修　　士　　論　　文

Evaluation of Mother-Child Interaction using Encoder-Decoder-Based Behavior Analysis

(エンコーダ・デコーダモデルを用いた
振る舞い解析による母子の関わり方評価)

指導教員　山﨑 俊彦　准教授

東京大学大学院
情報理工学系研究科
電子情報学専攻

氏 名　48-186460　林　遠

提 出 日　2020 年 1 月

Abstract

Childcare in early age would bring deep influence on children' s personality. For infants, their early education are most based on the interaction with mothers. ALthough there have already been some studies on analyzing the interaction of children and mothers in early-age, these researches are almost manual and therefore in small scale. With the great process of behavior analysis on computer version field, we apply behavior analysis to childcare researches in purpose to make large scale research possible. We built a dataset from large amount of video of mothers taking care of 15 months old children, all of which are annotated with childcare level labeled by professional analysts. We conducted experiments by our encoder-decoder model to rate childcare level automatically, and achieved acceptably high accuracy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Children's care and education are eternally important elements of the society. For infants, their early education are most based on the interaction with mothers. And this early-age education is critical in infants' character formation. There have already been some studies on analyzing the interaction of children and mothers in early-age.

Cohn Jeffrey F and Tronick Edward [1] found quality of mother's affective expression accounted for individual differences in the behavior of thirteen 7-month-old infants living in multiproblem families. Especially, withdrawn or intrusive maternal affective expression, together with lacking of contingent responsiveness, may in part be responsible for the risk-status of infants. Purhonen Maija's group [2] test the reaction of 4-month-old infants when hearing voice from their mothers and other unfamiliar female. As a result, they found the behaviorally well-documented mutual sensitization between infant and mother and the special importance of output from mother is seen as an enhanced arousal to mother's voice and as signs of a clear memory template for own mother's voice at very early age. Field Tiffany M [3] focused on the reaction of infants of mothers with depressed postpartum. They videotaped and analyzed 24 depressed and 24 non-depressed mothers in 3 face-to-face interactions with their 3-month-old infants. As a result, they found that infants are able to detect the affective qualities of their mothers' displays and appropriately modify their affective displays in response. Findings also suggest that depression or depressed affect emerges in infants as a function of early

interactions with their depressed mothers.

In Japan, there are also lots of researches about early-age child behavior analysis. Some studies used handheld cameras to record children's daily behavior [4]. Some studies used stationary cameras [5] but still requires significant effort for manual analysis. Additionally, some studies considered reducing the workload of manual analysis by attaching motion sensors to children and tracking their actions [6]. However, children may be uncomfortable with the motion sensors and not behave as usual. Besides, such sensor systems always suffer from safety problems and battery issues.

So far these behavior analysis of mother-child interaction is manual and therefore in small scale, in which only dozens of mother-child participate in experiments. Since these experiments are almost recorded in videos and the methods of behavior analysis in computer vision field has made great progress in this decade, we conducted behavior analysis on a new database of mother-child interaction automatically using computer vision technologies and achieved acceptably results.

1.2 Purpose

We acquired a large amount of videos of mothers taking care of 15-month-old children from The University of California, Davis. All of the videos are annotated with childcare level labeled by professional analysts. However, since rating the level of childcare requires significant manual effort for analysis, it is expected that the techniques of video processing to this rating task.

We built a dataset from the childcare videos, and proposed an encoder-decoder model to rate childcare level automatically. As a result, we achieved acceptably accuracy using our encoder-decoder model.

1.3 Organization of This Thesis

The organization of this thesis is as follows.

In Chapter 1, we introduce the importance of mother-child interaction analysis and the necessity of importing technology of computer vision into this field. In Chapter 2, we introduce the related work of behavior analysis. In Chapter 3, we introduce the dataset we built from childcare videos. In Chapter 4, we introduce the experiments based on image processing encoder. Finally, we draw conclusion and discuss future work in 5.

In Appendix A, we introduce another work about IoT sensor application on real estate evaluation.

Chapter 2

Related Work

In this section, we will introduce the related works about behavior analysis.

First, we will introduce some behavior analysis researches based on sensor system, which can be generally divided into wearable sensor system and non-attached sensor.

Next, we introduce image-based behavior analysis, which is the main method in current research. Image-based methods can be divided into traditional methods and deep learning methods. Traditional methods include space-time interest points [7] and dense trajectories [8, 9]. Meanwhile, deep learning methods contain spatio-temporal convolution [10], recurrent neural network (RNN) [11], and two-steam architectures [12].

Finally, we introduce skeleton-based behavior analysis. Also, we will introduce the related works about the method of extract skeleton from images or videos. In this subsection, posture estimation and object tracking are discussed.

2.1 Sensor-Based Behavior Analysis

2.1.1 Wearable Sensors

Before the dramatical improvement on computer vision, applying wearable sensors to analyze behavior is widely used in many applications such as medical, entertainment, security, and commercial fields. Wearable sensors can record the motion of the whole body or certain parts continuously and precisely [13, 14]. Therefore, it is believed that the smart wearable sensors will revolutionize human lifestyles and

social interactions in decades. In laboratory settings, the most prevalent everyday activities have been successfully recognized with accelerometers [15, 16, 17, 18, 19]. As for out-of-laboratory settings, Miikka Ermes et al. recognized behaviors by using a hybrid classifier. The hybrid classifier combined a tree structure containing a priori knowledge and artificial neural networks, and also by using three reference classifiers [20]. Also, Pietro Salvo et al. proposed a sweat monitoring sensor on the textile substrate, which made sensor system directly worn on the body possible [21].

2.1.2 Non-attached Sensors

On the other hand, non-attached sensors systems are also developed in behavior analysis, especially on child and elder care and security. Sensors like Microsoft Kinect sensor [22], Leap Motion [23], body mounted camera [24], 3D laser scanner [25] and infrared light source [26] can capture human poses and construct a human skeleton based on the captured body joints. Most of the sensors can record the deep information at same time. For example, N. Noury et al. [27] proposed a system for remotely monitoring human behavior in daily life at home aiming to improve safety and quality of life. Activity is monitored by infrared position sensors and magnetic switches. For fall detection, they had developed a smart sensor. Local communication was performed using RF wireless links to reduce wiring and allowed personnel to move. And for behavior analysis on sports, in considerate of reducing the affect from attached sensors, non-attached sensors are generally recommended. Per Wilhelm et al. [28] combined video tracking and wireless sensor and built a Sport Performance Analyzer (SPA) system that has three main modules, namely data acquisition, tracking and analysis-visualization.

2.2 Image-Based Behavior Analysis

2.2.1 Traditional Method

Space-time Interest Points

Ivan Laptev [7] pointed out that the key points in video images were usually the data which changes strongly in the space-time dimension, and these data reflect important information about the target movement. By extracting these change data and further analyzing their location information, the data can be used to distinguish other actions.

In the spatial domain, points with a significant local variation of image intensities have been extensively investigated [29, 30, 31, 32]. In Ivan Laptev's work, he extended Harris [30] and Förstner [29] interest point operators and detected local structures in space-time where the image values have significant local variations in both space and time. After obtaining the key points, based on the one to four partial derivatives of the points, they combined a 34-dimensional feature vector and clustered them using k-means.

Paul Scovanner et al. [33] used 3-dimensional SIFT instead of Harris interesting points to recognize behaviors in video. Every key point of 3D SIFT contains 3 values, amplitude and two angles. The histogram of the gradient around the key points in space and time can be used to form feature descriptors, and then k-means clustering is performed on all feature descriptors to divide the categories into vocabulary "word". All the different words constitute a vocabulary, and each video can be described by the number of words appearing in this vocabulary. Finally, an SVM [34] or perceptron are trained for action recognition.

Dense Trajectories

The space-time interest points are to encode the video information in spatio-temporal coordinates, and iDT (improved Dense Trajectories) [8, 9] is another very classic method that tracks the changes of the image along a given time along the coordinate.

The iDT algorithm consists of three steps: (1) densely sampling feature points, (2) feature trajectory tracking, and (3) trajectory-based feature extraction.

Dense sampling is the regular sampling of images at different scales, but not all points are really used for tracking, because the points in the smooth area have no tracking significance. By calculating the feature value of the autocorrelation matrix of each pixel and setting the threshold value feature points below the threshold are used to implement this selection.

The feature trajectory is tracked by optical flow. The optical flow rate of the image is calculated first, and then the image motion trajectory is described by this rate. Because the trajectory drifts over time, it may move far from the initial position. Therefore, they limited the trajectory tracking distance spatially and practically. If the tracked point is out of range, resampling and tracking are performed to ensure that the trajectory density will not be reduced. In this way they characterized the shape of behavior.

In addition to the shape, they also extracted features aligned with the trajectories to characterize appearance (histograms of oriented gradients) and motion (histograms of optical flow). Furthermore, they introduced a descriptor based on motion boundary histograms (MBH) which rely on differential optical flow. The MBH descriptor showed to consistently outperform other state-of-the-art descriptors, in particular on real-world videos that contain a significant amount of camera motion.

2.2.2 Deep learning Method

Before the application of Convolutional Neural Networks (CNNs) [35, 36, 37, 38], researches on behavior analysis focus on extracting hand-crafted features to recognize actions. However, with the rapid development of CNNs over the past decade, the hand-crafted features are gradually replaced.

The successful image classification architectures have been adapted to video processing in three ways: (1) with spatio-temporal convolutions or (2) with recurrent neural network, (3) by processing multiple streams such as motion representation in addition to RGB data.

Spatio-temporal Convolutions

Learning image representation using ConvNets by pre-training on ImageNet has proven useful in many visual understanding tasks including object detection, semantic segmentation, and image captioning. Although any image representation can be applied to video frames, a dedicated spatio-temporal representation is still important in order to incorporate motion patterns that can only be captured by appearance-based models.

Du Tran et al. [10, 39] proposed a simple, effective approach for spatio-temporal feature learning using deep 3-dimensional convolutional networks (3D ConvNets). Their finding has two points: (1) Compared to 2D ConvNets, 3D ConvNets are more suitable for spatio-temporal feature learning ; (2) A homogeneous architecture with small 3x3x3 convolution kernels in all layers is among the best performing architectures for 3D ConvNets. As a result, the feature they learned from C3D (Convolutional 3D), with a simple linear classifier, outperformed state-of-the-art methods in the year it proposed.

However, although spatio-temporal convolutions supply a easy way to combine spatial information and time information, the problem is the cost of space and time. Train a 3D CNN for a certain database would occupy lot of computing space and take large amount of time. Also, since 3D CNN regards spatial feature and time feature equally, it may not perform well in some specific tasks.

Recurrent Neural Networks

Recurrent neural networks (RNNs) have been explored in perceptual applications for decades, and the results were different. Although RNNs have proven successfully on tasks such as speech recognition [40] and text generation [41], it may be difficult to train them to learn long-term dynamics. That is because of the vanishing and exploding gradients problem [42] that can result from propagating the gradients down through the many layers of the recurrent network, each corresponding to a particular timestep.

To solve the vanishing and exploding gradients problem, Jeff Donahue et al. [11] developed Long Short-Term Memory (LSTM), a novel recurrent convolutional

architecture suitable for large-scale visual learning which is end-to-end trainable. Compared with RNNs, LSTM adds cell state and imports severals gates. This cell state carries the information of all previous states. At every new state, there are corresponding operations to decide what old information to discard and what new information to add. This state is different from the hidden layer state h. During the update process, its update is slow, while the hidden layer state h is updated quickly.

Two-stream Architectures

The two-stream method trains two independent CNNs, one using RGB data to operate on appearance, and the other is based on optical flow image processing.

The two-stream method has shown encouraging results in different video understanding tasks like video classification [43, 12, 43], video segmentation [44, 45] and action localization [46, 47]. In this case, the two classification streams are independently trained and combined during testing. The first one operates on the appearance by using RGB data as input. The second one is based on the motion, taking the optical flow calculated using off-the-shelf methods as input [48, 49], converting it to an image and stacking it over several frames. Feichtenhofer et al. [43] trained the two streams end-to-end by fusing different levels of streams instead of training them independently. The I3D method [50] also relies on the two-streaming method. This architecture processes video clips through spatio-temporal convolution and pooling operators, enriches them from an image classification network with spatial convolution and pooling layers.

2.3 Skeleton-Based Behavior Analysis

2.3.1 Posture Estimation

Human pose estimation is the process of estimating the 2D or 3D human body part positions from still images or videos. But generally, human pose estimation aims to detect 2D key-points linked to joints. Early work used robust image

features and sophisticated structured prediction: the former is used to produce local interpretations, whereas the latter is used to infer a globally consistent pose. This conventional pipeline, however, has been greatly reshaped by convolutional neural networks.

Single-Person Pose Estimation

The work by Newell et al. [51] introduces a novel "stacked hourglass" network design for predicting human pose. The network captures and consolidates information across all scales of the image. Recently, many work [52, 53, 54, 55] based on hourglass made great score on human pose estimation.

Yang et al. [53] argued that the residual unit in hourglass network can only capture visual patterns or semantics at one scale. Therefore, in their work, they used the proposed pyramid residual module as the building block for capturing multi-scale visual patterns or semantics. They also use the PRM at the beginning convolutional and max pooling layers, which are used to process features down to a very low resolution before Hourglass Module. Furthermore, they introduce score maps of body joint locations to produce at the end of each hourglass, and a squared-error loss is also attached in each stack of hourglass.

The work by Ke et al. [54] was also an improvement on Hourglass Network. Their framework, Multi-Scale Structure-Aware Network focus on the loss of Hourglass Network. Their work has three key points: (1) They propose the multi-scale supervision network (MSS-net) to learn deep features across multiple scales. (2) They use a fully convolutional multi-scale regression network (MSR-net) after the MSS-net convdeconv stacks to globally refine the multi-scale keypoint heatmaps to improve the structural consistency of the estimated poses. (3) they design a structure-aware loss function following a graph to model the human skeletal structure. Specifically, they introduced a human skeletal graph S for a visualization of the human skeletal graph to define the structure-aware loss.

Multi-Person Pose Estimation

As for multi-person pose estimation, there are mainly two kinds of approaches, top-down approach and bottom-up approach.

The top-down approach is also known as two-step framework, is to detect the multiple people first, get the bounding box, and then detect the key points of the human body in each bounding box, finally connect them into a human skeleton. It can be considered as a combination of multi-person detection and single-person pose estimation. The most difficulty of top-down approach is that the impact of the boundary box is too large, factors such as misalignment, Intersection over Union (IOU) size will significantly affect the results.

Bottom-up approach is also known as part-based framework. This approach detects all the joint points first, and then determine the attribution of each joint point. The main difficulty of bottom-up approach is splicing different parts of different people by one person. Newell et al. [56] integrate associative embedding with a stacked hourglass network, which produces a detection heatmap and a tagging heatmap for each body joint, and then group body joints with similar tags into individual people. Openpose [57] is the first open-source realtime system for multi-person 2D pose detection, including body, foot, hand, and facial keypoints. Openpose uses a nonparametric representation, which referred as Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. This bottom-up system achieves high accuracy and realtime performance, regardless of the number of people in the image.

Although almost all of approaches to estimate human poses was based on joint detection, the work by Guler et al. [58] aims at pushing further the envelope of human understanding in images by establishing dense correspondences from a 2D image to a 3D, surface-based representation of the human body.

2.3.2 Object Tracking

After detecting all skeleton from the videos, we need to attach IDs to these skeletons in purpose to divide them into different people. Applying object tracking,

we can track the skeletons across video. In this way, we can get the behavior of different people in one video.

In recent years, approaches to object tracking have been based on discriminative correlation filters (DCFs) or convolutional neural networks (CNN). Examples of DCF based approaches include minimum output sum of squared error (MOSSE) [59] and accurate scale estimation for robust visual tracking (DSST) [60]. CNN based approaches include TCNN [61], MDNet [62], and SANet [63]. Furthermore, there are approaches that employ both DCF and CNN such as C-COT [64] and ECO [65]. Although CNN based approaches can obtain high accuracy while sacrificing processing time, DCF based approaches can process quickly with relatively lower accuracy.

2.3.3 Action Recognition

After obtaining the skeletons with ID attached, we can finally use the skeletons to do behavior analysis.

There are lots of literature on behavior analysis from 3D skeleton data [66, 67, 68]. Most of these methods train a recurrent neural networks on the coordinates of human joints. However, this requires knowing the 3D coordinates of each joint of the actor in each frame. This does not apply to videos in the wild, which include occlusion, truncation, and multiple human actors.

The first attempts to use 2D poses were based on hand-crafted features [69, 70, 71]. Jhuang et al. [69] encoded the relative position and movement of joints relative to the center and scale of the human body. Wang et al. [70] proposed to group joints on body parts (such as the left arm) and use a bag-of-words to represent a series of poses. Xiaohan et al. [71] use a similar strategy to utilize the hierarchy of parts of the human body.

Several approaches have been proposed to use pose to guide CNNs. Most of them use joints to pool features [72, 73] or to define attention mechanisms [74, 75]. Chéron et al. [73] applied CNNs trained on patches around artificial joints. Cao et al. [72] pooled features according to joint positions. Du et al. [74] combined an end-to-end recurrent networks with a pose-attention mechanisms for behavior analysis.

Their methods requires pose keypoint supervision in training videos. Girdhar and Ramanan [75] proposed an attention module with a low-order second-order pooling method and presented that intermediate supervision based on estimated poses is helpful for behavior analysis.

Chapter 3

Dataset

The childcare level evaluation experiment is conducted in every mother-child' house across America during 1990s. All the children in video are 15-month-old and normally grown with ability of walking and making noise. In the video, mothers were asked to take care of their child for 15 minutes using three tools one by one. The tools were provided by organizers and included one book and two toys as shown in Figure 3.1.

Although most of the videos only contained the interaction between mother and child, because of the curiosity and vivacity of child, the observer might be recorded in some videos when the child run around. Also, the observer might appear in the beginning of some videos introducing the rules of the experiment to mothers.

Further more, although mothers were told to take care of their child using the tools provided, children of 15-month-old were unavoidably attracted by other thing. For example, in Figure 3.1a, the child was attracted by the house audio equipment.

In some videos like Figure 3.1b and Figure 3.1c, the timestamp was remained at bottom left corner.

Totally, we received 1252 videos of childcare, and 1116 of them are attached with childcare labels. All the videos are taken by hand-held cameras with image size of 640*480 and fps (frames per second) of 29.97. Although mothers are told to present childcare in 15 minutes, the length of videos varies from 9 minutes to 30 minutes.

(A) (B)

(C) (D)

FIGURE 3.1: Samples of childcare videos

3.1 Label

Every video is attached to 6 labels, 4 of them are labeled by professional analysts, and the rest 2 are calculated from the 4 existing labels. The 6 labels are SENSIO15, F15PQ2, F15PQ3, F15PQ6, SEN_HML and Intru_R. The description of these 6 labels is shown in Table 3.1.

F15PQ2, F15PQ3, F15PQ6 and Intru_R are labeled as integer varies from 1 to 4. SEN_MHL is labeled as integer varies from 1 to 3. SENSIO15, as the sum of F15PQ2, F15PQ6 and Intru_R, is labeled as integer varies from 3 to 12. Except F15PQ3, the higher score in all the labels presents the higher childcare level.

These labels are also applied in other childcare researches. Some researches compared SEN_HML of women who never reported symptoms of depression with those who reported symptoms sometimes or chronically [76, 77]. They found

TABLE 3.1: The Description of Labels

Labels	Description
SENSIO15	Maternal composite sensitivity score at 15 months (=F15PQ2+F15PQ6+Intru_R)
F15PQ2	Maternal sensitivity to non-distress
F15PQ3	Maternal intrusiveness
F15PQ6	Maternal positive regard for the child
SEN_HML	Maternal sensitivity
Intru_R	Maternal intrusiveness reverse coded (=5.0 - F15PQ3)

women with chronic symptoms of depression were the least sensitive when observed playing with their children from infancy through 36 months. Children whose mothers reported feeling depressed performed more poorly on measures of cognitive-linguistic functioning and were rated as less cooperative and more problematic at 36 months. Belsky Jay and Fearon RM Pasco [78] focused on the role of early experience in shaping development, and examined the hypothesis that the most competent 3-year-olds would be those with histories of secure attachment (at 15 months) who subsequently experienced (relatively) high-sensitive mothering (at 24 months), and that the least competent children would be those with histories of insecure attachment who subsequently experienced (relatively) low-sensitive mothering.

3.2 Distribution and Correlation

Figure 3.2 shows the distribution of all labels (except F15PQ3, because F15PQ3 = 5.0 - Intru_R). And Table 3.2 shows the correlation of all labels (except F15PQ3 and SEN_HML).

The distribution shows the number of worst performance in every label is extremely small. By viewing the videos, we find even the mother labeled with low level performs well in childcare from our point of view. It is speculated that since the mothers were told to take care of their children for a period of time and knew that their behaviors were recorded, they would perform their best to take care of

TABLE 3.2: The Correlation of Labels

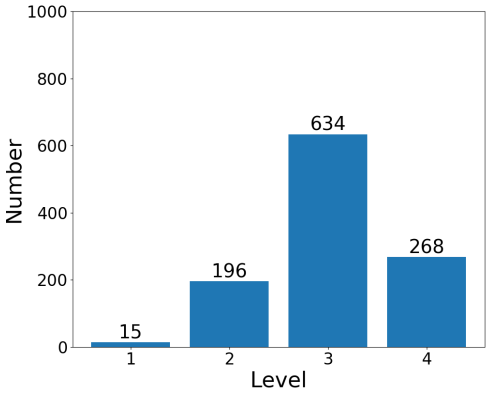
	F15PQ2	Intru_R	F15PQ6	SENSIO15
F15PQ2	1.00	0.57	0.49	0.86
Intru_R	0.57	1.00	0.28	0.79
F15PQ6	0.49	0.28	1.00	0.73
SENSIO15	0.86	0.79	0.73	1.00

their child. Thus, the childcare level of most mother in all labels are in middle rate or high rate.

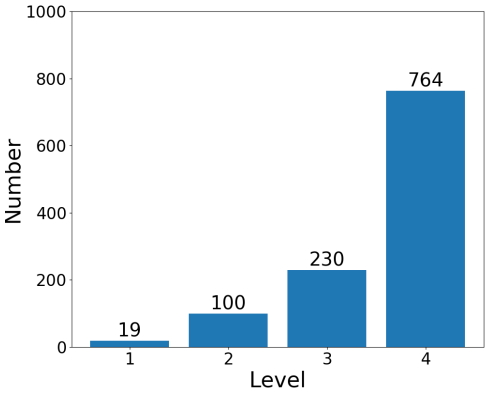
The correlation results shows that F15PQ2 is moderately related to Intru_R and F15PQ6, but correlation between Intru_R and F15PQ6 is low. SENSIO15, because it is the sum of F15PQ2, Intru_R and F15PQ6, is highly related with these three labels. Therefore, although SENSIO15 can present the integral level of childcare, other labels still need experiments individually.

3.3 Small Scale Dataset

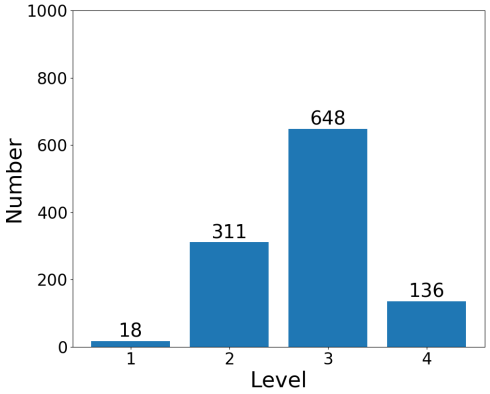
As described in Chapter 1.2, this work is a cooperation with The University of California, Davis. Since the videos are recorded in 1990s and stored in discs, we have not received all videos in the beginning. Actually, we only had 154 videos at first, and only 30 of them were labeled. Figure 3.3 shows the distribution of labels in the dataset. From the distribution, the number of every label is relatively balanced except label Intru_R, in which number of level 4 is extremely large compared with other levels.



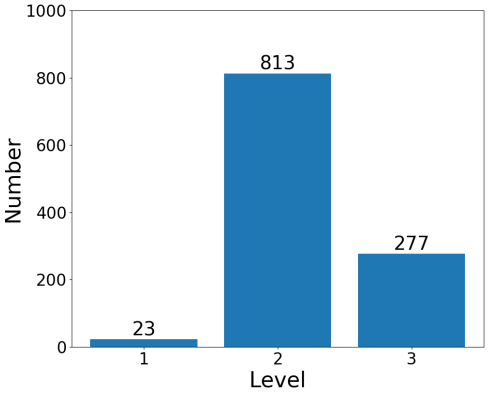
(A) F15PQ2



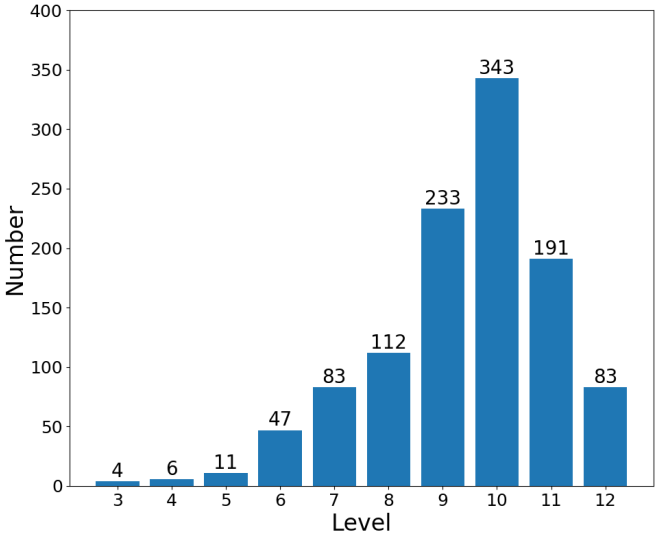
(B) Intru_R



(C) F15PQ6

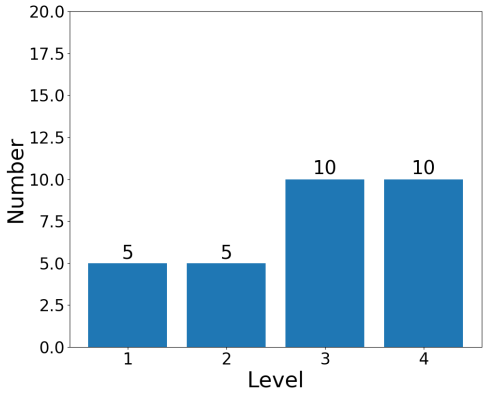


(D) SEN_HML

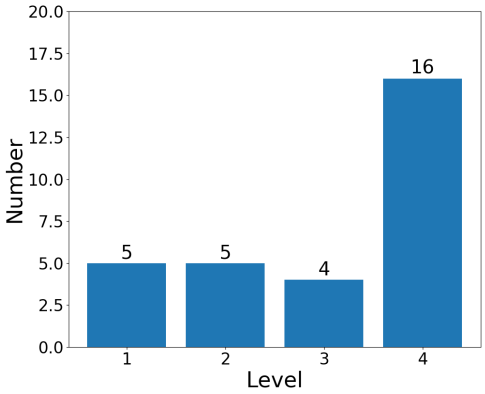


(E) SENSIO15

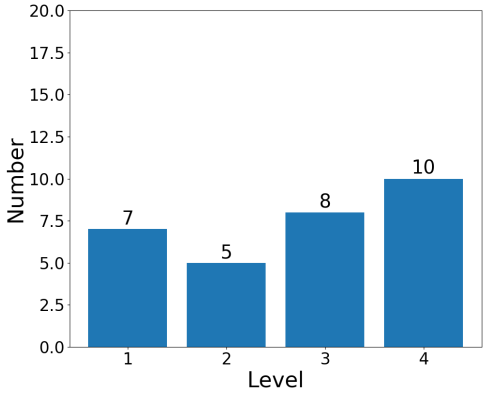
FIGURE 3.2: Distribution of Labels in the Whole Dataset



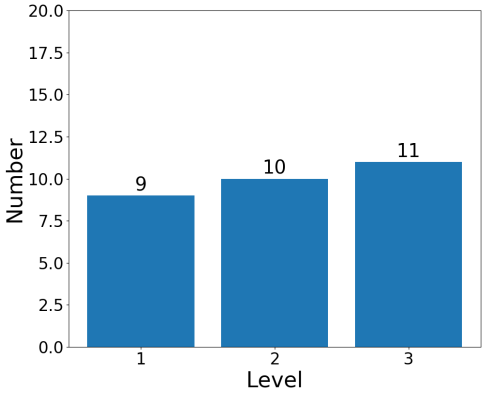
(A) F15PQ2



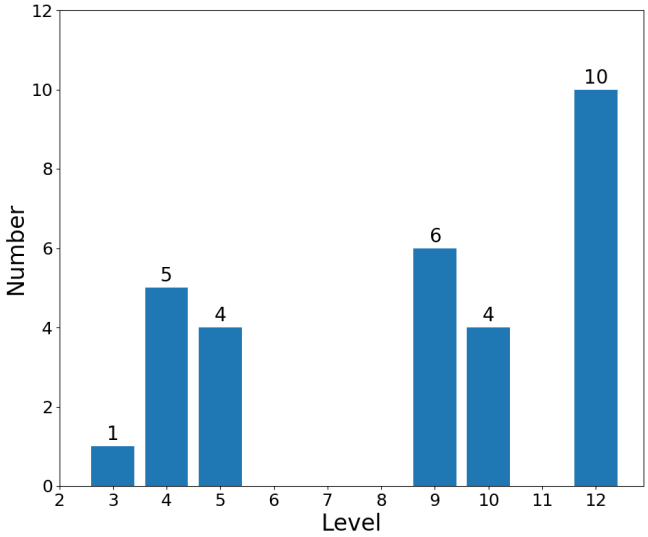
(B) Intru_R



(C) F15PQ6



(D) SEN_HML



(E) SENSIO15

FIGURE 3.3: Distribution of Labels in the Small Scale Dataset

Chapter 4

Encoder-Decoder-Based Behavior Analysis

4.1 Encoder-Decoder Model

Figure 4.1 shows the general structure of encoder-decoder model for behavior analysis from videos, c refers to feature extracted by encoder, and Y refers to the classification result. First, several frames are extracted from the whole video. Second, encoder is applied to extract feature of every frame. Finally, the decoder do classification with the features as input.

The widely used databases for behavior analysis include HMDB-51 [79], UCF-101 [80], Sports-1M [81], ActivityNet [82], Youtube-8M [83] and so on. However, the classes of these datasets vary different from each other. For example, HMDB-51 [79] has 51 classes including cartwheel, clap hands, climb, dive, fall on the floor, run, wave. The difference between classes is easy to distinguish and the class can be clearly divided. Similarly, UCF-101 [80] has 101 action categories varies in five types 1)Human-Object Interaction 2) Body-Motion Only 3) Human-Human Interaction 4) Playing Musical Instruments 5) Sports.

On the other hand, the childcare level in our dataset is hard to distinguish by non-professional analysts. Furthermore, since the length of videos varies from 9 minutes to 30 minutes and the action of childcare is limited, the simple feature extraction model for image classification may perform well in our case with little difference to feature extraction model for action recognition.

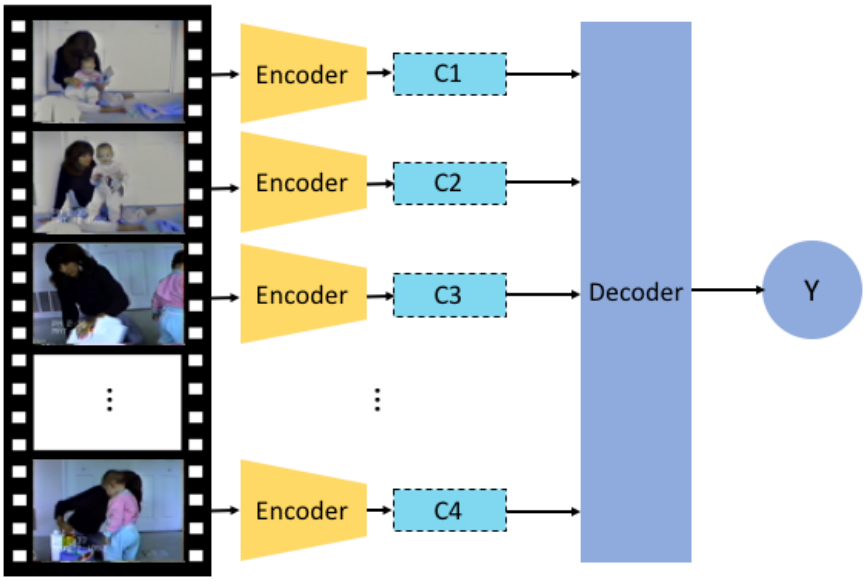


FIGURE 4.1: Encoder-Decoder Model for Behavior Analysis

Well-performed pretrained feature extraction model for image classification includes ResNet [35], VGG [37], AlexNet [84], SqueezeNet [85], DenseNet [86], GoogLeNet [38], MobileNet v2 [87], ResNeXt [88], Wide ResNet [89] and MNASNet [90]. We conducted experiment to evaluate the performance of every feature extraction model mentioned above on small scale dataset described in 3.3.

Empirically, we choose ResNet50 [35], VGG13 [37], AlexNet [84], and GoogLeNet [38] as encoder in our further experiment on Dataset II.

As for decoder, we conducted experiments to evaluate full connecting layer, LSTM [11] and attention LSTM. For full connecting layer, we test input of the mean value of all 450 features and the concatenation of all features. The structure of model with LSTM is showed as Figure 4.2. When using LSTM or attention LSTM as decoder, we use the final output (Yn) of LSTM or attention LSTM as the classification result. As a result, we selected LSTM in our further experiment on Dataset II.

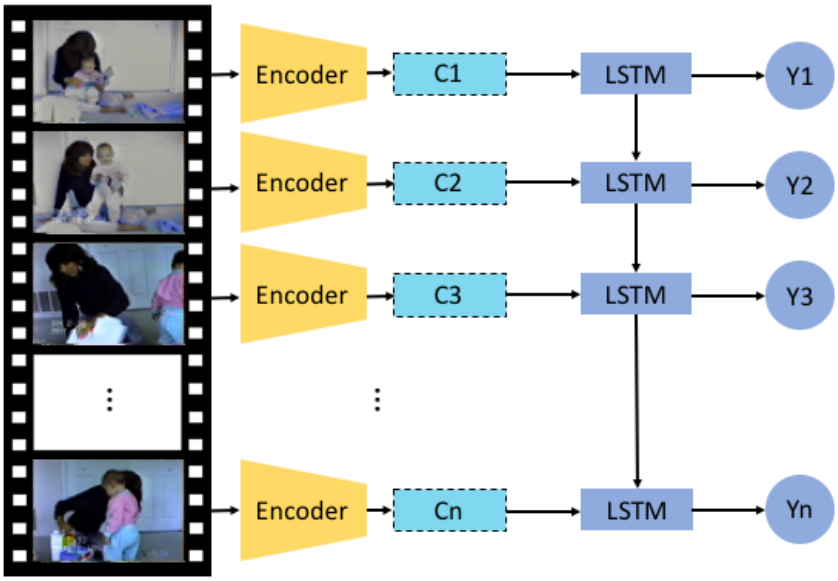


FIGURE 4.2: Encoder-Decoder Model with LSTM for Behavior Analysis

4.2 Experiment on Dataset I

4.2.1 Training Setting

Dataset I is consisted of 30 labeled videos as described in Chapter 3.3. To make full use of Dataset I, we divide the 30 labeled video into 6 groups to conduct cross validation. In every time of learning, one group is selected as test set and the rest 5 groups consist train set. Groups are manually divided in purpose of keeping the balance of every label in every group. Table 4.1 shows the label distribution in every group. According to the distribution of SENSIO15 in Dataset I 3.3e, we redefine the level of SENSIO15 with level 1 as original level < 8, level 2 as original level between 9 and 10, level 3 as original level $= 12$. Besides cross validation, since we only have 5 data for test in every time of learning, we trained the decoder model for 6 times to ensure universality of the results.

For every video, we extract one frame per second from 2 minutes since in some videos observer might appear at the beginning introducing the rules of the experiment to mothers. We extract 450 frames for every video according to the minimum length of videos. For every experiment, we set learning rate as 0.001

TABLE 4.1: The Label Distribution in Every Group

Label	Level	Number in Group					
		Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
F15PQ2	1	1	1	1	0	1	1
	2	1	1	1	2	0	0
	3	2	2	1	1	2	2
	4	1	1	2	2	2	2
Intru_R	1	0	1	1	2	1	0
	2	2	1	1	0	0	1
	3	1	1	1	1	0	0
	4	2	2	2	2	4	4
F15PQ6	1	2	1	1	1	1	1
	2	0	1	1	1	1	1
	3	2	2	1	1	1	1
	4	1	1	2	2	2	2
SEN_HML	1	1	2	2	2	1	1
	2	3	1	1	1	2	2
	3	1	2	2	2	2	2
SENSIO15	1	2	2	2	2	1	1
	2	2	2	1	1	2	2
	3	1	1	2	2	2	2

and number of train epochs as 200. The input images are in size of 256*192. We use Adam [91] as optimization and softmax cross entropy loss as loss function.

According to 3.2, we test all feature extraction model listed above on SENSIO15, which is the sum of F15PQ2, Intru_R and F15PQ6, and is highly related with these three labels. Level of SENSIO15 is divided into 3 classes as described in Chapter 3.3. Then we test 4 feature extraction model (ResNet50, VGG13, AlexNet and GoogLeNet) on all other labels because of their high accuracy in predicting SENSIO15. F15PQ2, Intru_R and F15PQ6 as divided into 4 classes while SEN_HML is divided into 3 classes according to their distribution.

When evaluating the performance of encoders, we use LSTM [11] as decoder. On the other hand, we use 4 encoders (ResNet50, VGG13, AlexNet, and GoogLeNet) to evaluate the performance of decoders.

TABLE 4.2: Experiment Result on Encoder on SENSIO15 in Small Dataset

Feature Extract Model	Accuracy
ResNet18	72.23%
ResNet34	75.0%
ResNet50	**75.56**%
ResNet101	75.56%
ResNet152	73.89%
VGG11	76.67%
VGG13	**77.78**%
VGG16	69.44%
VGG19	70.56%
AlexNet	**76.1**%
SqueezeNet	66.67%
DenseNet	66.11%
GoogLeNet	**81.11**%
MobileNet v2	71.67%
ResNeXt50	73.89%
ResNeXt101	73.33%
Wide ResNet	66.11%
MNASNet	72.22%

4.2.2 Result

Table 4.2 shows the result of experiment on all feature extraction model mentioned above on SENSIO15 in Dataset I. We find the performances of image classification models are already good in Dataset I with almost all accuracy reaches 70%. Especially, GoogLeNet acquires the highest accuracy over 80%. ResNet50, VGG13, and AlexNet also acquire high accuracy in predicting SENSIO15. Therefore, we conducted experiments on these 4 feature extraction models (ResNet50, VGG13, AlexNet and GoogLeNet) on all other labels in Dataset I. The result is shown in Table 4.3. We find GoogLeNet still performs best on predicting F15PQ2, Intru_R, and F15PQ6. However, AlexNet gets the best score on predicting SEN_HML. Also, the accuracy of predicting Intru_R using AlexNet is very close to it of GoogLeNet.

Figure 4.3 shows the result of experiments on decoder. Except SEN_HML, the best accuracy is acquired by LSTM using GoogLeNet as encoder. As for

TABLE 4.3: Experiment Result on Encoder on Other Labels in Small Dataset

Label	Feature Extract Model	Accuracy
F15PQ2	ResNet50	68.89%
	VGG13	68.37%
	AlexNet	66.67%
	GoogLeNet	**72.22**%
Intru_R	ResNet50	77.22%
	VGG13	74.44%
	AlexNet	79.44%
	GoogLeNet	**80.00**%
F15PQ6	ResNet50	62.22%
	VGG13	64.44%
	AlexNet	62.78%
	GoogLeNet	**74.44**%
SEN_HML	ResNet50	76.1%
	VGG13	80.56%
	AlexNet	**86.67**%
	GoogLeNet	81.67%

SEN_HML, the best model also using LSTM as decoder with AlexNet as encoder. Therefore, we choose LSTM as decoder for experiment in Dataset II.

4.3 Experiment on Dataset II

4.3.1 Training Setting

Dataset II contains the whole videos with labels. Different from Dataset I, according to the distribution of labels in Dataset II 3.2, for label F15PQ2, Intru_R, F15PQ6 and SEN_HML, the number of videos with level valuing 1 is extremely small. Therefore, different from the assignment in small scale dataset, we divide the label F15PQ2, Intru_R, F15PQ6 into 3 classes ,in which level 1 and level 2 are grouped in same class. And SEN_HML is divided into 2 classes with level 1 and level 2 in same class. As for SENSIO15, in considerate of the large number of videos in level 9 and level 10, we divide them into 4 classes with class 1 as original level $<= 8$, class 2 as original level $= 9$, class 3 as original level $= 10$ and level 4 as original level between 11 and 12.

TABLE 4.4: Experiment Result on the Big Dataset

Label	Feature Extract Model	Validation Accuracy	Test Accuracy
F15PQ2	ResNet50	57%	56%
	VGG13	58%	56%
	AlexNet	57%	59%
	GoogLeNet	61%	42%
Intru_R	ResNet50	68%	68%
	VGG13	68%	67%
	AlexNet	68%	68%
	GoogLeNet	68%	68%
F15PQ6	ResNet50	59%	47%
	VGG13	58%	47%
	AlexNet	59%	53%
	GoogLeNet	58%	58%
SEN_HML	ResNet50	77%	73%
	VGG13	77%	65%
	AlexNet	75%	73%
	GoogLeNet	75%	72%
SENSIO15	ResNet50	34%	27%
	VGG13	31%	31%
	AlexNet	36%	33%
	GoogLeNet	36%	25%

The number of labeled videos in Dataset II is 1113. For every label, we randomly selected 100 videos for valuation and 100 videos for test. The rest 913 videos consist train set. The distribution of labels in test and valuation set is same with the distribution of the in dataset. In other words, the number of every class in train set, valuation set, and test set is not balanced.

Based on the result from preliminary experiment, we conducted experiment in the whole dataset choosing ResNet50, VGG13, AlexNet and GoogLeNet as encoder. As for decoder, we use LSTM to do classification. The settings of details of experiment is same with experiment on Dataset I.

4.3.2 Result

Figure 4.4 shows the result on Dataset II. Although the accuracy seems to be relatively high accuracy, when we refer to confusion matrix showed in Figure 4.4, we find the model trained on Dataset II tends to classify videos to the class with

the largest amount in dataset. And the number of prediction to the largest class is much more than the number of truth. The result shows that the model trained on Dataset II can not be widely used.

4.4 Analysis

As the sample showed in Figure 4.5, overfitting occurred when train the LSTM in Dataset II. To solve overfitting, we tried several common method such as reducing learning rate, replacing optimization with SGD, and adjusting dropout. However, all the methods help little to solve overfitting.

Since we acquire high accuracy in Dataset I, the relatively low accuracy in Dataset II may be caused by the unbalanced distribution. Otherwise, Dataset I may have some noise dramatically influence the leaning process.

Focus on the results in small scale dataset as showed in Table. We find one video V was almost predict to be level 1 while the truth is level 3 in SENSIO15. By comparing the contain of this video and other videos which prediction is same with truth almost in all models. We find the videos predicted to be level 1 have high image contrast and the faces of mother and child is not clear. On the other hand, the image of videos predicted to be level 3 are bright with clear face.

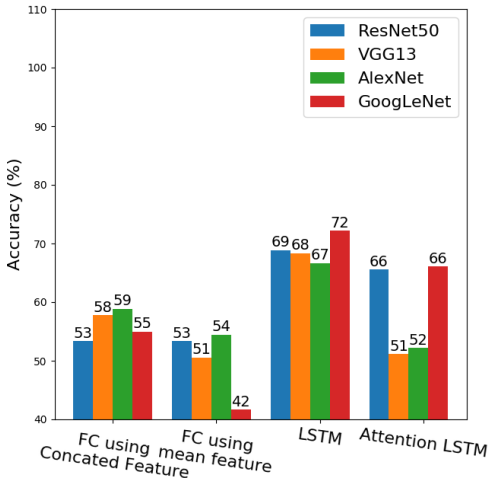
Therefore, we apply histogram equalization to the images. Figure 4.6 presents the different before and after the application of histogram equalization. The left images are original frames ,in which the face is not clear. And the right images are images after histogram equalization, we can see the face much more clearly. With the same setting to 4.2, we conducted experiment using 4 feature extraction models as encoder and LSTM as decoder only replace the input with image after histogram equalization.

Table 4.5 showed the result of image after histogram equalization as input compared with result of original image. We can see the accuracy drop down obviously in all encoders. On the other hand, the average accuracy of video V increased from 16.67% to 91.67%. From the result, the image contrast influence the learning process in Dataset I.

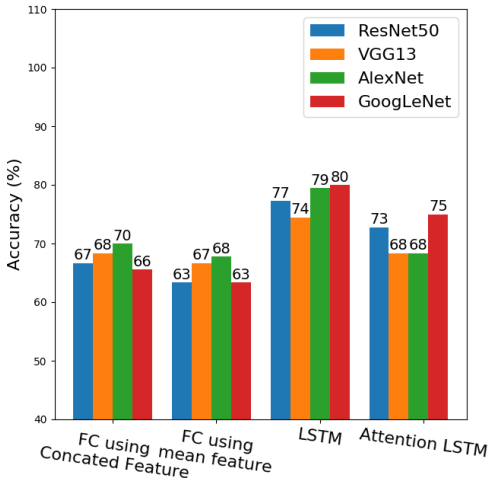
TABLE 4.5: Result of Application after Histogram Equalization on Dataset I

Feature Extract Model	Input Image	
	before histogram equalization	after histogram equalization
ResNet50	75.56%	69.68%
VGG13	77.78%	71.67%
AlexNet	76.1%	72.78%
GoogLeNet	81.11%	74.44%

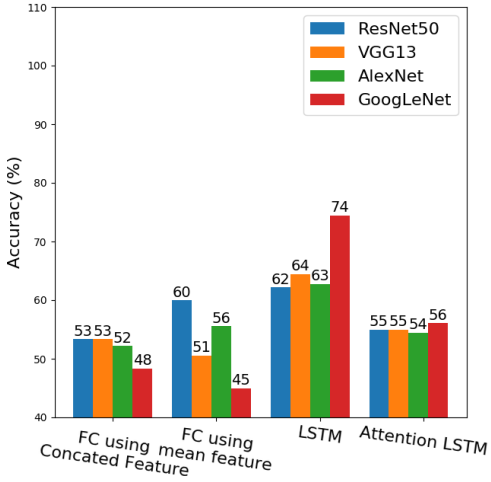
Taking the particularity of our dataset into consideration, rating childcare level is so different from action recognition that the general methods for behavior analysis may have difficult rating the level in high accuracy.



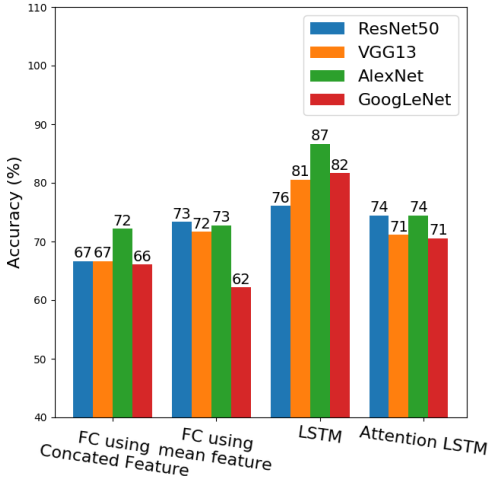
(A) F15PQ2



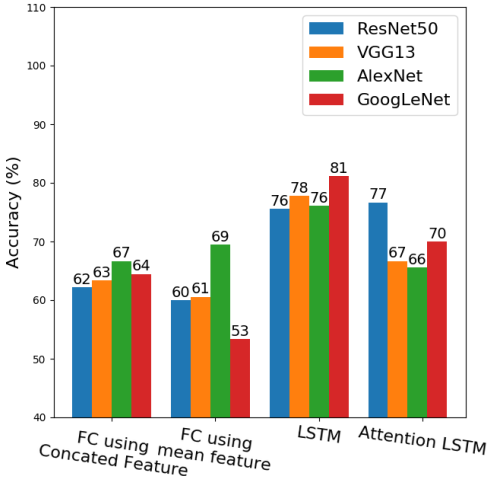
(B) Intru_R



(C) F15PQ6

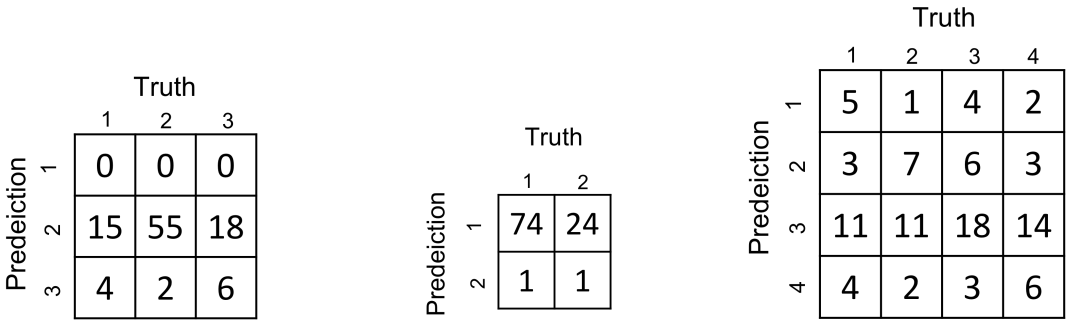


(D) SEN_HML



(E) SENSIO15

FIGURE 4.3: Experiment Result on Decoder in Small Dataset



(A) F15PQ2 (B) SEN_HML (C) SENSIO15

FIGURE 4.4: Confusion Matrix of Experiments using GoogLeNet as Encoder on Big Dataset

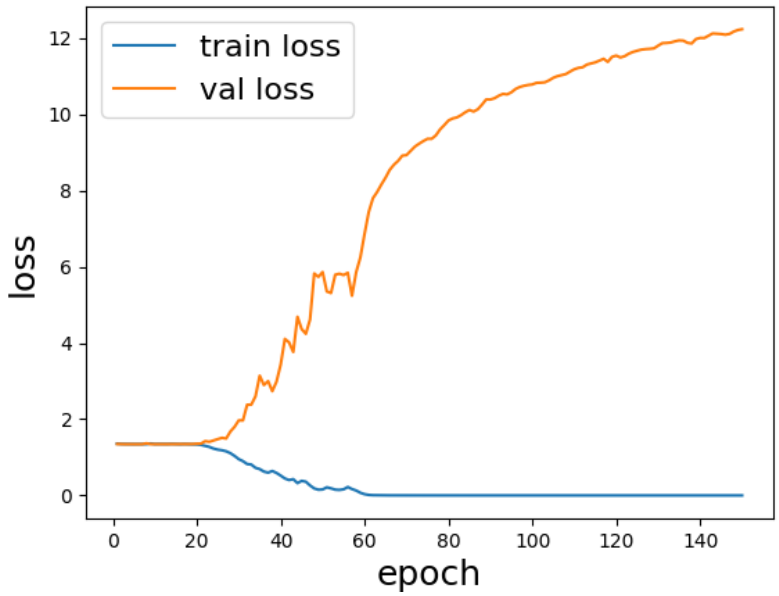
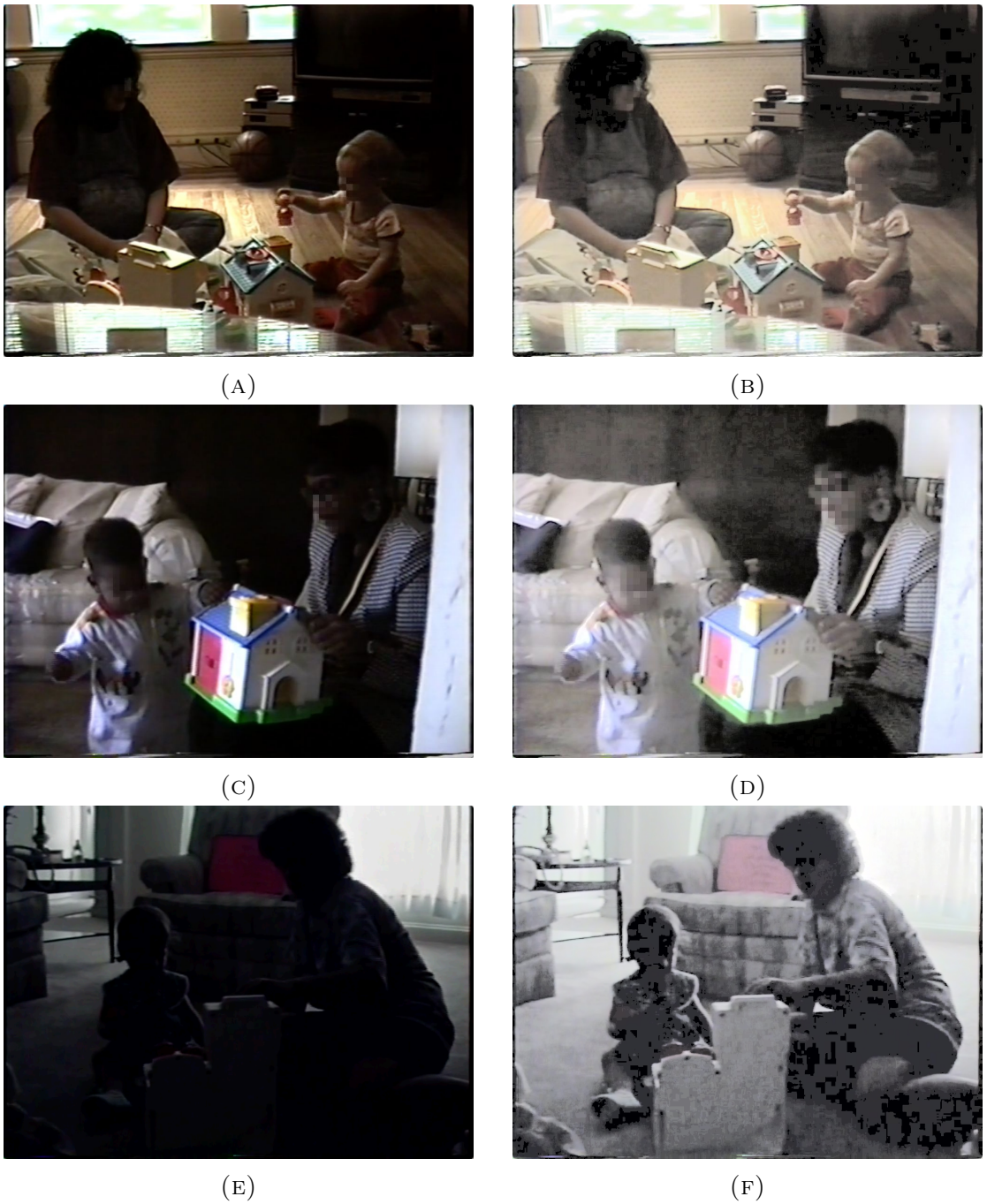


FIGURE 4.5: Loss Curve of Training using GoogLeNet as Encoder on SENSIO15 in Big Dataset



(A) (B)

(C) (D)

(E) (F)

FIGURE 4.6: Result of Histogram Equalization

Chapter 5

Conclusions and Future Works

5.1 Conclusions

In this thesis, we proposed to apply behavior analysis to childcare researches which are manual and therefore in small scale. With this attempt, we except to make large scale research on childcare possible.

We built a dataset in two scales from large amount of childcare videos. The videos are all annotated with childcare level labeled by professional analysts.

On Dataset I, we conducted experiments on different encoders and decoders. For encoder, we tested lots of feature extract model of image classification and finally selected ResNet50, VGG13, AlexNet and GoogLeNet as encoder for experiments in Dataset II. As for decoder, we tested full connection, LSTM and attention LSTM. As a result, we found LSTM performs best on our task.

On Dataset II, we use the encoder and decoder selected according to results of experiments on Dataset I. Although we acquired acceptable accuracy, the result on Dataset II is not as good as it on Dataset I. With analysis, we find some noise like image contrast influence a lot on Dataset I. And the application may be limited due to the particularity of childcare rating.

5.2 Future Works

In the future work, the current encoder-decoder method still has a lot of room for adjustment. Since the videos are recorded in 1990s, the relatively low quality caused by device and photography skill also have lots of direction for improvement.

FIGURE 5.1: Result of Openpose and Tracking

In addition, as described in 2.3, the behavior analysis method based on skeleton performances well in several general behavior analysis databases. We plan to use skeleton-based method on our task. Actually, we have already apply Openpose [57] to our database and achieved simple tracking as showed in Figure 5.1

Besides, we notice that the voice in video may matter to the childcare level. In videos of low childcare level, there are behaviors such as the mother calling the child loudly, repeating words to make child follow command. On the other hand, in videos attached with high childcare level, mother use tools to attract attention of children if they are appealed to cameraman or something else. Therefore, the usage of sound feature is also a considerate direction.

Appendix A

Real Estate Evaluation on Thermal Diffusivity and Noise Proof with IoT Sensors

A.1 Introduction

The real estate industry takes an important part in social daily life. Millions of people make great effort to find their suitable and contented houses or apartments. When selecting apartments, customers generally concentrate on information such as price, location, transportation, area size, room structure and orientation, which are provided by real estate agents and can be evaluated quantitatively and objectively.

Besides objective information, customers in current focus on information related to the house comfort, such as energy saving, noise insulation, air quality and daylight illumination. These additional information can not only help apartment searchers find apartments that meet their expectations, but also help owners to advertise their apartment in proper ways. For example, people living in cold areas may be very concerned about energy saving, and sensitive people may be strict with the performance of noise proof.

With the increasingly complex requirements of customers, the available information of apartments in present real estate agents is too limited to help customers make appropriate choices. However, information such as energy saving and noise

insulation is hard to quantify. To acquire the information, customers can only ask the real estate agents or go to visit the apartments in practice. Even though, customers can only ac-quire the information qualitatively.

We propose an approach for quantifying thermal diffusivity and noise proof performance of real estate properties. For this purpose, we use our IoT sensor system to collect multiple environmental data from 109 apartments in main cities across Japan. The IoT sensor system is based on a sensor developed by ourselves [92, 93, 94]. The self-developed sensor is also used in nursery school and nursing home sensing [95]. Through the experiments in real estate properties, we can quantitatively compare thermal diffusivity and noise proof of these apartments.

A.2 Related Works

A smart house provides various services according to the needs of consumers by optimally controlling home appliances and equipment. Yasumoto et al. proposed a method to save energy by controlling home appliances while minimizing the deterioration of comfort in smart houses [96]. In their method, the power consumption of each home appliance is changed ac-cording to the situation in order to achieve reduced consumption. For example, in the situation of reading, the power consumption of lighting equipment is reduced, and that of air conditioning is increased. According to their study, the comfort reduction rate could be reduced from 44.84% to 14.47%, which achieves 20% energy savings. In contrast to the case of smart houses, we consider the measurement of an apartment' s comfort level with no home appliances before the home is inhabited.

The Internet of Things (IoT) enables various objects to inter-operate (connect and exchange data) within the existing Internet infrastructure. IoT has potential applications in a wide array of studies, therefore, it is currently receiving considerable attention. IoT devices are already being used in various fields, and the extent of its application is expected to be continually in-creasing.

IoT devices have also been developed for observing the living environment. Examples are OMRON' s environment sensor 2JCIE-BL01 [97] and Netatmo' s personal weather station [98]. These devices measure environmental aspects as temperature, relative humidity, atmospheric pressure, and noise. Acer air monitor [99] can measure particulate matter (PM) 2.5, PM10, total concentration of volatile organic compounds (TVOC), CO2 concentration, temperature, and relative humidity. Furthermore, it provides a real-time indoor air quality index. SenStick [100] has an acceleration sensor, a gyro sensor, a geomagnetic sensor, a temperature sensor, a relative humidity sensor, an air pressure sensor, an illuminance sensor, and an ultraviolet (UV) sensor. It is very small and can be mounted on objects as small as chop-sticks, toothbrush, and glasses. Awair Glow [101] can measure temperature, relative humidity, CO2 concentration, chemicals, and dust in air. This device can inform the user of the quality of the environment with different colors.

Some studies have addressed environmental issues such as air and water pollution using IoT. For example, Ray proposed a novel technique to monitor the level of PM2.5 in the atmosphere using IoT [102]. Although IoT has been used in various fields, its application to the determination of the comfort level of houses has not been considered.

A.3 Proposal

A.3.1 Sensor System

We proposed a sensor system based on a sensor developed by ourselves (Figure A.1). Our sensor is controlled by I2C interface with Raspberry Pi 3 Model B and can measure temperature, humidity, illuminance, atmospheric pressure, noise, UV intensity, acceleration and PM2.5 at the same time. We also use OMRON environment sensor 2JCIE-BL01 (Figure A.2) to consist our sensor system. Data measured by OMRON environment sensor is collected by Raspberry Pi with Bluetooth. All measured data is collectively managed using Future Standard Co., Ltd. SCORER platform. The data could be uploaded to cloud automatically in real

FIGURE A.1: Proposed IoT sensor developed by ourselves

FIGURE A.2: OMRON environment sensor

time in an environment with the Internet. Thus, it does not require human labor to collect data from the sensors and can keep working for a long period with power supplement.

As we aim at quantitatively calculating thermal diffusivity and noise proof, we need to measure temperature and noise level both inside and outside the apartments as showed in Figure A.3 and Figure A.4. Therefore, we prepare at least two sensors for every apartment. Furthermore, since the floor temperature changes drastically outside, the sensors would lose heat faster if exposed to floor. We set the sensor in a contain to keep it away from floor. In this way, we can also avoid the influence of bad weather like rain and strong wind.

FIGURE A.3: Sensor setting inside the apartment

A.3.2 Thermal Diffusivity

The energy saving performance of the apartment can be evaluated by the Q factor (heat loss coefficient) and the UA (average U-value). However, in consideration of the inadequate maintenance record and the measurement conditions required for evaluation, it is not realistic to obtain Q or UA. Therefore, we use thermal diffusivity instead to evaluate the ability to con-serve heat. Apartments with high thermal diffusivity are prone to transfer heat to the outside. In other words, apartments with high thermal diffusivity would cost more on adjusting temperature.

The thermal diffusivity can be calculated as follows. $T_i(t)$ and $T_o(t)$ represent the temperature inside and outside an apartment at time t. We define the thermal conductivity as λ, wall thick-ness as d and heat capacity as C. The heat quantity Q can be calculated by the following two equations.

$$-\frac{\mathrm{d}Q(t)}{\mathrm{d}t} = \frac{\lambda}{d}(T_i(t) - T_o(t)), \tag{A.1}$$

$$\frac{\mathrm{d}Q(t)}{\mathrm{d}t} = C\frac{\mathrm{d}T_i(t)}{\mathrm{d}t}. \tag{A.2}$$

FIGURE A.4: Sensor setting outside the apartment

From these equations, we can acquire thermal diffusivity D as the following equation.

$$D = \frac{\lambda}{dC} = \frac{-\frac{\mathrm{d}T_i(t)}{\mathrm{d}t}}{T_i(t) - T_o(t)}. \tag{A.3}$$

Due to the influence of sunlight, the varied temperature during daytime cannot reflect thermal conductivity correctly. Therefore, we only use data after sunset to calculate the mean value and standard deviation of D values.

A.3.3 Noise Proof

Since the noise proof performance of an apartment depends on the noise level and type, it is difficult to generalize. Therefore, we compare and visualize the noise level measured indoors and outdoors at same time to evaluate the noise proof performance.

Besides the noise from outside, the noise from neighbor is also an inevitable problem. Since the type of noise from neighbor is limited, we prepared several common noises such as instrument sounds, and loud vocals. As shown in Figure A.5, we use speakers to play audio in one room and use a smartphone to record sound in

FIGURE A.5: Equipment for neighbor noise experiment

neighbor room. To compare the sound level of speaker and receiver, we evaluated the neighbor noise proof perfor-mance of apartments.

A.4 Experiments

A.4.1 Experiment Scale

We put our IoT sensor system in 109 apartments in Tokyo, Osaka, Fukushima, Nagoya, Aichi, and Hokkaido. In consideration of human effect, the apartments we selected are empty or unmanned in a period of time. For every apartment, we collect data for at least 3 days.

The materials of apartments we investigated include wood, steel, and reinforced concrete. In addition, age of the apartments also spread over a wide range from several months to 50 years.

However, because of the difficulty of obtaining neighbor rooms, we only conducted experiment on neighbor noise in two pairs of rooms. Every pair of rooms is horizontally adjacent.

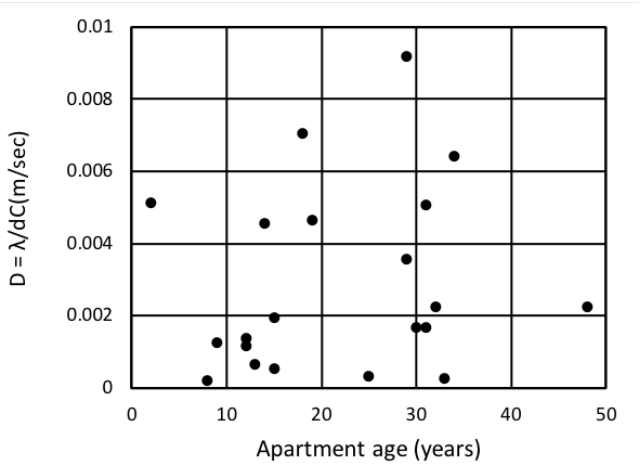


FIGURE A.6: Thermal diffusivity of reinforced concrete frame apartments

A.4.2 Thermal Diffusivity Result

Figure A.6 and Figure A.7 show the result of D (representing thermal diffusivity) of apartments built by reinforced concrete and wood. Overall, thermal diffusivity of reinforced concrete frame apartments is lower than that of wooden apartments. Thermal diffusivity of wooden apartments is obviously influenced by the age of the apartment. Although the age also influences reinforced concrete frame apartments, there are several old apartments remaining low thermal diffusivity.

In Figure A.8, the left column is a lightweight steel frame apartment (1996) and the right column is a wooden frame apartment (2015). The upper row shows the time-series change of the temperature measured, and the lower row shows the D value. We can see that the value of D is much smaller in the wooden house (2015), which means the thermal diffusivity performance is better.

With the thermal diffusivity performance, the size of the room, and the air conditioning equipment, we can set a simple indicator to describe energy saving performance of the apartment.

A.4.3 Noise Proof Result

Figure A.9 shows the result of outdoor noise proof performance. Horizontal axis shows indoor noise level, and vertical axis shows the outdoor noise level at the

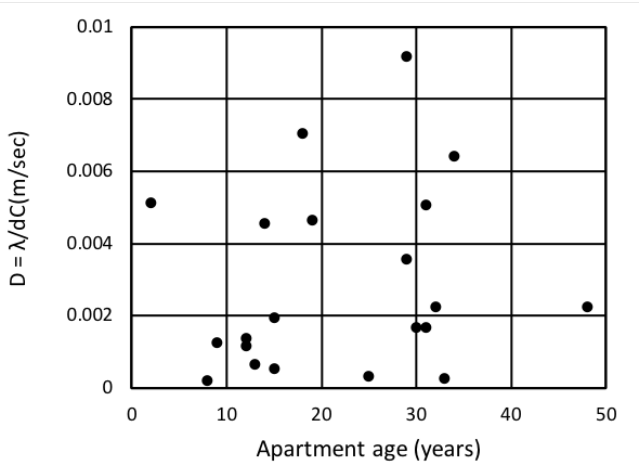


FIGURE A.7: Thermal diffusivity of reinforced concrete frame apartments

same time. The distance from every point to the red line presents the noise proof performance. For points above the red line, the further distance away from the red line means the better noise proof performance. In Figure A.9a, for points with outdoor noise level between 35dB to 45dB, the indoor noise level is almost between 32dB to 38dB. On the other hand, in Figure A.9b, for points with outdoor noise level between 30dB to 55dB, the indoor noise level is almost be-tween 31dB to 35dB. Since the Figure A.9b has further average distance from the red line and performances better with high out-door noise level, the outdoor noise proof performance of (B) is better than (A).

Figure A.10, Figure A.11, Figure A.12, and Figure **??** show the result of neighbor noise proof performance. Figure A.10 and Figure A.11 show the noise proof performance of instrument sound. We conduct experiment on sound of piano, guitar and drum. Figure A.10 is the result in new-build apartment and Figure A.11 is the result in old apartment. In new-build apartment, the receiver room is not influence from speaker room. However, in old apartment the influence from speaker room is obvious. Figure A.12 and Figure A.13 show the noise proof performance of vocals. We conduct experiment on sound of cry, laugh and talk, and we test two type of laugh sound and talk sound. Figure A.12 is the result in new-build apartment and Figure A.13 is the result in old apartment. In both new-build apartment and old apartment are hardly influence by vocals from speaker room.

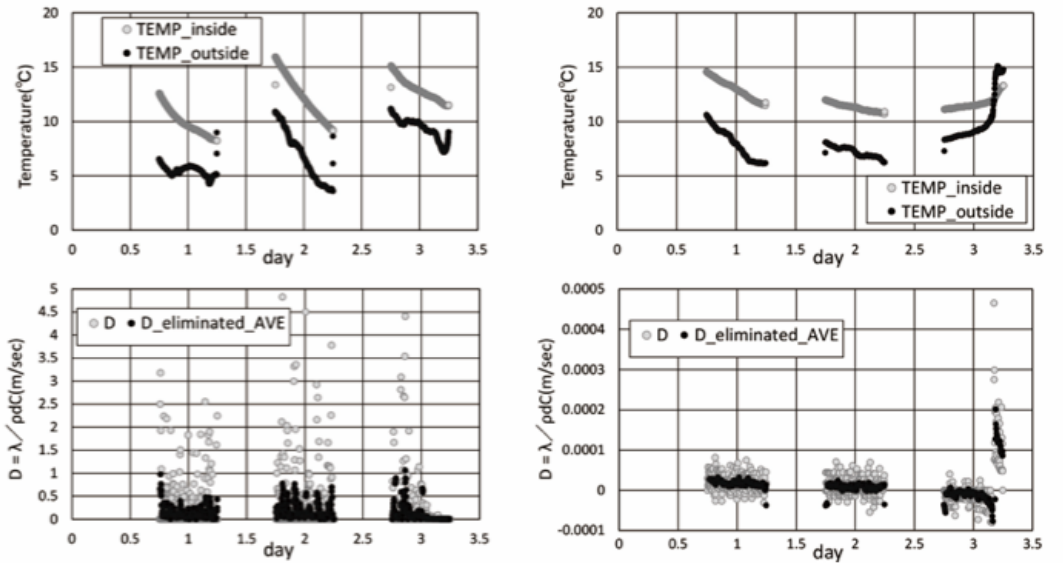


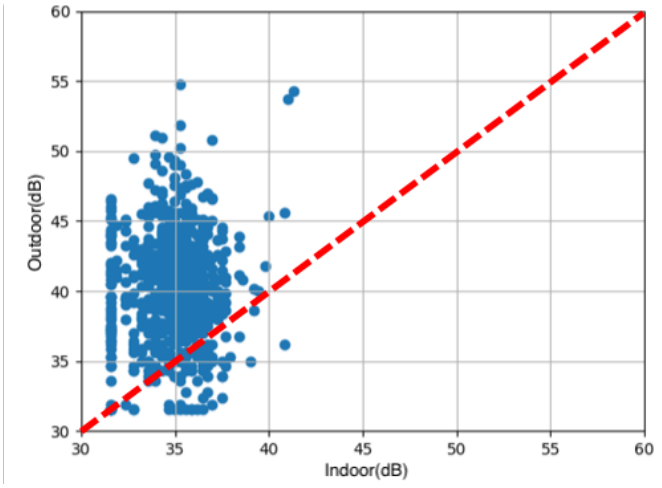
FIGURE A.8: Calculation result of thermal diffusivity

A.5 Conclusion

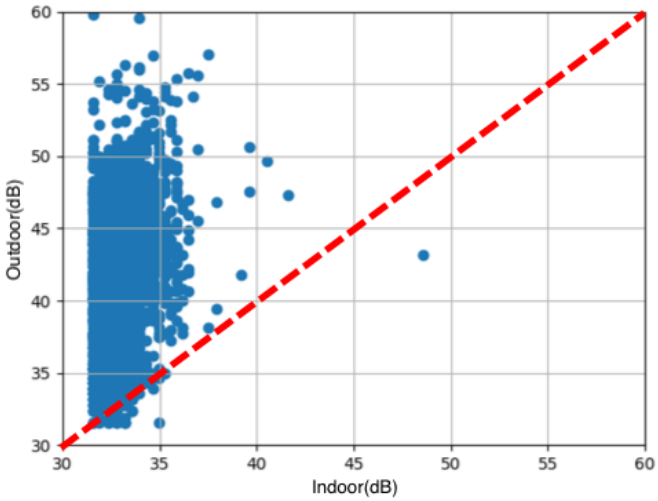
We focus on an IoT sensor based approach to quantify thermal diffusivity performance and noise proof performance for real estate evaluation. The sensors we developed can measure multiple environmental data such as temperature, humidity, illuminance and so on. We collected large amount of data in main cities across Japan using our sensor system.

We proposed the value D to quantitatively evaluate thermal diffusivity performance of apartments. Instead of other existing factors, value the D is easy to be acquired from inside and outside temperature change. We conduct comparison experiments on different materials of apartments and apartments ages. As a result, the thermal diffusivity performance of reinforced concrete frame apartments is good and is hardly influenced by building ages.

For noise proof performance, we find a way to present the outdoor noise proof performance apartment by apartment. As for neighbor noise proof performance, because of the difficulty of obtaining usable neighbor rooms, the experiment scale is small. From the experiment result, new apartment holds good neighbor noise proof performance on instrument sounds and loud vocals. Old apartment holds good performance on insulating vocals, but is not good at insulating instrument



(A)



(B)

FIGURE A.9: Outdoor Noise Proof Performance

sounds. In the future, we plan to search for cooperation to obtain more neighbor rooms or find methods to test neighbor noise proof performance inside one room.

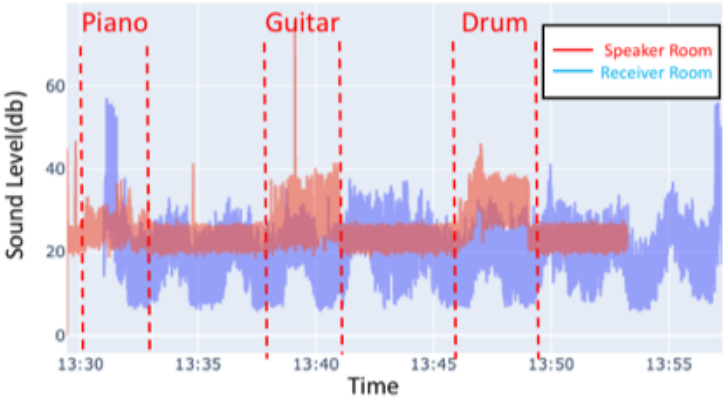


FIGURE A.10: Neighbor Noise Proof Performance on Instrument in New-build Apartment

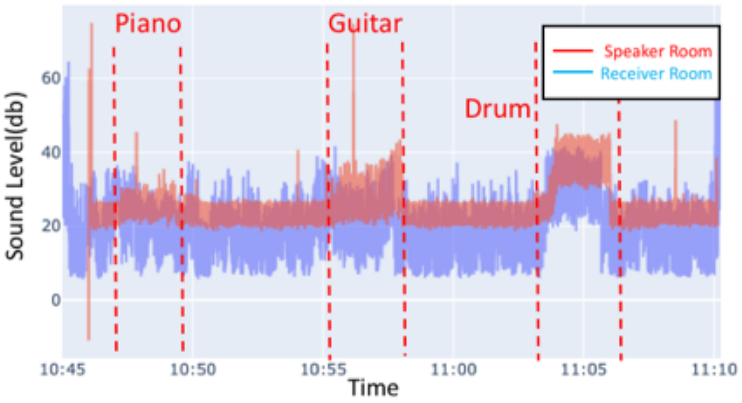


FIGURE A.11: Neighbor Noise Proof Performance on Instrument in Old Apartment

FIGURE A.12: Neighbor Noise Proof Performance on Vocal in Newbuild Apartment

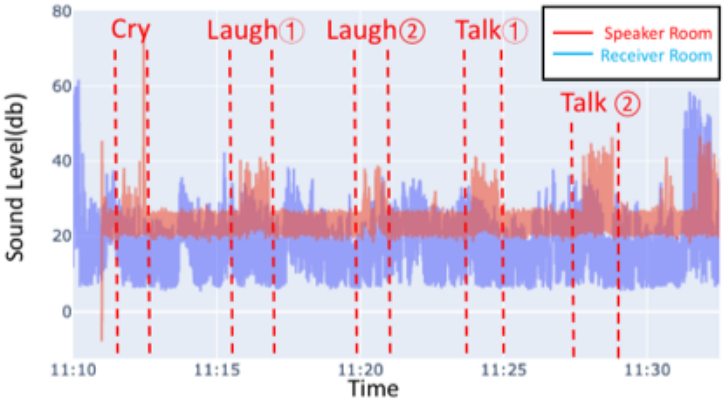


FIGURE A.13: Neighbor Noise Proof Performance on Vocal in Old Apartment

References

[1] J. F. Cohn and E. Tronick, "Specificity of infants' response to mothers' affective behavior," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 28, no. 2, pp. 242–248, 1989.

[2] M. Purhonen, R. Kilpeläinen-Lees, M. Valkonen-Korhonen, J. Karhu, and J. Lehtonen, "Cerebral processing of mother's voice compared to unfamiliar voice in 4-month-old infants," *International Journal of Psychophysiology*, vol. 52, no. 3, pp. 257–266, 2004.

[3] T. M. Field, "Early interactions between infants and their postpartum depressed mothers." *Infant behavior & development*, 1984.

[4] K. Noriko, "Toddlers' refusalbehaviors and caregivers' interventionat lunchtime: An analysisof interactionpatterns," vol. 42, no. 2, pp. 112–120, 2004.

[5] N. Yumiko and K. Kiyomi, "The acquisition process of "daily work" in the kindergarten for preschoolers: Focus on the clean-up scene," vol. 62, no. 11, pp. 735–741, 2011.

[6] S. Kimio, K. Shigeo, and E. Teiji, "An automatic record making system for nurserly schools -location detection for infants using passive sensros-," vol. 2003, no. 31 (2002-IS-083), pp. 73–80, 2003.

[7] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[8] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[9] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[11] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[13] B. A. CV C and S. M. V. J. Janssen, "Effects of placementand orientation of body-fixed accelerometers on the assessment of energy expenditure during walking," *Medical & biological engineering & computing*, 1997.

[14] M. L. Fruin and J. W. Rankin, "Validity of a multi-sensor armband in estimating rest and exercise energy expenditure," *Medicine & Science in Sports & Exercise*, vol. 36, no. 6, pp. 1063–1069, 2004.

[15] M. Mathie, A. Coster, N. Lovell, and B. Celler, "Detection of daily physical activities using a triaxial accelerometer," *Medical and Biological Engineering and Computing*, vol. 41, no. 3, pp. 296–301, 2003.

[16] K. Aminian, P. Robert, E. Buchser, B. Rutschmann, D. Hayoz, and M. Depairon, "Physical activity monitoring based on accelerometry: validation and comparison with video observation," *Medical & biological engineering & computing*, vol. 37, no. 3, pp. 304–308, 1999.

[17] J. Ng, A. V. Sahakian, and S. Swiryn, "Accelerometer-based body-position sensing for ambulatory electrocardiographic monitoring," *Biomedical instrumentation & technology*, vol. 37, no. 5, pp. 338–346, 2003.

[18] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," *IEEE transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 156–167, 2006.

[19] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*. IEEE, 2006, pp. 4–pp.

[20] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled

conditions," *IEEE transactions on information technology in biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[21] P. Salvo, F. Di Francesco, D. Costanzo, C. Ferrari, M. G. Trivella, and D. De Rossi, "A wearable sensor for measuring sweat rate," *IEEE Sensors Journal*, vol. 10, no. 10, pp. 1557–1558, 2010.

[22] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[23] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[24] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," in *ACM SIGGRAPH 2011 papers*, 2011, pp. 1–10.

[25] N. Werghi, "Segmentation and modeling of full human body shape from 3-d scan data: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1122–1136, 2007.

[26] A. Boyali, M. Kavakli, and J. Twamley, "Real time six degree of freedom pose estimation using infrared light sources and wiimote ir camera with 3d tv demonstration," in *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 2010, pp. 137–148.

[27] N. Noury, T. Hervé, V. Rialle, G. Virone, E. Mercier, G. Morey, A. Moro, and T. Porcheron, "Monitoring behavior in home using a smart fall sensor and position sensors," in *1st Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Proceedings (Cat. No. 00EX451)*. IEEE, 2000, pp. 607–610.

[28] P. Wilhelm, E. Monier, P. Thomas, and U. Ruckert, "Spa—a system for analysis of indoor team sports using video tracking and wireless sensor network," in *2009 Proceedings of 6th International Symposium on Image and Signal Processing and Analysis*. IEEE, 2009, pp. 237–241.

[29] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Proc.*

ISPRS intercommission conference on fast processing of photogrammetric data. Interlaken, 1987, pp. 281–305.

[30] C. G. Harris, M. Stephens *et al.*, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[31] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision*, vol. 30, no. 2, pp. 79–116, 1998.

[32] S. Rea, F. Eisenhaber, D. O'Carroll, B. D. Strahl, Z.-W. Sun, M. Schmid, S. Opravil, K. Mechtler, C. P. Ponting, C. D. Allis *et al.*, "Regulation of chromatin structure by site-specific histone h3 methyltransferases," *Nature*, vol. 406, no. 6796, pp. 593–599, 2000.

[33] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 357–360.

[34] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[39] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[40] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust asr," in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4085–4088.

[41] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 1017–1024.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[44] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 2117–2126.

[45] P. Tokmakov, K. Alahari, and C. Schmid, "Learning motion patterns in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3386–3394.

[46] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4405–4413.

[47] S. Saha, G. Singh, M. Sapienza, P. H. Torr, and F. Cuzzolin, "Deep learning for detecting multiple space-time action tubes in videos," *arXiv preprint arXiv:1608.01529*, 2016.

[48] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European conference on computer vision*. Springer, 2004, pp. 25–36.

[49] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.

[50] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[51] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[52] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1212–1221.

[53] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *proceedings of the IEEE international conference on computer vision*, 2017, pp. 1281–1290.

[54] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 713–728.

[55] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 190–206.

[56] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Advances in neural information processing systems*, 2017, pp. 2277–2287.

[57] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[58] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[59] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2544–2550.

[60] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5, 2014*. BMVA Press, 2014.

[61] H. Nam, M. Baek, and B. Han, "Modeling and propagating cnns in a tree structure for visual tracking," *arXiv preprint arXiv:1608.07242*, 2016.

[62] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[63] H. Fan and H. Ling, "Sanet: Structure-aware network for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 42–49.

[64] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 472–488.

[65] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "Eco: efficient convolution operators for tracking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6638–6646.

[66] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[67] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[68] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[69] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[70] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 915–922.

[71] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.

[72] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3d deep convolutional descriptors." in *IJCAI*, vol. 1, 2016, p. 3.

[73] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[74] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3725–3734.

[75] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Advances in Neural Information Processing Systems*, 2017, pp. 34–45.

[76] M. Depression, "Chronicity of maternal depressive symptoms, maternal sensitivity, and child functioning at 36 months," *Developmental Psychology*, vol. 35, no. 5, pp. 1297–1310, 1999.

[77] S. B. Campbell, P. Matestic, C. von Stauffenberg, R. Mohan, and T. Kirchner, "Trajectories of maternal depressive symptoms, maternal sensitivity, and children's functioning at school entry." *Developmental psychology*, vol. 43, no. 5, p. 1202, 2007.

[78] J. Belsky and R. P. Fearon, "Early attachment security, subsequent maternal sensitivity, and later child development: Does continuity in development depend upon continuity of caregiving?" *Attachment & human development*, vol. 4, no. 3, pp. 361–387, 2002.

[79] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2556–2563.

[80] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[81] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[82] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[83] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[84] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

[85] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[86] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[87] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[88] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[89] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[90] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[92] 大渕友暉, 山崎俊彦, 相澤清晴, 鳥海哲史, and 林幹久, "不動産物件の快適度評価のための iot センサ実装と評価." ITE 冬期大会, 2016, pp. 11C–4.

[93] 大渕友暉, 山崎俊彦, 相澤清晴, 鳥海哲史, and 林幹久, "Iot センサを用いたマンション物件計測と快適度評価." JSAI, 2017, pp. 1H2–OS–15a–4.

[94] 大渕友暉, 山崎俊彦, 鳥海哲史, 林幹久, 野澤祥子, 高橋翠, 遠藤利彦, and 秋田喜代美, "保育施設における iot カメラを用いた環境・行動解析," in **信学技報**, *vol. 117, no. 217, MVE2017-15*. IE-IEICE-MVE, 2017, pp. 7–11.

[95] Y. Obuchi, T. Yamasaki, K. Aizawa, S. Toriumi, and M. Hayashi, "Measurement and evaluation of comfort levels of apartments using iot sensors," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2018, pp. 1–6.

[96] K. Yasumoto, K. Ogura, S. Yamamoto, and M. Ito, "Device control method for energy-saving with minimal degradation of users' comfort level," *IPSJ SIG Technical Reports (in Japanese)*, vol. 2011, no. 9, pp. 1–8, 2011.

[97] "Omron environment sensor," http://www.omron.co.jp/ecb/products/sensor/special/environmentsensor/.

[98] "Netatmo personal weather station," https://www.netatmo.com/product/weather/weatherstation.

[99] "Acer, acer air monitor," https://home.cloud.acer.com/airmonitor/.

[100] "Senstick," http://senstick.com/r/.

[101] "Awair, awair glow," https://getawair.com/pages/awair-glow.

[102] P. P. Ray, "Internet of things cloud based smart monitoring of air borne pm2. 5 density level," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, 2016, pp. 995–999.

Publications

International Conferences and Workshops

[1] Yuan Lin, Yuki Obuchi, Xueting Wang, Toshihiko Yamasaki, Ryoma Kitagaki, Satoshi Toriumi, Mikihisa Hayashi, Ai Sakai, Nobuhito Haga, Shimpei Nomura, and Yoichi Ikemoto. "Real Estate Evaluation on Thermal Diffusivity with IoT Sensors", *IEEE GCCE 2019 OS-RET, pp.456-457, Oct. 16 2019, Senri Life Science Center, Osaka, Japan.*

[2] Yuan Lin, Yuki Obuchi, Xueting Wang, Toshihiko Yamasaki, Satoshi Toriumi, Mikihisa Hayashi, Sachiko Nozawa, Midori Takahashi, Toshihiko Endo, Kiyomi Akita. "Human Tracking for Children Behavior Analysis in Nursery Schools", *IEEE SITIS 2019 - The 15th International Conference on Signal Image Technology & Internet based Systems, Nov. 26-29 2019, Sorrento (NA), Italy.*

[3] Yuan Lin, Yuki Obuchi, Toshihiko Yamasaki. "Behavioral Analysis in Childcare Facilities with IoT Cameras", *J-TMM2019, April 13-14 2019, Tainan, Taiwan.*

Domestic Conferences and Symposia

[4] 山崎 俊彦, 大渕 友暉, 林 遠, 北垣 亮馬, 鳥海 哲史, 林 幹久, 酒井 藍, 芳賀 宣仁, 野村 眞平, 池本 洋一. IoT センシングによる不動産物件の断熱・防音性能評価. *2019* 年度 人工知能学会全国大会 *JSAI2019, 2019* 年 *JSAI2019* 巻 *1D2-OS-10a-01, 2019* 年 *6* 月 *4* 日～*7* 日, 新潟県新潟市

[5] 林遠, 大渕友暉, 汪雪婷, 山﨑俊彦, 鳥海哲史, 林幹久, 野澤祥子, 高橋翠, 遠藤利彦, 秋田喜代美. スマート保育への挑戦: センシング技術を活用した保育環境の調査. 東京大学大学院教育学研究科附属発達保育実践政策学センター主催 *2019* 年度公開シンポジウム「発達と保育の本質の探究～人の育ちとそれを支える営みを見つめて～」, 東京大学・安田講堂, 東京都文京区, *Aug. 3, 2019.*

[6] 林 遠, 大渕 友暉, 汪 雪婷, 山崎 俊彦, 北垣 亮馬, 鳥海 哲史, 林 幹久, 三田 涼介, 芳賀 宣仁, 野村 眞平, 池本 洋一. IoT センシングを用いた不動産物件の断

熱・遮音性能評価. 画像工学研究会 *(IE)*, 信学技報, *2020* 年 *2* 月 *27* 日～*28* 日 北海道. (Submitted)

Acknowledgements

I came to Japan as a student of the University of Tokyo two and a half years ago. At first, I was accepted as a research student by my supervisor, Prof. Toshihiko Yamasaki. During the 30 months in Yamasaki Lab, though the international conference and large number of cooperation meetings with companies nearly filled his schedule, he held meeting with all students one by one almost every week. Even when I was still a research student, I had the opportunity to share my research progress with him and got lots of wise advice. Besides, he would check the publications and presentations of all students with generous and responsibility. He provided cooperation research in varied fields, and thanks to that I have learned a lot not only about computer science but also knowledge of advertising, architecture and so on. Furthermore, except research, he cared about my future career and introduced some ways to famous companies. I really feel fortunate to meet such a kind and patient supervisor in my master life. I sincerely appreciate my supervisor, Prof. Toshihiko Yamasaki.

I express gratitude to Prof. Kiyoharu Aizawa. His suggestions in laboratory meetings are always inspiring and encouraging. In addition, he provided some advice to me about career selection.

Many thanks to Ms. Matsubayashi and Ms. Egawa, the secretaries of our laboratory. They are so friendly and patient with foreign student like me. Every time I meet school affairs, they would offer convenience to students and preserve the interest of students.

Thanks to all lab members, making this laboratory full of happiness and atmosphere of research. Thanks to project researcher of our laboratory, Dr. Wang. She is like another supervisor giving lots of suggestions and always follows a lot in weekly meeting. Thanks to all phD students, Furuta, Miyata, Sourav, Inoue, Ikuta, Tao, Yiwei, Zhong, showing me about what is a researcher and the attitude about research. Also thanks to senior master students who have already graduated. Thanks to Nakamura and Shen, offering lots of help at the beginning when I came to our lab. Thanks to Obuchi, whose work helped a lot in my current research.

Thanks to the students of the same year, Yu, Kato, Kaneko, Kosugi, Tanaka, Tsubota, Yi, Zhang, Kawarada, Chen. On one hand, they shared lots of helpful information about master life and shows the models as a master student of the University of Tokyo. On the other hand, they often hold some entertainment like have dinner together. Actually, the talk in dinner help me a lot to know the native life and find my own position.

Especially thanks to Chinese student in our lab, Zhang, Tao, Chen, Yi, Zhong, Yiwei. Talking in Chinese sometime and gathering in holidays really bring me great relaxation.

Finally, I sincerely appreciate my dear parents, who support me studying in a foreign country.

Yuan Lin

January 30, 2020