2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)

Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

Sinan Chen [1], Sachio Saiki [1], Masahide Nakamura [1,2]
[1]Graduate School of System Informatics, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe, 657-0011, Japan
[2]RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan
Email: chensinan@ws.cs.kobe-u.ac.jp, sachio@carp.kobe-u.ac.jp, masa-n@cs.kobe-u.ac.jp

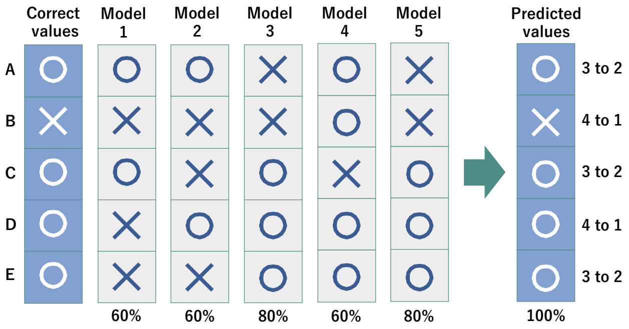
***Abstract*—To implement fine-grained context recognition affordable for general households, we are studying techniques that integrate image-based cognitive API and light-weight machine learning. Specifically, our method first captures images of a target space in different context, then sends them to the cognitive API. For each image, the API returns a set of words, called tags, representing concepts recognized in the picture. Regarding these tags as features of the target context, we apply the supervised machine learning. Our preliminary results with a commercial API showed that the overall accuracy was more than 90%, however, the accuracy decreased for contexts with multiple people (e.g., "General meeting", "Dining together" and "Play games"). The goal of this paper is to improve the recognition accuracy of such difficult contexts, with preserving the affordability to general households. In the proposed method, we use multiple cognitive APIs. For each API, we construct an independent recognition model. Then, the context is determined by majority voting among results of the independent models. Experimental evaluation with five commercial APIs shows that the ensemble of the five independent models achieved 98% of overall accuracy, where the individual models complement mutual limits of recognition.**

***Index Terms*—context recognition, cognitive APIs, machine learning, majority voting, smart home**

I. Introduction

Recognizing *fine-grained contexts* within individual houses is a key technology for next-generation smart home services, such as elderly monitoring [1] [2] [3], autonomous security [4], and personalized healthcare [5] [6]. It has been studied for many years in the field of *ubiquitous computing*. The traditional ubiquitous computing employs ambient sensors, wearable sensors, and indoor positioning systems are installed at home to retrieve various contexts. In recent years, the emerging *deep learning* [7] allows the system to recognize *multimedia*. Since image, voice, and text usually contain richer information than the conventional sensor data, it is promising to use such multimedia data for recognizing fine-grained home contexts.

However, these existing technologies are yet far from practical use in general households, since they usually require expensive devices and resources at home. It is difficult for ordinary users to operate and maintain complex ubiquitous devices at home on a daily basis. One may try to recognize home contexts via image recognition based on deep learning. However, constructing a custom recognition model dedicated for a single house requires a huge amount of labeled datasets and computing resources [7] [8]. Thus, there is still a big gap between the research and real life.



	Correct values	Model 1	Model 2	Model 3	Model 4	Model 5	Predicted values	
A	○	○	○	×	○	×	○	3 to 2
B	×	×	×	×	○	×	×	4 to 1
C	○	○	×	○	×	○	○	3 to 2
D	○	×	○	○	○	○	○	4 to 1
E	○	×	×	○	○	○	○	3 to 2
		60%	60%	80%	60%	80%	100%	

Fig. 1. Example of the majority voting with ensemble learning

Our long-term goal is to implement fine-grained home context recognition that can adapt to custom contexts in every single house, and can achieve accurate recognition with a small amount of resources affordable by general households. To achieve the goal, we are currently investigating techniques that integrate inexpensive camera devices, image-based cognitive API, and light-weight machine learning. The cognitive API is application program interface to cloud services that provide various recognition features as a service. Famous APIs include Microsoft Azure Computer Vision API [9] and IBM Watson Visual Recognition API [10].

In our preliminary study [11], we have developed a method based on supervised machine learning. Using the camera device, the proposed method first captures images of a target space in different contexts. It then sends the images to the cognitive API. For each image, the API returns a set of words, called *tags*, which represent concepts that the API recognized within the image. Considering these tags as *features* of the target context, the proposed method constructs a multi-class classifier using an ordinary machine learning algorithm. We conducted an experiment where seven kinds of contexts within our laboratory were recognized from images. The classifier was constructed by Multiclass Neural Network from the tags derived by Microsoft Azure Computer Vision API. The overall accuracy achieved more than 90%. However, the accuracy significantly decreased for contexts with multiple people (e.g., "General meeting", "Dining together", "Play games").

978-1-7281-2542-8/19/$31.00 ©2019 IEEE
DOI 10.1109/CyberC.2019.00068

The goal of this paper is to improve the recognition accuracy of such difficult contexts. For this purpose, we propose a new method that uses multiple cognitive APIs for *ensemble learning* [12]. Specifically, for each cognitive API, we first construct an independent recognition model based on the tags derived from the API. We then improve the accuracy of recognizing contexts by *majority voting* among results of the multiple independent models. Figure 1 shows an intuitive illustration of the mechanism. Since different APIs observe the same image from different perspectives, they would be able to complement mutual limits of recognition capability.

In order to evaluate the performance of the proposed method, we have conducted an experiment to recognize the seven contexts within our laboratory: “Dining together”, “General meeting”, “Nobody”, “One-to-one meeting”, “Personal study”, “Play games”, and “Room cleaning”. For each context, we selected 100 representative images. Each of these images was sent to the following five commercial APIs: Microsoft Azure Computer Vision API [9], IBM Watson Visual Recognition API [10], Clarifai API [13], Imagga REST API [14], and Paralleldots API [15]. Considering a set of tags derived from each image as a *document*, we convert the tags into a vector representation using *TF-IDF (Term Frequency - Inverse Document Frequency)* [16].

The vectorized tag sets and the corresponding context labels were imported to Microsoft Azure Machine Learning Studio [17], where five independent recognition models were constructed with Multiclass Neural Network. The final recognition result was determined by the majority voting among results of the five models. The experimental results showed that the overall accuracy of the five models varied between 0.77 to 0.94. In contrast, the overall accuracy by the majority voting of the five models reached 0.98. Through context-wise analysis, it was shown that the recognition accuracy of “General meeting”, “Nobody”, “One-to-one meeting”, “Play games” were 1.00, and that the one “Dining together”, “Personal study”, “Room cleaning” were 0.96. Thus, the recognition accuracy was significantly improved by the proposed method.

II. Preliminaries

A. Recognizing Fine-Grained Home Contexts

The *home context* refers to any situational information at home, including daily activities of residents, the environment in the house, and the status of the room. Typical home contexts include, for example, “residents are in the dining room”, “it is warm in the dining room”, “the dining room is clean”, and “the light in the dining room is on”.

We use the term *fine-grained* home context to represent a home context that is more concrete and is specifically defined by individual houses, residents, and environment, for a special purpose of application. For example, suppose that a son is worried about his old parents living in a remote place. Then, the contexts like “parents are eating breakfast in a dining room”, “father is taking medicine in a dining room”, “mother is cleaning a dining room”, are crucial information for the son. If these fine-grained contexts can be recognized by an elderly monitoring system, and the information is regularly sent to the son, it would be a great value for the son.

B. Technical Challenges

There exist a lot of research and development of home context recognition. However, the technology is not yet widely spread within general households. For this, we consider that there are two big challenges. The first challenge lies in *individuality*. As seen in the above example, the fine-grained contexts are defined by every user depending on a special purpose. Also, the layout, the environment, and the configuration of the target space vary from one house to another. Therefore, it is hard to construct a *universal* recognition model.

The second challenge is *acceptability*. Most existing technologies are developed and tested on research labs or dedicated smart homes, and few of them are actually operated on general households. In order for the technology to be accepted, it should be easy to operate and maintain, should be affordable enough, and should not be intrusive for daily life. Of course, the security and privacy issues influence the acceptability. The user of the technology must be fully convinced what data is collected for what purpose and consumed by whom.

C. Image Recognition API of Cognitive Services

The cognitive service is a cloud service that provides the capability to understand multimedia data, based on sophisticated machine-learning algorithms powered by big data and large-scale computing. The cognitive API is application program interface, with which developers can easily integrate powerful recognition features in their own applications. Among various APIs, we especially focus on *image recognition* API.

An image recognition API receives an image from an external application, extracts specific information from the image, and returns the information. The information usually contains a set of words called *tags*, representing objects and concepts that the API recognized in the given image. An example of tags is like: [living, room, indoors, classroom, basement, support, supporting structure]. The information of interest and the way of recognizing the image vary among individual services.

Currently, various kinds of image recognition APIs are available, such as Microsoft Azure Computer Vision API [9], IBM Watson Visual Recognition API [10], Clarifai API [13], Imagga REST API [14], and Paralleldots API [15].

D. Preliminary Study

Our interest is to apply these image recognition APIs to implement affordable context sensing at home. More specifically, we aim to realize a system, where a simple edge system just capturing and pre-processing images is deployed at home, and all heavy tasks of image recognition are delegated to the cognitive service on the cloud. Note, however, that the existing cognitive APIs are trained for general-purpose image recognition. Therefore, the API is not optimized for individual fine-grained home contexts.

In our previous study [11], we developed a method of fine-grained home context recognition based on supervised

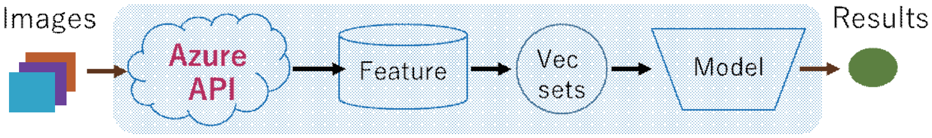


Fig. 2. Recognizing fine-grained home contexts [11]

machine learning. Figure 2 depicts the outline of the method. The key idea was to use the tags extracted by the image recognition API (i.e., Microsoft Azure Computer Vision API) as *features* for machine learning. Since every image was converted to a set of words by the API, the expensive deep learning was no more needed. In the preliminary study, we vectorized the extracted tag sets using the *TF-IDF* method [16], and constructed a multi-class classifier using Multiclass Neural Network on Microsoft Azure Machine Learning Studio [17].

We conducted an experiment recognizing seven kinds of contexts in our laboratory: "Dining together", "General meeting", "Nobody", "One-to-one meeting", "Personal study", "Play games", "Room cleaning". The experimental results showed that the overall accuracy achieved more than 90%. However, the accuracy significantly decreased for contexts with multiple people. For instance, the classifier sometime could not distinguish "General meeting", "Dining together", and "Play games". Improving the recognition accuracy for these difficult contests is the main concern of this paper.

III. Proposed Method

A. Outline

Figure 3 shows the overview of the proposed method. The key idea to improve the accuracy is to use *multiple* cognitive APIs, introducing the concept of ensemble learning. The way of image recognition is different from one API to another. Therefore, we can construct multiple classifiers with different perspectives. These classifiers may return different prediction results for the same image. However, taking *majority voting* derives the context with maximum likelihood.

Although there are various methods known for the ensemble learning, the proposed method constructs an independent recognition model per a single cognitive API, as shown in Figure 3. In the following sections, we first describe the method that constructs a recognition model from a single API. Then, we present the method of majority voting that integrates multiple models.

B. Constructing Context Recognition Model

This section describes how to construct a recognition model of fine-grained home contexts using a given image recognition API. The procedure consists of the following five steps.

STEP 1: Acquiring Data
A user of the proposed method first defines a set $C = \{c_1, c_2, ..., c_l\}$ of home contexts to be recognized. Then, the user deploys a camera device in the target space to observe. The user configures the device so as to take a snapshot of the space periodically with an appropriate interval.

STEP 2: Creating datasets
For each context $c_i \in C$, the user manually selects representative n images $IMG(c_i) = \{img_{i1}, img_{i2}, ..., img_{in}\}$ that well expose c_i from all images obtained in Step 1. At this time, the total $l \times n$ images are sampled as datasets. Then, the n images in $IMG(c_i)$ are split into two sets $train(c_i)$ an $test(c_i)$, which are the training dataset with α images and the test dataset with $n - \alpha$ images, respectively.

STEP 3: Extracting tags as features
For every image img_{ij} in $train(c_i)$, the method sends img_{ij} to an image recognition API, and obtains a set $xTag(img_{ij}) = \{t_1, t_2, ...\}$, where $t_1, t_2, ...$ are tags that the API extracted from img_{ij}. The method performs the same process for $test(c_i)$ and obtain $yTag(img_{i'j'})$. At this step, the total $l \times n$ tag sets.

STEP 4: Converting tags into vectors
Regarding every $xTag(img_{ij})$ as a document, and the whole tag sets as a document corpus, the method transforms $xTag(img_{ij})$ into a *vector representation* $xVec(img_{ij}) = [v_1, v_2, ...]$, where v_r represents a numerical value characterizing r-th tag. Famous document vectorization techniques include TF-IDF [16], Word2Vec [18] and Doc2Vec [19]. The selection of the vector representation is up to the user. Similarly, the method converts $yTag(img_{i'j'})$ into $yVec(img_{i'j'})$ using the same vector representation.

STEP 5: Constructing a classifier
Taking $xVec(img_{ij})$ $(1 \leq i \leq l, 1 \leq j \leq \alpha)$ as predictors and c_i $(1 \leq i \leq l)$ as a target label, the method executes a supervised machine learning algorithm to generate a *multi-class classifier* CLS. For a given vector $v = [v_1, v_2, ...]$, if CLS returns a context c_i, it means that the context of the original image of v is recognized as c_i. The accuracy of CLS can be evaluated by $yVec(img_{i'j'})$ to see if CLS returns the correct context $c_{i'}$.

C. Integrating Models Generated from Different APIs

This section describes how to construct a whole recognition model by integrating multiple recognition model.

STEP 6: Constructing multiple classifiers
By repeating STEP 3 to STEP 5 for different image recognition APIs, the proposed method constructs m independent recognition models. Note that training and test datasets created in STEP 1 and STEP 2 can be reused and shared among different models. As a result of the model construction, we have a set of classifiers $CLS_1, CLS_2, ..., CLS_m$.

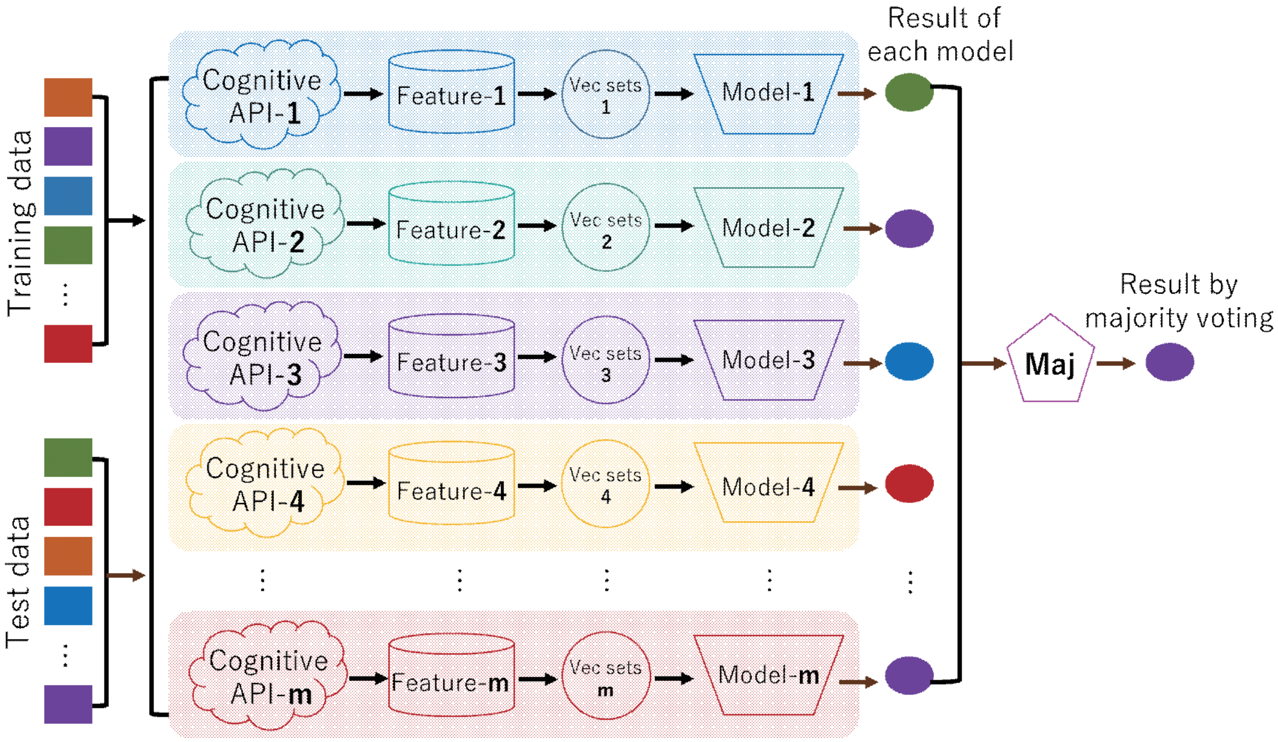


Fig. 3. The overview of proposed method using multiple cognitive APIs

STEP 7: Add vectorizer for new images
For each CLS_q, the method generates a vectorizer VEC_q, which transforms a given image img into a vector representation $xVec(img)$ through q-th cognitive API. Now, if we input any new image of the target space, the concatenation $VEC_q + CLS_q$ outputs c_i as a predicted context class.

STEP 8: Integrate multiple models
The method adds a fork module F which sends a given image simultaneously to m recognition models $VEC_q + CLS_q$ $(1 \leq q \leq m)$. Also, the method adds a majority voting module Maj, which receives m outputs $c^1, c^2, ..., c^m$ from $VEC_q + CLS_q$ $(1 \leq q \leq m)$, and returns $mode(c^1, c^2, ..., c^m)$. This completes the model construction.

IV. Experimental Evaluation

A. Experimental Setup

We have conducted an experiment recognizing fine-grained contexts in a shared space of our laboratory. First, we installed an USB camera in a fixed position to acquire images of the space. We then developed a program that takes a snapshot with the camera every 5 seconds, and uploads the image in a server. The images have been accumulated since July 2018.

The target shared space is used by members of our laboratory for various activities. In this experiment, we have chosen the seven kinds of fine-grained contexts: “Dining together”, “General meeting”, “Nobody”, “One-to-one meeting”, “Personal study”, “Play games”, and “Room cleaning”.

For each context, we selected and labeled 100 representative images from the server, taken on different date. Figure 4 shows one of the representatives for each context and USB camera. We then randomized the order of a total of 700 image data, and split them into half as training data and test data.

B. Building Recognition Model

Based on the proposed method, we built a recognition model for the seven contexts. The following five cognitive APIs were used to extract tags from each image: Microsoft Azure Computer Vision, IBM Watson, Clarifai API, Imagga REST API, and Paralleldots API. Table I shows example of the tags extracted from an image of “Room cleaning”. We can see that different APIs see the same image from different perspectives. Each tag sets extracted from an image was then transformed into a vector representation using the TF-IDF method.

We imported the datasets and the corresponding context labels on Microsoft Azure Machine Learning Studio. For each cognitive API, we trained a classifier using Multiclass Neural Network with default setting. Each of the five trained models was evaluated by the test data to see the performance of individual models. Finally, the five models were combined by a majority voting function. For every image given, it produces the majority of contexts recognized by the five models.

C. Results

Table II summarizes the results listing the overall accuracy and context-wise accuracy. The first five rows represent the

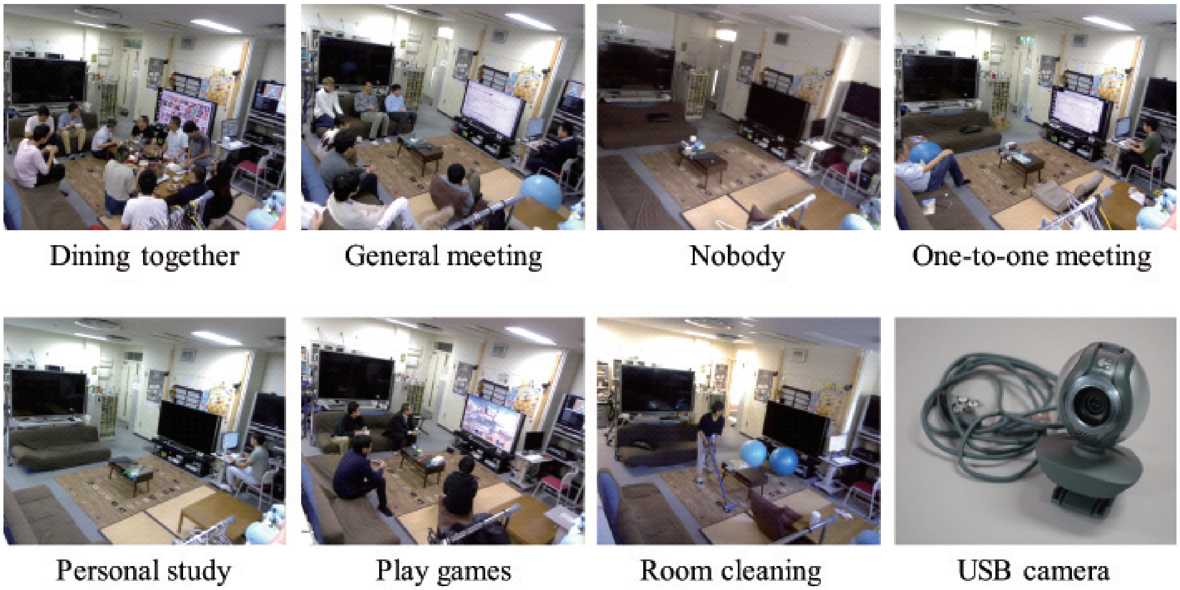


Fig. 4. Representative images of fine-grained contexts and USB camera

TABLE I
TAG SETS THAT DIFFERENT APIS EXTRACTED FROM AN IMAGE

Names of API used	The tags extracted from an image of "Room cleaning"
Microsoft Azure Computer Vision API	indoor,living,room,table,television,furniture,sitting,messy,cluttered,area,computer,fireplace,filled,fire,bedroom,large,flat,view,screen,desk,video,woman,young,playing,bed,game,man,standing,dog,people
IBM Watson Visual Recognition API	living,room,indoors,classroom,basement,support,supporting structure
Clarifai API	room,furniture,indoors,table,desk,seat,chair,trading,floor,interior,design,home,hospital,medicine,technology,window,mirror,business,computer,people,production
Imagga REST API	room,interior,furniture,table,home,house,modern,floor,decor,chair,sofa,design,wood,window,luxury,living,lamp,apartment,indoors,light,home,theater,building,office,architecture,comfortable,wall,residential,couch,inside,theater,carpet,desk,fireplace,kitchen,living,room,pillow,3d,structure,relax,seat,lighting,decoration,estate,glass,furnishings,style,bedroom,indoor,empty,domestic,real,decorate,relaxation,cozy,chairs,residence,family,rest,area,space,contemporary,comfort,equipment,monitor,leather,render,television,classroom,hardwood,nobody,vase,hotel,bed,business,elegance,clean,upscale,lifestyle,computer,studio,apartment,rug,new,plant,elegant,furnishing,stylish,guest,spacious,cabinet,ceiling,armchair,device,marble,restaurant,work,mirror,dining,ottoman,shelf,fixtures,wooden,suburbs,suite,suburban,dwelling,lounge,tile,display,fashion,place,book
ParallelDots API	Room,Interior,design,Property,Vehicle,Building,Home,Sport,venue,Screenshot,Furniture

results of the five individual models whose features were extracted by individual cognitive APIs. The last row represents the result obtained by the majority voting of the five models.

With regard to the overall accuracy, the majority voting achieved the accuracy of 0.98. Among the five models, the Imagga API-based model was the best (0.94), while the Paralledots API-based model was the lowest (0.77).

As for the context-wise accuracy, the performance of the five models was all different. For instance, let us compare the Watson API-based model and the Paralledots API-based model. The Watson was bad at recognizing “General meeting” (0.67), compared to the Paralledots did (0.89). Interestingly, however, the Watson was better at recognizing “One-to-one meeting” (0.80) than the Paralledots (0.45).

These limitations of the individual models were mutually complemented by the majority voting. The recognition accuracy of “Dining together”, “Personal study”, “Room cleaning” were 0.96, while the accuracy of “General meeting”, “Nobody”, “One-to-one meeting”, “Play games” were 1.00.

D. *Discussion*

In the proposed method, the recognition accuracy heavily depends on the *quality of tags* extracted by the cognitive API. The reason why the Paralledots-based model was bad at “One-to-one meeting” (0.45) was that (1) no distinctive word characterizing the context was found, and that (2) the number of words in the tag sets was relatively small.

The accuracy also depends on the *nature of context*. We found that contexts where people are dynamically moving (e.g., “Dining together”, “Room cleaning”) were relatively difficult to recognize. In such contexts, observable features are frequently changed from one image to another, for instance, positions of people, visible furniture and background. Therefore, the API may produce variable tag sets for the same

TABLE II
THE RECOGNITION ACCURACY RESULTS OF EACH COGNITIVE API-BASED MODEL AND MAJORITY VOTING IN THIS EXPERIMENT

Model names	Overall accuracy	Dining together	General meeting	Nobody	One-to-one meeting	Personal study	Play games
Azure API-based model	0.85	0.96	0.89	1.00	0.66	0.92	0.84
Watson API-based model	0.80	0.89	0.67	0.82	0.80	0.94	0.80
Clarifai API-based model	0.91	0.91	0.98	0.91	0.84	0.92	0.92
Imagga API-based model	0.94	0.96	0.93	1.00	0.89	0.96	0.92
Paralledots API-based model	0.77	0.80	0.89	0.93	0.45	0.88	0.67
Majority voting	0.98	0.96	1.00	1.00	1.00	0.96	1.00

context, which decreases the internal cohesion of the feature vectors.

Taking the majority voting was a great solution to improve the accuracy. In the typical ensemble learning, the individual classifiers should be weak to avoid overfitting. This is because the classifiers use the same features for the training. However, in our case we extract different features by different APIs. Since the individual models are trained by different features, it does not cause the overfitting problem.

V. Related Work

The activity recognition in a smart house has been widely studied in the field of ubiquitous computing. Nakamura et al. [20] proposed a system that recognizes activities of residents using big data accumulated within a smart house. Ueda et al. [21] also proposed an activity recognition system using ultrasonic sensors and indoor positioning systems within a smart house. Although the performance of these systems is great, they are yet too expensive for general households.

The activity recognition with deep learning becomes a hot topic recently (e.g., [7] [8]). Although the deep learning is a powerful approach to recognize image data, a huge amount of data is required to build a high-quality model. Therefore, it is unrealistic for individual households to prepare a huge amount of labeled datasets for custom fine-grained contexts.

Menicatti et al. [22] proposed a framework that recognizes indoor scenes and daily activities using cloud-based computer vision. Their concept and aim are similar to our method. However, the way of encoding tags is based on a Naive Bayes model where each word is present or not. Also, the method is supposed to be executed on a mobile robot, where the image is dynamically changed. Thus, the method and the premise are different from ours.

Research in [23] investigates the influence of person's cultural information towards vision-based activity recognition at home. The accuracy of the fine-grained context recognition would be improved by taking such personal information into machine learning. We would like investigate this perspective in our future work.

VI. Conclusion

We have presented a method that recognizes fine-grained home contexts using multiple cognitive APIs. To achieve affordable context recognition in general households, the proposed method delegates the image recognition task to cognitive APIs in the cloud, and use retrieved tags (words) as features of the supervised machine learning. Since different APIs return different tag sets, the proposed method constructs an independent classifier for each API. These independent classifiers are integrated with a majority voting function, which achieves a very accurate context recognition model.

We also conducted an experiment with five commercial cognitive APIs. Employing the TF-IDF method as the vector representation, and Multiclass Neural Network as the learning algorithm, the proposed method achieved 0.98 of overall accuracy for recognizing seven contexts in our laboratory.

As future work, we should evaluate the performance limitations in practical scenes with more difficult contexts. Also, we are planning the integration of the method with actual systems (e.g., elderly monitoring system [24]). Investigating cultural information to improve the accuracy is also an interesting challenge.

Acknowledgments: This research was partially supported by JSPS KAKENHI Grant Numbers JP19H01138, JP17H00731, JP18H03242, JP18H03342, JP19K02973.

References

[1] L. Vuegen, B. Van Den Broeck, P. Karsmakers, H. Van hamme, and B. Vanrumste, "Automatic monitoring of activities of daily living based on real-life acoustic sensor data: a preliminary study," in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*. Grenoble, France: Association for Computational Linguistics, August 2013, pp. 113–118. [Online]. Available: https://www.aclweb.org/anthology/W13-3918

[2] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring Activities of Daily Living in Smart Homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, pp. 81–94, March 2016.

[3] A. Marjan, R. Jennifer, K. Uwe, K. Annica, K. L. B. Eva, T. Nicolas, V. Thiemo, and L. Amy, "An ontology-based context-aware system for smart homes: E-care@home," *Sensors*, vol. 17, no. 7, 2017.

[4] Y. Ashibani, D. Kauling, and Q. H. Mahmoud, "A context-aware authentication framework for smart homes," *CCECE 2017*, May 2017.

[5] K. Deeba and R. K. Saravanaguru, "Context-aware healthcare system based on iot - smart home caregivers system (shcs)," *ICICCS 2018*, June 2018.

[6] S. C. Joo, C. W. Jeong, and S. J. Park, "Context based dynamic security service for healthcare adaptive application in home environments," *Software Technologies for Future Dependable Distributed Systems*, March 2009.

[7] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *CoRR*, vol. abs/1707.03502, 2017. [Online]. Available: http://arxiv.org/abs/1707.03502

[8] A. Brenon, F. Portet, and M. Vacher, "Context Feature Learning through Deep Learning for Adaptive Context-Aware Decision Making in the Home," in *The 14th International Conference on Intelligent Environments*, Rome, Italy, June 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01802747

[9] Microsoft Azure, "Visual recognition," https://azure.microsoft.com/ja-jp/services/cognitive-services/computer-vision/, visited on April 15, 2019.

[10] IBM Watson, "Computer vision," https://cloud.ibm.com/apidocs/visual-recognition, visited on February 1, 2019.

[11] S. Chen, S. Saiki, and M. Nakamura, "Proposal of home context recognition method using feature values of cognitive api," *IEICE technical report*, vol. 118, no. 511, SC2018-38, pp. 7–12, August 2019.

[12] CodExa, "Explanation of the mechanism of ensemble learning for three types," https://www.codexa.net/what-is-ensemble-learning/, visited on February 1, 2019.

[13] Clarifai, "Transforming enterprises with computer vision ai," https://clarifai.com/, visited on April 15, 2019.

[14] Imagga, "Imagga api," https://docs.imagga.com/, visited on April 15, 2019.

[15] ParallelDots, "Image recognition," https://www.paralleldots.com/object-recognizer, visited on April 15, 2019.

[16] "Vectorize documents with TF-IDF," http://ailaby.com/tfidf/, August 2016, visited on February 1, 2019.

[17] M. Azure, "Azure machine learning studio," https://azure.microsoft.com/ja-jp/services/machine-learning-studio/, visited on February 1, 2019.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[19] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1188–II–1196. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045025

[20] S. Nakamura, A. Hiromori, H. Yamaguchi, T. Higashino, Y. Yamaguchi, and Y. Shimoda, "Activity sensing, analysis and recommendation in smarthouse," *Multimedia, Distributed Collaboration and Mobile Symposium 2014 Proceedings*, vol. 2014, pp. 1557–1566, July 2014.

[21] K. Ueda, M. Tamai, and K. Yasumoto, "A system for daily living activities recognition based on multiple sensing data in a smart home," *Multimedia, Distributed Collaboration and Mobile Symposium 2014 Proceedings*, vol. 2014, pp. 1884–1891, July 2014.

[22] R. Menicatti and A. Sgorbissa, "A cloud-based scene recognition framework for in-home assistive robots," *RO-MAN 2017*, December 2017.

[23] R. Menicatti, B. Bruno, and A. Sgorbissa, "Modelling the influence of cultural information on vision-based human home activity recognition," *CoRR*, vol. abs/1803.07915, 2018. [Online]. Available: http://arxiv.org/abs/1803.07915

[24] K. Tamamizu, S. Sakakibara, S. Saiki, M. Nakamura, and K. Yasuda, "Capturing activities of daily living for elderly at home based on environment change and speech dialog," in *Digital Human Modeling 2017 (DHM 2017)*, no. LNCS 10287. Springer International Publishing AG 2017, July 2017, pp. 183–194, vancouver, Canada.