Proposal of Home Context Recognition Method Using Feature Values of Cognitive API

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*—The emerging deep learning technology is a promising means for context recognition with multimedia data. We are interested in using the deep learning with images for context recognition in smart homes. In the home context recognition, the room layout, the environment, and the contexts to be recognized are different from one household to another. Therefore, a unique recognition model is required for every different household. For this, if we take a naive approach that uses the deep learning directly, a huge amount of labeled images are required, which is practically impossible for general households. The goal of this research is to develop an image-based context recognition method that is affordable at home. In the proposed method, we exploit a cognitive API which performs general image recognition, and retrieve the information within the image as text. By using the text as features, we classify the context with ordinal supervised machine learning. Compared with the expensive approach with deep learning, the proposed method uses generic image recognition of the cognitive API, and light-weight machine learning. As a result, the context recognition customized for every household can be achieved with much less effort.**

***Index Terms*—context recognition, image, cognitive API, machine learning**

I. Introduction

With the rapid progress of IoT (Internet of Things) technologies, make it possible to acquire various information in the physical space, and use it for value-added services. In the smart homes, research and development of recognize various *contexts* about users and home environments have been actively conducted. Examples of home contexts include situations about ADL (Activities of Daily Living) of users, such as eating, sleeping, watching tv, reading books, and situations about home environments, such as lights off, no people, messy room.

The mainstream of the conventional home context recognition is to use the numerical data obtained from ambient and/or wearable sensors and home appliances. Typical researches include a daily living activity sensing using power consumption of home appliances and location information of users [1], and a same sensing utilizing sensors of the smart phones [2]. Also, there are researches to learn and estimate situations from measurement of environmental change values such as temperature, humidity and illuminance at home [3].

In recent years, the emerging development of the deep learning has greatly advanced the learning and recognition technology of *multimedia data* such as images, voices, videos, and texts. We consider that the multimedia data include rich information than conventional sensor data, and home context recognition using multimedia data is promising. Therefore, our interest is to develop a new home context recognition method utilizing multimedia data (especially, *image data*) [4] [5].

However, the main difficulty of home context recognition using image data is that *individual differences from one household to another*. Since the room layout, the existing objects, and the environment are different from one household to another even though the same context (e.g., eating), the information shown in the image is largely different. Also, the contexts to be recognized are different from one household to another. Therefore, a unique recognition model is required for every different household. As a simple approach, although the recognition model with high accuracy can be constructed by acquiring images at home with the deep learning directly, this approach requires a huge amount of labeled images and highest computing power. Therefore, it is not realistic to implement it in general households.

The goal of this paper is to propose a method of home context recognition using image data that can be realized in general households. In the proposed methods, we first proposed a *framework for home context recognition* with the machine learning. In this framework, we define arbitrary contexts at home, acquire data, and recognize contexts with the machine learning. At the same time, we do not specifying the type of data and algorithms of the machine learning, and using the deep learning is also possible. As an implementation of the above framework, we then proposed a new home context recognition method utilizing the feature values of *cognitive API*. The cognitive API is a cloud service API (Application Programming Interface) that highly recognizes multimedia data such as images, voices, videos, and texts.

Our key idea is to acquire images at home and send them to the general purpose *image-based cognitive API*, retrieve information (*tag sets*) included in each image from the API results. We then conduct text mining to all tag sets, and vectorize them. We finally use these vectors to construct a multi-valued classification model with the *supervised machine learning*. Since we use feature values and the light-weight machine learning instead of image data and the conventional deep learning, the context recognition customized for every household can be achieved with much less effort.

Based on the proposed method, we have conducted an experiment to acquire images and recognize the contexts in our

978-1-7281-1651-8/19/$31.00 ©2019 IEEE
SNPD 2019, July 8-11, 2019, Toyama, Japan

laboratory. We first installed an USB camera to take a snapshot every five seconds, and the images are cumulated in a server during two weeks. We then defined seven contexts: General meeting, Cleaning, Eating, No people, Personal discussion, Gaming, and Studying. For each context, we selected 100 representative images considered to expose the context well.

Based on the key idea, we sent the selected 700 images to *Microsoft Azure Computer Vision API* [1] and retrieved the tag sets from the API results. We then regarded the tag sets (obtained from an image) as a document (corpus), and vectorized each tag sets using *TF-IDF (Term Frequency - Inverse Document Frequency)* [2]. We finally introduced each vectorized tag sets and corresponding the context labels to *Microsoft Azure Machine Learning Studio* [3], and constructed a recognition model with *Multiclass Neural Network*.

The experimental results showed that the overall accuracy of the recognition model constructed was 0.929, and the average accuracy was 0.980. Then, the accuracy of reliably recognized data in each context label was 0.929, and the accuracy of reliably recognized data in all context labels was 0.924. According to the results using *Confusion Matrix*, the recognition accuracy of general meeting was 95.3%, cleaning was 90.9%, eating was 83.3%, no people was 100.0%, personal discussion was 96.0%, gaming was 82.2%, studying was 100.0%. In addition, the recognition accuracy of no people and studying was the highest, and the recognition accuracy of eating and gaming was lowest in the 7 contexts.

II. Preliminary

A. Home Context Recognition

The home contexts refer to all situation information on users and home environments. What kind of the situation be existed in the home at the moment is great importance to the contents and the timing of the service. Accordingly, the important research topics has been studied for many years in the field of ubiquitous computing about how to improve the accuracy of home context recognition and how to use contexts to provide the smart service (context-aware service).

The mainstream of the ubiquitous computing is to recognize the contexts using the numerical data obtained from ambient and/or wearable sensors and smart phones, etc. A living activity recognition system based on power consumption of appliances and inhabitant's location information [1] and living activity recognition technology using sensors in smartphone [2] and capturing activities of daily living for elderly at home based on environment change and speech dialog [3] are concrete examples.

The above examples apply numerical data obtained from various sensors to rules and machine learning to estimate and determine ADL (Activities of Daily Living) of users and situations of home environments. Unfortunately, in many conventional researches, the home context recognition has not been widely used in general households since the necessity of dedicated sensors and the complexity of operation.

Nowadays, home context recognition is promised to be applied to smart homes where practical use is remarkable. Typical application examples include the monitoring system for elderly living alone and improvement for the rhythm of life of users.

B. Home Context Recognition using Image Data

In recent years, camera devices (e.g., web camera) are becoming more easy to introduce and install even in the general households due to the low cost, and the miniaturization. Also, the information amount of the image data obtained from the camera is larger than the numerical data of the sensor. Therefore, the home context recognition is potential to realize more powerful and easy to introduce using image data.

Generally, the advanced image recognition technology is required for recognize and understand the image data. With the recent development and spread of the *deep learning*, recognition with high precision that can withstand practical use has become possible. However, from the viewpoint of the training data preparation and computation resources, it is unrealistic to construct context recognition models for each household with the deep learning.

C. Image-Based Cognitive Service

Cognitive service is a cloud service that recognizes multimedia data such as images, voices, and texts. It is implemented by the trained machine learning models which are generally built using abundant cloud computing resources. The cognitive API (Application Programming Interface) are the APIs for calling and using cognitive service from external applications. This makes it easy to incorporate large-scale and complex recognition processing into applications.

The *image-based cognitive service* recognizes and retrieves various information from given images, and returns them. Famous services include Microsoft Azure Computer Vision, IBM Watson Visual Recognition, Google Cloud Vision, and Amazon Rekognition. The APIs recognized and retrieved information include face, age, sex, hair style, object, text, background, category, place, and color.

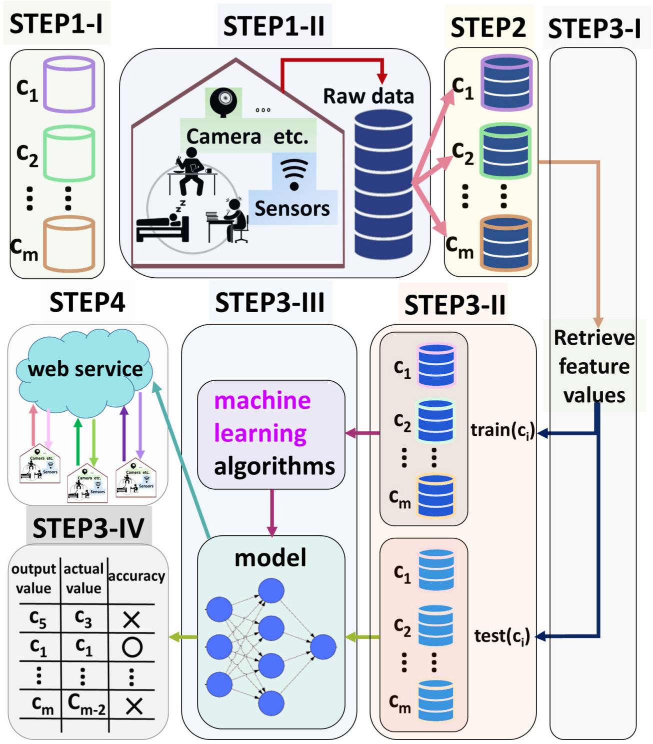
D. Previous Study: Evaluating Feasibility of Image-Based Cognitive APIs for Home Context Sensing [4] [5]

In the previous study [4] [5], we examined whether the home context recognition can be realized using the commercial image-based cognitive APIs. Specifically, we first installed an USB camera to acquire images of the daily activities of members of the laboratory such as meeting, eating, and gaming. We then sent the images to the APIs and examined whether the contexts can be estimated using the information retrieved from the API recognition results. In the experiment, we used three different APIs include Microsoft Azure Computer Vision API, IBM Watson Visual Recognition API, and Google Cloud Vision API. We finally analyzed the set of tags describing the images output from each the API results.

[1] https://azure.microsoft.com/ja-jp/services/cognitive-services/computer-vision/

[2] http://ailaby.com/tfidf/

[3] https://azure.microsoft.com/ja-jp/services/machine-learning-studio/



output value	actual value	accuracy
c_5	c_3	×
c_1	c_1	○
⋮	⋮	⋮
c_m	c_{m-2}	×

Fig. 1. The framework of home context recognition using machine learning

In the analysis, we evaluated whether set of tags reflects the original context, and we checked if the cohesion of the same context and the isolation of the different contexts are possible by *document similarity measures*. As a result, we found that too much performance can not be obtained if the set of tags output by the APIs are applied to the context recognition.

E. Machine Learning Platform

The cloud platform has appeared that can construct and deploy machine learning models depending on requirements of users. Users can upload data according to their own purpose, then experiment and construct their own machine learning models using combining various kinds of algorithms. Such as *Microsoft Azure Machine Learning Studio*, *Google Cloud Machine Learning Engine* [4], and *Amazon Sage Maker* [5] have been known. We mainly use Azure Machine Learning Studio in this paper.

III. The Framework of Home Context Recognition using Machine Learning

A. Goal

The individual differences from one household to another such as the room layout and home environment need to be considered when the home context recognition is performing. Therefore, it is necessary to construct a customized recognition model for each household when implementing the recognition using the machine learning.

In this section, we proposed a general flow of the home context recognition as a framework. Also, users of each household can freely select data, define contexts, and use algorithms of the machine learning. By doing so, we aim to flexibly respond to various requirements and constraints which differ one household to another.

B. Overall Flow

More specically, the proposed framework consists of the following four steps (Fig. 1):

STEP1: Acquiring data

I. Defining home contexts to recognize: We define a set $C = \{c_1, c_2, ..., c_m\}$ of home contexts to be recognized.

II. Acquiring data used for the context recognition: We first deploy an device for acquiring data such as sensor and camera in the target space to observe. Using the device, we then acquire data of the space periodically with an appropriate interval in order to accumulate data.

STEP2: Creating datasets

For each context $c_i \in C$, we manually select representative n data $D(c_i) = \{data_{i1}, data_{i2}, ..., data_{in}\}$ that well expose c_i from all data obtained in Step 1. Therefore, we create totally $m \times n$ data sets in this step.

STEP3: Constructing the model using the machine learning

I. Retrieving feature values: We retrieve feature values useful for the context recognition. Here, retrieving feature values with the deep learning are automated.

II. Splitting data: We split training data and test data from created data sets. Specifically, we split α data and $n - \alpha$ data as $train(c_i)$ and $test(c_i)$ from n data of each $data(c_i)$. Regarding splitting data sets, there are various methods such as splitting randomly, Hold-out, and Cross Validation.

III. Training the model: We apply algorithms of supervised machine learning A with training data $train(c_i) = \{train_{c1}, train_{c2}, ..., train_{cm}\}$ as input values, and we construct the recognition model M. Here, M is a multilevel classifier that output category c_i for input $d_{ij} (1 \leq j \leq n)$. Regarding A, there are various algorithms of the deep learning, and typical algorithms include NN (Neural Network), SVM (Support Vector Machine), and Decision Tree.

IV. Evaluating the model: We input test data $test(c_i) = \{test_{c1}, test_{c2}, ..., test_{cm}\}$ to M, and we evaluate the recognition accuracy of M by checking if it outputs category c_i corresponding to test data. If the recognition accuracy of M is low or unable to meet our requirements, we will reconstruct the model in previous step.

STEP4 : Deploying and operating the model

I. Deploying the model: We save the trained model M and make it online accessible from the target space. As for this, there are the functions in the machine learning platform to deploy the models as web services on the cloud and make it accessible by the APIs.

II. Operating the model: In this way, we input the data acquired in the target space to M, and let the obtained output c as the recognized home context. Ultimately, we can use c

[4] https://cloud.google.com/ml-engine/?hl=ja
[5] https://aws.amazon.com/jp/machine-learning/

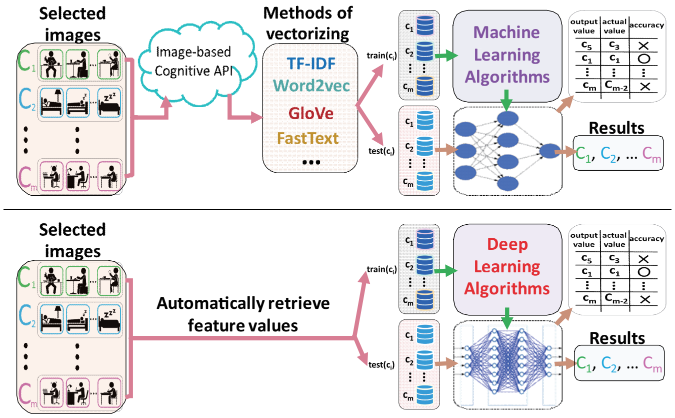


Fig. 2. The comparison of the machine learning and the deep learning

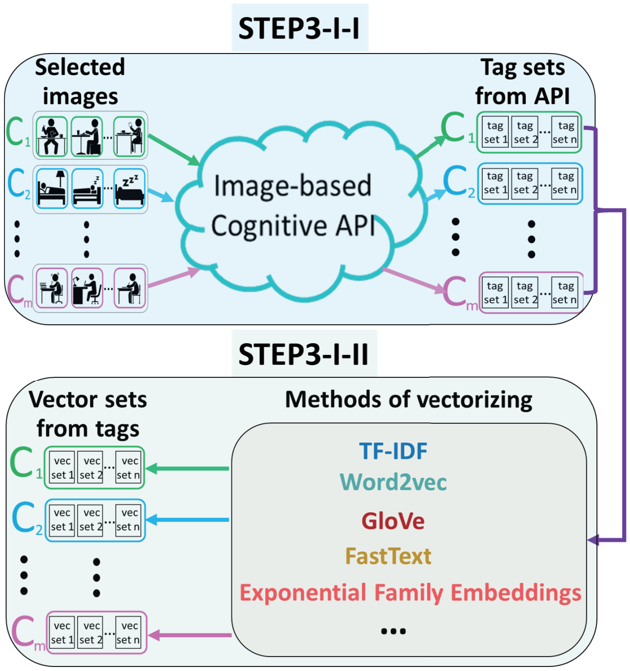


Fig. 3. The flow of the proposed method

obtained to develop new value-added services corresponding to requirements and situations from one household to another.

IV. Proposed Method

A. Key Idea

We consider implementing the framework of section III using image data obtained from the camera. Generally, when constructing a high-precision machine learning model with image data as input values, the most powerful method is to use deep learning. That is, construct M using algorithms of deep learning into A of STEP3 of section III (Fig. 2). However, since this method requires a huge amount of training data and computing resources to construct the model, it is not realistic to implement in general households.

In this section, we proposed a new home context recognition method. Regarding our key idea, we first retrieve information (tag sets) included in the image as feature values from image-based cognitive API, we then construct the home context recognition model using the light-weight machine learning. As described in section II-D, the set of tags itself could not be used for the context recognition. That is, general purpose image-based cognitive API did not sufficiently characterize the household specific context.

In the proposed new method, we retrieve feature values using image-based cognitive API, and we apply these to the light-weight machine learning in order to create a new classifier of the household specific context recognition. This section aims at constructing a model with much less effort than when using the deep learning.

B. Flow of the Proposed Method

We apply the above key idea in STEP3-I of the framework of section III. The proposed new method consists of following two steps (Fig. 3) :

STEP3-I-I: Retrieving feature values

We send each image $data_{ik}(1 \leq i \leq m)(1 \leq k \leq n)$ in n data $D(c_i)$ to image-based cognitive API, and we obtain set of tags $Tag(data_{ik}) = \{w_1, w_2, w_3, ...\}$ from the API results corresponding to each image $data_{ik}$.

STEP3-I-II: Vectorizing feature values

We first regard the total of all tag sets $\bigcup_{ik} Tag(data_{ik})$ (obtained from STEP3-I-I) as a document (corpus), and we vectorize each set of tags $Tag(data_{ik})$ to set of vectors $V_{ik} = [v_1, v_2, ...]$. We then associate each context c_i as a label with V_k. As the methods of vectorizing documents, there are TF-IDF, Word2Vec, GloVe, and FastText, ect.

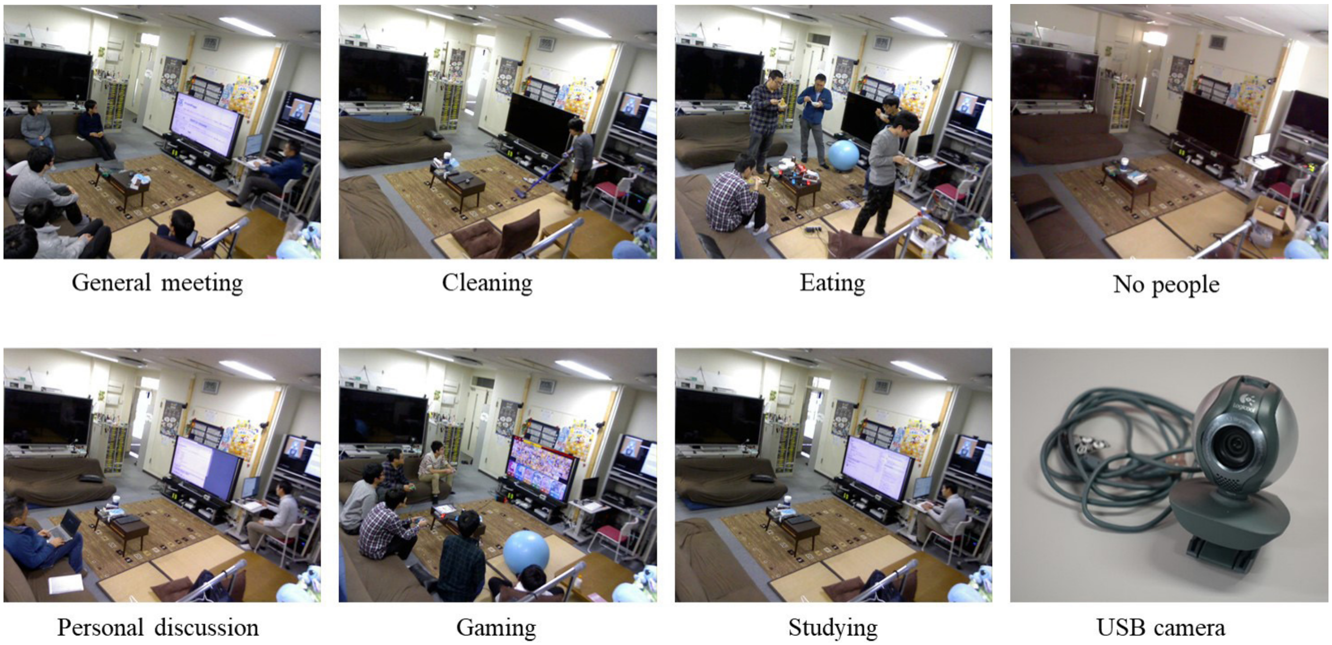
C. Constructing the recognition model with this method

We first regard set of vectors V_k and labels c_i corresponding to them (obtained from STEP3-I-II) as created data sets, and we then apply STEP3 of the framework in order to construct the recognition model M using algorithms of general supervised machine learning. Here, M is a multilevel classifier that classifies input set of vectors to $c_1, c_2, ..., c_n$. Just in STEP4, it is necessary that sending to image-based cognitive API, and vectorizing feature values plus $\bigcup_{ik} Tag(data_{ik})$ together when inputting the value to this model every time.

V. Experimental Evaluation

A. Preparing Data

In this experiment, we set the target space to be a smart home space, which is a part of our laboratory. In STEP1-I, we define seven contexts: General meeting, Cleaning, Eating, No people, Personal discussion, Gaming, and Studying. In STEP1-II, we install an USB camera to acquire images of the daily activities of members of the laboratory. We develop a program that takes a snapshot with the USB camera every five seconds, and the images are cumulated in a server during two weeks. In STEP2, for each context, we select 100 representative images considered to expose the context well, and the selection is done by visual inspection.


General meeting

Cleaning

Eating

No people

Personal discussion

Gaming

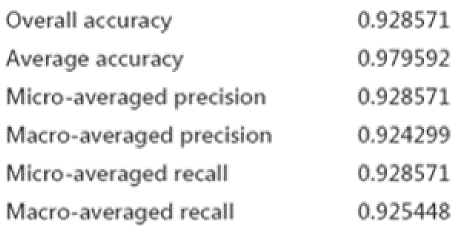
Studying

USB camera

Fig. 4. The representative images for each context and USB camera

TABLE I
THE EXAMPLES OF EXTRACTED TAG SETS FROM THE API

Context Label	Tag Results
Cleaning	indoor, living, room, table, television, fire, fireplace, man, standing, filled, video, playing, woman, furniture, large, people, wii, dog, game
Eating	indoor, person, room, table, living, man, sitting, food, filled, luggage, people, standing, suitcase, television, young, large, fire, kitchen
Gaming	indoor, person, room, table, sitting, living, people, man, food, standing, large, group, woman, playing, computer, kitchen, game



Overall accuracy	0.928571
Average accuracy	0.979592
Micro-averaged precision	0.928571
Macro-averaged precision	0.924299
Micro-averaged recall	0.928571
Macro-averaged recall	0.925448

Fig. 5. The results of this experiment with metrics

B. Execution of the Proposed Method

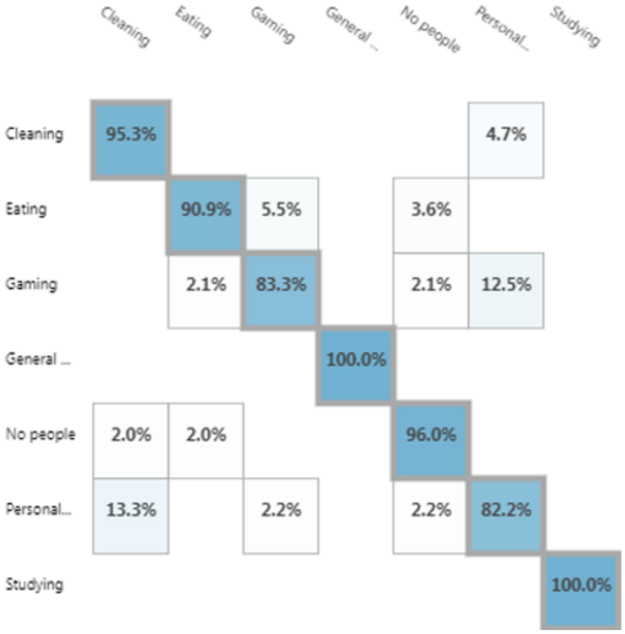
We execute STEP3 of framework according to added two substeps of section IV-B. We used the *Microsoft Azure Computer Vision API* to retrieve feature values in STEP3-I-I. We first sent all selected images data to API and retrieved tag sets as feature values from the API recognition results. Table I shows the examples of extracted tag sets from the API recognition results. We then vectorized tag sets using TF-IDF in STEP3-I-II. TF-IDF is a method of numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. By this method, we can get a vector of each tag present in a document (corpus). In Step 3-II, we randomized all created data sets, and we splited them into half as training data and test data. Fig. 4 shows the representative images for each context and USB camera. We finally constructed a model for the home context recognition using training data and the algorithm of Multiclass Neural Network on machine learning platform of Microsoft Azure Machine Learning Studio in Step 3-III.

C. Evaluate the Model

Regarding evaluating the model in STEP3-IV, we input test data to the trained model (constructed from STEP3-III), and we evaluate the accuracy of the model by compare output values from the model with actual values. We used the following metrics to evaluating the model. By the results of metrics, we can evaluate the overall accuracy, average accuracy, micro-averaged precision, macro-averaged precision, micro-averaged recall, and macro-averaged recall of the model. In addition, we can also evaluate the recognition accuracy for each context using Confusion Matrix.

D. Results

The results of this experiment shown in Fig. 5 using metrics and Fig. 6 using confusion matrix. From the results of metrics in Fig. 5, we found that the constructed model achieves high



	Cleaning	Eating	Gaming	General ...	No people	Personal...	Studying
Cleaning	95.3%					4.7%	
Eating		90.9%	5.5%		3.6%		
Gaming		2.1%	83.3%		2.1%	12.5%	
General ...				100.0%			
No people	2.0%	2.0%			96.0%		
Personal...	13.3%		2.2%		2.2%	82.2%	
Studying							100.0%

Fig. 6. The results of this experiment with confusion matrix

recognition accuracy of 0.92 or more for any metrics. Also, from the results of confusion matrix in Fig. 6, we knew that the proportion of correct and/or incorrect recognition of each context as following.

The proportion of correct recognition of “General meeting” was 95.3%, and the remaining 4.7% was incorrectly recognized as “Gaming”. The proportion of correct recognition of “Cleaning” was 90.9%, the remaining 5.5% was incorrectly recognized as “Eating”, and the remaining 3.6% was incorrectly recognized as “Personal discussion”. The proportion of correct recognition of “Eating” was 83.3%, the remaining 4.2% was incorrectly recognized as “Cleaning, and the remaining 12.5% was incorrectly recognized as “Gaming. The proportion of correct recognition of “No people” was 100.0%. The proportion of correct recognition of “Personal discussion” was 96.0%, the remaining 2.0% was incorrectly recognized as “General meeting”, and the remaining 2.0% was incorrectly recognized as “Clearning”. The proportion of correct recognition of “Gaming” was 82.2%, the remaining 13.3% was incorrectly recognized as “General meeting”, the remaining 2.2% was incorrectly recognized as “Eating”, and the remaining 2.2% was incorrectly recognized as “Personal discussion”. The proportion of correct recognition of “Studying” was 100.0%.

From the above results, we found that the proportion of correct recognition of “No people” and “Studying” were highest, “Eating” and “Gaming” were lowest. Also, we found that the proportion of incorrect recognition of “Eating” as “Gaming” and “Gaming” as “General meeting” were highest.

E. Discussion

Regarding the evaluating results of constructed the model in this experiment, we found that the accuracy of context recognition with the small group (especially, one person) or nobody was the highest. In contrast, we found that the context recognition with the big group such as “Eating”, “Gaming”, and “General meeting” were difficult.

Generally, the context recognition with the big group is associated the feature values using interactions or actions between people in one certain period of time. However, in the context recognition using image data, since the contexts are clipped as instantaneous snapshots, it is sometimes hard to distinguish them even by human eyes. We consider that these were possibly expressed by numbers.

VI. Conclusion

In this paper, we first proposed the framework of the home context recognition using the machine learning, and we then proposed a new method of the context recognition using image data that can be introduced in general households. Regarding the new method, we retrieve the information (tag sets) from image data using cognitive API, and we construct the model using the feature values and the light-weight machine learning. By doing so, we can realize the home context recognition with much less effort than the deep learning.

In experimental evaluation, we vectorized tag sets (obtained from cognitive API) using TF-IDF, and we have been conducted the experiment of the recognition for seven contexts in the laboratory. As a result, we constructed the model that the recognition accuracy was 0.92 or more. And in the recognition with confusion matrix, we found that the proportion of incorrect recognition of “Eating” as “Gaming” and “Gaming” as “General meeting” were highest.

As future work, we will focus on the experimental evaluation using various methods of splitting data sets and algorithms of the machine learning. Also, we will deploy the model constructed from this study in the laboratory, and we will evaluate its performance and effectiveness by conduct the context recognition on the cloud. In finally, we would like to study the reconstruct and reuse of the model with the addition of new contexts or the operation of other environments.

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

References

[1] K. Ueda, M. Tamai, Y. Arakawa, H. Suwa, and K. Yasumoto, “A living activity recognition system based on power consumption of appliances and inhabitant’s location information,” *Trans.IPS.Japan*, vol. 416-425, no. 57, p. 2, Feb 2016.

[2] K. Ouchi and M. Doi, “Living activity recognition technology using sensors in smartphone,” *Toshiba review*, vol. 68, no. 6, pp. 40–43, Jun 2013.

[3] 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,” *IEICE technical report Japan*, vol. 116, no. 405, pp. 7–12, Jan 2017.

[4] S. Chen, S. Saiki, and M. Nakamura, “Feasibility study of image-based cognitive api for home sensing,” *IEICE technical report Japan*, no. SC-19, pp. 31–36, August 2018.

[5] ——, “Evaluating feasibility of image-based cognitive apis for home context sensing,” *ICSPIS2018*, November 2018.