Master’s Thesis

Recognizing Fine-Grained Contexts at Home with Image-based Cognitive API

189X226X Sinan Chen

Department of Computational Science
Graduate School of System Informatics
Kobe University

Supervisor
Assoc. Masahide Nakamura

Primary Examiner
Prof. Kuniaki Uehara

Examiner
Prof. Mitsuo Yokokawa
Assoc. Masahide Nakamura

August 19, 2019
Recognizing Fine-Grained Contexts at Home with Image-based Cognitive API

Sinan Chen

Abstract

With the rapid progress of ICT and IoT (Internet of Things) technologies, research and development of smart homes have been actively conducted. The key challenge in the smart home is how smartly the system is able to recognize various contexts at home, automatically. 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. Moreover, 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.

The technologies for home context recognition have been studied for many years in the field of ubiquitous computing. The traditional ubiquitous computing employs ambient sensors (e.g., temperature, humidity, presence) [1], wearable sensors (e.g., accelerometer, heart rate), and indoor positioning systems [2]. They are installed at rooms or worn by residents to collect essential data characterizing various 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.

The goal of research 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. We extensively utilize the image-based cognitive API throughout the research. An image-based cognitive API receives an image from an
external application, recognizes specific information within the image, and returns the information as a set of words called tags. Our key idea is to consider these tags as features of the image, and apply light-weight machine learning techniques to infer the target context. We first present a method that evaluates the feasibility of the cognitive APIs for the home context recognition. We then develop 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 tags. Considering these tags as features of the target context, the proposed method constructs a multi-class classifier using an ordinary machine learning algorithm. Finally, we present a technique that uses multiple cognitive APIs for ensemble learning [3], in order to improve the recognition accuracy of difficult contexts.

**Keywords**  Context recognition, Cognitive APIs, Images, Machine learning, Majority voting, Smart home, General households
# Contents

Chapter 1  Introduction  
1.1  Background ................................................. 1  
1.2  Achievements ................................................. 3  
1.3  Overview of the Thesis ........................................ 5  

Chapter 2  Evaluating Feasibility of Image-Based Cognitive APIs for Home Context Sensing  
2.1  Introduction ................................................. 7  
2.2  Proposed Method ............................................ 10  
2.3  Experimental Evaluation ........................................ 12  
   2.3.1  Experimental Setup ....................................... 12  
   2.3.2  Encoding Output Tags by TF-IDF ......................... 13  
   2.3.3  Results ................................................ 14  

Chapter 3  Proposal of Home Context Recognition Method Using Feature Values of Cognitive API  
3.1  Introduction ................................................. 16  
3.2  Preliminary ................................................ 19  
   3.2.1  Home Context Recognition ............................. 19  
   3.2.2  Home Context Recognition Using Image Data .......... 19  
   3.2.3  Image-Based Cognitive Service ....................... 20  
   3.2.4  Previous Study ......................................... 20  
   3.2.5  Machine Learning Platform ........................... 21  
3.3  The Framework of Home Context Recognition Using Machine Learning ........................................ 21
3.3.1 Goal ........................................... 21
3.3.2 Overall Flow .................................. 23

3.4 Proposed Method .................................. 25
   3.4.1 Key Idea ...................................... 25
   3.4.2 Flow of The Proposed Method ............... 26
   3.4.3 Constructing the recognition model with this
         method ........................................ 27

3.5 Experimental Evaluation ............................. 27
   3.5.1 Preparing Data .................................. 27
   3.5.2 Execution of The Proposed Method ............ 28
   3.5.3 Evaluating The Model ........................... 29
   3.5.4 Results ........................................ 29
   3.5.5 Discussion .................................... 31

Chapter 4  Recognizing Fine-Grained Home Contexts

Using Multiple Cognitive APIs .......................... 32

4.1 Introduction ........................................ 32

4.2 Preliminaries ....................................... 35
   4.2.1 Recognizing Fine-Grained Home Contexts .. 35
   4.2.2 Technical Challenges ........................... 35
   4.2.3 Image Recognition API of Cognitive Services 36
   4.2.4 Preliminary Study .............................. 37
   4.2.5 Ensemble learning ............................... 37

4.3 Proposed Method .................................... 38
   4.3.1 Outline ........................................ 38
   4.3.2 Constructing Context Recognition Model .. 39
   4.3.3 Integrating Models Generated from Different
         APIs .............................................. 40

4.4 Experimental Evaluation ............................. 41
   4.4.1 Experimental Setup .............................. 41
   4.4.2 Building Recognition Model .................... 42
List of Figures

2.1 Usage of image recognition APIs of cognitive services . . . 8
2.2 The flow from context label setting to analysis of results . . 11
3.1 The framework of home context recognition using machine
learning ............................................................... 22
3.2 The comparison of the machine learning and the deep learning 24
3.3 The flow of the proposed method ............................. 25
3.4 The representative images for each context and USB camera 28
3.5 The results of this experiment with metrics .................. 28
3.6 The results of this experiment with confusion matrix ....... 30
4.1 Recognizing fine-grained home contexts .................... 33
4.2 Example of the majority voting with ensemble learning . . 34
4.3 The overview of proposed method using multiple cognitive
APIs ................................................................. 38
4.4 Representative images of fine-grained contexts and USB
camera ............................................................ 41
List of Tables

2.1 Experiment settings ............................................ 13
2.2 Result representing the internal cohesion of output tags ... 14
2.3 Result representing the external isolation of output tags (pre-processing applied) ............................................ 14
3.1 The examples of extracted tag sets from the API ............... 27
4.1 Tag sets that different APIs extracted from an image ........... 42
4.2 The recognition accuracy results of each cognitive API-based Model and majority voting in this experiment .................. 43
Chapter 1

Introduction

1.1 Background

With the rapid progress of ICT and IoT (Internet of Things) technologies, research and development of smart homes have been actively conducted. The smart home enables to collect various data within the house, and use the data for value-added services. The value-added services include elderly monitoring [4] [5] [6], autonomous security [7], and personalized healthcare [8] [9].

The key challenge in the smart home is how smartly the system is able to recognize various contexts at home, automatically. 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 are “residents are in the dining room”, “it is warm in the dining room”, “the dining room is clean”, “the light in the dining room is on”, and so on. Such contexts are recognized by based on various data gathered within a house.

Moreover, 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 the dining room”, “father is taking medicine in the dining room”, “mother is cleaning the dining room”, are crucial information for the son. If these fine-grained contexts can be recognized by an elderly monitoring system, and the infor-
mation is regularly sent to the son, it would be a great value for the son.

The technologies for home context recognition have been studied for many years in the field of _ubiquitous computing_. The traditional ubiquitous computing employs ambient sensors (e.g., temperature, humidity, presence) [1], wearable sensors (e.g., accelerometer, heart rate), and indoor positioning systems [2]. They are installed at rooms or worn by residents to collect essential data characterizing various contexts. Using such sensors and devices, many systems have been proposed. They include an elderly watching system using human motion sensors [10], the context labeling and recognition system based on environment change [11], the recognition of daily activities with time-series environmental data [12], the activity recognition with power consumption of home appliances and user’s location [13], and context sensing with smart phones[14].

In recent years, the emerging _deep learning_ [15] allows the system to recognize _multimedia_ data. Since image, voice, and text data usually contain richer information than the conventional sensor data, it is promising to use such multimedia data for recognizing fine-grained home contexts. Indeed, for general residents, deploying a Web camera at home is much easier than introducing a proprietary sensor system or an intrusive wearable device. Thus, an image-based context recognition with the deep learning technologies would be an interesting approach.

Unfortunately, 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 [15] [16]. Thus, there is still a big gap between the research and real life.

More specifically, 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.

The goal of research 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 investigate image-based approaches with machine learning for fine-grained home context recognition. Although the image recognition with deep learning is emerging in recent years, the feasibility to the fine-grained home context recognition is still a question. A naive application of the deep learning will not be acceptable for general households. Thus, it is worth to devise more affordable and light-weight methods.

1.2 Achievements

In order to implement the fine-grained home context recognition, we extensively utilize the image-based cognitive API throughout the research. In recent years, world’s cloud companies such as Microsoft, IBM, and Google, release cognitive services. A cognitive service provides a capability of understanding multimedia data based on sophisticated machine-learning algorithms powered by big data and large-scale computing resources. Typical services include image recognition, speech recognition, and natural language processing.
The image-based cognitive API is application program interface to cloud services that provide various image recognition features as a service. Famous APIs include Microsoft Azure Computer Vision API [17] and IBM Watson Visual Recognition API [18]. An image-based cognitive API receives an image from an external application, recognizes specific information within the image, and returns the information as a set of words called tags. Our key idea is to consider these tags as features of the image, and apply lightweight machine learning techniques to infer the target context.

In this master’s thesis, the following three achievements performed during the master course are presented.

Achievement A1: Feasibility Evaluation of Image-Based Cognitive APIs for Home Context Sensing

We first present a method that evaluates the feasibility of the cognitive APIs for the home context recognition. Basically, the existing cognitive APIs are trained for general-purpose image recognition, and they are not tuned for fine-grained contexts of individual houses. The proposed method evaluates how these APIs can recognize or distinguish the home contexts from given images, without specific training. The key idea of evaluation is to regard a set of tags (obtained from an image) as a document (corpus), and to apply document similarity measures [19] to see how clusters of contexts are constructed. The document clusters are evaluated with respect to the internal cohesion and external isolation. That is, we see if images belonging to the same (or different) context(s) are associated with similar tags (or dissimilar tags, respectively).

Achievement A2: Home Context Recognition Method Using Feature Values of Cognitive API

We then develop 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 tags. Considering these tags as features of the target context, the proposed method constructs a multi-class classifier using an ordinary machine learning algorithm. We conduct an experiment where seven kinds of contexts within our laboratory were recognized from images. The classifier was constructed by Multi-class Neural Network from the tags derived by Microsoft Azure Computer Vision API [17].

Achievement A3: Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

Finally, we present a technique that uses multiple cognitive APIs for ensemble learning [3], in order to improve the recognition accuracy of difficult contexts. 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. Since different APIs observe the same image from different perspectives, they would be able to complement mutual limits of recognition capability. In the experiment, we implement the majority voting of five models based on the following commercial APIs: Microsoft Azure Computer Vision API [17], IBM Watson Visual Recognition API [18], Clarifai API [20], Imagga REST API [21], and Paralleldots API [22].

1.3 Overview of the Thesis

The rest of the thesis is organized as follows: In Chapter 2, we present Achievement A1: “Feasibility Evaluation of Image-Based Cognitive APIs for Home Context Sensing”. Starting with the brief idea of how to use the cognitive API, we describe the proposed method by five steps. Then, we conduct an experiment where adopts three commercial APIs to recognize eleven contexts within our laboratory.

In Chapter 3, we present Achievement A2: “Home Context Recognition Method Using Feature Values of Cognitive API”. We first propose a frame-
work, which prescribes a workflow of context definition, data acquisition, model construction, and operation. We then develop an implementation with the image-based cognitive API.

In Chapter 4, we present Achievement A3: “Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs”. Based on the experimental results in Achievement A2, we first review the limitations of the method. Then, we extend the method by introducing multiple cognitive APIs. We conduct an experimental evaluation to see how the majority voting improves the recognition performance.

Finally, in Chapter 5, we conclude the thesis with a summary and future work.
Chapter 2

Evaluating Feasibility of Image-Based Cognitive APIs for Home Context Sensing

2.1 Introduction

With the rapid progress of ICT and Internet of Things (IoT) technologies, research and development of smart homes have been actively conducted. In smart homes, it is common to use ambient and/or wearable sensors such as temperature, humidity, motion, and accelerometer in order to retrieve contexts of users and homes. Typical systems include an elderly watching system using human motion sensors [10], and a daily living activity sensing system for elderly people using environmental sensors [23].

Using multimedia data, such as image and audio, for home context sensing is promising for value-added smart services, since the multimedia data contains richer information than the conventional sensor data. However, recognizing multimedia data generally requires massive computation. It was thus unrealistic for general households to install and maintain such an expensive system at home.

In recent years, world’s cloud companies such as Microsoft, IBM, and Google, released cognitive services. A cognitive service provides the capability to understand multimedia data based on sophisticated machine-learning algorithms powered by big data and large-scale computing resources. Typical services include image recognition, speech recognition, and natural language
processing. A cognitive service usually provides cognitive APIs (Application Program Interface), with which developers can easily integrate powerful recognition features in their own applications. We consider that the cognitive APIs make full use of multimedia data, therefore, they have great potential to improve smart homes since the user would no longer need to maintain an expensive system.

Although various kinds of cognitive APIs exist, we especially focus on image recognition APIs in this paper. An image recognition API receives an image from an external application, extracts specific information from the image, and returns the information as a set of words called tags. Figure 2.1 represents the usage of image recognition APIs. The information of interest varies between services. For example, MS-Azure Face API [24] estimates age, sex, and emotional values from a given human face image. IBM Watson Visual Recognition [25] recognizes items in the image such as home appliances, furniture, and tools. Google Cloud Vision API [26] outputs concept labels associated to recognized objects.
Chapter 2  Evaluating Feasibility of Image-Based Cognitive APIs for Home Context Sensing

Our interest is to apply these image recognition APIs to implement smart and 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 in the cloud. Note, however, that the existing cognitive APIs are trained for general-purpose image recognition. Therefore, the API may not be of practical use for our specific purpose of the home context sensing.

The goal of this paper is to propose a method that evaluates the feasibility of cognitive APIs for a home context sensing. Since the detailed configuration varies from one house to another, the feasibility would be different among individual smart homes. However, using the proposed method, one can reproduce the experiment with different configurations. Also, one can understand the coverage and limitation of APIs towards specific home contexts.

In the proposed method, we first capture images of different contexts. Afterward, we send the image to cognitive APIs to retrieve tags from the images. Finally, we evaluate the performance of the APIs by checking if the tags can sufficiently characterize (or distinguish) the context shown in the original image.

Our key idea of evaluation is to regard a set of tags (obtained from an image) as a document (corpus), and to apply document similarity measures [19] to see how clusters of contexts are constructed. More specifically, we evaluate the document clusters, with respect to the internal cohesion and external isolation. That is, we see if images belonging to the same (or different) context(s) are associated with similar tags (or dissimilar tags, respectively).

Based on the proposed method, we have conducted an experiment. In the smart home space of our laboratory, we collected images of 11 different contexts: general meeting, reading, cleaning, eating, gaming, no people, personal meeting, studying, sleeping, touching smartphone, and watching TV. We then sent the images to three cognitive APIs to retrieve tags from
the images: Microsoft Azure Computer Vision API [24], IBM Watson Visual Recognition [25], and Google Cloud Vision API [26]. Finally, we analyzed the retrieved tags using Term Frequency - Inverse Document Frequency (TF-IDF) [27] and cosine similarity.

The experimental results showed that among the three cognitive APIs, there was no significant difference in the performance of the internal cohesion. Tags produced by Google Cloud Vision API tend to be more similar with each other, compared to tags produced by Microsoft Azure Computer Vision (or IBM Watson Visual Recognition). As for the external isolation, we found that background objects irrelevant to the context would produce a steady bias component. It was shown that removing the bias by subtracting tags produced by the context “no people” improved the performance of the external isolation.

2.2 Proposed Method

In this section, we present a method that evaluates and compares the capability of multiple image recognition APIs, for a given set of home contexts.

Figure 2.2 depicts the essential part of the proposed method. In the figure, \( \{c_1, c_2, \ldots, c_m\} \) represent a given set of home contexts. For each context, we collect \( n \) images at home, and then send the images to cognitive APIs. Finally, we evaluate the performance of the APIs, by analyzing the output tags. More specifically, the proposed method consists of the following six steps:

**Step1: Acquiring images**

A user of the proposed method deploys an image capturing device (e.g., USB camera) in the target space, and configures the device to take snapshots of the space periodically with an appropriate interval.

**Step2: Defining home contexts to recognize**

The user defines a set \( C = \{c_1, c_2, \ldots, c_m\} \) of home contexts to be recognized by the cognitive API.

**Step3: Selecting representative images**
For each context $c_i \in C$, the user manually selects representative $n$ images $IMG(c_i) = \{img_{1i}, img_{2i}, ..., img_{ni}\}$ that well expose $c_i$, from all images obtained in Step 1.

**Step4: Calling cognitive API**

The user designates a set $API = \{api_1, api_2, \ldots, api_q\}$ of cognitive APIs to be evaluated. For every $c_i \in C$, $img_{ij} \in IMG(c_i)$, and $api_k \in API$, $api_k(img_{ij})$ is invoked, and a set $Tag(img_{ij}, api_k) = \{w_1, w_2, w_3, ...\}$ of output tags is obtained. $Tag(img_{ij}, api_k)$ represents a recognition result for cognitive API $api_k$ for an image $img_{ij}$ belonging to a context $c_i$. The size of $Tag(img_{ij}, api_k)$ varies for $img_{ij}$ and $api_k$. Since there are $m$ contexts, $n$ images for each context, and $q$ APIs, this step creates totally $m \times n \times q$ sets of output tags.

**Step5: Analyzing output tags**

Regarding every set $Tag(img_{ij}, api_k)$ of output tags as a document, the method calculates the similarity, which is denoted as ‘$\approx$’, between any two of documents using a certain method of document similarity measure.
For each $api_k$, we evaluate the performance of $api_k$ of context recognition, with respect to internal cohesion and external isolation. The internal cohesion represents a capability that $api_k$ can produce similar output tags for images in the same context. That is, for $c_i \in C$, we evaluate $Tag(img_{ij}, api_k) \approx Tag(img_{ij'}, api_k)$. On the other hand, the external isolation represents capability that $api_k$ can produce dissimilar output tags for images in different contexts. That is, for $c_x \neq c_y$, we evaluate $Tag(img_{xj}, api_k) \neq Tag(img_{yj'}, api_k)$.

Regarding the calculation of document similarity, there exists a variety of methods in the field of natural language processing. One of the most basic methods is to use TF-IDF and the cosine similarity [27]. Modern techniques include Word2Vec[28] and Doc2Vec [29]. The selection of the similarity measure is left for the user of the proposed method.

2.3 Experimental Evaluation

2.3.1 Experimental Setup

Using the proposed method, we evaluate the feasibility of Microsoft Azure Computer Vision (Azure) API [24], IBM Watson Visual Recognition (Watson) [25], and Google Cloud Vision API (Google) [26], for context sensing of a smart space in our laboratory. Table 2.1 summarizes settings of the experiment.

In this experiment, we set the target space to be a smart home space, which is a part of our laboratory. For Step 1, we install a 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 5 seconds, and the images are cumulated in a server during one week. In Step 2, we define 11 contexts: general meeting, reading, cleaning, eating, gaming, no people, personal meeting, studying, sleeping, touching smartphone, and watching TV. In Step 3, for each context, we select 10 representative images considered to expose the context well. The selection is done by visual inspection so that the 10 images are chosen from as different date and time as possible. In Step
Table 2.1. Experiment settings

<table>
<thead>
<tr>
<th></th>
<th>Smart space in CS27 Nakamura Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target space</td>
<td></td>
</tr>
<tr>
<td>Image accumulation period</td>
<td>7 Days</td>
</tr>
<tr>
<td>Shooting method</td>
<td>USB camera</td>
</tr>
<tr>
<td>Image Resolution</td>
<td>1280 × 1024</td>
</tr>
<tr>
<td>Number of contexts (m)</td>
<td>11</td>
</tr>
<tr>
<td>Number of selected images (n)</td>
<td>10</td>
</tr>
<tr>
<td>Number of APIs (q)</td>
<td>3</td>
</tr>
<tr>
<td>Document vectorization</td>
<td>TF-IDF</td>
</tr>
<tr>
<td>Document similarity metrics</td>
<td>Cosine similarity</td>
</tr>
</tbody>
</table>

4, the images are sent to the three different APIs, and the total 330 sets of output tags (= 11 contexts × 10 images × 3 APIs) are obtained. In Step 5, we use TF-IDF to encode each set of output tags to a vector, and the cosine similarity to calculate the similarity.

2.3.2 Encoding Output Tags by TF-IDF

As mentioned in Step 5 in Section 2.2, we evaluate the internal cohesion and external isolation among the sets of output tags. For the internal cohesion of API $api_k$ for context $c_i$, we want to see the similarity between output tags $Tag(img_{ij}, api_k)$ ($j = 1, 2, ..., 10$). Therefore, TF-IDF is calculated among 10 documents for each context $c_i$, regarding each $Tag(img_{ij}, api_k)$ ($j = 1, 2, ..., 10$) as a unique document.

For the external isolation of API $api_k$, we want to see how far a context $c_x$ is from another context $c_y$. Therefore, TF-IDF is calculated among 11 contexts $c_i$ ($i = 1, 2, ..., 11$), where we join $Tag(img_{ij}, api_k)$ ($j = 1, 2, ..., 10$) into a single document $Tag(c_i, api_k)$ that characterizes context $c_i$. Once a pair of documents is encoded as TF-IDF vectors $x$ and $y$, we define the similarity between the documents as

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$
which expresses the similarity by the angle formed by the vectors. When the angle is small, the cosine value is close to 1.0 (or -1.0), meaning that two documents are similar (or dissimilar, respectively).

2.3.3 Results

Table 2.2 shows the result of the internal cohesion of output tags. The values in the table show the average of cosine similarities of output tags within each of the 11 contexts, produced by the three APIs. According to the definition, the higher value represents better performance, meaning that the API can produce similar tags for images belonging to the same contexts. Although the difference was not so significant, Google marked a slightly better performance for the
internal cohesion.

Table 2.3 shows the result of the external isolation. Each entry shows a cosine similarity between $Tag(c_x, api_k)$ and $Tag(c_y, api_k)$. According to the definition, the lower value represents better performance, meaning that the API can produce dissimilar tags for images belonging to different contexts. In the experiment, we found that background objects irrelevant to the context produced a steady bias component. To remove the bias, we applied a pre-processing which subtracts $Tag(\text{"No people"}, api_k)$ from every $Tag(c_i, api_k)$. As the result of the pre-processing, some output tags produced by Google became empty, which means that the API could not distinguish the context from “No people”. Azure and Watson marked similar performance for the external isolation.

From the experiment, we obtained the following findings:

- Individual APIs have their own strong (or week) contexts.
- For home sensing with arbitrary contexts, we cannot expect too much performance for the general-purpose APIs without training.
- Depending on the target context, we should consider appropriate pre-processing of the image to improve the recognition performance.
Chapter 3

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

3.1 Introduction

With the rapid progress of IoT (Internet of Things) technologies, they 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, and reading books, and situations about home environments, such as lights off, no people, and 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[13], and the same sensing utilizing sensors of the smart phones[14]. Also, there are researches to learn and estimate situations from measurement of environmental change values such as temperature, humidity and illuminance at home[23].

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)[30][31].

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 \textit{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 laboratory. We first installed a 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 \textit{Microsoft Azure Computer Vision API}[17] 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 \textit{TF-IDF} (\textit{Term Frequency - Inverse Document Frequency})[32]. We finally introduced each vectorized tag sets and corresponding the context labels to \textit{Microsoft Azure Machine Learning Studio}[33], and constructed a recognition model with \textit{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 \textit{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 the lowest in the 7 contexts.
3.2 Preliminary

3.2.1 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 smartphones, etc. A living activity recognition system based on power consumption of appliances and inhabitant’s location information[13] and living activity recognition technology using sensors in smartphone[14] and capturing activities of daily living for elderly at home based on environment change and speech dialog[23] 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.

3.2.2 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.

3.2.3 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.

3.2.4 Previous Study

In the previous study[30][31], 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.

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.

3.2.5 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 [34], and Amazon Sage Maker[35] have been known. We mainly use Azure Machine Learning Studio in this paper.

3.3 The Framework of Home Context Recognition Using Machine Learning

3.3.1 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.
3.3.2 Overall Flow

More specifically, the proposed framework consists of the following four steps (Figure 3.1):

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 $\alpha$ 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 $test(c_i) = \{test_{c1}, test_{c2}, ..., \}$
Fig. 3.2. The comparison of the machine learning and the deep learning

\[ \text{test}_{cm} \] 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 \( c \) obtained to develop new value-added services corresponding to requirements and situations from one household to another.
3.4 Proposed Method

3.4.1 Key Idea

We consider implementing the framework of section 3.3 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 3.3 (Figure 3.2). 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 3.2.4, 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.

### 3.4.2 Flow of The Proposed Method

We apply the above key idea in STEP3-I of the framework of section 3.3. The proposed new method consists of following two steps (Figure 3.3):

**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 $Tag(data_{ik}) = \{w_1, w_2, w_3, \ldots\}$ 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 $V_{ik} = [v_1, v_2, \ldots]$. 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, etc.
Chapter 3  Proposal of Home Context Recognition Method Using Feature Values of Cognitive API

Table 3.1. The examples of extracted tag sets from the API

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

3.4.3 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 $c_1, c_2, \ldots, c_n$. 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.

3.5 Experimental Evaluation

3.5.1 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.

3.5.2 Execution of The Proposed Method

We execute STEP3 of framework according to added two substeps of section 3.4.2. 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 3.1 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 split them into half as training data and test data. Figure 3.4 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.

3.5.3 Evaluating 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.

3.5.4 Results

The results of this experiment shown in Figure 3.5 using metrics and Figure 3.6 using confusion matrix. From the results of metrics in Figure 3.5, we found that the constructed model achieves high recognition accuracy of 0.92 or more for any metrics. Also, from the results of confusion matrix in Figure 3.6, 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 propor-
Fig. 3.6. The results of this experiment with confusion matrix.

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 “Cleaning”. 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.

3.5.5 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.
Chapter 4

Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

4.1 Introduction

Recognizing fine-grained contexts within individual houses is a key technology for next-generation smart home services, such as elderly monitoring [4] [5] [6], autonomous security [7], and personalized healthcare [8] [9]. 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 [15] 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 data and computing resources[15] [16]. Thus, there is still a big gap between the research and real life.
Our long-term goal is to implement fine-grained home context recognition that can adapt to custom contexts in every single house, and achieve accurate recognition with 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 of any cloud service that provides various recognition features as a service. Famous APIs include Microsoft Azure Computer Vision API [17] and IBM Watson Visual Recognition API [18].

In our preliminary study [36], 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 (Figure 4.1). 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”).

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 [3]. Specifically, for each cognitive API, we first construct an independent recognition model based on the tags de-
Chapter 4  Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

We then improve the accuracy of recognizing contexts by **majority voting** among results of the multiple independent models. Figure 4.2 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 [17], IBM Watson Visual Recognition API [18], Clarifai API [20], Imagga REST API [21], and Paralleldots API [22]. Regarding a set of tags derived from each image as document corpus, we vectorized each tag set using **TF-IDF (Term Frequency - Inverse Document Frequency)** [32].

The vectorized tag sets and the corresponding context labels were imported to Microsoft Azure Machine Learning Studio [33], where five independent recognition models were constructed with Multiclass Neural Net-
work. 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 multiple models reached 0.98. By context-wise analysis, the recognition accuracy of “General meeting”, “Nobody”, “One-to-one meeting”, “Play games” were 1.00, and “Dining together”, “Personal study”, “Room cleaning” were 0.96. From these results, it was shown that the recognition accuracy was significantly improved by the proposed method.

4.2 Preliminaries

4.2.1 Recognizing Fine-Grained Home Contexts

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

We use the term fine-grained home context to represent 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.

4.2.2 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, environment, configuration of the target space are 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, 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.

### 4.2.3 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 as a set of words called *tags*, representing objects and concepts that the API recognized. 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 [17], IBM Watson Visual Recognition API [18], Clarifai API [20], Imagga REST API [21], and Paralleldots API [22].
4.2.4 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 [36], we developed a method of fine-grained home context recognition based on supervised machine learning. Figure 4.1 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 document 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 [32], and constructed a multi-class classifier using Multiclass Neural Network on Microsoft Azure Machine Learning Studio [33].

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.

4.2.5 Ensemble learning

Ensemble learning mean that the accuracy of the model is greatly improved by combining the multiple models to generate the single learning model (Fig. 4.2). The basic ensemble learning include that the three method of bagging,
boosting, and stacking. In this paper, we used the majority voting (stacking) method of ensemble learning to piled up the models and the final recognition result was generated.

4.3 Proposed Method

4.3.1 Outline

Figure 4.3 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 4.3. 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.

### 4.3.2 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 created. Then, $n$ images in $IMG(c_i)$ are split into two sets $train(c_i)$ and $test(c_i)$, which are the training dataset with $\alpha$ 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 $yTag(img_{j'j'})$. 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 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 [32], Word2Vec [28] and Doc2Vec [29]. The selection of the vector re-
representation is up to the user. Similarly, the method converts \( yTag(img_{ij'}) \) into \( yVec(img_{ij'}) \) 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 \( yVec(img_{ij'}) \) to see if \( CLS \) returns the correct context \( c_{ij'} \).

### 4.3.3 Integrating Models Generated from Different APIs

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

**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 \( CLS_1, CLS_2, ..., CLS_m \).

**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 \( \text{mode}(c^1, c^2, ..., c^m) \). This completes the model construction.
Chapter 4  Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

4.4 Experimental Evaluation

4.4.1 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 4.4 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.

Fig. 4.4. Representative images of fine-grained contexts and USB camera
Table 4.1. Tag sets that different APIs extracted from an image

<table>
<thead>
<tr>
<th>Names of API used</th>
<th>The tags extracted from an image of &quot;Room cleaning&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Azure Computer Vision API</td>
<td>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</td>
</tr>
<tr>
<td>IBM Watson Visual Recognition API</td>
<td>living, room, indoors, classroom, basement, support, supporting structure</td>
</tr>
<tr>
<td>Clarifai API</td>
<td>room, furniture, indoors, table, desk, seat, chair, trading, floor, interior, design, home, hospital, medicine, technology, window, mirror, business, computer, people, production</td>
</tr>
<tr>
<td>Imagga REST API</td>
<td>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, renderer, television, classroom, hardwood, nobody, vase, hot el, 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</td>
</tr>
<tr>
<td>ParallelDots API</td>
<td>Room, Interior, design, Property, Vehicle, Building, Home, Sport, venue, Screenshot, Furniture</td>
</tr>
</tbody>
</table>

4.4.2 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
Chapter 4 Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

Table 4.2. The recognition accuracy results of each cognitive API-based Model and majority voting in this experiment

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

each image: Microsoft Azure Computer Vision, IBM Watson, Clarifai API, Imagga REST API, and Paralledots API. Table 4.1 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.

4.4.3 Results

Table 4.2 summarizes the results listing the overall accuracy and context-wise accuracy. The first five rows represent the results of the five individual models whose features were extracted by individual cognitive APIs. The last row represents the result achieved 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 is 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.

4.4.4 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 saw 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. In such case, the API may produce variable tag sets for the same 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.

4.5 Related Work

The activity recognition in a smart house has been widely studied in the field of ubiquitous computing. Nakamura et al. [37] proposed a system that recognizes activities of residents using big data accumulated within a smart
Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs

house. Ueda et al. [38] also proposed an activity recognition system using ultrasonic sensors and indoor positioning systems within a smart house. Although the performance of these systems are great, they are yet too expensive for general households.

The activity recognition with deep learning becomes a hot topic recently (e.g., [15] [16]). Although the deep learning is 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 data for custom fine-grained contexts.

Menicatti et al. [39] 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 [40] 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.
Chapter 5

Conclusion

In this paper, we first presented a method that evaluates the feasibility of image-based cognitive APIs towards the home context sensing of smart home. Applying the document similarity measure to the output tags produced from the image, the proposed method evaluates the performance of image-based context recognition with respect to the internal cohesion and the external isolation. In the experiment, we evaluated the feasibility of three different APIs towards context sensing within our laboratory. The experimental evaluation shows that we cannot expect too much performance for those general-purpose APIs without training.

We then proposed the framework of the home context recognition using the machine learning, and based on the proposed framework, we presented 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 a 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 proportions of incorrect recognition of “Eating” as “Gaming” and “Gaming” as “General meeting” were highest.
We finally have shown a method that recognizes fine-grained home contexts using multiple cognitive APIs. To achieve affordable context recognition in general households, the method delegates the image recognition task to cognitive APIs in the cloud, and uses 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 [23]). Investigating cultural information to improve the accuracy is also an interesting challenge.
Acknowledgements

This research was conducted in the CS24 laboratory, Department of Computational Science, Graduate School of System Informatics, Kobe University.

My deepest heartfelt appreciation goes to Professor Kuniaki Uehara, and Professor Mitsuo Yokokawa of Kobe University for valuable guidance in the process of examination.

I would like to show my greatest appreciation to Associate Professor Masahide Nakamura of Kobe University for providing me the great opportunity to study in his laboratory, and in personal meeting weekly with discussion to guidance for this research. I am deeply grateful as well to Assistant Professor Sachio Saiki of Kobe University, who gave me effective guidance on rules of academic papers for writing, use of phrases, etc. I am also indebted to the members of CS24 for my daily helps.

I would particularly like to thank Kobe University for tuition reduction and scholarships.

Finally, I owe my deepest gratitude to my parents for encouraging my study and supporting my living abroad all along.
Bibliography


Bibliography


A

List of Publications

A.1 Presentations in International Conferences


A.2 Presentations in Domestic Conferences

