Master's Thesis

Title

Bayesian Estimation for Enhancing 3D-point Object Recognition using Spatio-temporal Information

Supervisor

Associate Professor Shin'ichi Arakawa

Author

Hiroaki Sato

February 2nd, 2023

Department of Information Networking

Graduate School of Information Science and Technology

Osaka University

Master's Thesis

Bayesian Estimation for Enhancing 3D-point Object Recognition using Spatio-temporal Information

Hiroaki Sato

Abstract

In recent years, there has been expected to understand situations in real space and to make use of information. In order to understand accurately situations in real space, it requires exact sensing and highly accurate object recognition. In much research on object recognition, the machine learning method in 2D image is used. However, it has problems that 2D image cannot manage 3D information such as overlapping objects. In terms of the machine learning method in directly 3D, lacks of data due to as blind spot of the sensor make a loss of accuracy in object recognition. Moreover, there is no way for applications to determine whether lacks of the 3D data, characteristics of the method, or the amount of training, when the score is low. Thus, it is essential for understanding real space to verify confidence of predictions and correcting and augmenting as necessary by a different method from object recognition. In this thesis, towards understanding situation in real space, we developed a way to represent recognition including ambiguity, that is, what the object is and where the object is located in real space. As a different method from the machine learning, we represent probabilistically real space, focusing on spatio and temporal knowledge that humans potentially equipped with. Using nuScenes dataset for autonomous driving, we acquired the spatio knowledge describes that similar and related objects are often located close together. In addition, we acquired the temporal knowledge based on temporal sequence of objects. Moreover, in order to verify confidence of predictions from the machine learning method and enhance as necessary, we proposed the method of Bayesian estimation. As a result, in using the spatio knowledge, we improved accuracy from 75.7% to 80.2% than the machine learning method. In using spatio-temporal knowledge, we increased confidence of predictions with incorporating temporal dependency.

Keywords

Bayesian estimation

Object recognition

Cyber physical system

Deep learning

Point cloud

Digital twin

Contents

List of Figures

List of Tables

1 Introduction

In recent years, there has been expected to understand situations in real space and to make use of information. For example, the digital twin [1] is a reproduction of real space in virtual space by collecting data through sensors and projecting and reproducing information. Digital twin enables analysis and simulation in urban planning and factory monitoring and is expected to be realized for Digital Transformation (DX) and/or Industry 4.0. One of the most important tasks in realizing digital twin is to understand situations in real space [2, 3]. In the case of digital twin of a city, it is necessary to understand accurately real space situations, such as how the city is organized, where people, cars, and other objects are located, with approaches on sensing and object recognition.

In the field of sensing and object recognition, many researches are being conducted to recognize object classes and locations using 2D video images and the machine learning techniques [4, 5]. However, since the real space is 3D, sensing through video images lacks depth information and cannot manage 3D information such as overlapping objects. Therefore, in recent years, object recognition using the machine learning techniques on 3D data is investigated [6, 7]. Although the machine learning techniques on 3D data captures the depth information to some extend, the 3D data may not be accurate due to blind spots from the sensor, or color and material. Especially, LIDAR sensors using near-infrared light are known to appear areas where the reflections cannot be adequately focused [8]. Specifically, black areas, such as hair and paint, absorb and glass and mirrors reflect and transmit. In addition, object recognition methods using the machine learning techniques require a large amount of data for training because computers automatically learn features [9]. Therefore, inaccuracy of 3D data makes a loss of accuracy in object recognition.

In general, object recognition methods have predicted an object label and a score value that used for the prediction along with, referred to as a predicted label and a predicted score, and some applications treat predicted scores as confidence. However, there is no way for applications to determine whether lacks of the 3D data, characteristics of the method, or the amount of training, when the score is low. Improving the accuracy of the predicted label and score is an important in the machine learning techniques. However,

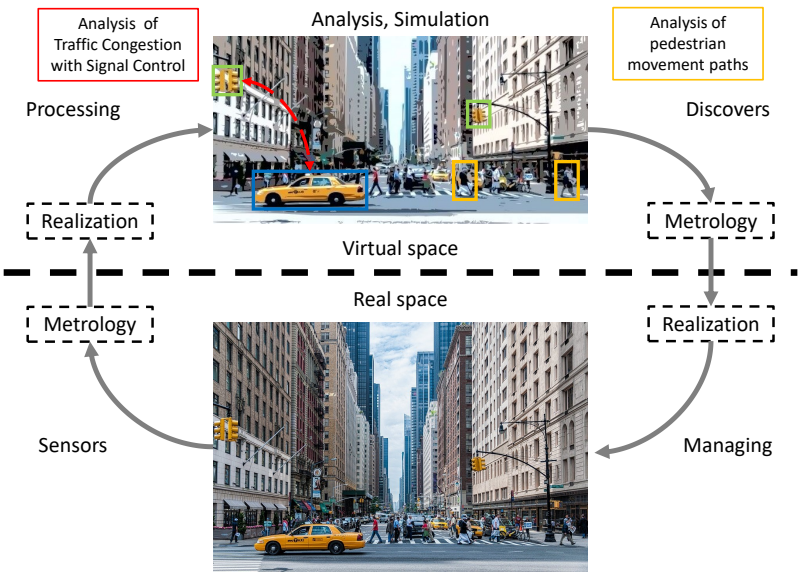


Figure 1: Image diagram of digital twin

the essential problem of understanding real space is how to verify confidence of predicted labels and scores from the machine learning techniques.

In this thesis, towards understanding situation in real space, we developed a way to represent recognition including ambiguity, that is, what the object is and where the object is located in real space. Representing recognition with ambiguity means that express how legitimate or trustworthy they have, rather than predict objects directly. As a different method from the machine learning, we represent probabilistically real space, focusing on spatio and temporal knowledge that humans potentially equipped with. Spatio knowledge is the information based on the spatio location of objects, for example, how often desks and chairs are used next to each other, or how often beds are not near a kitchen. Temporal knowledge is the information based on the temporal relationship of objects that exists in real space though cannot be continuously observed on data due to blind spots from the sensor. Using the knowledge, with enhancing recognition by verifying confidence and correcting and augmenting as necessary, we recognize accurately situations in real space. In order to obtain the knowledge, we extract the above information for objects from a large amount of data of real space, and statistically determine what kind of correlation. Furthermore, by treating the correlation as a probability, it is given as prior knowledge

about objects that may occur in real space. We propose the method for enhancing object recognition using Bayesian estimation method and demonstrate the effectiveness using datasets of real space.

This thesis is organized as follows: We organized what we need in understanding real space information and described knowledge in Section 2. We described the method of Bayesian estimation using spatio knowledge and evaluate in Section 3. We described the method of Bayesian estimation using spatio-temporal knowledge and evaluate in Section 4. We present conclusions of this thesis in Section 5.

2 Understanding Real-space

In order to understand situations in real space, it is necessary sensing real space and understanding what kind of objects based on acquired data.

2.1 Sensing Real-space

It is important to collect accurate data on real space environment. The data of real space uses in various applications, such as CAD, Robotics, Behavior analysis and prediction [10]. In architecture, we treat the model of object in real space, such as bridges, houses, and furniture. In order to measure the object and simulate in computers, we capture the object as CAD data in 2D or 3D. Incorrect data in CAD makes wrong design and simulation. Therefore, we use calibration techniques to fuse data to more correctly. In robotics, we use real space data in moving, grasping, assembling, and so on. In behavior analysis and prediction, we use data under a variety of conditions. The behavior prediction is that predicts the location of objects in future frame using the data in previous frames. The behavior analysis is that analyzes the movement of objects using the data in a range of frames. We use various sensors to collect data from real space, for example, cameras and LIDAR. Each sensor has its own strengths and weaknesses.

A camera is a sensor that captures still or video images by light to project real space. The data is represented on computers as a bitmap image, which stored with its own color information in each pixel. Because of getting by projecting light, it can perceive real space in the same way as humans. On the other hand, it has disadvantaged that information is reduced to two dimensions and cannot sense accurately at night or in bad weather. In terms of amount of information, there are some types of 3D scanner using cameras [11]. One of 3D scanner is a stereo camera that uses two or more lenses and captures in each. This allows cameras to capture 3D information because of parallax, means difference in the apparent position. In addition, a ToF-camera is a camera with light to measure distance between the camera and the object by time-of-flight. These cameras takes an image with optical system, cannot take in bad weather.

In capturing 3D data, LIDAR is a sensor that measures distance and a shape of the target object by laser based on directions and time of reflected. The data is expressed as

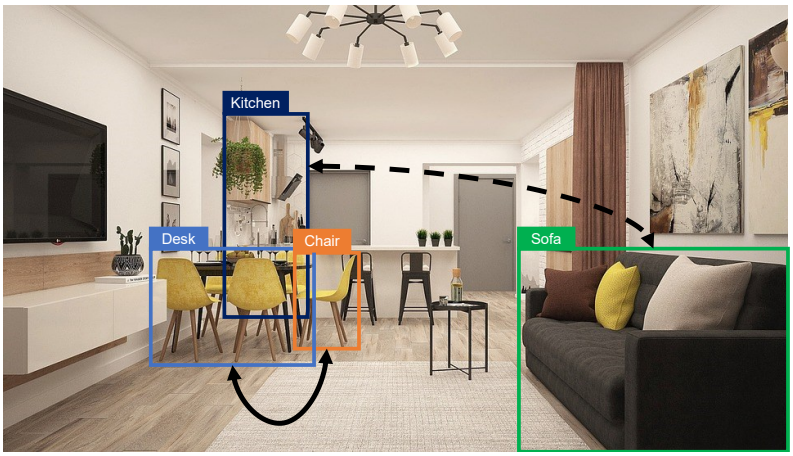
point cloud, which stored with its own coordinates in each point. The similar sensor is RADAR that uses radio waves to directions and time. RADAR cannot capture non-metal objects because cannot reflect radio waves. Because of getting by laser or radio waves, it can perceive real space as 3D and can collect even in bad weather. On the other hand, it takes higher cost than by cameras.

2.2 Recognizing Objects from Sensor Information

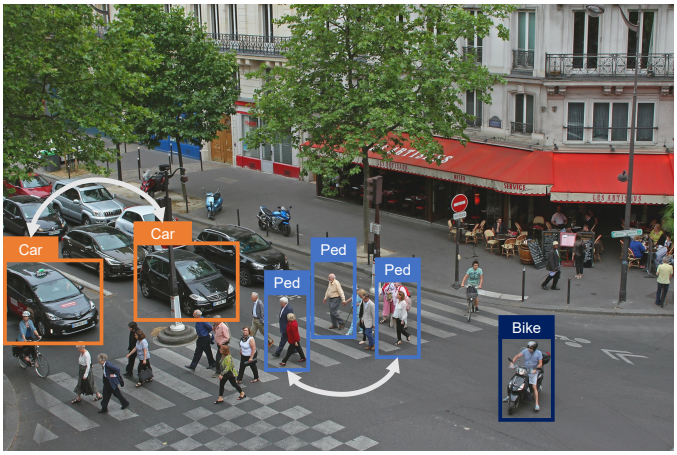
It is important to understand how a city is organized based on acquired data from real space. Recently, there has been an increase in research on object recognition, that recognize what an object is and where it is, such as people, cars, etc. By recognizing objects, we can process information, for example, where people congregate and how many cars are in real space. Many methods in research use the machine learning techniques, especially deep learning, to perform object recognition. YOLO [12] is the one of the approaches to object recognition in 2D image. These methods predict the bounding box, means the location of objects, and the class, means what the object is, and the confidence score, means how accurate it thinks the box. Recently, many methods are proposed that predicts object recognition towards to 3D point cloud, such as PointNet [13] and VoxelNet [14]. One of the problems using deep learning method is black box, means that we do not know how the method came up with the answers. This problem causes much research Explainable AI, means that presents the description we make sense. However, that cannot verify confidence of predictions, etc.

2.3 Enhancing Object Recognition: Approach

In this thesis, we focus on perception about objects in real space, which humans potentially equipped with, and acquire information as knowledge by statistically determining correlations between objects. Using this knowledge, we probabilistically recognize in real space through enhancing predictions by object recognition. For enhancing, we verify confidence of predicted labels and scores and correct and augment, as necessary. We deal with two types of knowledge: Spatio knowledge and Temporal knowledge.



(a) Indoor



(b) Outdoor

Figure 2: Image diagram of spatio knowledge

2.3.1 Spatio Knowledge

As spatio knowledge, we focus on information based on spatio positional relationship of objects. We detect adjacency of objects by positions within space and statistically calculate frequencies between object classes.

We construct relationship information about objects that we potentially recognize in our daily lives such as Figure 2, for example, desks and chairs are frequently used next to each other, sofas are not located near the kitchen, etc.

2.3.2 Temporal Knowledge

As temporal knowledge, we focus on information based on temporal sequence of objects.

Object Tracking is a task that predicts locations of the same object through the multiple scans in a scene and assigns a unique ID to the predictions considered to be the same. One of approaches for the object tracking is Tracking-by-Detection that predicts the tracking from the multiple predictions of object detection in a scene. It is based on information and features between the frames and associate the objects in some scans. CenterTrack [15] is one of the object tracking approaches that uses an estimate velocity of centers of objects through multiple frames.

With our approach, given the information on the current frame, we detect presences of objects and calculate frequency that the object was present at previous time.

3 Bayesian Estimation Method using Spatio Knowledge

The aims in this section are, towards predicted labels and scores by existing object recognition methods, to verify confidence and to correct and augment as necessary, by a different method from object recognition.

By a different method, we use Bayesian Estimation [16] which infers the predicted score of a target object by using acquired knowledge. Surrounding circumstances allows the target to compute prediction under observation, not just information about the target.

Bayesian Estimation is given by

$$P(x|z) = \frac{G(z|x) \times P(x)}{\sum_{x' \in X} G(z|x') \times P(x')} \tag{1}$$

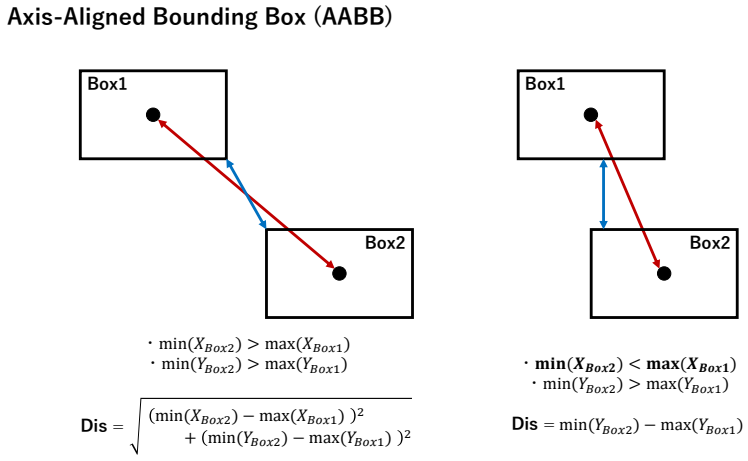
where $P(x)$ is a predicted score of each class $x \in X$ by object recognition, as a prior probability. By obtaining surrounding circumstances in predictions, treating as a feature z, we get corresponding probability $P(x|z)$, referred as a posterior probability. $G(z|x)$ is a conditional probability that z given x is true, referred as a likelihood. In this section, we focus on the surrounding circumstance as the feature z, especially the object around the target, and treat the spatio knowledge as the likelihood.

3.1 Spatio Knowledge as Positional Relationship among Objects

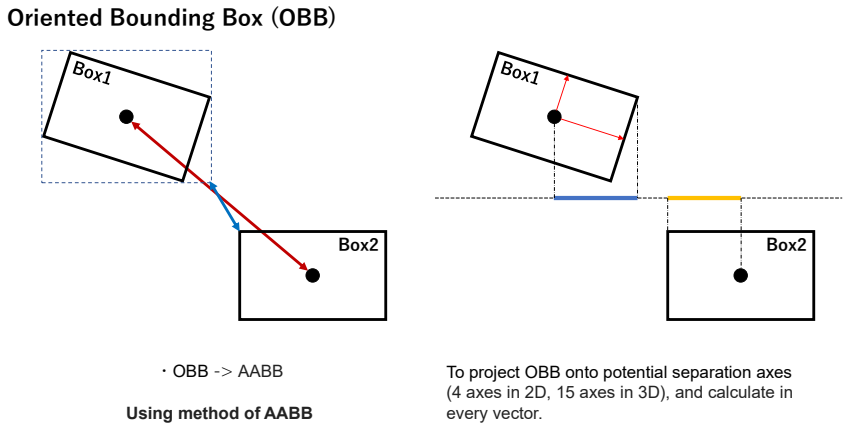
In this section, we acquire spatio knowledge, which is used as likelihood in formula 1. Described in Section 2.3, we treat adjacencies detected by distance between objects as the knowledge.

We use information about where the object is located. Many datasets for the machine learning methods use bounding boxes to represent the position of objects. The bounding box is categorized into Axis-Aligned Bounding Box (AABB) and Oriented Bounding Box (OBB) [17]. The AABB is a bounding box that is parallel to coordinate axes. It is simple because it is not rotated but may have wasted space. The OBB is a bounding box that is not parallel to coordinate axes. It has rotation and can enclose without waste.

As methods of detecting adjacencies between bounding boxes, we use Euclidean distance between centers of boxes and the distance between boxes. The distance of AABB-to-AABB is calculated by using the corners and faces because it does not rotate, as illustrated



(a) Distance of AABB-to-AABB



(b) Distance of OBB-to-OBB

Figure 3: Image diagram of calculating distance between bounding boxes

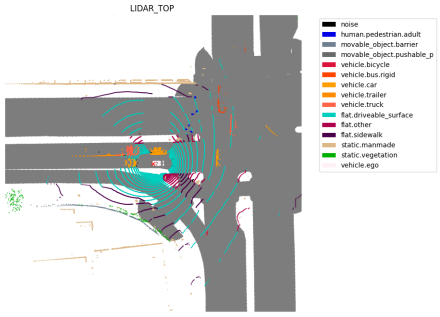
in Figure 3a. In the distance of OBB-to-OBB, there is a way to use the AABB formula after transforming, as shown in Figure 3b. While this has the advantage of a concise calculation, it has the disadvantage of not providing accurate information. However, calculating distance between OBBs requires a lot of computation because of principal component analysis, that considered in 15 axes in 3D [18]. In this thesis, in terms of acquiring statistical knowledge, we obtain distance after converting OBB to AABB.

Table 1: List of main class in nuScenes dataset

Index	Class
1	car
2	truck
3	construction_vehicle
4	bus
5	trailer
6	barrier
7	motorcycle
8	bicycle
9	pedestrian
10	traffic_cone

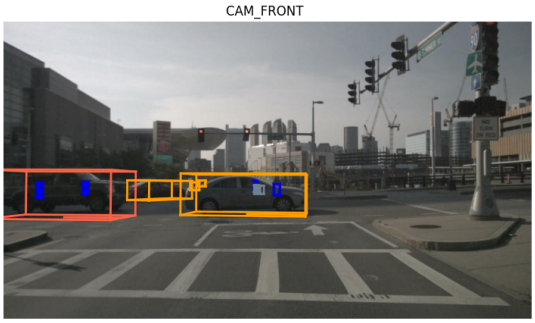
In the following, we describe the example of spatio knowledge using the dataset for the machine learning. We use nuScenes [19], a dataset for autonomous driving, acquired data of real space using images, LIDAR, and RADAR. It consists of 1000 scenes, each constructed with 40 frames, and is annotated with 3D boxes in each data.

In the dataset, we use 3D information based on LIDAR point cloud. As the annotation data, it is given 3D OBBs and 10 object classes in Table 1. Using the tools provided in the dataset, we visualize the actual data, Figure 4a shows LIDAR point cloud from bird's-eye view, Figure 4b shows LIDAR point cloud in images, and Figure 4c shows 3D boxes in images.



(a) Bird's-eye view images

(b) Front camera (point cloud)



(c) Front camera (bounding box)

Figure 4: The actual data in nuScenes

Table 2: Acquired spatio knowledge

target \ adjacent	1	2	3	4	5	6	7	8	9	10
1	0.730	0.085	0.004	0.013	0.008	0.057	0.010	0.004	0.057	0.033
2	0.438	0.217	0.012	0.010	0.099	0.100	0.005	0.003	0.047	0.069
3	0.086	0.053	0.214	0.003	0.023	0.350	0.007	0.003	0.062	0.198
4	0.416	0.061	0.005	0.141	0.009	0.113	0.005	0.010	0.165	0.075
5	0.153	0.373	0.019	0.006	0.217	0.105	0.000	0.003	0.075	0.050
6	0.050	0.017	0.013	0.003	0.005	0.789	0.001	0.001	0.037	0.085
7	0.365	0.035	0.010	0.005	0.000	0.025	0.494	0.020	0.040	0.005
8	0.202	0.030	0.008	0.015	0.008	0.053	0.030	0.224	0.379	0.052
9	0.084	0.013	0.004	0.007	0.006	0.062	0.002	0.011	0.784	0.029
10	0.066	0.027	0.017	0.005	0.005	0.192	0.000	0.002	0.039	0.648

Table 2 shows the acquired spatio knowledge using 850 scenes for training data in nuScenes dataset that obtained in Boston. We make the threshold of adjacent distance of bounding boxes to 3[m] and detect which object classes are adjacent to the target. We treat the number of adjacencies between classes divided by the total as an adjacent probability between each class. In Table 2, the adjacent probability that the target car, in Column, is adjacent to the class pedestrian, in Row, is 5.7%. The properties of the adjacency probabilities in Table 2 is;

- High probability of same-class adjacency, 73% for cars, 78% for pedestrians.
- As the adjacent class to the traffic cone, the associated class follows, traffic cone (65%) and barrier (19%).
- As the adjacent class to the construction vehicle, the associated class follows, barriers (35%), construction vehicle (21%), and traffic cone (20%).
- Different adjacencies can be obtained even between two-wheeled vehicle: the adjacency of motorcycle is motorcycle and car, the adjacency of bicycle is pedestrian and bicycle.
- There are 0 cases of between trailers and motorcycles.

Thus, we have the acquired spatial knowledge that extracts features that occur in real space.

3.2 Method for Bayesian Estimation

In this section, we propose the method calculating the posterior probability in Bayesian estimation, which adds spatio knowledge to the prediction from existing object recognition methods. Since every object has a predicted score, the score of the maximum likelihood class can be said to be certainty of the object. That is, the object is more likely to be the class if the predicted score is high, and the object is less likely to be the class if the predicted score is low. Therefore, we consider Bayesian estimation, which uses not only spatio knowledge but also the predicted score of the adjacent objects as the likelihood.

We define the likelihood of the formula 1 using information of the set of detected adjacent objects Z. We focus on the predicted class of one object $O \in Z$ as the surrounding circumstance, defined as $z \in X$. In addition, we use the predicted score of z in the object O, defined as $P_O(z)$, and the adjacent probability $g(z|x)$ that class x is adjacent to z. Furthermore, compared to directly information of the adjacent object O, we use indirect. This means that we use the probability that the object is not predicted class, and the adjacent probability that class x is not adjacent to z. Therefore, we use the likelihood $G(z|x)$ in Bayesian estimation is

$$G(z|x) = P_O(z) \times g(z|x) + (1 - P_O(z)) \times (1 - g(z|x)). \tag{2}$$

In the case that there are multiple adjacent objects in Z, Bayesian estimation is repeated by treating the posterior probability as a new prior probability.

As an example, it considers its use in in the binary classification of car and pedestrian $\in X$. We treat a target object, predicted as a predicted score of car, P(car), is 0.9 and of pedestrian, P(pedestrian), is 0.1. And an adjacent object $Z = \{O\}$, predicted as $P_O(car)$ is 0.6 and $P_O(pedestrian)$ is 0.4. In addition, as acquired spatio knowledge for simplicity, the adjacent probability that car is next to car: $g(car|car)$ is 0.7 and the adjacent that pedestrian is next to car: $g(pedestrian|car)$ is 0.2. In these conditions, as the corresponding evidence, we treat that the predicted class of the adjacency, z, is car and of the predicted score $P_O(z)$ is 0.6. We calculate a likelihood and a posterior probability

of the target in:

$$
\begin{aligned}
G(z|car) &= P_O(car) \times g(car|car) + (1 - P_O(car)) \times (1 - g(car|car)) \\
&= 0.6 \times 0.7 + (1 - 0.6) \times (1 - 0.7) = 0.54 \\
G(z|pedestrian) &= P_O(car) \times g(car|pedestrian) \\
&\quad + (1 - P_O(pedestrian)) \times (1 - g(car|pedestrian)) \\
&= 0.6 \times 0.2 + (1 - 0.6) \times (1 - 0.2) = 0.44
\end{aligned}
$$

$$
\begin{aligned}
P(car|z) &= \frac{G(z|car) \times P(car)}{\sum_{x' \in X} G(z|x') \times P(x')} \\
&= \frac{0.54 \times 0.9}{\sum_{x' \in X} G(z|x') \times P(x')} = 0.917 \\
P(pedestrian|z) &= \frac{G(z|pedestrian) \times P(pedestrian)}{\sum_{x' \in X} G(z|x') \times P(x')} \\
&= \frac{0.44 \times 0.1}{\sum_{x' \in X} G(z|x') \times P(x')} = 0.083.
\end{aligned}
$$

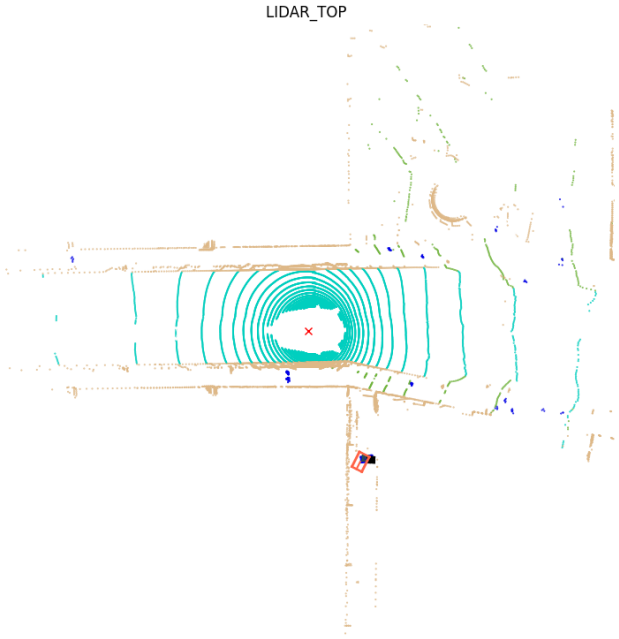
Through adding information of adjacent objects Z, the score of car recalculate from 0.9 to 0.917 and of pedestrian recalculate from 0.1 to 0.083.

3.3 Example of Applying Bayesian Estimation

An example of applying Bayesian estimation using actual data is shown in this section.

We use the TransFusion model [20] as an object recognition method using a machine learning. The TransFusion model is a model for object recognition using the deep learning method and is listed as one of the top accuracy rankings for the Object Detection task on the nuScenes dataset [21]. The Object Detection task for 3D point cloud is a task that predicts bounding boxes of objects in space and a predicted score for each object class.

Figure 5 shows an example of predictions from TransFusion. We set the distance of detecting adjacent objects of the target to 3[m] same as in acquiring spatio knowledge. The target is a prediction that a predicted class is pedestrian, same as the ground truth, and a predicted score of the class is 0.3373, shown as a black box in Figure 5a. In addition, as adjacent objects, 3 pedestrians, which are correct predictions, and 1 car, which is an incorrect prediction, are detected as show 3 blue boxes and 1 red box in Figure 5a. In this situation, we calculate a posterior probability using Bayesian estimation. As a result, shown in Table 3, the probability of pedestrian is increased from 0.3373 to 0.9155 and the other decreased through adding information of adjacent objects. By Bayesian estimation, we can verify the confidence of prediction and increase the score of the correct class. Moreover, it suggests that we make wrong predictions from the machine learning methods change the maximum likelihood class through adding information of adjacent objects.



(a) Prediction from TransFusion

(b) Annotaion data in nuScenes

Figure 5: An example of applying Bayesian estimation using spatio knowledge

Table 3: Prior and posterior probabilities of the target object in the example applying using spatio knowledge

Class	Prior Prob	Posterior Prob
car	0.0310	0.0010
truck	0.0030	0.0001
construction_vehicle	0.0133	0.0005
bus	0.0080	0.0013
trailer	0.0142	0.0005
barrier	0.2725	0.0058
motorcycle	0.0531	0.0012
bicycle	0.1030	0.0708
pedestrian	0.3373	0.9155
traffic_cone	0.1647	0.0034

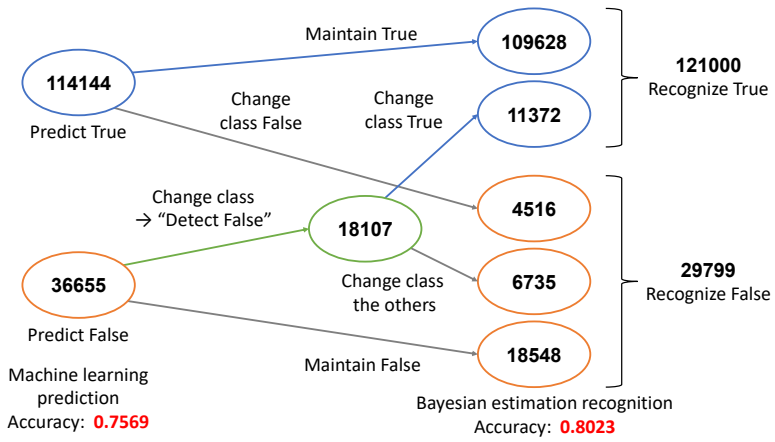


Figure 6: Evaluation of Bayesian estimation using spatio knowledge

3.4 Evaluation of Bayesian Estimation Method using Spatio Knowledge

In this section, it is shown the evaluation when applied to validation data of the dataset. In below, we refer to a predicted class by TransFusion method and a maximum likelihood class of the posterior probability from Bayesian estimation to a prior class and a posterior class.

Figure 6 shows the flow of numbers of true-false between a prior class and a posterior class. The left part in the figure show the results of TransFusion method; 114144 objects were correctly predicted and 36655 objects were wrong. The other part in the figure show the number of objects applying Bayesian estimation. Among 114144 objects, there were 109628 objects that maintained the correct class after applying Bayesian estimation, and 4516 objects that led to the wrong class. For the 36655 objects where a machine learning made wrong predictions, the predicted class did not change in 18548 cases, but the predicted class changed in 18107 cases, of which 11372 could be corrected. Therefore, there are 121000 predictions that were correctly recognized by Bayesian estimation, indicating that accuracy is improved from 75.7% to 80.2%. Here, in the 18107 cases where the predicted class changed, we can say that the prediction by the machine learning was not appropriate, i.e., we were able to detect False predictions. From this, we were able to do so with about half of the predictions that were wrong by the machine learning.

In terms of verifying confidence of predictions, for the correct prediction in the machine learning, we expect that maintain the maximum likelihood class, blue-to-blue in Figure

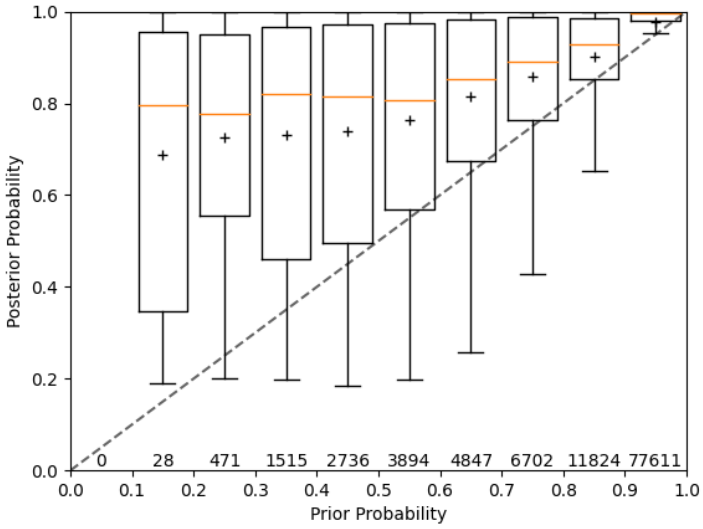


Figure 7: Boxplot of predicted score using spatio knowledge

6, and raise the predicted score. Therefore, we examine change in predicted scores before and after applying Bayesian estimation in 109628 predictions in Figure 6. Figure 7 is a boxplot shows a correlation between score of a prior class in a horizontal axis and of a posterior class in a vertical axis. In Figure 7, each box includes the prediction in the range of predicted scores, each value below the box shows the number of predictions in the range, and the dashed line represents the position where a prior and a posterior score are equal. Figure 7 shows that means of the predictions are above the dashed line in all bands and most of the third quartiles are above.

In the other hand, on the pattern of a predicted class changes, it shows a histogram about predicted score in Figure 8. In this histogram, bin in a horizontal axis is separated the predicted score in 0.05 and a vertical axis shows the number of patterns in each bin. The pattern of a predicted class changes is 11372 cases that a posterior class correct from being a prior class incorrect and 4516 cases that a posterior class incorrect from being a prior class correct. Figure 8 shows that the former pattern tends to appear more often in low predicted scores and the latter tends to appear in high.

As shown above, the method can verify the confidence for many objects. Nevertheless, there are still many predictions for which Bayesian estimation cannot be used to correct

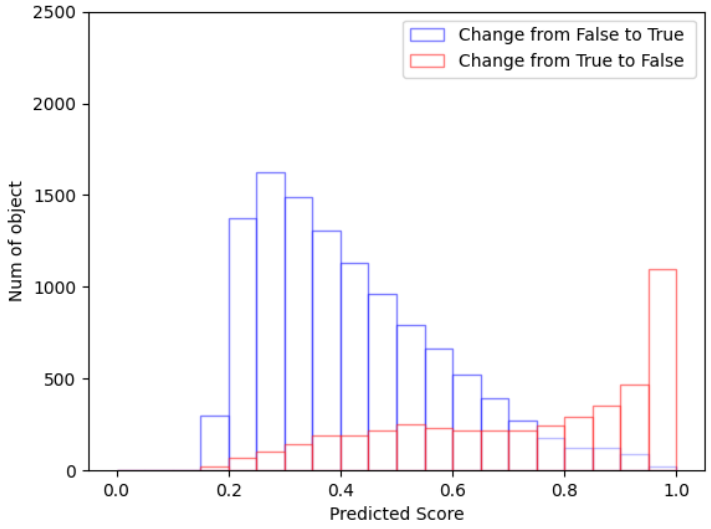


Figure 8: Histogram on the pattern of predicted class changes using spatio knowledge

for errors. In addition, shown as Figure 7, in terms of increased confidence, Bayesian estimation using spatio knowledge has limitations. Therefore, we will try to apply Bayesian estimation by adding not only spatio knowledge but also temporal knowledge.

4 Bayesian Estimation Method using Spatio-temporal Knowledge

Compared to Bayesian estimation using spatio knowledge in Section 3.2, we use Bayesian estimation using both spatial and temporal knowledge in this section.

4.1 Spatio-temporal Knowledge from Streaming Point Cloud

In Section 3.1, we obtained surrounding circumstances within the frame where the target object exists, but in data with a time axis, there are dependency between frames, such as continually existing. Therefore, in addition to the adjacency in space, consider Bayesian estimation in temporal dependence as a likelihood.

We use information based on temporal sequence of objects to acquire temporal knowledge, as described in Section 2.3. In the nuScenes dataset, images and LIDAR and RADAR point clouds are acquired simultaneously with real space. Therefore, even if an object cannot be observed on LIDAR, there is information on the location of the object on LIDAR, because it can be observed on an image or a radar. For this reason, we can track objects on the point cloud in the corresponding frame even if the points constituting an object do not exist. Hence, as temporal knowledge, we handle information that exists in the previous frame and continuously exists in the current frame, and that does not exist in the previous frame and exists in the current frame, as shown in Figure 9.

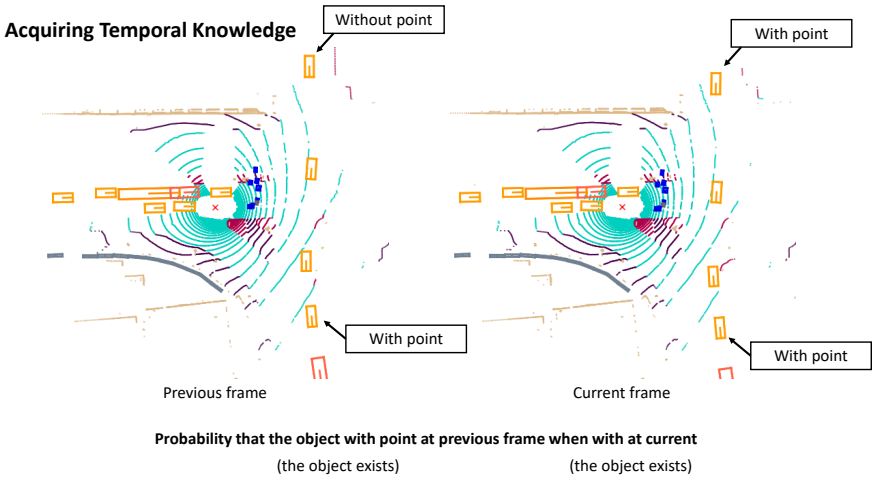


Figure 9: An example of acquiring temporal knowledge

Table 4: Acquired temporal knowledge

Index	Prob
car	0.8932
truck	0.9355
construction_vehicle	0.9477
bus	0.9562
trailer	0.9530
barrier	0.9218
motorcycle	0.9169
bicycle	0.9077
pedestrian	0.9201
traffic_cone	0.8546

Table 4 shows the acquired temporal knowledge using 850 scenes for training data in nuScenes dataset that obtained in Boston, as well as spatio knowledge. As adjacent probability in each category, we treat the number of observing objects in the previous frame divided by the number of objects exists in the frame. Under these conditions, as a temporal adjacent, it cannot happen that an object present in the frame is a different class in the previous frame. Therefore, unlike Table 2, we show the adjacent probabilities only between the same class in Table 4. The properties of adjacent probability in Table 4 is that the adjacent probability is high which indicates that most objects will continue to remain as long as they do not enter a blind spot. The probability is lower for car and traffic cone, compared to the other class. Thus, we have acquired temporal knowledge that extracts features that occur in real space.

4.2 Modification on the Bayesian Estimation Method

In this section, we propose the method calculating a posterior probability in Bayesian estimation, which adds temporal knowledge to the result using spatio knowledge. As a temporal adjacent dependency, we detect the objects that appear to be the same of the target in the previous frame. As an implementation, we detect objects located within a radius of 1[m] from coordinates of the target in the previous frame. We use Bayesian estimation to obtain a posterior score using the maximum likelihood class of detected object and temporal knowledge as the likelihood. In other words, we use the adjacent probability which the object class detected in the previous frame and the the probability for the extra event that the object class not detected, defined as $G_p(x)$. In this Bayesian estimation, we treat a posterior probability calculated using spatio knowledge as a prior using spatio-temporal knowledge. Therefore, the method of Bayesian estimation using the previous frame is

$$P_p(x|z) = \frac{G_p(x) \times P(x|z)}{\sum_{x' \in X} G_p(x') \times P(x'|z)}$$

$$G_p(x) = \begin{cases} g(x) & \text{(exists x class in the previous)} \\ 1 - g(x) & \text{(do not exist x class in the previous)} \end{cases}$$

, where the $g(x)$ is adjacent probability of the class x. We recalculate predicted scores using the spatio-temporal knowledge through using results using spatio knowledge to a prior probability.

By incorporating information from the previous frame, we increase confidence in predicted score incorporating temporal dependency for continuous data. For data that is not real-time and has some time width, we can treat information from the next frame as well as the previous. Therefore, for predictions in the current frame, we can calculate Bayesian estimation using information from before and after the frame in time by detecting objects that are considered the same. We recalculate predicted scores using information of the previous and the next frame, through applying a posterior probability using information of the previous frame to a prior. This means that the method of Bayesian estimation using

Table 5: Posterior probability of the target object in the example applying using spatio-temproal knowledge

Class	Prior	Posterior		
		Spatio	Spatio-temporal (previous frame)	Spatio-temporal (previous/next frames)
car	0.0310	0.0010	0.0001	0.0000
truck	0.0030	0.0001	0.0000	0.0000
construction_vehicle	0.0133	0.0005	0.0000	0.0000
bus	0.0080	0.0013	0.0001	0.0000
trailer	0.0142	0.0005	0.0000	0.0000
barrier	0.2725	0.0058	0.0005	0.0000
motorcycle	0.0531	0.0012	0.0001	0.0000
bicycle	0.1030	0.0708	0.0077	0.0008
pedestrian	0.3373	0.9155	0.9908	0.9990
traffic_cone	0.1647	0.0034	0.0006	0.0001

the previous and the next frame is

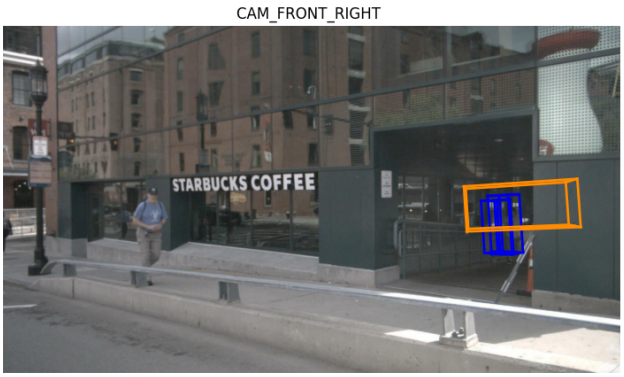
$$P_{pn}(x|z) = \frac{G_n(x) \times P_p(x|z)}{\sum_{x' \in X} G_n(x') \times P_p(x'|z)}$$

$$G_n(x) = \begin{cases} g(x) & \text{(exists x class in the next)} \\ 1 - g(x) & \text{(do not exist x class in the next)} \end{cases}$$

Because of this, in addition to calculation of a posterior probability only using spatio knowledge in Section 3.2, we evaluate the method of Bayesian estimation by comparing using the previous and both the previous and the next to only machine learning.

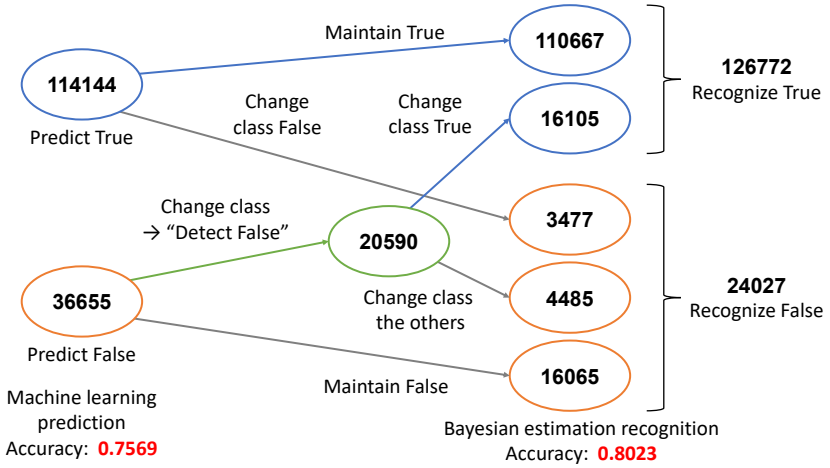
We consider adding temporal knowledge to the result of recognition using spatio in Figure 5. In the previous frame, we find a pedestrian prediction, same as the ground truth, as Figure 10a. In the next frame, we find a pedestrian and a trailer prediction as Figure 10b. Table 3 shows an example of applying Bayesian estimation using both knowledge. Using spatio-temporal knowledge, the posterior probability of pedestrian in the target increases the posterior probability of pedestrian 0.9990 from 0.3373 in the prior than 0.9155 using spatio. Even included the incorrect prediction of trailer in the next frame, we can increase the posterior probability. Therefore, Table 3 shows that it is helpful to recalculate predicted scores using spatio and temporal knowledge.

(a) Prediction in the previous frame

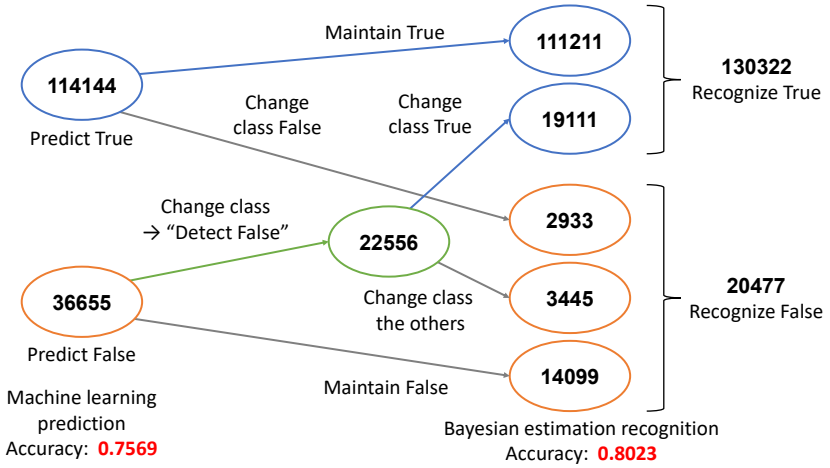


(b) Prediction in the next frame

Figure 10: An example of applying Bayesian estimation using spatio-temproal knowledge



(a) Using the previous frame



(b) Using the previous and the next

Figure 11: Evaluation of Bayesian estimation using spatio-temporal knowledge

4.3 Evaluation

In this section, it is shown the evaluation when applied to validation data of the dataset. We set the distance searching objects considered to be the same to 1[m].

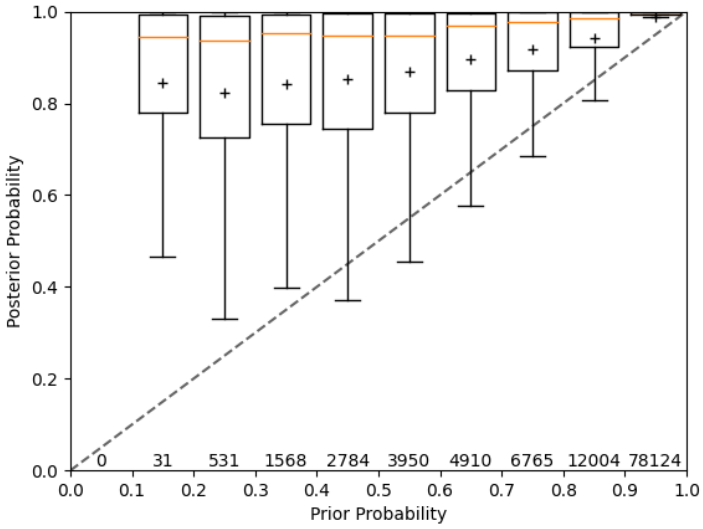
Figure 11 shows the flow of numbers of true-false between a prior class and a posterior class, same as in Figure 6. In using the previous frame, Figure 11a shows that, of the 114144 objects that were correctly predicted by machine learning, there were 110667 objects that maintained the correct class after applying Bayesian estimation, and 3477 objects that led

Table 6: Notable difference between using spatio and spatio-temporal

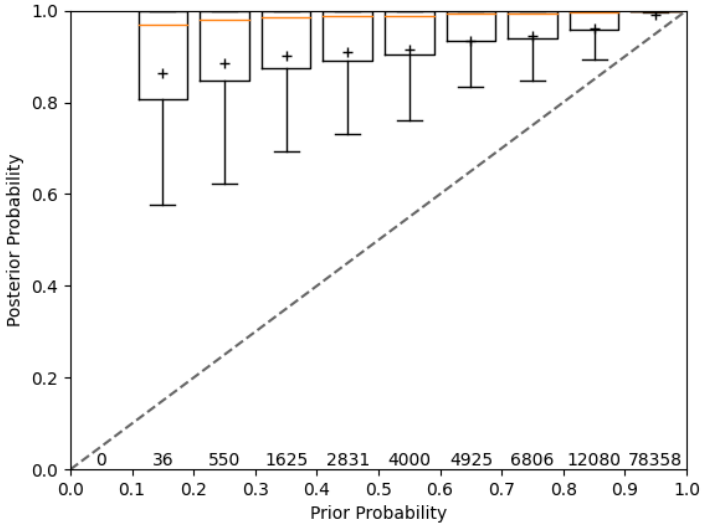
	Unchanged	False detected	Corrected	Accuracy
Spatio	109628	18107	11372	80.23%
Spatio-temporal (previous frame)	110667	20590	16105	84.06%
Spatio-temporal (previous/next frames)	111211	22556	19111	86.42%

to the wrong class. For the 36655 objects where machine learning made wrong predictions, the predicted class did not change in 16065 cases, but the predicted class changed in 20590 cases, of which 16105 could be corrected. Therefore, there are 126772 predictions that were correctly recognized by Bayesian estimation, indicating that accuracy is improved from 75.7% to 84.1%. Similarly, In using the previous and the next frame, Figure 11b shows that, of the objects correctly predicted by machine learning, there were 111211 objects that maintained the correct class after applying Bayesian estimation, and 2933 objects that led to the wrong class. For the predictions machine learning made wrong, the predicted class did not change in 14099 cases, but the predicted class changed in 22556 cases, of which 19111 could be corrected. Therefore, there are 130322 predictions that were correctly recognized by Bayesian estimation, indicating that accuracy went from 75.7% to 86.4%.

Table 6 shows the notable difference of the Figure 6 and 11. In comparison with using spatio knowledge, the number of changing the maximum likelihood class, red-to-green in Figure 11, means can find out what is wrong prediction, increases from 18107 to 22556. This leads to that the number that maintain False, red-to-red in Figure 11, decraeses more. Moreover, using information of the previous and the next frames, we correct 19111 predictions, red-to-blue in Figure 11, than 11372 using spatio knowledge. Additionally, the number that maintain True, blue-to-blue in Figure 11, increases from 109628 with spatio knowledge to 111211 with spatio-temporal knowledge of the previous and the next frames. This means the number that changes the class False from True, blue-to-red in Figure 11, decreases. We can capture dependencies over time in predictions because of not only using the previous frame but also the next. From the all, the accuracy, based on 75.69% from



(a) Using the previous frame



(b) Using the previous and the next frames

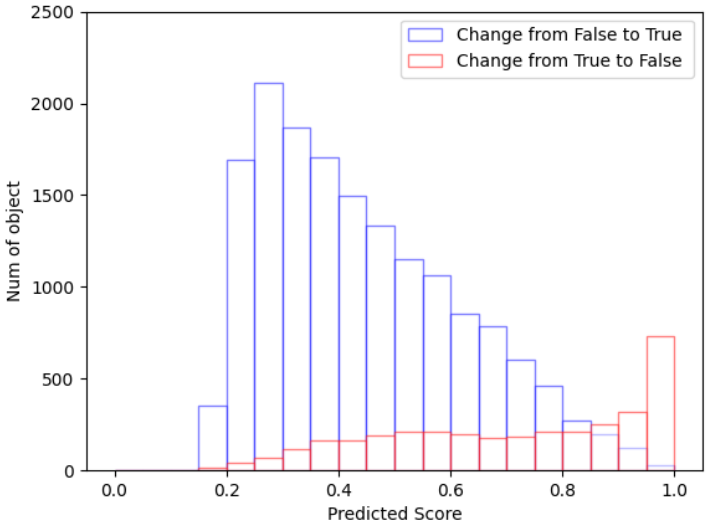
Figure 12: Boxplot of predicted score using spatio-temporal knowledge

the method of machine learning, is more increased 86.42% with spatio-temporal knowledge of the previous and the next frames than 80.23% in spatio knowledge.

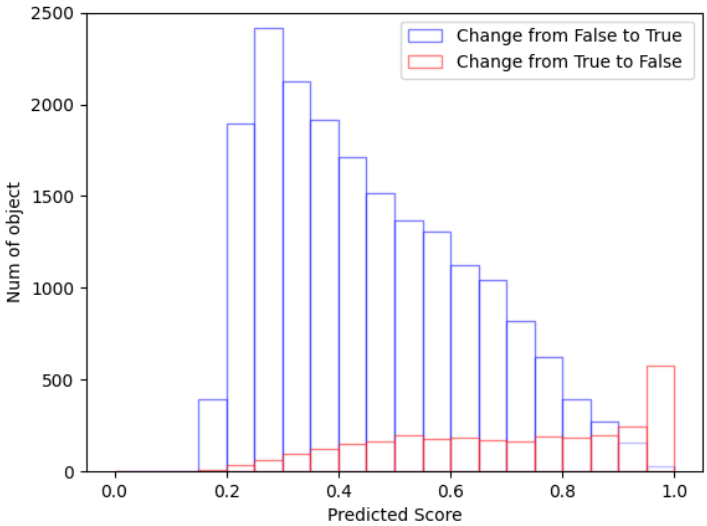
Compared to Figure 7, in order to verify confidence of predictions, Figure 12 is boxplot

shows correlations between scores. In Figure 12a, many predicted objects in all predicted probability bands are above the dashed line, in other words, it shows that an increase in confidence can be achieved for correct predictions. In particular, in Figure 12b, a significant increase occurred in all bands, suggesting that the confidence of the predictions is increased. In addition, through increasing the number of predictions maintaining True, the number of the predictions in all bins, number showed below the box, increased.

Compared to Figure 7, on the pattern of predicted class changes, it shows a histogram about predicted scores in Figure 13. The pattern of predicted class changes is that a posterior class correct from being a prior class incorrect and that a posterior class incorrect from being a prior class correct. Same as Figure 7, Figure 13 shows that the former pattern tends to appear more often in low predicted score and the latter tends to appear in high. In using spatio-temporal knowledge, the number of the former pattern increases, and the number of the latter pattern decreases. In terms of verifying confidence of predictions, it is more useful that use the spatio-temporal knowledge and the information of the previous and the next frames. because of the better accuracy and the higher posterior score.



(a) Using the previous frame



(b) Using the previous and the next frames

Figure 13: Histogram on the pattern of predicted class changes using spatio-temporal knowledge

5 Conclusion

In this thesis, we proposed a Bayesian estimation method in terms of verifying confidence of predictions in different methods for object recognition using the machine learning. As Bayesian estimation method using spatial knowledge, we proposed an example of constructing knowledge for use as a likelihood and applying to actual predictions of the dataset representing real space. Moreover, we evaluated the method using validation data of the dataset and increased accuracy than the method using machine learning. Furthermore, as using spatio-temporal knowledge, we proposed an example observed an increase in confidence of the prediction in evaluate.

As a method independent of the machine learning, we confirmed an increase in accuracy using the method compared to by only machine learning. This means that it is useful to use the proposed method to predict object recognition regardless of existing machine learning methods. We suggest that more increasing accuracy from the method of machine learning can make more reliable predictions through this method.

As possible improvements to the proposed method, although using the previous and the next frame, we will use the previous and the second previous or more previous as temporal dependency. It would also be possible to use information not considered in this thesis as a surrounding situation or to apply in different situations.

Acknowledgments

I would like to express my gratitude to all those who supported this master's thesis. I would like to express the deepest appreciation to Associate Professor Shin'ichi Arakawa of Osaka University. My deepest appreciation also goes to Professor Masayuki Murata of Osaka University for his insightfull and kindfull advice. I would like to express my gratitude to gives me constructive comments for Associate Professor Yuichi Ohsita, Assistant Professor Daichi Kominami, Assistant Professor Tatsuya Otoshi, and Specially Appointed Assistant Professor of Osaka University Masaaki Yamauchi. I would like to thank all of the member of Advanced Network Architecture Research Laboratory.

References

[1] W. Kritzinger, M. Karner, G. Traar, J. Henjes, and W. Sihn, "Digital twin in manufacturing: A categorical literature review and classification," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022, Jun. 2018.

[2] D. Jones, C. Snider, A. Nassehi, J. Yon, and B. Hicks, "Characterising the digital twin: A systematic literature review," *CIRP Journal of Manufacturing Science and Technology*, vol. 29, pp. 36–52, May 2020.

[3] C. Boje, A. Guerriero, S. Kubicki, and Y. Rezgui, "Towards a semantic construction digital twin: Directions for future research," *Automation in Construction*, vol. 114, pp. 1–16, Jun. 2020.

[4] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, pp. 1–18, Jun. 2022.

[5] Y. Liu, P. Sun, N. Wergeles, and Y. Shang, "A survey and performance evaluation of deep learning methods for small object detection," *Expert Systems with Applications*, vol. 172, pp. 1–12, Jun. 2021.

[6] D. Fernandes, A. Silva, R. Névoa, C. Simões, D. Gonzalez, M. Guevara, P. Novais, J. Monteiro, and P. Melo-Pinto, "Point-cloud based 3D object detection and classification methods for self-driving applications: A survey and taxonomy," *Information Fusion*, vol. 68, pp. 161–191, Apr. 2021.

[7] S. Qi, X. Ning, G. Yang, L. Zhang, P. Long, W. Cai, and W. Li, "Review of multi-view 3D object recognition methods based on deep learning," *Displays*, vol. 69, pp. 1–12, Sep. 2021.

[8] S.-W. Yang and C.-C. Wang, "On solving mirror reflection in lidar sensing," *IEEE/ASME Transactions on Mechatronics*, vol. 16, no. 2, pp. 255–265, Apr. 2011.

[9] W. Ma, J. Chen, Q. Du, and W. Jia, "Pointdrop: Improving object detection from sparse point clouds via adversarial data augmentation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Jan. 2021, pp. 10 004–10 009.

[10] T. Raj, F. Hashim, A. B. Huddin, M. F. Ibrahim, and A. Hussain, "A survey on lidar scanning mechanisms," *Electronics*, vol. 9, no. 5, pp. 741–765, Apr. 2020.

[11] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (tof) cameras: A survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, Sep. 2011.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788.

[13] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 652–660.

[14] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 4490–4499.

[15] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Oct. 2020, pp. 474–490.

[16] B. Ristic, C. Gilliam, M. Byrne, and A. Benavoli, "A tutorial on uncertainty modeling for machine reasoning," *Information Fusion*, vol. 55, pp. 30–44, Mar. 2020.

[17] C. Tu and L. Yu, "Research on collision detection algorithm based on aabb-obb bounding volume," in *Proceedings of the International Workshop on Education Technology and Computer Science (ETCS)*, Mar. 2009, pp. 331–333.

[18] A. Hu and Y. He, "Research on hybrid collision detection algorithm based on separation distance," *Journal of Physics: Conference Series*, vol. 2258, pp. 1–7, Apr. 2022.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 11 621–11 631.

[20] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 1090–1099.

[21] "nuScenes detection task Benchmark," https://www.nuscenes.org/object-detection, (Accessed: 2022-04-05).