Dual Track Multimodal Automatic Learning through Human-Robot Interaction

Shuqiang Jiang[1,2], **Weiqing Min**[1,2], **Xue Li**[1,3], **Huayang Wang**[1,2], **Jian Sun**[1,3], **Jiaqi Zhou**[1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, China
[2]University of Chinese Academy of Sciences, China
[3]Shandong University of Science and Technology, China
sqjiang@ict.ac.cn

Abstract

Human beings are constantly improving their cognitive ability via automatic learning from the interaction with the environment. Two important aspects of automatic learning are the visual perception and knowledge acquisition. The fusion of these two aspects is vital for improving the intelligence and interaction performance of robots. Many automatic knowledge extraction and recognition methods have been widely studied. However, little work focuses on integrating automatic knowledge extraction and recognition into a unified framework to enable jointly visual perception and knowledge acquisition. To solve this problem, we propose a Dual Track Multimodal Automatic Learning (DTMAL) system, which consists of two components: Hybrid Incremental Learning (HIL) from the vision track and Multimodal Knowledge Extraction (MKE) from the knowledge track. HIL can incrementally improve recognition ability of the system by learning new object samples and new object concepts. MKE is capable of constructing and updating the multimodal knowledge items based on the recognized new objects from HIL and other knowledge by exploring the multimodal signals. The fusion of the two tracks is a mutual promotion process and jointly devote to the dual track learning. We have conducted the experiments through human-machine interaction and the experimental results validated the effectiveness of our proposed system.

1 Introduction

The cognitive ability of human beings is constantly updated and improved through the interaction with the environment [Gotts, 2016], including the enhancement of recognition ability and the growth of the new knowledge. These two aspects are parallel but interrelated. Because of the open and dynamic properties of the dataset from the interaction: new object samples and new object classes increase continuously, the enhancement of recognition ability requires the incremental learning, which can learn both the new instance of known objects and new object classes. The growth of the new knowledge involves constructing the multimodal knowledge graph by recognizing, extracting and summarizing the multimodal knowledge based on multiple input signals. In addition, the learned new objects help the growth of the knowledge by adding new nodes and their relations into the multimodal knowledge graph. Meanwhile, the knowledge items in the multimodal graph are helpful for visual recognition.

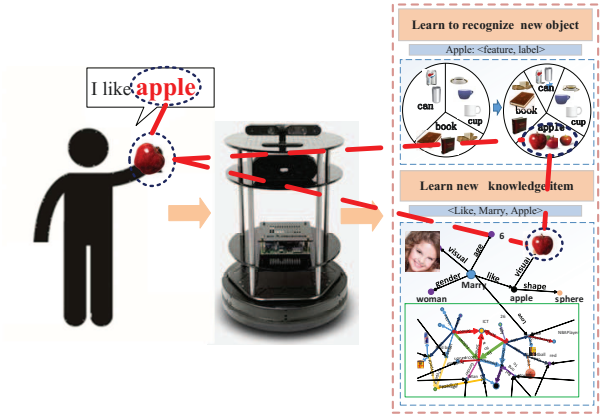


Figure 1: An example of dual track multimodal automatic learning

These two aspects lead to the continuously growing capability of automatic learning and enable many applications in human-machine interaction. Fig.1 shows a toy example. The robot learned the new object "apple" through incremental learning and thus improved its recognition ability. Meanwhile, the robot can fuse the multimodal information to add the "apple" and its relations with other objects to the knowledge graph. In addition, the enhanced recognition ability of the object "apple" can facilitate the multimodal knowledge graph construction, especially when only the visual information is available. Similarly, the constructed knowledge items including objects and their relations are helpful for object recognition. Therefore, in this paper, we jointly study incremental learning and multimodal knowledge extraction, namely dual track automatic learning.

Existing work mainly focuses on single track learning for either the improvement of recognition ability or automatic knowledge extraction from different perspectives. For vision-based recognition, there are two different types of incremental learning, namely data-incremental learning [Bor-

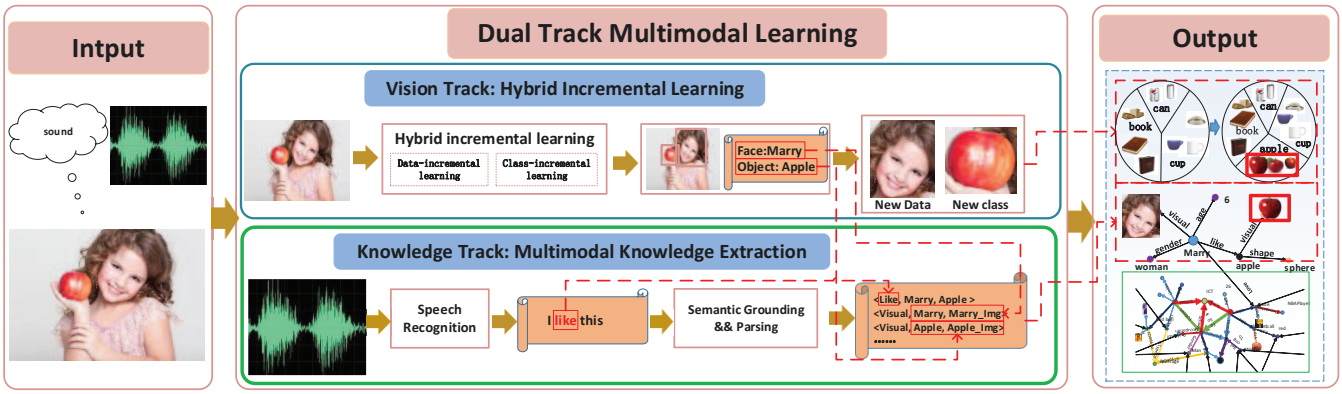


Figure 2: The proposed system of Dual Track Multimodal Automatic Learning (DTMAL)

des *et al.*, 2005; Cauwenberghs and Poggio, 2000; Ruping, 2001][Murata *et al.*, 2002] for new samples of known objects and class-incremental learning [Kuzborskij *et al.*, 2013; Lampert *et al.*, 2009; Ristin *et al.*, 2015] for unknown objects. These two types of incremental learning are indispensable for recognizing new concepts and enhancing the recognition ability of known concepts, and work jointly as hybrid incremental learning. For the construction of knowledge base, some methods utilized the language information to construct text based knowledge graph from the dialog [Hakkani-Tür *et al.*, 2014; Hixon *et al.*, 2015] while others pay more attention to the representation and analysis of the multimodal knowledge graph [Saxena *et al.*, 2014; Johnson *et al.*, 2015; Zhu *et al.*, 2015]. Such construction enables more personalized and task-oriented applications. However, little work has investigated the problem of dual track learning with both multimodal knowledge extraction and hybrid incremental learning.

In addition, some work [Thomason *et al.*, 2016; Al-Omari *et al.*, 2017] used the multimodal data to learn grounded linguistic semantics for language grounding. Moritz *et al.* [Tenorth and Beetz, 2013] proposed a knowledge processing system KnowRob for acquiring and representing knowledge by fusing multi-source information. The difference is that we aim to fuse visual perception and knowledge acquisition for automatic learning while they attempt to the language grounding or knowledge representation.

The extraction of the knowledge items relies on the improvement of the recognition ability by adding new objects and their relations to update the knowledge graph. Meanwhile, the objects, attributes and their relations from the constructed multimodal knowledge graph is helpful for online learning. Therefore, they work together as a whole. In order to solve this problem, we propose a Dual Track Multimodal Automatic Learning (DTMAL) system, which consists of two components: the vision track and knowledge track (Fig. 2). For the vision track, Hybrid Incremental Learning (HIL) adds the new classification-planes and adjusts existing classification-planes under the setting of Support Vector Machine (SVM) to learn both the new instance of known objects and new object classes. For the knowledge track, Multimodal Knowledge Extraction (MKE) utilizes the symbol grounding and parsing methods to extract the multimodal knowledge items via the interaction with humans. These two tracks are fused into a unified framework: the extraction of the multimodal knowledge relies on the improved recognition ability while the objects and their relations from the multimodal knowledge graph are helpful for improving the incremental learning. As an example in Fig. 2, the user says: "I like apple" with an apple in her hand. Our system can extract the visual features of apple and the apple label < visual features, apple>, the visual features of person and the person name < visual features, Scarlett>, and a knowledge item <like, Scarlett, apple> from the interaction process. The visual features and labels are used for the improvement of recognition ability. The extracted multimodal knowledge items are added into the knowledge graph.

2 Our Framework

As shown in Fig. 2, DTMAL mainly consists of two components: The vision track and the knowledge track. For the vision track, HIL is to learn new objects and new information of existing objects from visual information. Based on the learned new objects, new users and recognized speech, MKE from the knowledge track is mainly to utilze the symbol grounding and parsing techniques to extract rich triple knowledge. Finally, the extracted knowledge items and recognized objects are used to generate and update the multimodal knowledge graph. Meanwhile, the recognition ability of the system is improved from the HIL.

2.1 Hybrid Incremental Learning (HIL)

To enhance the ability of the object recognition, our HIL method adds the new classification-planes and adjusts existing classification-planes under the setting of SVM. Therefore, it can simultaneously improve the recognition quality of known concepts by minimizing the prediction error and transfer the previous model to recognize unknown objects.

For a visual representation pair (x, y), where x is the visual feature and y is the label. At step t, we modify the source classification-planes $W^t = [w_1^t, w_2^t, \cdots, w_M^t]$ to preserve the performance on known concepts and find a new group of classification-planes w_{M+1}^t, which is close to source classification-planes to make the M-class source classifier

transfer to a $(M+1)$-class target classifier. We minimize the following objective function to realize the hybrid incremental learning when the new data is available at step t:

$$\min_{W^t,b^t,e^t} J(W^t,b^t) = \frac{1}{2}\left\|W^t - W^{t-1}\right\|_F^2 + \frac{1}{2}\left\|w_{M+1}^t - W^{t-1}\beta\right\|_F^2 + \frac{C^t}{2}\sum_{m=1}^{M+1}\Big(\sum_{i\in I^t}(e_{mi}^t)^2 + L^t\sum_{j\in S^{t-1}}(e_{mj}^t)^2\Big) \quad (1)$$

s. t.

$$(w_m^t)^T\varphi(x_i) + b_m^t - y_i \le e_{mi}^t,\ i\in I^t,\ m = 1, ..., M+1$$

$$(w_m^t)^T\varphi(x_j) + b_m^t - y_j \le e_{mj}^t,\ j\in S^{t-1},\ m = 1, ..., M+1$$

where I^t denotes the new data at step t, and S^{t-1} denotes the old support vectors at step $t-1$. L, β, C are parameters, which need to be optimized. The first term $\frac{1}{2}\left\|W^t - W^{t-1}\right\|_F^2$ aims at preserving previous classification model. The second term $\frac{1}{2}\left\|w_{M+1}^t - W^{t-1}\beta\right\|_F^2$ incorporates new concepts into the current model. The last two terms $\frac{C}{2}\sum_{m=1}^{M+1}(\sum_{i\in I^t}(e_{mi}^t)^2 + L\sum_{j\in S^{t-1}}(e_{mj}^t)^2)$ define the loss. The first one is to minimize the prediction error of the new information and the second is to minimize the prediction error of support vectors.

Note that for incremental face learning, compared with incremental object recognition, the difference of faces from different persons is small, and the technique of face recognition is relatively mature now. Therefore, the mechanism of incremental object recognition is not suitable. We can directly use matching based method to achieve incremental face learning. Particularly, for the first time, the user should say "I am XX", we use the face detection method VIPLFaceNet [Liu *et al.*, 2016] to store the detected face image, its deep visual features and his name for registration. Then next time, the system recognizes him via the matching between the detected faces with the faces from the dataset. Similarly, the incremental face learning can support both data-incremental and class-incremental learning.

Through HIL, the system can automatically learn new objects and the new information of existing objects. After HIL, the visual recognition ability of the system is enhanced, which is helpful in the following MKE.

2.2 Multimodal Knowledge Extraction (MKE)

The knowledge extraction mainly includes: 1) symbol grounding and 2) Combinatory Categorial Grammar (CCG) parsing.

The purpose of symbol grounding is to establish the connection between the object included in the dialogue and specific context [Barsalou, 1999; Thomason *et al.*, 2016; Parde *et al.*, 2015]. The grounding set contains a set of semantically related symbols such as objects and persons. It is obvious to associate with noun phrases and objects. Based on the learned new objects o from HIL, the user p from incremental face recognition, and recognized speech s, we first use rule-based grammar tree to extract the grammatical structure from the recognized speech to obtain a set of phrases $\lambda = \{\lambda_1, \lambda_2, .., \lambda_k\}$. The grounding problem [Paul *et al.*, 2016] is then posed as estimating the likely set of groundings $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_n\}$ for the input: $\Lambda = \{\lambda, o, p\}$:

$$\max_{\gamma_1,\gamma_2,...\gamma_n} p(\Gamma|\Lambda) \quad (2)$$

After symbol grounding, we get the sentence S. We then extract the triple information from S using EasyCCG [Lewis and Steedman, 2014]. The CCG parser y of the sentence S is defined as a list of lexical categories $c_1, c_2 \ldots c_n$and a derivation. We compute the optimal parser $\hat{y}$ by the following formula:

$$\hat{y} = \arg\max_y \prod_{i=1}^{n} p(c_i|S) \quad (3)$$

We use the A^* algorithm to search [Klein and Manning, 2003] for the most probable complete CCG derivation of a sentence. In A^* parsing, the items on the agenda are sorted by their cost. If two agenda items have the same cost, we prefer to the one with longer dependencies.

We choose the core nouns from the most important component of the grammar dependent tree as the object word for the triple relation extraction, and finally obtain the triple information. For example, we firstly get the recognized speech "I like this", recognized person name of the speaker "Scarlett" and the recognized object "apple" in her hand from HIL. Then the symbol grounding process can change "I" and "this" to "Scarlett" and "apple", respectively. The new sentence S is "Scarlett like Apple". Then through CCG parsing, we can obtain the triple <like, Scarlett, apple>.

Through HIL and MKE, our DTMAL system can automatically improve the visual recognition ability and learn new knowledge items simultaneously according to user's multimodal inputs.

To store these parsed triple information and images (the detected face or objects), we use a non-relational SQL document storage database. We will update the multimodal knowledge graph by inserting unknown information into the database to learn new knowledge items. Some measures are taken to eliminate conflicts and generate representative images.

The conflict eliminating is to change knowledge items with lower confidence into ones with high confidence. We design multiple strategies for conflict eliminating. For example, the designed first person priority is that if the recognized sentence is the first person "I", the confidence of the parsed knowledge items is higher than the parsed ones from other person like "Tom". Besides the first person priority, there are also other ones, such as majority priority and friend relationship priority. Furthermore, the priority level of different rules is also specified.

For representative image generation, considering the storage limitation of the knowledge graph and representativeness of images, we should remove both wrong recognized images and redundant images. We utilize the extracted deep visual features VGG features [Simonyan and Zisserman, 2014] to calculate the mean of all image features. We then use a

clustering method and removed the image farthest from the clustering center, as this image is the wrong-recognized images with higher probability in the interaction process. Meanwhile, we calculate pair-wise similarity matrix to remove much similar images to reduce the redundancy.

Besides these two components of DTMAL, the key of DTMAL is the fusion and integration of these two tracks as a unified system.

3 Dual Track Fusion

During the interaction process, the obtained information is simultaneously used for both knowledge item learning and visual recognition improvement. Our system fuses the language analysis, visual recognition and the representation of multimodal information to obtain relevant information about the user, objects and speech. The knowledge extraction module utilizes the symbol grounding and semantic parsing techniques to extract rich triple knowledge items and visual representation pairs for the dual track learning process. For example, if the user Scarlett says: "I like apple" with an apple in his hand, the system obtains one knowledge item in the triple form<like, Scarlett, apple> and the visual representation of the apple <feature, label> simultaneously.

The fusion of the two tracks is a mutual promotion process. In our system, on the one hand, learning more relevant knowledge items will strengthen the understanding of the object which is helpful for object recognition. The knowledge items in our multi-modal graph improve the recognition ability of the system. For example, when the system recognizes an apple, it not only knows it is an apple, but also utilizes the symbol grounding (Eqn. (2)) and CCG parsing (Eqn. (3)) to know who likes apple. On the other hand, visual recognition is directly used for multimodal knowledge item learning. The recognized object is usually an element of the knowledge item. Thus the improvement of visual recognition ability can lead to more accurate and reliable knowledge item extraction. For example, if one person says: "I like this" with an apple in his hand, the visual object recognition track will recognize the visual feature of "this" as an apple and the people as "Scarlett" via HIL. And then the system obtains one knowledge item in the triple form <like, Scarlett, apple>. During the dual track learning process, the system will find out whether the information is learned or not. For a new object, according to Eqn.(1), the system will learn new related knowledge items and improve the recognition performance for this object class. For the new information of known objects, the system will update relevant knowledge items and improve the recognition performance for the new available instances.

The system's interaction performance is based on the fusion of both tracks of cognitive abilities. The fusion of the dual track learning process makes the interaction more intelligent. When you have a talk with the intelligent system, it can not only recognize your face, the object you hold and your words, but also know some attributes of the objects and the relation between you and the object. The improvement of two tracks of automatic learning can lead to better experience of interaction. The system finally can "see", "hear" and "think" like human beings through the dual track fusion.

4 Experiments

4.1 Experimental Setup

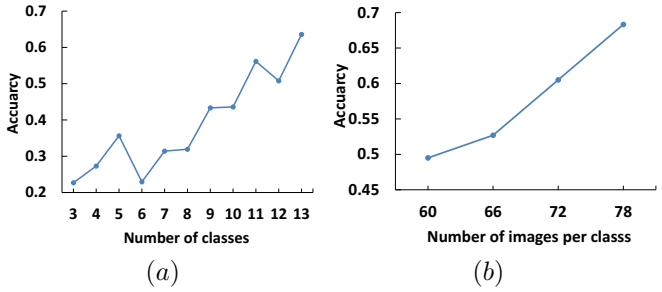
In order to validate the effectiveness of our system, there are 25 subjects in the study. We use 13 kinds of hand-held objects and define 16 kinds of relations. The objects are common ones, including "apple", "volleyball", "book", "bottle", "toothpaste", "stapler","keyboard", "flashlight", "wallet", "neck pillow", "the bag of milk", "the packet of biscuits" and "racket". The relations are also common ones, including common interpersonal relations and human-object relations, such as Friends, Like and BelongTo. Our system is implemented in an online and interactive way: these subjects propose a dialogue based on the 16 kinds of relations to our system, and the system learns new information in the interaction. The examples of interaction dialogues are these like "Wang *Like* play basketball" and "Li and Wang are *Friends*". Our system is run on a personal computer with an Intel Core (4 CPU) and 3.1GHz processor. We select the Kinect-1.0 device to capture objects and GPU with NVIDIA GeForce GTX 770 to extract deep features. The camera with Logitech HD 720P is used to capture the face.

4.2 Evaluation of The Vision Track:HIL

As manipulating objects with hands is a straight way for human-machine interaction , we therefore focus on incremental learning on the hand-held objects. Our HIL track utilizes both RGB information, the depth and skeletal information from Kinect. During interaction, we collect the object label and RGB-D information automatically by recognizing user's voice and capturing images of the object. We then follow the segmentation method [Lv *et al.*, 2015] to segment objects, and fuse the CNN features from RGB and depth features, which are extracted separately from the AlexNet network [Krizhevsky *et al.*, 2012], leading to the fused 8196-D features as the final feature representation. Accuracy is used to evaluate the classification performance.

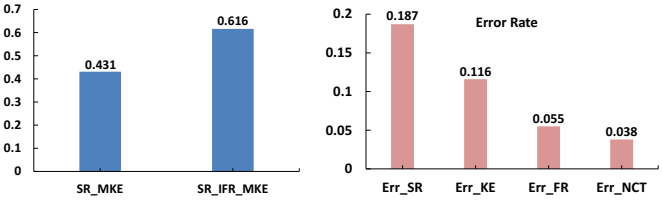
This evaluation of HIL consists of two parts: class-incremental learning and data-incremental learning. For class-incremental learning, we randomly select 3 classes as the source concepts to train a 3-class source model. And then the model learns the remaining 10 classes. We introduced a new class for class-incremental learning at a time. For each time, we capture 60 images per class as the training data for each incremental learning. The 3-class model turns into a 13-class target model after class-incremental learning. We repeat this process 3 times to average the results. Note that when we train the classifier of this new class, we need to say "This is an XX", where "XX" is the object. For data-incremental learning, we use the 13-class model as the source model. The model then learns new samples of the 13 classes via data-incremental learning. We conduct 3 groups of this experiment. For each group, we introduced new samples of one class at a time and repeated 13 times to learn new samples of the 13 classes. For each time, we take 6 images as the training data for each incremental learning.

Fig. 3 shows the results of HIL . From Fig. 3(a), we can see that the accuracy on the test dataset shows steadily increasing performance, indicating the algorithm is able to learn the



(a) (b)

Figure 3: Performance of HIL: (a) class-incremental learning and (b) data-incremenatl learning



(a) (b)

Figure 4: Performance of MKE: (a) the accuracy from two different settings in MKE and (b) the error rate from different modules among all extracted knowledge items

new classes. The slight drop may be caused that some new concepts is difficult to learn, especially in our interaction of the real-world scenarios. For data-incremental learning, as shown in Fig. 3(b), the source model is the 13-class classifier. To avoid the influence of the data imbalance, we add a fixed amount of new data to every source concept. For each time we capture 6 new images per class as the training data for each data-incremental learning. After three steps, the amount of data for each source class increases from 60 to 78 images. From Fig. 3(b), we can see that the accuracy of data-incremental learning grew up because of increasing data volume.

4.3 Evaluation of The Knowledge Track:MKE

In this evaluation, for each subject, the source model, the new object class or instances of the source concepts are in random order. Each subject interacts 25 times. For each interaction, he/she is holding or not holding one object from 13 kinds of objects. Then he/she proposes a dialogue based on the hand-held object and pre-defined 16 kinds of relations. The system learns new information and responds to the subject. For evaluation, our system can extract the knowledge items from each interaction, we label correctly extracted items. Accuracy is used as the performance metric, which is defined as the ratio between correctly extracted items and all the extracted items. Our MKE track consists of two cases: (1) Speech Recognition + Multimodal Knowledge Extraction (SR_MKE) and (2) Speech Recognition + Incremental Face Recognition + Multimodal Knowledge Extraction (SR_IFR_MKE). Therefore, we compare these two cases.

Fig. 4 shows the results. We can see that the performance of SR_IFR_MKE significantly outperforms SR_MKE. There is about 43% improvement. The reason is that in the interaction, there are many cases involving face recognition. Therefore introducing the incremental face recognition increases the accuracy. For example, in common scenarios, the subject often says sentences, such as "I like XX", "I am XX", where XX is the name of the person. Without incremental face recognition, these sentences are not succesfully parsed into the triple knowledge. Furthermore, we analyze the errors caused by all possible factors in SR_IFR_MKE, from the following four aspects: the error of speech recognition (Err_SR) , the error of knowledge extraction (Err_KE), the error of face recognition (Err_FR) and network connection timeout (Err_NCT). The results are shown in Fig. 4 (b). In all the interaction, there are total 138 wrong extracted knowledge items. We can see that the main error sources are caused by speech recognition and knowledge extraction. The reason is that for speech recognition, besides the algorithm, features and pronunciation accuracy of the different subjects are probably the main factors affecting the accuracy of speech recognition. In contrast, because of the fast development of the face recognition, the error caused by face recognition is very low.

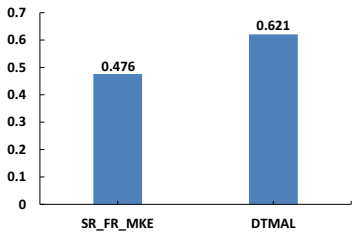


Figure 5: The accuracy from the baseline and DTMAL

4.4 Evaluation of DTMAL

In order to verify the effectiveness of HIL in DTMAL, similar to the experimental setting for the MKE track, the average times of interaction for each subject is about 25. Different from the MKE track, the faces from all the subjects have been registered. In addition, the interaction involves many cases of object recognition. For each subject, the source model is randomly selected. For example, the trained source classes for subject A is "apple","volleyball" and "book", while "toothpaste", "stapler" and "keyboard" for subject B. In the process of interaction, the hand-held new object class or new object instances of the source concepts appear in random order. Since there is little work on dual track multimodal automatic learning, we cannot compare our method with existing methods. Therefore, we consider the following baseline for comparison: Speech Recognition + Face Recognition + Multimodal Knowledge Extraction (SR_FR_MKE). The difference between this baseline and the DTMAL is that DTMAL introduced the HIL.

As shown in Fig. 5, we can see that after introducing the incremental object recognition, DTMAL has 30% increase than MKE. This verified that the incremental object recognition is capable of improving the extraction of the knowledge items. Further analysis shows that there are totally

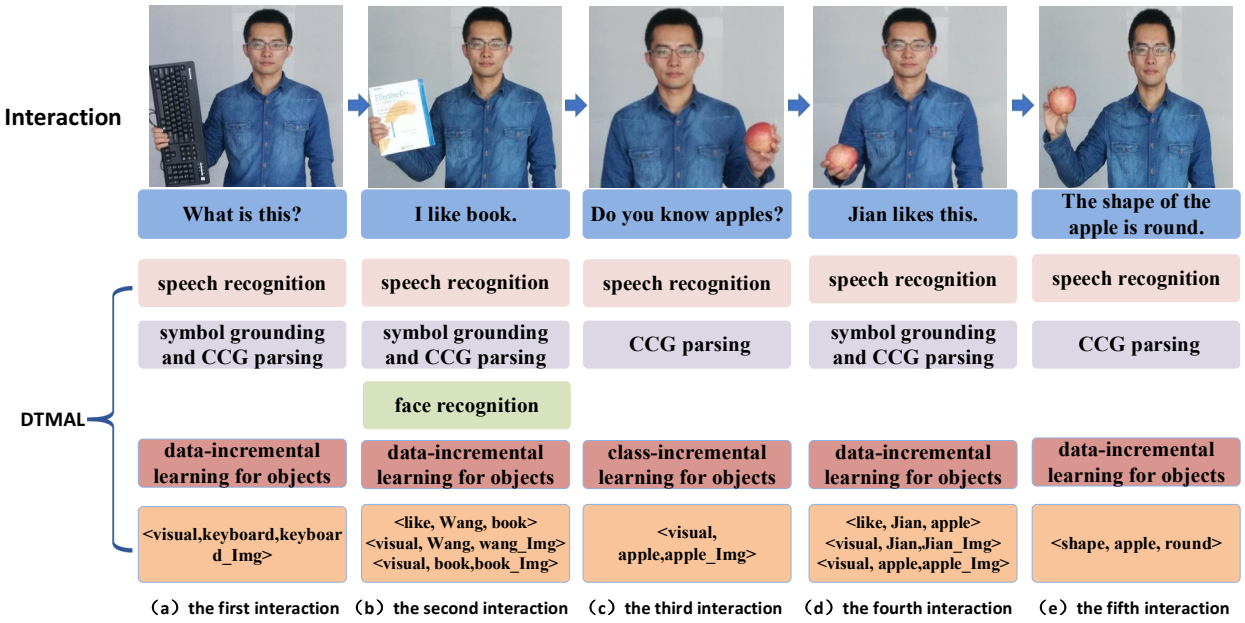


Figure 6: The case study for one process of interaction in DTMAL. Note that the source concepts for HIL includes "book","bottle" and "keyboard".

148 knowledge items, which are successfully extracted from incremental object recognition, where 78 knowledge items from class-incremental learning and 70 items from data-incremental learning.

Fig. 6 shows one process of interaction in DTMAL. We can see that DTMAL utilizes different technologies, such as incremental face recognition, HIL and MKE in the process. For example, in the second interaction, DTMAL utilizes the face recognition to recognize the face, and then uses the symbol grounding and CCG parsing method to replace "I" with the person "Wang". In the third interaction, since the object class "apple" is not in DTMAL, DTMAL utilizes the apple image and corresponding labels to conduct the class-incremental learning. Therefore, in the fourth interaction, DTMAL can recognize the object "apple", and then uses the symbol grounding and CCG parsing method to replace "this" with "apple". As a result, DTMAL sucessfully extracted the knowledge items, such as <like, Jian, Apple>. Note that our knowledge graph is multimodal. Therefore, in the third interaction process, the knowledge item<visual, apple, apple_Img> is also extracted and it represents that there is a "visual " relationship between apple entity and its image. The case study further verified the effectiveness of DTMAL in fusing HIL and MKE to enable the incremental learning of visual recognition and automatic growth of knowledge items.

5 Conclusions

In this paper, we proposed a Dual Track Multimodal Automatic Learning (DTMAL) system, which enables both the incremental learning of visual recognition and automatic growth of knowledge items by utilizing multimodal knowledge extraction and hybrid incremental learning methods. Furthermore, different recognition methods (e.g., speech recognition and face recognition) and fusing strategies are used for strengthening the automatic learning process. The experimental results have demonstrated the effectiveness of the proposed system.

This work is an effort in improving the automatic learning abilities of robots. The learning mechanism of DTMAL is reasonable for intelligent human-machine interaction system. We hope this work could serve as a good chance to further the agenda of intelligent human-machine interaction systems in this community. Our system is scalable and flexiable. Therefore, our future work can be extended in many directions. For example, we plan to continuely improving the system including supporting the synonyms in incremental learning and more complex interaction.

Acknowledegements

This work was supported in part by the National Natural Science Foundation of China (61532018, 61322212 and 61602437), in part by the Beijing Municipal Commission of Science and Technology (D161100001816001), in part by Beijing Natural Science Foundation (4174106), in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals, and in part by China Postdoctoral Science Foundation (2016M590135).

References

[Al-Omari *et al.*, 2017] M. Al-Omari, P. Duckworth, D. C. Hogg, and A. G. Cohn. Natural language acquisition and grounding for embodied robotic systems. In *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4349–4356, 2017.

[Barsalou, 1999] L. W. Barsalou. Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22:577–660, 1999.

[Bordes *et al.*, 2005] Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, December 2005.

[Cauwenberghs and Poggio, 2000] Gert Cauwenberghs and Tomaso A. Poggio. Incremental and decremental support vector machine learning. In *Conference on Neural Information Processing Systems*, pages 409–415, 2000.

[Gotts, 2016] Stephen J. Gotts. Incremental learning of perceptual and conceptual representations and the puzzle of neural repetition suppression. *Psychonomic Bulletin & Review*, 23(4):1055–1071, 2016.

[Hakkani-Tür *et al.*, 2014] Dilek Hakkani-Tür, Asli Çelikyilmaz, Larry P. Heck, Gökhan Tür, and Geoffrey Zweig. Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding. In *Annual Conference of the International Speech Communication Association*, pages 2113–2117, 2014.

[Hixon *et al.*, 2015] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. Learning knowledge graphs for question answering through conversational dialog. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 851–861, 2015.

[Johnson *et al.*, 2015] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Image retrieval using scene graphs. In *In Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.

[Klein and Manning, 2003] Dan Klein and Christopher D. Manning. A* parsing: Fast exact viterbi parse selection. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Conference on Neural Information Processing Systems*, pages 1106–1114, 2012.

[Kuzborskij *et al.*, 2013] I. Kuzborskij, F. Orabona, and B. Caputo. From n to n+1: Multiclass transfer incremental learning. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365, June 2013.

[Lampert *et al.*, 2009] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, June 2009.

[Lewis and Steedman, 2014] Mike Lewis and Mark Steedman. A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 990–1000, 2014.

[Liu *et al.*, 2016] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. VIPLFacenet: An open source deep face recognition SDK. *Computing Research Repository*, abs/1609.03892, 2016.

[Lv *et al.*, 2015] Xiong Lv, Shuqiang Jiang, Luis Herranz, and Shuang Wang. Rgb-d hand-held object recognition based on heterogeneous feature fusion. *Journal of Computer Science and Technology*, 30(2):340–352, 2015.

[Murata *et al.*, 2002] Noboru Murata, Motoaki Kawanabe, Andreas Ziehe, Klaus-Robert Müller, and Shun ichi Amari. On-line learning in changing environments with applications in supervised and unsupervised learning. volume 15, pages 743–760, 2002.

[Parde *et al.*, 2015] Natalie Parde, Adam Hair, Michalis Papakostas, Konstantinos Tsiakas, Maria Dagioglou, Vangelis Karkaletsis, and Rodney D Nielsen. Grounding the meaning of words through vision and interactive gameplay. In *International Conference on Artificial Intelligence*, pages 1895–1901, 2015.

[Paul *et al.*, 2016] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M. Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Robotics: Science and Systems*, 2016.

[Ristin *et al.*, 2015] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool. Incremental learning of random forests for large-scale image classification. *Proceedings IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2015.

[Ruping, 2001] S. Ruping. Incremental learning with support vector machines. In *Proceedings IEEE International Conference on Data Mining*, pages 641–642, 2001.

[Saxena *et al.*, 2014] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K. Misra, and Hema S. Koppula. Robobrain: Large-scale knowledge engine for robots, April 12 2014.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository*, abs/1409.1556, 2014.

[Tenorth and Beetz, 2013] Moritz Tenorth and Michael Beetz. Knowrob: A knowledge processing infrastructure for cognition-enabled robots. *International Journal of Robotics Research*, 32(5):566–590, 2013.

[Thomason *et al.*, 2016] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3477–3483, July 2016.

[Zhu *et al.*, 2015] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base for visual question answering. *Computing Research Repository*, abs/1507.05670, 2015.