Article

Integrating Multiple Models Using Image-as-Documents Approach for Recognizing Fine-Grained Home Contexts [†]

Sinan Chen [1],[*], **Sachio Saiki [1]** and **Masahide Nakamura [1,2]**

[1] Graduate School of System Informatics, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan; sachio@carp.kobe-u.ac.jp (S.S.); masa-n@cs.kobe-u.ac.jp (M.N.)

[2] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[*] Correspondence: chensinan@ws.cs.kobe-u.ac.jp; Tel.: +81-78-803-6295

[†] This paper is an extended version of the conference paper: Sinan, C.; Sachio, S.; Masahide, N. Recognizing Fine-Grained Home Contexts Using Multiple Cognitive APIs. In Proceedings of the CyberC 2019, Guilin, China, 17–19 October 2019.

Abstract: To implement fine-grained context recognition that is accurate and affordable for general households, we present a novel technique that integrates multiple image-based cognitive APIs and light-weight machine learning. Our key idea is to regard every image as a document by exploiting "tags" derived by multiple APIs. The aim of this paper is to compare API-based models' performance and improve the recognition accuracy by preserving the affordability for general households. We present a novel method for further improving the recognition accuracy based on multiple cognitive APIs and four modules, fork integration, majority voting, score voting, and range voting.

Keywords: context recognition; image; cognitive APIs; machine learning; majority voting; score voting; range voting; smart home

1. Introduction

As Internet of Things (IoT) and Artificial Intelligence (AI) technologies continue to develop, people have increasing expectations about smart home services. Recognizing fine-grained contexts within individual houses is a key technology for next-generation smart home services. 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 the environment for special purposes of application, such as elderly monitoring [1–3], autonomous security [4], and personalized healthcare [5,6]. It has been studied for many years in the field of ubiquitous computing [7,8]. Traditional ubiquitous computing employs ambient sensors [9,10], wearable sensors [11], and indoor positioning systems [12] that are installed at home to retrieve various data.

In recent years, the emerging deep learning [13–16] allows the system to recognize multimedia. Since image, voice, and text usually contain richer information than conventional sensor data, it is promising to use such multimedia data for recognizing fine-grained home contexts. In our study, we especially focus on image data with human activities at home for recognising home contexts. For this, one may try to recognize home contexts via image recognition based on naive deep learning. However, constructing a custom recognition model dedicated to a single house requires a huge amount of labeled datasets and computing resources [17,18]. It is not only hard to construct a universal recognition model from one house to another, but also the security and privacy issues that come with a large amount of data influence acceptability for the users. Thus, there is still a big gap between research and real life.

Our interest is to use little image data, applying cognitive services to implement affordable context sensing that can adapt to custom contexts in every single house. The cognitive service is a cloud

service with cognitive computing functions that provide the capability to understand multimedia data (i.e., vision (object recognition) [19], speech recognition [20], natural language processing [21], etc.), based on sophisticated machine-learning algorithms powered by the offered big data and large-scale computing. They are offered by cloud companies, but the data processing algorithms behind them are not public. They are also widely used in various fields of research, such as modern knowledge management solutions [22] and criminal detection and recognition [23]. The cognitive API is an application program interface via the HTTP/REST protocol [24], with which developers can easily integrate powerful recognition features in their own applications. An image-based cognitive API receives an image from an external application, i.e., extracts specific information from the image, and returns the information in JavaScript Object Notation (JSON) [25] format from the cloud server rather than the local. The information usually contains a set of words called "tags", representing objects and concepts that the API has recognized in the given image. Examples of tags from the API are: [Living, room, indoors, classroom, basement, supporting structure]. The information of interest and the way of recognizing the image vary among individual cognitive services. Related work uses image tagging technology with deep learning, as in [26–28], but the implementation is more complex. In our future realistic implementation, for security and privacy of the users the images are not saved after sending the APIs over.

The main contribution of this paper is to present a novel method which is not only affordable but also has higher accuracy in recognizing fine-grained home contexts. For this purpose, we are currently investigating techniques that integrate inexpensive camera devices, multiple image-based cognitive APIs, and light-weight machine learning. We previously encoded the tags of a single API to document vectors, then applied them into machine learning for the model construction [29]. However, we found that the accuracy significantly decreased for contexts with multiple people (e.g., "General meeting", "Dining together", "Play games"). In this paper, we define a concept called "image as documents", which uses different cognitive APIs for receiving the same image. As the proposed method, we present four modules, fork integration and three voting approaches (i.e., majority [30], score [31], range [32]), to integrate multiple models generated from different APIs. In this way, we can not only compare the difference in API-based models' performance, but improve the recognition accuracy. Furthermore, we also discuss the potential implementation of the model simplification as in [33], of the more efficient process as in [34], and of the other techniques as in [35] and [36], integrated into the proposed method.

In order to evaluate the proposed method, we experimented to recognize the seven contexts of our laboratory, which use little image data and five cognitive APIs. Based on the proposed method, we completed the process from each independent model construction to multiple model integration, and implemented the above four modules. As a result, for each API-based model, the Imagga API-based model performed best within the five models, and the ParallelDots API-based model was the worst. Meanwhile, the overall accuracy by majority voting reached 0.9753. Furthermore, the overall accuracy by score voting reached 0.9776. We also checked the accuracy distribution by range voting, which was meant to solve the problem of recognition instability in contexts with multiple people. In this way, we found that the top of overall accuracy reached 0.9833 when the value of the lower limit was between 0.5 and 0.6. Thus, the recognition accuracy was significantly improved by the method proposed in our experiment.

2. Related Work

The problem of recognizing fine-grained home contexts with human activities [37–40] has been widely studied in the field of ubiquitous computing. As described in the introduction, it is defined by every user depending on a special purpose. However, for realizing a technique that applies for general households, we consider that it should have several advantages at least: (1) Low cost of devices and systems in both purchase and maintenance, (2) light-weight and a high-accuracy approach for data processing, (3) a stable and secure approach for data communication. In this section, we introduce some related works from recent years around the above three points.

Nakamura et al. [41] proposed a system that recognizes the activities of residents using big data accumulated within a smart house. Ueda et al. [42] also proposed an activity recognition system using ultrasonic sensors and indoor positioning systems within a smart house. While the performance of these systems is great, they are still too expensive for general households. Sevrin et al. [43] contributed to the effort of creating an indoor positioning system based on low cost depth cameras (Kinect). However, for retrieving more specific information on human activities at home (e.g., cleaning, dining, etc.), embracing the position of users is obviously not enough. In recent years, activity recognition with deep learning has become a hot topic. Research in [44] built on the idea of 2D representation of an action video sequence by combining the image sequences into a single image called the Binary Motion Image (BMI) to perform human activity recognition. Asadi-Aghbolaghi et al. [45] presented a survey on current deep learning methodologies for action and gesture recognition in image sequences. While deep learning is a powerful approach for recognizing image data, a huge amount of data is required to build a high-quality model. Therefore, it is unrealistic for individual households to prepare a huge amount of labeled datasets for custom fine-grained contexts.

Using cloud services to recognize human activities at home is key for implementing light-weight data processing. Pham et al. [46] presented a Cloud-Based Smart Home Environment (CoSHE) for home healthcare. While the effect of the system is good, various basic sensors and devices must be installed, which is not ideal for implementation and long-term maintenance. Menicatti et al. [47] proposed a framework that recognizes indoor scenes and daily activities using a cloud-based computer vision. Their concept and aim are similar to our method. However, the way of encoding tags is based on a the naive Bayes model where each word is present or not. Moreover, 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 [48] investigates the influence of a 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 to investigate this perspective in future work.

Regarding the security of uploading images to cloud services, Qin et al. [49] studied the design targets and technical challenges that lie in constructing a cloud-based privacy-preserving image-processing system. There are also some the related works that focus on the security and privacy of smart homes. Dorri et al. [50] proposed a blockchain-based smart home framework with respect to the fundamental security goals of confidentiality, integrity, and availability. Geneiatakis et al. [51] employed a smart home IoT architecture to identify possible security and privacy issues for users. Through the above articles, we also plan, in future work, to focus more on making computation and communication practical for the encrypted data of smart homes.

3. Methodology

This section produces a complete description on the preliminary study, proposed method, and discussion of the related techniques.

3.1. Preliminary Study

Constructing a single classifier model is a basic and essential part of realizing fine-grained home context recognition. Unlike naive deep learning, we previously dedicated the features of images extracted from a single cognitive API, to apply to light-weight supervised machine learning [29]. The key step for building the context recognition model is to make every image document covert for a set of numerical values, which is document vectorization processing (see step 4). In this section, we describe the method that constructs a recognition model from a single API. It also can be used for performance comparison among the different cognitive APIs, which has been developed from our other preliminary study using unsupervised learning [52].

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 effectively expose c_i from all images obtained in step 1. At this time, the total $l \times n$ images are sampled as datasets. Then, the n images in $IMG(c_i)$ are split into two sets, $train(c_i)$ an $test(c_i)$, which are the training dataset with α images and the test dataset with $n - \alpha$ images, respectively.

Step 3: Extracting tags as features

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

Step 4: Converting tags into vectors

Regarding every $xTag(img_{ij})$ as a document, and the whole tag set 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 the r-th tag. Famous document vectorization techniques include TF-IDF [53], Word2Vec [54], and Doc2Vec [55]. The selection of the vector representation is up to the user. Similarly, the method converts $yTag(img_{i'j'})$ into $yVec(img_{i'j'})$ using the same vector representation.

Step 5: Constructing a classifier

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

3.2. Proposed Method

To improve recognition accuracy (especially in the case of Figure 1) based on Section 3.1, this section describes the most important question in this paper, which is how to construct a whole recognition model by integrating multiple recognition models (see Figure 2). As we mentioned in Section 1, the core of our method is to use the image-as-documents concept, which operates with a fork integration module and a choice among three voting modules. For ensuring that many results of multiple cognitive APIs accurately correspond to every image, the preliminary stage of the training classifier is also very important.
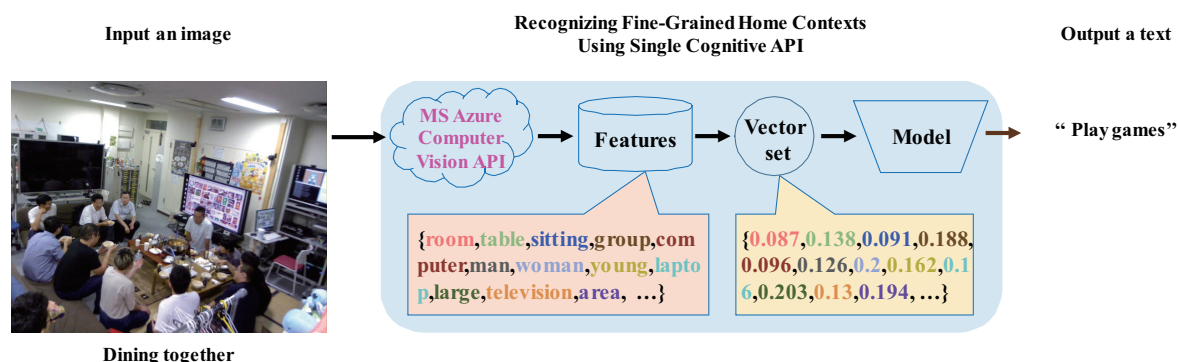

Figure 1. Example of an image with multiple people misrecognized using a single cognitive API.
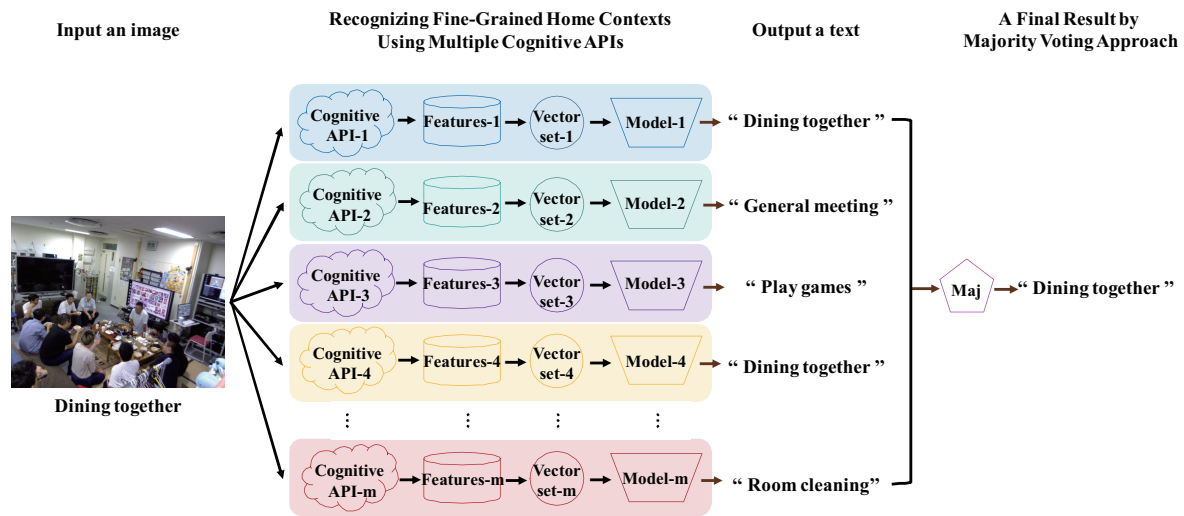
The specific whole steps are as follows.

Figure 2. Example of recognizing an image using multiple cognitive APIs and majority voting.

Step 6: Constructing multiple classifiers

By repeating steps 3 to 5 of Section 3.1 for different image recognition APIs, the proposed method constructs m independent recognition models. Note that training and test datasets created in steps 1 and 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 the 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

To complete the model construction, the method first adds a fork integration module F, which sends a given image simultaneously to m recognition models $VEC_q + CLS_q$ ($1 \leq q \leq m$), corresponding to the image-as-documents concept. Then, the method adds three voting modules, which users choose in different home contexts, as follows.

- **Majority voting**: It receives m outputs $c^1, c^2, ..., c^m$ from $VEC_q + CLS_q$ ($1 \leq q \leq m$), and returns $mode(c^1, c^2, ..., c^m)$ (see Figure 3).
- **Score voting**: It receives m outputs $c^1, c^2, ..., c^m$ and scores $s^1, s^2, s^3, ..., s^m$ of each output from $VEC_q + CLS_q$ ($1 \leq q \leq m$), and returns $\max \sum_{i=1}^{m} s^i(c^i)$ by comparing the total of scores with the same output c^i (see Figure 4).
- **Range voting**: It receives m outputs $c^1, c^2, ..., c^m$ and scores $s^1, s^2, s^3, ..., s^m$ of each output from $VEC_q + CLS_q$ ($1 \leq q \leq m$), and sets a lower limit for scores s^i to be used. The output c^i will be used in the score voting if the corresponding score s^i is above the lower limit (compare Figures 4 and 5).
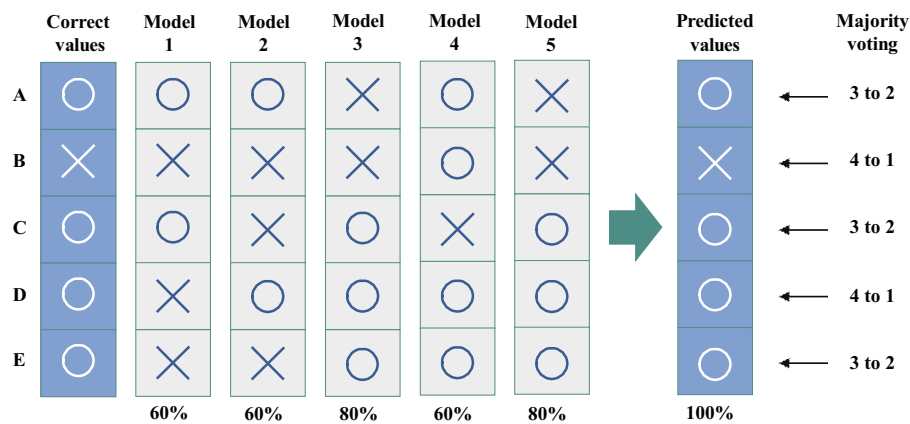
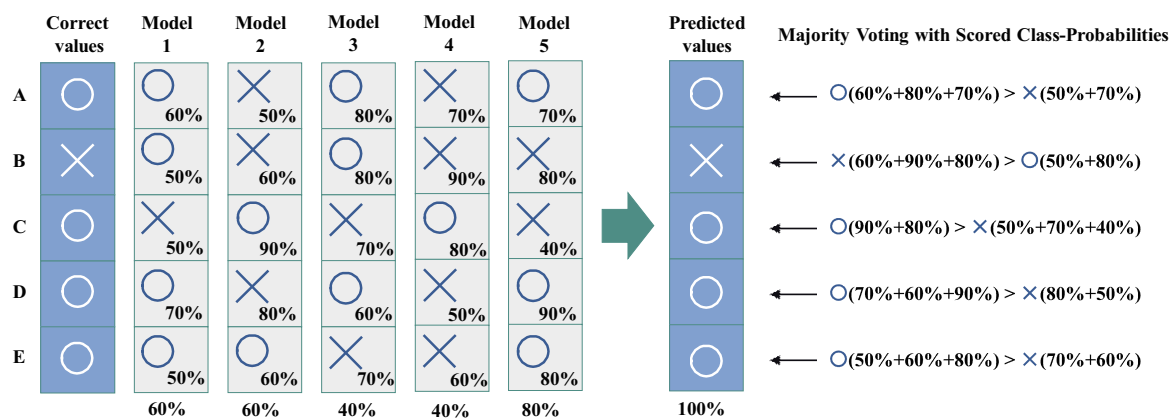**Figure 3.** Example of majority voting with ensemble learning.

Figure 4. Example of score voting using the total of each class probability.
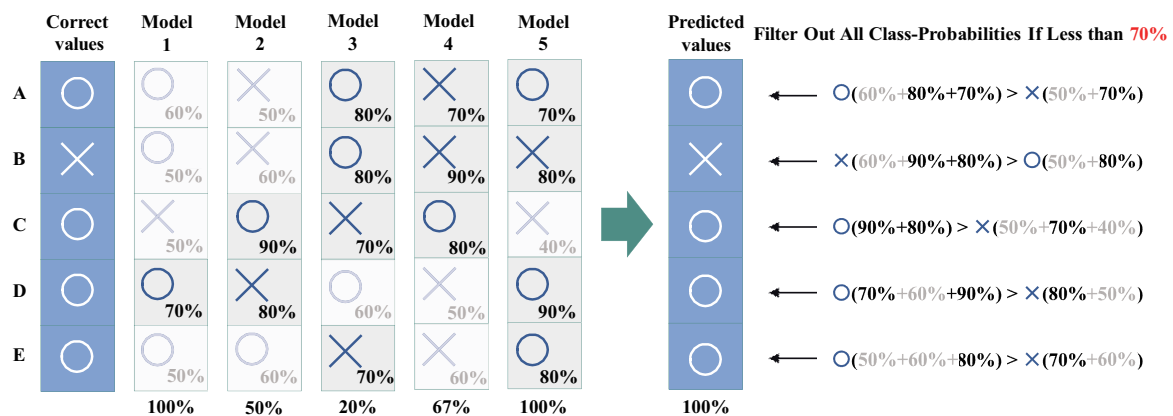
Figure 5. Example of range voting using all class probabilities that scored above 70% based on Figure 4.

3.3. Discussion

To construct a model using multiple cognitive APIs, another method is combining the tags of different cognitive API models, which come from each image (see Figure 6). However, this will increase the dimension of the input data of the built model. The related techniques for dimensionality reduction include Principal Component Analysis (PCA) [33], Locally Linear Embedding (LLE) [56], Latent Semantic Analysis (LSA) [57] and so on. We would like to experiment with them for model simplification and accuracy in our future work. In addition, for dimension reduction [58–60] of

document vectors, the Restricted Boltzmann Machines (RBMs) [61,62] is a good method. In the existing research, many models only use a small number of features as input; hence, there may not be enough information to classify documents accurately. Conversely, if more features are input, as we discussed earlier, it will increase the dimension of input data, resulting in a large increase in the training time of the model, and its recognition accuracy may also lower. Therefore, we also would like to use RBMs to extract highly distinguishable features from the combined input features, and use them as input to the corresponding model in our future work. We believe that this will greatly improve the efficiency of model construction. Furthermore, conducting the proposed method using the other machine learning techniques (e.g., Hidden Markov Model (HMM) [63], regression, graphical models [64], etc.) is also feasible. We will conduct experiments to compare these technologies combined with the proposed method in future work.
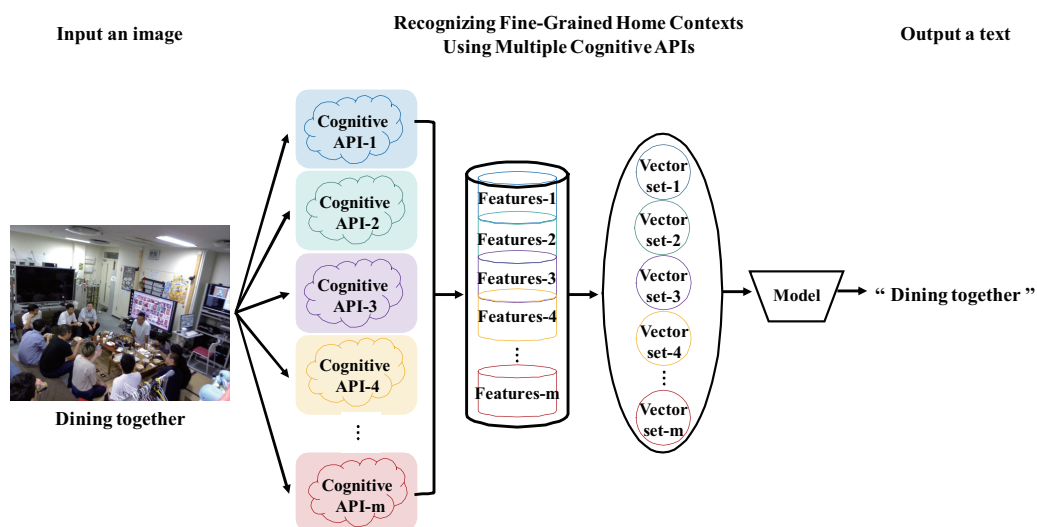

Figure 6. Example of constructing a model by combining the features of multiple cognitive APIs.

4. Experimental Evaluation

This section introduces results, discussion, and an experiment conducted for comparing the difference in API-based models' performance and improving the recognition accuracy.

4.1. Experimental Setup

The experiment was conducted in a shared space of our laboratory. First, we installed a 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 five seconds, and uploads the image to a server. The image resolution is 1280×1024. The images were accumulated from July 2018 to March 2019. The target shared space is used by members of our laboratory for various activities. In this experiment, we chose the seven kinds of fine-grained contexts: "Dining together", "General meeting", "Nobody", "One-to-one meeting", "Personal study", "Play games", and "Room cleaning". The detail of each context is as in Table 1. For each context, we selected and labeled 100 representative images from the server, taken on different dates. We then randomized the order of a total of 700 image data, and split them into half, as training data and test data.

Table 1. The detail of each defined context in this experiment.

Context labels	The contents of what the images of each context represent
Dining together	We often cook by ourselves to **dining together** in our laboratory
General meeting	We are sitting together in a **general meeting** every Monday
Nobody	There is also the **nobody** situation during the weekend or holidays
One-to-one meeting	We often have a **one-to-one meeting** for the study discussion
Personal study	Sometimes the public computer is used for **personal study**
Play games	We often gather around and **play games** to relax in our spare time
Room cleaning	The staff twice a week come for **room cleaning** in our laboratory

4.2. Building and Combining API-Based Models

We first built a recognition model for the seven contexts. The following five cognitive APIs were used to extract tags from images referring to each context: Microsoft Azure Computer Vision API [65], IBM Watson Visual Recognition API [66], Clarifai API [67], Imagga REST API [68], and ParallelDots API [69]. Tables 2 and 3 show the representative images of seven contexts (including tag results) and USB camera. We can see that the different APIs recognized the same image from different perspectives. Each tag set extracted from an image was then transformed into a vector representation using TF-IDF (Term Frequency–Inverse Document Frequency) [53]. We then imported the datasets and the corresponding context labels to Microsoft Azure Machine Learning Studio [70]. For each cognitive API, we trained a classifier using the Multiclass Neural Network with the default setting. Each of the five trained models was evaluated by the test data to observe the performance of individual models. We finally used four modules to build a whole recognition model by integrating five individual models. To check accuracy distribution, we adjusted the lower limit of scored class probabilities to between 0 and 0.9.

4.3. Results

Table 4 shows the accuracy results of each cognitive API-based model and three voting modules. Among the five models, the Imagga API-based model was the best (0.9429), while the ParalleDots API-based model scored the lowest (0.7718). These results can be used as reference values for model performance comparison. As for the context-wise accuracy, the performances of the five models were each different (see Table 4). For instance, let us compare the Watson API-based model and the ParalleDots API-based model. The Watson model was bad at recognizing "General meeting" (0.6730), compared to the ParalleDots model (0.8910). Interestingly, however, the Watson model was better at recognizing "One-to-one meeting" (0.8040) than the ParalleDots model (0.4460).

With regard to the overall accuracy with three voting modules, the majority voting achieved an accuracy of 0.9753, the score voting achieved an accuracy of 0.9776, and the range voting achieved the top accuracy of 0.9833 with the range 0.5 to 0.6. Regarding the context-wise accuracy with the majority voting, these limitations of the individual models were mutually complemented. The recognition accuracy of "Dining together" was 0.9565, "Personal study" was 0.9561, and "Room cleaning" was 0.9572, while the accuracy of "General meeting", "Nobody", "One-to-one meeting", and "Play games" were 1.0000. Regarding the context-wise accuracy with score voting, it further made up for the shortage of simply obtaining the final result by the quantity. The recognition accuracy of "General meeting" was 0.9685, "Personal study" was 0.9751, and "Room cleaning" was 0.9720, while the accuracy of "Dining together", "Nobody", "One-to-one meeting", and "Play games" were 1.0000. Regarding the context-wise accuracy with the range voting, it excludes some low-score API results before voting, which further promotes the improvement of the accuracy. The recognition accuracy of "General meeting" was 0.9836 and of "Personal study" was 0.9800, while the accuracy of "Dining together",

Table 2. The representative images of the four contexts (including tag results from different APIs).

Images				
Contexts	Dining together	General meeting	Play games	One-to-One meeting
Tag results of Microsoft Azure Computer Vision API	indoor, person, ceiling, table, room, living, people, food, sitting, filled, items, cluttered, group, woman, man, large, place, several, television, fire, many, kitchen, fireplace, pizza, crowded, plate, bed	indoor, room, table, living, computer, sitting, cluttered, laptop, desk, woman, man, dog, people, office, kitchen, filled, food, playing, furniture, standing, television, large, wooden, young, group, cat, holding, video, fire	indoor, ceiling, person, room, living, table, sitting, child, young, computer, small, cluttered, woman, food, filled, boy, man, desk, kitchen, television, playing, people, laptop, little, standing, girl, furniture, large, fireplace, fire, video, bed, game, holding, group, bedroom	indoor, table, room, ceiling, computer, desk, living, cluttered, office, sitting, laptop, television, area, filled, monitor, equipment, screen, large, people, several, video, standing, playing, woman, game, keyboard, man, desktop, bed, group
Tag results of IBM Watson Visual Recognition API	control room, indoors, newsroom, office, building, television equipment, control center, workstation, digital computer, computer, machine, equipment, ultramarine color	newsroom, office, building, classroom, indoors, beauty salon, shop, retail store, war room, workroom, control room, sage green color	television room, indoors, control room, workstation, digital computer, computer, machine, audiovisual aid, television equipment, newsroom, office, building, equipment, system, electronic equipment, ultramarine color	control room, indoors, microfiche, photographic film, photographic equipment, control center, building, workstation, digital computer, computer, machine, television equipment, system, electronic equipment, ultramarine color
Tag results of Clarifai API	room, furniture, education, indoors, school, people, exhibition, adult, desk, group, production, computer, commerce, class, vehicle, technology, election, classroom, healthcare	room, indoors, desk, furniture, table, exhibition, technology, computer, business, chair, interior design, office, production, education, industry, people, commerce, seat, classroom	room, computer, technology, desk, indoors, furniture, education, group, table, people, exhibition, business, adult, school, television, medicine, commerce, production, industry	room, furniture, indoors, desk, table, chair, seat, technology, education, interior design, office, hospital, trading floor, school, business, university, computer, classroom, exhibition
Tag results of Imagga REST API	office, business, corporate, teamwork, building, businessman, people, man, meeting, team, work, group, men, happy, businesswoman, professional, male, executive, success, working, women, center, job, businesspeople, adult, person, indoors, shop, hall, communication, modern, restaurant, passenger, training, attractive, interior, smiling, colleagues, lifestyle, suit, career, room, table, successful, worker, manager, handshake, businessmen, computer, partnership, company, portrait, smile, indoor, partner, workplace, barbershop, place of business, strength, education, structure, occupation, laptop, corporation, adults, equipment, diverse, support, handsome, exercising, diversity, gym, boss, pretty, commerce, power, light, workers, standing, sitting, two, 20s, hand, confident, fitness, employee, associate, 20 24 years, coworkers, determination, conference, 40s, hands, mercantile establishment, happiness, counter, ethnic, black, health, chair, health spa, looking, together, seminar, healthy lifestyle, agreement, leadership, establishment, talking, desk, horizontal, presentation, lady, exercise, bright, day, architecture	room, classroom, office, center, interior, table, chair, furniture, computer, modern, desk, indoors, business, home, work, house, design, working, meeting, floor, monitor, professional, empty, corporate, group, people, decor, laptop, man, businessman, education, businesswoman, sitting, light, executive, restaurant, person, indoor, window, teamwork, building, seat, team, male, inside, smiling, communication, technology, wood, contemporary, success, sofa, keyboard, screen, learning, luxury, school, residential, lamp, display, mouse, confident, structure, equipment, board, chairs, engineer, job, class, businesspeople, living, nobody, couch, together, workplace, women, wall, glass, teacher, manager, coffee, adult, living room, lighting, comfortable, architecture, talking, career, pen, showing, study, occupation, successful, style, happy, kitchen, worker, decoration, associate, conference, discussion, partners, colleagues, student, cooperation, corporation, employee, alcove, presentation, training, book, suit	office, room, table, interior, furniture, desk, monitor, modern, computer, chair, center, home, business, indoors, classroom, work, house, decor, working, design, floor, window, light, wood, lamp, inside, corporate, laptop, people, technology, seat, education, empty, professional, sofa, living, equipment, indoor, luxury, comfortable, television, executive, screen, keyboard, display, chairs, apartment, nobody, meeting, sitting, school, 3d, worker, mouse, residential, person, alcove, occupation, architecture, businesswoman, decoration, man, job, businessman, success, adult, restaurant, learning, wall, smiling, relax, building, pillow, workplace, furnishing, contemporary, glass, place, teamwork, communication, network, style, carpet, vase, couch, male, structure, elegance, training, board, group, women, notebook, living room, conference, class, desktop, businesspeople, career, pen, relaxation, presentation, book, data, confident, stylish, lifestyle, electronic equipment, suit, team, kitchen	room, interior, monitor, furniture, office, table, modern, computer, desk, home, house, decor, indoors, design, lamp, living, mouse, floor, light, sofa, 3d, chair, work, luxury, center, window, business, apartment, equipment, display, comfortable, wall, wood, keyboard, television, screen, architecture, technology, classroom, working, desktop computer, residential, decoration, living room, seat, carpet, building, liquid crystal display, inside, home theater, corporate, device, personal computer, structure, theater, empty, domestic, contemporary, electronic equipment, elegance, relaxation, nobody, education, indoor, electronic device, glass, lifestyle, armchair, school, pillow, vase, render, sitting, place, laptop, relax, digital computer, furnishings, ceiling, chairs, reflection, couch, lighting, desktop, people, learning, alcove, fashion, network, bedroom, parquet, estate, space, lifestyles, style, success, smiling, pillows, life, residence, class, workplace, hand, bed, communication, pen, horizontal, training, executive, data, new, board, machine, family
Tag results of ParallelDots API	Room, Sport venue, Person, Machine, Clothing, Interior design, Physical fitness, Sports, Vehicle	Room, Person, Convention, Sport venue, Clothing, Classroom, Academic conference, Man, Interior design	Room, Sport venue, Person, Clothing, Machine, Interior design, Classroom, Furniture, Physical fitness	Room, Sport venue, Person, Interior design, Clothing, Machine, Furniture, Classroom, Physical fitness

Table 3. The representative images of the other three contexts (including tag results) and USB camera.

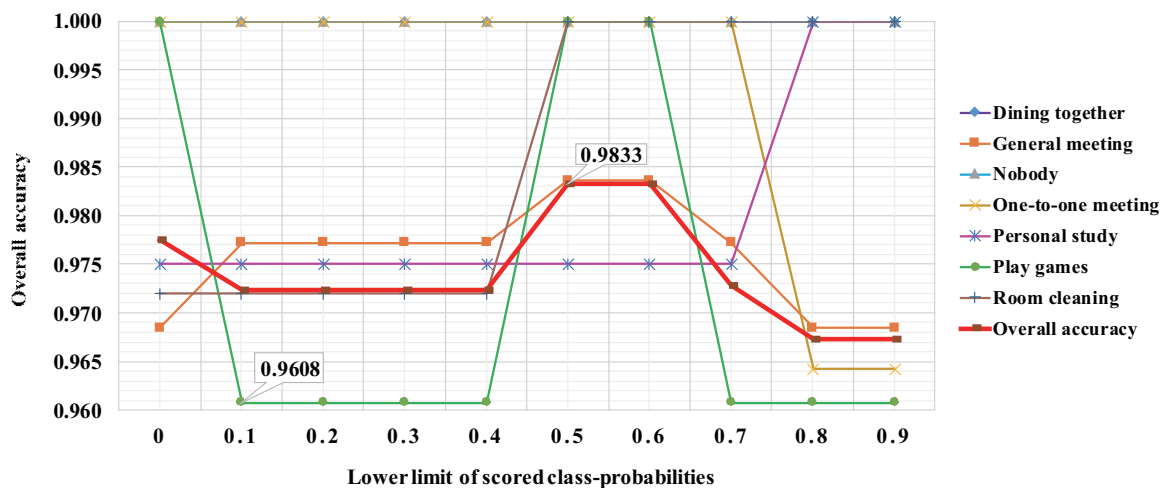|	Personal study	Room cleaning	Nobody	USB camera
Images				
Contexts	Personal study	Room cleaning	Nobody	USB camera
Tag results by Microsoft Azure Computer Vision API	indoor, living, table, room, television, furniture, items, sitting, cluttered, filled, computer, desk, bed, fire, large, kitchen, several, screen, laptop, wooden, fireplace, standing, suitcase, refrigerator, white, luggage, bedroom	indoor, table, living, room, computer, desk, sitting, window, laptop, television, large, office, wooden, monitor, furniture, screen, keyboard, video, man, fireplace, game, fire, people, kitchen, white, playing, plate	indoor, table, room, desk, computer, living, monitor, sitting, small, cluttered, office, area, filled, food, laptop, furniture, wooden, keyboard, television, home, large, video, game, kitchen, screen, mouse, standing, desktop, white, fire, playing, remote, refrigerator, plate, man	
Tag results by IBM Watson Visual Recognition API	workroom, indoors, office, building, workstation, digital computer, computer, machine, control room, microfiche, photographic film, photographic equipment, equipment, sage green color	control room, indoors, workstation, digital computer, computer, machine, television equipment, beauty salon, shop, retail store, building, control center, equipment, system, electronic equipment	control room, indoors, living room, television room, workstation, digital computer, computer, machine, office, building, electronic equipment, gray color	
Tag results by Clarifai API	indoors, room, furniture, desk, table, technology, chair, trading floor, computer, business, production, seat, hospital, television, exhibition, industry, interior design, office, cabinet	room, furniture, indoors, table, seat, chair, desk, interior design, sofa, trading floor, computer, window, technology, office, contemporary, business, television, inside, hospital	indoors, room, furniture, trading floor, table, chair, desk, hospital, cabinet, window, home, seat, interior design, inside, business, production, exhibition, people, medicine	
Tag results by Imagga REST API	furniture, room, interior, table, home, monitor, house, modern, decor, desk, office, lamp, floor, sofa, light, living, 3d, apartment, design, home theater, wood, chair, comfortable, luxury, indoors, inside, theater, center, wall, living room, window, carpet, pillow, building, decoration, architecture, computer, equipment, relax, television, residential, furnishings, vase, structure, rest, glass, seat, comfort, work, bedroom, lifestyle, relaxation, bed, reflection, electronic equipment, domestic, technology, couch, display, furnishing, estate, indoor, ceiling, entertainment center, mouse, render, business, fireplace, parquet, lighting, working, mirror, space, drawing room, spacious, chairs, residence, plant, illumination, objects, rendering, lifestyles, contemporary, elegance, family, niche, situation, pillows, cabinet, armchair, area, decorating, keyboard, shade, sitting, nobody, laptop, book, style, stylish, new, minimalism, blind, tables, cozy, corporate, empty, real, horizontal, fashion, device, leisure, screen	interior, room, table, furniture, office, modern, home, house, decor, lamp, light, sofa, design, indoors, floor, chair, desk, living, wood, 3d, center, apartment, comfortable, window, luxury, monitor, wall, inside, computer, architecture, glass, residential, home theater, building, seat, decoration, structure, living room, business, carpet, pillow, theater, relax, equipment, work, domestic, comfort, lifestyle, empty, chairs, vase, indoor, couch, lighting, furnishings, ceiling, relaxation, space, television, reflection, rest, style, classroom, elegance, working, lifestyles, technology, render, display, contemporary, fireplace, stylish, area, spacious, cozy, armchair, plant, keyboard, rendering, mirror, estate, corporate, mouse, desktop computer, nobody, drawing room, residence, illumination, furnishing, kitchen, fashion, situation, tables, parquet, urban, device, objects, screen, shade, bed, horizontal, personal computer, restaurant, broadcasting, laptop, family, electronic equipment, bedroom, life	table, furniture, interior, room, pool table, game equipment, equipment, home, house, modern, decor, lamp, wood, indoors, light, apartment, luxury, design, kitchen, furnishing, 3d, chair, floor, comfortable, inside, office, sofa, glass, living, residential, architecture, window, wall, decoration, indoor, domestic, render, chairs, living room, relaxation, seat, building, bedroom, comfort, relax, pillow, style, vase, lighting, nobody, bed, contemporary, rest, area, empty, carpet, lifestyle, steel, stove, mirror, dining, structure, refrigerator, furnishings, cabinet, ceiling, bowling pin, couch, tile, restaurant, clean, stylish, shelf, reflection, oven, sink, business, monitor, theater, center, desk, home theater, dinner, counter, device, elegance, parquet, fashion, new, tables, hotel, estate, lifestyles, brown, bowling equipment, wooden, blind, television, food, mansion, granite, real estate, residence, marble, illumination, health spa, expensive, decorate, real, computer, plant, lights, appliance, mouse, family, night	
Tag results by ParallelDots API	Room, Interior design, Design, Machine, Person, Art, Furniture, Sport venue, Classroom	Room, Vehicle, Transport, Interior design, Machine, Sport venue, Furniture, Person, Public transport	Room, Vehicle, Interior design, Property, Transport, Building, Furniture, Public transport, Sport venue	

Table 4. The accuracy results of each cognitive API-based model and three voting modules.

Model or Voting Names	Overall Accuracy	Dining Together	General Meeting	Nobody	One-to-one Meeting	Personal Study	Play Games	Room Cleaning
Azure API – model	0.8543	0.9550	0.8910	1.0000	0.6610	0.9170	0.8430	0.7650
Watson API – model	0.8000	0.8860	0.6730	0.8230	0.8040	0.9380	0.8040	0.7060
Clarifai API – model	0.9143	0.9090	0.9820	0.9110	0.8390	0.9170	0.9220	0.9220
Imagga API - model	**0.9429**	0.9550	0.9270	1.0000	0.8930	0.9580	0.9220	0.9610
ParalleDots API – model	**0.7718**	0.7950	0.8910	0.9330	**0.4460**	0.8750	0.6670	0.8040
Majority voting	0.9753	0.9565	1.0000	1.0000	1.0000	0.9561	1.0000	0.9572
Score voting	0.9776	1.0000	0.9685	1.0000	1.0000	0.9751	1.0000	0.9720
Range voting (0.5 to 0.6)	**0.9833**	1.0000	0.9836	1.0000	1.0000	0.9800	1.0000	1.0000

"Nobody", "One-to-one meeting", "Play games", and "Room cleaning" were 1.0000. Figure 7 presents the distribution of accuracy results using range voting within the range 0 to 0.9, and includes the overall accuracy and the context-wise accuracy. We can see the top of the overall accuracy was 0.9833 when the lower limit was between 0.5 and 0.6. The entire accuracy of "Dining together" stabilized at 1.0000. However, the accuracy of "Play games" and "General meeting" were unstable, especially for "Play games", which had the lowest accuracy, 0.9608.

Figure 7. The distribution of accuracy results using range voting within the range 0 to 0.9.

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" (0.4460) was that (1) no distinctive word characterizing of the context was found, and (2) the number of words in the tag sets was relatively small. The accuracy also depends on the nature of the context. We

found that contexts where people are dynamically moving (e.g., "Dining together", "Room cleaning") were relatively difficult to recognize. In such contexts, observable features are frequently changed from one image to another; for instance, positions of people, visible furniture and background. Therefore, the API may produce variable tag sets for the same context, which decreases the internal cohesion of the feature vectors.

Including 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. 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. It was seen from the results of naive majority voting that the accuracy of "Dining together", "Personal study", and "Room cleaning" were not perfect. The reason is that some situations of the majority results of an image were wrong. The recognition accuracy of "Personal study" and "Room cleaning" improved significantly using score voting. However, there was greater instability in the contexts with multiple people (e.g., "Dining together", "General meeting") compared with the results of majority voting. On adjusting the lower limit of scored class probabilities to between 0 and 0.9, there was instability in the accuracy of "Play games" and "General meeting" but not for "Dining together". One of the reasons for this is that the context richness of "Dining together" was prominent compared to others. This means the output tags of "Dining together" were many, whether by the total or the semantic (see Table 2). The other reason is there were some difficulties in recognizing the contexts with no big change in the number of persons and objects (e.g., "Play games", "General meeting"). With regard to the top of the overall accuracy, 0.9833, by range voting with the range 0.5 to 0.6, it means that some situations in the majority results of an image were wrong when the scored class probabilities were less than 0.5 or above 0.6.

5. Conclusions

In this paper, a method that integrated models based on multiple cognitive APIs and four presented modules for improving the recognition accuracy is proposed. From experimental evaluation, the difference in API-based models' performance is compared, confirming the advantage that the recognition accuracy is improved by the proposed method.

The image recognition method is different from one API to another. By constructing multiple classifiers with different perspectives, taking majority voting derives the context with a maximum likelihood for the same image. However, there are also some images with recognition difficulty; hence, the case of the false results being output by the majority APIs. Using score voting, we could reduce false results determined by only the number of outputs within the same context to some extent. Furthermore, by setting the different range of lower limits, we deeply understood the recognition difficulty for each context by finding the range with the highest accuracy. This is of great significance for improving the proposed method in our future study. As to the topic of fine-grained home context recognition for general households, this paper has some points that need to be improved. While the different cognitive APIs were used for building models and performance evaluation, the different experimental spaces have not yet been used. Moreover, to build the model of context recognition in future different households, how to select the representative data of context more scientifically is still a task directly related to recognition accuracy. These are the directions of our future work. In addition, to achieve effective application, it is necessary to consider both the retrieval of more features from the data, and the development of various algorithms. Therefore, investigating more ways to use cloud resources to retrieve feature values of local images will also be a topic of future work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vuegen, L.; Van Den Broeck, B.; Karsmakers, P.; Van hamme, H.; Vanrumste, B. Automatic Monitoring of Activities of Daily Living based on Real-life Acoustic Sensor Data: A preliminary study. In Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics: Grenoble, France, 2013; pp. 113–118.

2. Debes, C.; Merentitis, A.; Sukhanov, S.; Niessen, M.; Frangiadakis, N.; Bauer, A. Monitoring Activities of Daily Living in Smart Homes: Understanding human behavior. *IEEE Sig. Process. Mag.* **2016**, *33*, 81–94. doi:10.1109/MSP.2015.2503881.

3. Marjan, A.; Jennifer, R.; Uwe, K.; Annica, K.; Eva, K.L.B.; Nicolas, T.; Thiemo, V.; Amy, L. An Ontology-based Context-aware System for Smart Homes: E-care@home. *Sensors* **2017**, *17*, doi:10.3390/s17071586.

4. Ashibani, Y.; Kauling, D.; Mahmoud, Q.H. A context-aware authentication framework for smart homes. *CCECE 2017* **2017**. doi:10.1109/CCECE.2017.7946657.

5. Deeba, K.; Saravanaguru, R.K. Context-Aware Healthcare System Based on IoT - Smart Home Caregivers System (SHCS). *ICICCS 2018* **2018**. doi:10.1109/iccons.2018.8663211.

6. Joo, S.C.; Jeong, C.W.; Park, S.J. Context Based Dynamic Security Service for Healthcare Adaptive Application in Home Environments. *Softw. Tech. Future Dependable Distrib. Sys.* **2009**. doi:10.1109/STFSSD.2009.14.

7. Sharpe, V.M. Issues and Challenges in Ubiquitous Computing. *Tech. Commun.* **2004**, *51*, 332–333.

8. Greenfield, A. *Everyware: The Dawning Age of Ubiquitous Computing*; Peachpit Press, Berkeley, CA, USA, 2006.

9. Chen, Y.H.; Tsai, M.J.; Fu, L.C.; Chen, C.H.; Wu, C.L.; Zeng, Y.C. Monitoring elder's living activity using ambient and body sensor network in smart home. In Proceedings of 2015 IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, China, 9–12 October 2015; pp. 2962–2967.

10. Sprint, G.; Cook, D.; Fritz, R.; Schmitter-Edgecombe, M. Detecting health and behavior change by analyzing smart home sensor data. In Proceedings of 2016 IEEE International Conference on Smart Computing (SMARTCOMP), St Louis, MO, USA, 18–20 May 2016; pp. 1–3.

11. Cook, D.J.; Schmitter-Edgecombe, M.; Dawadi, P. Analyzing activity behavior and movement in a naturalistic environment using smart home techniques. *IEEE J. Biomed. Health Informat.* **2015**, *19*, 1882–1892.

12. Bergeron, F.; Bouchard, K.; Gaboury, S.; Giroux, S.; Bouchard, B. Indoor positioning system for smart homes based on decision trees and passive RFID. In Proceedings of 20th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Auckland, New Zealand, 19–22 April 2016; pp. 42–53.

13. Yadav, U.; Verma, S.; Xaxa, D.K.; Mahobiya, C. A deep learning based character recognition system from multimedia document. In Proceedings of 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 21–22 April 2017; pp. 1–7.

14. Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H.G.; Ogata, T. Audio-visual speech recognition using deep learning. *Appl. Intell.* **2015**, *42*, 722–737.

15. Saito, S.; Wei, L.; Hu, L.; Nagano, K.; Li, H. Photorealistic facial texture inference using deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5144–5153.

16. Li, R.; Si, D.; Zeng, T.; Ji, S.; He, J. Deep convolutional neural networks for detecting secondary structures in protein density maps from cryo-electron microscopy. In Proceedings of 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 41–46.

17. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep Learning for Sensor-based Activity Recognition: A Survey. *Pattern Recog. Lett.* **2019**, *119*, 3–11.

18. Brenon, A.; Portet, F.; Vacher, M. Context Feature Learning through Deep Learning for Adaptive Context-Aware Decision Making in the Home. In Proceedings of The 14th International Conference on Intelligent Environments, Rome, Italy, 25–28 June 2018. doi:10.1109/IE.2018.00013.

19. Microsoft Azure. Object detection - Computer Vision - Azure Cognitive Services | Microsoft Docs. Available online: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-object-detection (accessed on 9 January 2020).

20. IBM Cloud. Speech to Text - IBM Cloud API Docs. Available online: https://cloud.ibm.com/apidocs/speech-to-text/speech-to-text (accessed on 9 January 2020).

21. Google Cloud. Cloud Natural Language API documentation. Available online: https://cloud.google.com/natural-language/docs/ (accessed on 9 January 2020).

22. Tadejko, P. Cloud Cognitive Services Based on Machine Learning Methods in Architecture of Modern Knowledge Management Solutions. In *Data-Centric Business and Applications*; Springer: Cham, Switzerland, 2020; pp. 169–190.

23. Shirsat, S.; Naik, A.; Tamse, D.; Yadav, J.; Shetgaonkar, P.; Aswale, S. Proposed System for Criminal Detection and Recognition on CCTV Data Using Cloud and Machine Learning. In Proceedings of the 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN). Vellore, India, 30–31 March 2019; pp. 1–6.

24. Mumbaikar, S.; Padiya, P. Web services based on soap and rest principles. *Inter. J. Sci. Res. Publ.* **2013**, *3*, 1–4.

25. Crockford, D. Introducing JSON. https://www.json.org/json-en.html (accessed on 9 January 2020).

26. Fu, J.; Mei, T. Image Tagging with Deep Learning: Fine-Grained Visual Analysis. In *Big Data Analytics for Large-Scale Multimedia Search*; Wiley: Hoboken, NJ, USA, 2019, p. 269.

27. Jocic, M.; Obradovic, D.; Malbasa, V.; Konjo, Z. Image tagging with an ensemble of deep convolutional neural networks. In Proceedings of the 2017 International Conference on Information Society and Technology, ICIST Workshops, Beijing, China, 17–20 September 2017; pp. 13–17.

28. Nguyen, H.T.; Wistuba, M.; Grabocka, J.; Drumond, L.R.; Schmidt-Thieme, L. Personalized deep learning for tag recommendation. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2017, pp. 186–197.

29. Chen, S.; Saiki, S.; Nakamura, M. Towards Affordable and Practical Home Context Recognition: - Framework and Implementation with Image-based Cognitive API -. *International Journal of Networked and Distributed Computing(IJNDC)* **2019**, *8*, 16 – 24. https://doi.org/10.2991/ijndc.k.191118.001.

30. Kalech, M.; Kraus, S.; Kaminka, G.A.; Goldman, C.V. Practical voting rules with partial information. *Autonomous Agents and Multi-Agent Systems* **2011**, *22*, 151–182.

31. Umamaheswararao, B.; Seetharamaiah, P.; Phanikumar, S. An Incorporated Voting Strategy on Majority and Score-based Fuzzy Voting Algorithms for Safety-Critical Systems. *International Journal of Computer Applications* **2014**, *98*.

32. Filos-Ratsikas, A.; Miltersen, P.B. Truthful approximations to range voting. International Conference on Web and Internet Economics. Springer, 2014, pp. 175–188.

33. Cao, L.; Chua, K.S.; Chong, W.; Lee, H.; Gu, Q. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* **2003**, *55*, 321–336.

34. Mousas, C.; Anagnostopoulos, C.N. Learning Motion Features for Example-Based Finger Motion Estimation for Virtual Characters. *3D Research* **2017**, *8*, 25.

35. Abdel-Hamid, O.; Mohamed, A.r.; Jiang, H.; Penn, G. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP). IEEE, 2012, pp. 4277–4280.

36. Bilmes, J.A.; Bartels, C. Graphical model architectures for speech recognition. *IEEE signal processing magazine* **2005**, *22*, 89–100.

37. Mousas, C.; Newbury, P.; Anagnostopoulos, C.N. Evaluating the covariance matrix constraints for data-driven statistical human motion reconstruction. Proceedings of the 30th Spring Conference on Computer Graphics. ACM, 2014, pp. 99–106.

38. Chéron, G.; Laptev, I.; Schmid, C. P-cnn: Pose-based cnn features for action recognition. Proceedings of the IEEE international conference on computer vision, 2015, pp. 3218–3226.

39. Li, Z.; Zhou, Y.; Xiao, S.; He, C.; Li, H. Auto-conditioned lstm network for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* **2017**, *3*.

40. Rekabdar, B.; Mousas, C. Dilated Convolutional Neural Network for Predicting Driver's Activity. In Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3245–3250.

41. Nakamura, S.; Hiromori, A.; Yamaguchi, H.; Higashino, T.; Yamaguchi, Y.; Shimoda, Y. Activity Sensing, Analysis and Recommendation in Smarthouse. *Multimedia, Distributed Collaboration and Mobile Symposium 2014 Proceedings* **2014**, *2014*, 1557–1566.

42. Ueda, K.; Tamai, M.; Yasumoto, K. A System for Daily Living Activities Recognition Based on Multiple Sensing Data in a Smart Home. *Multimedia, Distributed Collaboration and Mobile Symposium Proceedings* **2014**, *2014*, 1884–1891.

43. Sevrin, L.; Noury, N.; Abouchi, N.; Jumel, F.; Massot, B.; Saraydaryan, J. Characterization of a multi-user indoor positioning system based on low cost depth vision (Kinect) for monitoring human activity in a smart home. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 5003–5007.

44. Dobhal, T.; Shitole, V.; Thomas, G.; Navada, G. Human activity recognition using binary motion image and deep learning. *Procedia Comput. Sci.* **2015**, *58*, 178–185.

45. Asadi-Aghbolaghi, M.; Clapes, A.; Bellantonio, M.; Escalante, H.J.; Ponce-López, V.; Baró, X.; Guyon, I.; Kasaei, S.; Escalera, S. A survey on deep learning based approaches for action and gesture recognition in image sequences. In Proceedings of the 12th IEEE international conference on automatic face & gesture recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 476–483.

46. Pham, M.; Mengistu, Y.; Do, H.; Sheng, W. Delivering home healthcare through a cloud-based smart home environment (CoSHE). *Future Gener. Comput. Syst.* **2018**, *81*, 129–140.

47. Menicatti, R.; Sgorbissa, A. A Cloud-Based Scene Recognition Framework for In-Home Assistive Robots. *RO-MAN 2017* **2017**. doi:10.1109/ROMAN.2017.8172472.

48. Menicatti, R.; Bruno, B.; Sgorbissa, A. Modelling the Influence of Cultural Information on Vision-Based Human Home Activity Recognition. *CoRR* **2018**, *abs/1803.07915*, [1803.07915].

49. Qin, Z.; Weng, J.; Cui, Y.; Ren, K. Privacy-preserving image processing in the cloud. *IEEE Cloud Comput.* **2018**, *5*, 48–57.

50. Dorri, A.; Kanhere, S.S.; Jurdak, R.; Gauravaram, P. Blockchain for IoT security and privacy: The case study of a smart home. In Proceedings of the IEEE international conference on pervasive computing and communications workshops (PerCom workshops), Big Island, HI, USA, 13–17 March 2017; pp. 618–623.

51. Geneiatakis, D.; Kounelis, I.; Neisse, R.; Nai-Fovino, I.; Steri, G.; Baldini, G. Security and privacy issues for an IoT based smart home. In Proceedings of the 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 22–26 May 2017; pp. 1292–1297.

52. Chen, S.; Saiki, S.; Nakamura, M. Evaluating Feasibility of Image-Based Cognitive APIs for Home Context Sensing. *ICSPIS 2018* **2018**, pp. 5–8. doi:10.1109/CSPIS.2018.8642772.

53. Maklin, C. TF IDF | TFIDF Python Example - Towards Data Science. https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76. (accessed on 27 November 2019).

54. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, **2013**.

55. Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, June 21–26 June 2014; pp. II–1188–II–1196.

56. Polito, M.; Perona, P. Grouping and dimensionality reduction by locally linear embedding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; pp. 1255–1262.

57. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Tech.* **2004**, *38*, 188–230.

58. Tenenbaum, J.B.; De Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323.

59. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326.

60. Belkin, M.; Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **2003**, *15*, 1373–1396.

61. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of The 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.

62. Nam, J.; Herrera, J.; Slaney, M.; Smith, J.O. Learning Sparse Feature Representations for Music Annotation and Retrieval. In Proceedings of ISMIR 2012, Porto, Portugal, 8–12 October 2012; pp. 565–570.

63. Yamato, J.; Ohya, J.; Ishii, K. Recognizing human action in time-sequential images using hidden markov model. In Proceedings of 1992 IEEE Computer Society conference on computer vision and pattern recognition, Champaign, IL, USA, 15–18 June 1992; pp. 379–385.

64. Chun, H.; Lee, M.H.; Fleet, J.C.; Oh, J.H. Graphical models via joint quantile regression with component selection. *J. Multivariate Anal.* **2016**, *152*, 162–171.

65. Microsoft. Computer Vision | Microsoft Azure. Available online: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/ (accessed on 27 November 2019).

66. IBM. Watson Visual Recognition. Available online: https://www.ibm.com/watson/services/visual-recognition/ (accessed on 27 November 2019).

67. Clarifai. TRANSFORMING ENTERPRISES WITH COMPUTER VISION AI. Available online: https://clarifai.com/ (accessed on 15 April 2019).

68. Imagga. Imagga API. Available online: https://docs.imagga.com/ (accessed on 15 April 2019).

69. ParallelDots. Image Recognition. Available online: https://www.paralleldots.com/object-recognizer (accessed on 15 April 2019).

70. Microsoft. Azure Machine Learning | Microsoft Azure. Available online: https://azure.microsoft.com/en-us/services/machine-learning/ (accessed on 27 November 2019).