaches have also been proposed and validated on various video QA datasets. Some examples are appearance motion networks [12, 11], temporal attention mechanisms [12, 11], and memory networks [19, 20].

While these deep learning approaches have shown their powerful learning capability to capture the complex relations between visual and text features, existing solutions that are trained for video QA dataset are inadequate for direct application in TC-QA due to a lack of task-oriented inference. End-to-end learning for our dataset is also very challenging due to the requirement of immense training examples. We investigate to extend the learning capacity of existing deep learning methods to apply on our TC-QA dataset.

In visual perception, deep learning performs mapping from signals to basic semantic representations for image recognition. From the semantic representations, symbolic reasoning can be applied to infer high-level purpose-driven understanding. In this paper, we propose to combine the advantages of these two approaches to build a hybrid system that is able to recognize visual patterns while reasoning on its goal purpose. This hybrid system serves as the baseline method for TC-QA, and does not require direct training on the QA pairs.

In our hybrid system, deep neural networks perform object detection, hand detection and gesture recognition, while the symbolic module performs reasoning on the questions and observations to provide the correct answer. The overview of the method is illustrated in Fig. 2. The details are described in the following.

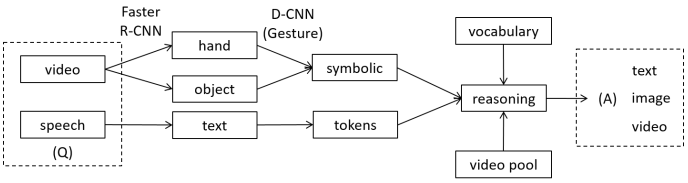

Fig. 2. The block diagram of the proposed baseline method.

Bottom-up perception: The bottom-up process performs vision detection, recognition, and perception tasks. First, we employ a Faster R-CNN with Resnet-101 network pre-trained on COCO dataset and fine-tune it for our object and hand detection. When applied on a QA task, it provides the semantic representations of the related objects and their bounding boxes. If a hand is detected, a trained D-CNN is applied to recognize the hand gesture and predict hand pointing direction. From the distance of objects' center and hand pointing direction, we can compute the probability of the potential objects and select the object with the highest probability as the pointed object. We also design a visual working memory to record the pointed objects at frames where the pointing gesture is stable.

Top-down reasoning: The top-down process is driven by the question. First, the question text is tokenized and parsed based on pre-defined task related vocabularies. Two clusters of vocabularies are defined for task-relevant objects and question keywords. With the parsing algorithm and pre-defined vocabularies, we generate a set of question related objects and text grounding based on semantic vision observations. A rule-based algorithm is added to provide answers to the question based on its text grounding. The rules of grounding includes: 1) symbol matching between objects in question and vision detection, 2) spatial association between hand pointed objects and mentioned objects in the question, and 3) operation retrieval based on tracking observed objects mentioned in the question in the pool of potential operation videos provided in

Table 3. Examples of TC-QA, with quantities and purposes (A: acknowledgement, C: correction, T: attention, I: information).

Type	# QA pairs	Purpose	Question	Text answer	Image	Video
what	471	A, C	Is this/that O_A?	yes, this/that is O_A	O_A	
				no, this/that is O_B	O_B	
		T, I	What are these/those?	these/those are O_A, O_B	O_A, O_B	
where	120	T, I	Where is O_A?	O_A is here	O_A	
		T, I	Which is O_A?	O_A is here	O_A	
		T, I	Can you find O_A?	yes, O_A found	O_A	
how	400	A, C, T, I	How to load O_A onto O_B?	O_A and O_B found, load operation can be performed.	O_A O_B	✓
		A, C, T, I	How to load this onto this?	O_A and O_B found, load operation can be performed.	O_A O_B	✓

the question.

Our symbolic reasoning is designed based on the common knowledge of assembly task. No training on QA pairs is required for specific tasks, but the domain-specific vocabularies for the task, such as gearbox assembly, have to be provided.

4. EXPERIMENTS

In this section, we provide the protocol, metrics, and baseline performance for benchmarking on the TC-QA dataset.
Experimental Setup: When end-to-end learning method is used, one can randomly split the dataset into training and test sets for evaluation. If no training on QA pairs is required, one can use the whole dataset for evaluation. TC-QA provides a new answer format including text, image and video replies.

The performance of the OE text replies is evaluated by matching the semantic representations of the predicted and ground truth answers, similar to the evaluation methods in image captioning [21, 22]. One can extract keywords from the answers and represent them in semantic tuples that consist of three components - (*Semantic*, *Object* and *Action*). The F-measure of the tuples is used as the evaluation metric. The F-measure of detected objects is employed to evaluate the image reply. For the video reply, the temporal retrieval and localization scores are measured using accuracy and temporal mean intersection-over-union (mIoU) following previous works [13] and [3] respectively. The average scores on the test set represent the performance for benchmarking.
Experimental Results: The baseline performance on TC-QA by our method is presented in Table 4. From the scoring of each question type, the "where" question type obtained highest score in both text and image replies, as its probe videos have minimal hand movement that distracts the detection process. The "how" question type performed averagely, as it involves a mixture of visual and text referencing. For the video reply, the retrieval score is 0.59, which is mostly due to object detection errors which lead to failure to trigger operation search. For the "what" question type, the answers are mainly relied on hand pointing gestures to a single object or multiple objects. This question type is challenging as there are ambiguities in the finger pointing direction, distance to target, occlusion, and gesture tracking. The Faster R-CNN has limitations in detecting smaller objects such as *oil_level_indicator*, and identifying similar objects such as *casing_top* and *casing_base*. Detailed analysis on the performance of image and video replies are presented in Tables 5 and 6.

Table 4. Baseline performance on TC-QA by our method.

Type	Text score	Image score	Video retrieval score	Video mIoU score
what	0.5366	0.4515	NA	NA
where	0.7033	0.7056	NA	NA
how	0.6913	0.6736	0.5875	0.4557
Total	**0.6193**	**0.5719**	**0.5875**	**0.4557**

Table 5. Scores of image replies for "what" and "where" questions with single or multiple object pointed.

Question	Single object	Multiple objects
what (this/these)	0.6417	0.3738
what (that/those)	0.5391	0.2484
where	0.7125	0.6918

Table 6. Temporal grounding analysis for "how" questions with mixture of visual-linguistic referencing.

Question	Spatial score	Temporal retrieval score	Temporal mIoU score
obj - obj	0.8944	0.6500	0.5057
obj - this	0.6542	0.6150	0.4736
this - this	0.4918	0.4700	0.3699

5. CONCLUSION

Advancements in video QA, although encouraging, are inadequate for task-oriented collaborative applications. We propose the TC-QA task, which extends the capacity of existing QA systems with spatio-temporal grounding and task understanding. We then introduce the TC-QA dataset, a task-oriented multi-modal dataset designed to investigate QAs for grounding. A hybrid system is proposed as the baseline method. There remains significant room for improvement. In the future, we will expand our dataset for different tasks and investigate various learning approaches on the dataset.

6. REFERENCES

[1] Pierre Dillenbourg, David Traum, and Daniel Schneider, "Grounding in multi-modal task-oriented collaboration," Oct 1996.

[2] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Movieqa: Understanding stories in movies through question-answering," in *IEEE CVPR*, 2016.

[3] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal, "TVQA+: spatio-temporal grounding for video question answering," *CoRR*, vol. abs/1904.11574, 2019.

[4] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," *CoRR*, vol. abs/1906.02467, 2019.

[5] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel, "Visual question answering: A survey of methods and datasets," *CoRR*, vol. abs/1607.05910, 2016.

[6] Mengye Ren, Ryan Kiros, and Richard S. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *CoRR*, vol. abs/1505.02074, 2015.

[7] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg, "Visual madlibs: Fill in the blank image generation and question answering," *CoRR*, vol. abs/1506.00278, 2015.

[8] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei, "Visual7w: Grounded question answering in images," *CoRR*, vol. abs/1511.03416, 2015.

[9] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *IEEE CVPR*, 2017.

[10] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel, "Explicit knowledge-based reasoning for visual question answering," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1290–1296.

[11] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim, "TGIF-QA: toward spatio-temporal reasoning in visual question answering," *CoRR*, vol. abs/1704.04497, 2017.

[12] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 1645–1653.

[13] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim, "Tutorialvqa: Question answering dataset for tutorial videos," 2019.

[14] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your neurons: A neural-based approach to answering questions about images," *CoRR*, vol. abs/1505.01121, 2015.

[15] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.

[16] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, "Deep compositional question answering with neural module networks," *CoRR*, vol. abs/1511.02799, 2015.

[17] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, "Learning to compose neural networks for question answering," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June 2016.

[18] Caiming Xiong, Stephen Merity, and Richard Socher, "Dynamic memory networks for visual and textual question answering," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 2016.

[19] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia, "Motion-appearance co-memory networks for video question answering," *CoRR*, vol. abs/1803.10906, 2018.

[20] Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu, "Multi-turn video question answering via multi-stream hierarchical attention context network," Jul 2018, pp. 3690–3696.

[21] Lily D Ellebracht, Arnau Ramisa, Pranava Swaroop Madhyastha, Jose Cordero-Rama, Francesc Moreno-Noguer, and Ariadna Quattoni, "Semantic tuples for evaluation of image to sentence generation," in *Proceedings of the Fourth Workshop on Vision and Language*, 2015, pp. 18–28.

[22] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.