Driving Video Prediction based on Motion Estimation of 3D Objects using a Stereo Camera System

Takuya Umemura, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Gokiso, Show, Nagoya 466-8555, Japan

Keywords: Future Image Prediction, Driving Assist, Deep Image Prior.

Abstract: In this paper, we propose a method to synthesize future images in a driving scene using a stereo camera system fitted on vehicles. In this method, three-dimensional (3D) objects in a driving scenario, such as vehicles, buildings, and humans, are reconstructed by a stereo camera system. The reconstructed objects are separated by semantic image segmentation based on 2D image information. Furthermore, motion prediction using a Kalman filter is applied to each object. 3D objects in future scenes are rendered using this motion prediction. However, some regions, which are occluded in the input images, cannot be predicted. Therefore, an image inpainting technique is used for the occluded regions in the input image. Experimental results show that our proposed method can synthesize natural predicted images.

1 INTRODUCTION

In recent years, various information processing technologies, such as automated driving and driving assistance, have been installed in vehicles, and these techniques have been researched and developed extensively. Typically, in such technologies, various data are obtained from different kinds of sensors, such as cameras and 3D sensors, and the system assists drivers based on this data.

These systems inherently have a delay because it takes some time to acquire and process the data. Thus, we cannot obtain “real” real-time information through the sensors. Moreover, recent systems obtain data from not only installed sensors but also surrounding sensors, e.g., sensors on the other vehicles and street cameras, through the network. Similarly, remote driving systems developing recently receive information from the vehicle through the network to drive the vehicle from far away. In these cases, the delay of the data becomes larger since data transmission through the network requires lots of time. Therefore, the delay cannot be ignored for achieving safety operations. If the system has an information delay, it is difficult to replicate the acquired information in real-time operations. For example, in cases where the vehicle should prevent collision with a pedestrian who appears suddenly, a system with a time delay can lead to grave consequences. Therefore, the delay in the feedback time from acquisition to processing of information should be considered as a critical problem to be solved.

To solve this problem, it is vital to minimize the response time of the system. However, the acquisition, processing and transmission of information takes a specific duration in principle, which cannot be 0 s. Therefore, we attempt to solve this problem by predicting the transition of the whole scene instead of shortening the delay time. If such a scene prediction system is developed, the delay problem can be avoided. That is, even when the information from the sensor has a delay, the predicted data can correspond to the real-time data, and the predicted data can be utilized in real-time driving. In this research, we focus on images taken by the camera fitted on the vehicles. In this paper, we propose an image prediction technique, using images taken by an in-vehicle camera of the driving scene.

Recently, several video image prediction methods have been proposed (Vondrick et al., 2016; Vondrick and Torralba, 2017; Finn et al., 2016). In these methods, deep neural networks are used for predicted image synthesis. Notably, generative adversarial network (GAN) is one of the most effective architectures to synthesize images. In the image generating network, a model that describes scene transitions is implicitly trained in the network, which synthesizes the future image from the current input image, based on this implicit model. Although the network predicts the future image when the model fits the input

Umemura, T., Sakaue, F. and Sato, J.
Driving Video Prediction based on Motion Estimation of 3D Objects using a Stereo Camera System.
DOI: 10.5220/0009349108590866
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 859-866
ISBN: 978-989-758-402-2; ISSN: 2184-4321

scene, many images are required to train the network correctly. Furthermore, these methods mainly focus on the partial transition of the scene. Therefore, the process may not be suitable in predicting a driving scene because the entire image significantly changes depending on the camera motion in the driving scene. Furthermore, such case-based methods are not suitable for predicting unusual events because it is not easy to collect training data of unusual scenes.

In contrast, our proposed method has an explicit scene transition model for image prediction. The model is based on a physical model as opposed to a training case, which we can control and explain; thus, our proposed method does not require a training image set to optimize the image prediction. Therefore, our proposed method can predict future images even in unusual events, e.g., traffic accidents.

2 IMAGE PREDICTION IN DRIVING SCENE

2.1 System Construction

For image prediction in a driving scene, we consider the type of scene taken by the cameras fitted on the vehicles. When the camera runs on the road, the images contain pedestrians, vehicles driving, and buildings around the road. The objects typically move independently, and should be separated to estimate and predict their movement. Because the motion of the object is not 2D but 3D, 3D measurement of the objects is required to predict their movement.

Therefore, we first measure the 3D road scenes using a stereo camera system and predict the scene using the 3D information obtained from them. We separate the measured map into multiple objects, and each object's motion is estimated. From the estimated motions, the object's future position is predicted. Finally, we render a future image from the expected state and measure its 3D shape.

In recent studies, a method using a laser sensor such as light detection and ranging (LiDAR) has been considered for acquiring accurate 3D information. However, we choose a stereo camera system because the cameras can measure dense depth maps in real-time. Additionally, the stereo system can receive 3D shapes and a 2D image, which have much information. Therefore, the stereo camera system is suitable for predicting the driving scene in our research.

Note that the distance measurement results obtained by the stereo camera are less accurate than the result from LiDAR. However, in the synthesis of a future image, it is not a significant issue. As the distant 3D points that are less accurate are projected to a small region in the image, the effect of the accuracy degradation is suppressed in the predicted image.

2.2 3D Reconstruction by the Stereo Camera

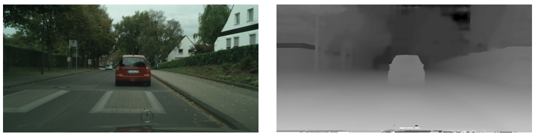
We first explain 3D scene reconstruction using the stereo camera system. The 3D reconstruction of the image is based on epipolar geometry. Let us consider the case when the stereo cameras are calibrated, and camera projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ are computed in advance. When correspondences $\mathbf{m}_1$ and $\mathbf{m}_2$ are detected from each camera image, the 3D point $\mathbf{X}$, which corresponds to the image points is measured as follows:

$$\mathbf{X} = \arg\min_{X} \sum_{i=1}^{2} \|\mathbf{m}_i - \mathbf{P}_i(\mathbf{X})\|^2 \qquad (1)$$

where $\mathbf{P}_i(\mathbf{X})$ is the projection of the point $\mathbf{X}$ to the i-th camera. As shown in the equation, the 3D points are reconstructed by minimizing the reprojection error.

As mentioned above, 3D points are reconstructed when correspondences are detected from stereo images. Therefore, accurate 3D shape reconstruction from the stereo image depends on the accurate detection of corresponding feature points. The correspondence detection is achieved by epipolar geometry using a 1D search as the correspondences are on epipolar lines. Therefore, image rectification is applied to stereo images, which makes the epipolar lines parallel to the horizontal axis. In this case, finding correspondences is equivalent to the estimation of the disparity in each pixel.

In recent studies, the estimation of the disparity is optimized in the entire image, i.e., a 2D regularizing constraint is applied to achieve stable estimation. Especially, semi-global matching (SGM)(Hirschmuller, 2005) is widely used because the method provides a better disparity map at a low computational cost. The technique uses several 1D constraints for reducing computational cost rather than a 2D constraint. Thus, the required computational cost is much smaller than in the case when 2D optimization is used. In our method, the computational cost should be low because our method needs to operate in real-time. Therefore, we use the SGM to compute the disparity map from the stereo images efficiently. Figure 1 shows the computed disparity by SGM. Using the SGM, a disparity map in the driving scene can be estimated, and a dense 3D map is reconstructed from the map.



(a) input image (b) disparity map

Figure 1: Disparity map by SGM.

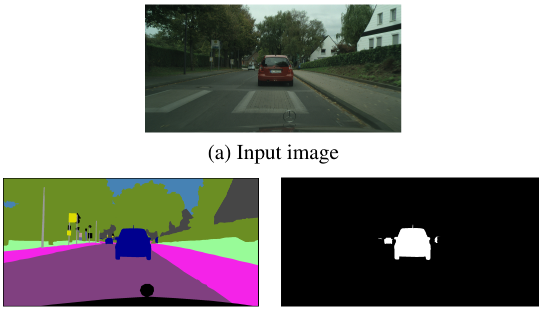
2.3 Object Segmentation

Next, we consider the object segmentation of the measured 3D map. As mentioned, the 3D point cloud from the stereo camera includes different kinds of 3D objects, which move independently. Therefore, it is necessary to segment 3D points to multiple objects to predict the scene transition accurately. The segmentation of the 3D points is often based on the motion of the points. However, when the object motion is inaccurate, this segmentation is challenging. For example, when two or three frames are taken in a short duration are used to estimate the object's motion, the estimated motion has many estimation errors, and it cannot provide enough information to separate the objects. Therefore, a method based on motion is not suitable for this research. Consequently, we classify the object based on 3D image data instead of motion because camera images can be used in our method.

To achieve image segmentation in the driving scene, we use semantic segmentation of images(Sharma et al., 2018). Semantic segmentation adds image labels pixel by pixel, and the pixels are classified into people, roads, and vehicles. In recent years, convolutional neural networks (CNN) are often used for semantic segmentation(Sharma et al., 2018). The CNN can label the pixels at high speed and accuracy. As objects with different labels are considered to perform different motions, it is possible to separate the motions using the semantic segmentation results.

In this research, conditional GAN(Isola et al., 2016) is utilized to achieve semantic segmentation in the driving scene. Using the conditional GAN, natural input images are converted to image labels, and thus image segmentation is accomplished. In this segmentation, a limited number of labels are used as objects, which appear in the road scene are limited. Furthermore, a detailed label is not required in our method because we aim to separate the objects by their motion. Therefore, the labels are integrated into two labels, dynamic and static objects.

The dynamic objects can move independently, and they include humans and vehicles. In the dynamic objects, there are multiple objects, which have different motions. Therefore, detailed segmentation results are preserved because the motion of these objects should be estimated independently. The static objects include buildings and roads. They cannot move by themselves; therefore, the relative motion of the objects is caused by camera motion. Thus, their motion is estimated concurrently as a camera motion. Since the 3D points in the scene correspond to the pixels in the input image, it is possible to label the recovered 3D point group and obtain a separate point group for each object because of the segmentation. An example of image segmentation using CNN is shown in Fig. 2. This example shows that vehicles, pedestrians, and road surfaces are given different labels, and the objects in the scene can be properly separated in Fig. 2 (b). Furthermore, the labels are integrated into static and dynamic objects, as shown in Fig. 2(c).



(a) Input image

(b) Semantic segmentation (c) Segmentation of results into dynamic and static objects

Figure 2: Example of object segmentation by semantic segmentation: (a) shows the input image, (b) shows the semantic segmentation result, and (c) shows the segmentation result of static and dynamic objects. In image (c), dynamic objects are shown by white pixels and static objects by black pixels.

3 MOTION ESTIMATION AND PREDICTION OF 3D OBJECT

3.1 Object Corresponding

Next, we consider a method to estimate the motion data of each separated 3D object. Using image segmentation, it is possible to obtain a 3D position of the separated objects at each time. However, it is necessary to find the corresponding objects for estimating their motion. In this research, SIFT(Lowe, 2004) is used for object matching. The SIFT is a local image feature, and it is often used for detecting matching points. The SIFT feature is robust against illumination change, image rotation, and scaling. In the driving scene captured by vehicle cameras, illumination of the scene changes given the sunshine conditions. Moreover, the scale and rotation of the object image change with the object's movement. Therefore, we expect that these differences can be suppressed using SIFT.

For motion estimation, SIFT points are detected from the images in each frame. The corresponding points are found in the neighboring frame based on the SIFT feature. The corresponding object can be identified using SIFT matching because identical objects have corresponding feature points.

Note that, we obtain not only 2D point correspondences but also 3D point correspondences since the image has 3D information at each point by the stereo method. These 3D correspondences are utilized for motion estimation described in the next section.

3.2 Motion Estimation based on Kalman Filter

Here, we discuss motion estimation. In this estimation, we assume that objects in the scene are rigid. Under this assumption, the motions of the objects are only rotation and translation. Therefore, we only estimate the rotation and the translation of objects from input images.

For this estimation, corresponding SIFT points in the object are used. Let $\mathbf{X}_{t-1}^{j,i}$ denote the i-th 3D point with label j at time $t-1$. The point $\mathbf{X}_{t-1}^{j,i}$ corresponds to the point $\mathbf{X}_t^{j,i}$ at time t. In this case, a rotation matrix $\mathbf{R}_t^{j\prime}$ and translation vector $\mathbf{T}_t^{j\prime}$ between t and $t-1$ can be estimated as follows:

$$(\mathbf{R}_t{}^{j\prime}, \mathbf{T}_t{}^{j\prime}) = \arg\min_{\mathbf{R},\mathbf{T}} \sum_i^n ||\mathbf{X}_t^{j,i} - (\mathbf{R}\mathbf{X}_{t-1}^{j,i} + \mathbf{T})|| \quad (2)$$

Using this estimation, the motion of each object can be predicted.

However, because the measured 3D point is affected by noise, such as the reconstruction error in the stereo method, a reliable estimation cannot be expected using only two frames. Therefore, a Kalman filter(Kalman, 1960) is used to achieve reliable motion estimation and prediction.

The Kalman filter is an iterative estimator that takes an observed value of the current time as an input. The filter sequentially updates the state of the model while providing an estimated value of the current time and a predicted value of the time ahead. In this research, when the transition time interval is 1 min, it is assumed that the transition of each motion parameter can be approximated by a linear form, and this is expressed using a linear Kalman filter.

Let $\mathbf{T}_t^j$ and $\mathbf{a}_t^j$ denote the translation and acceleration of the j-th object at time t. In this case, we assume that the object moves with uniform acceleration approximately, and the prediction model of motion from $\mathbf{T}_{t-1}^j$ to the predicted translation $\hat{\mathbf{T}}_t^j$ is defined as follows:

$$\begin{bmatrix} \hat{\mathbf{T}}_t^j \\ \mathbf{a}_t^j \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{T}_{t-1}^j \\ \mathbf{a}_{t-1}^j \end{bmatrix} \quad (3)$$

where $\mathbf{I}$ denotes a 3×3 identity matrix and $\mathbf{0}$ denotes a 3×3 all zero matrix. In addition, it assumes that the transition of the prediction error covariance matrix $\mathbf{P}_t^j$ is defined as follows:

$$\hat{\mathbf{P}}_t^j = \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{P}_{t-1}^j \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^T + \mathcal{N}(\mathbf{0}, \mathbf{Q}_t) \quad (4)$$

where $\mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ is a zero means normal distribution with a covariance matrix $\mathbf{Q}_t$. According to the estimated translation vector $\mathbf{T}_t^{j\prime}$, all parameters for estimation are updated as follows:

$$\begin{aligned} \mathbf{e}_t &= \mathbf{T}_t^{j\prime} - \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{T}}_t^j \\ \mathbf{a}_t^j \end{bmatrix} && (5) \\ \mathbf{S}_t &= \mathbf{Q}_t + \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \hat{\mathbf{P}}_t^j \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}^T && (6) \\ \mathbf{K}_t &= \hat{\mathbf{P}}_t^j \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}^T \mathbf{S}_t^{-1} && (7) \\ \mathbf{T}_t^j &= \hat{\mathbf{T}}_t^j + \mathbf{K}_t \mathbf{e}_t && (8) \\ \mathbf{P}_t^j &= (\mathbf{I} - \mathbf{K}_t \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix}) \hat{\mathbf{P}}_t^j && (9) \end{aligned}$$

The updated $\mathbf{T}_t^j$ is a motion of the j-th object considering past frames and $\hat{\mathbf{T}}_{t+1}^j$ is the predicted translation in the next frame.

Similar to the estimation of the translation, a rotation matrix $\mathbf{R}_t^j$ and $\hat{\mathbf{R}}_{t+1}^j$ is estimated. In this estimation, the matrix $\mathbf{R}_t^j$ is converted to a rotation vector $\mathbf{r}_t^j$ by Rodrigues' formula. Then, the Kalman filter is applied to the rotation vector, and the rotation $\mathbf{R}_t^j$ is estimated.

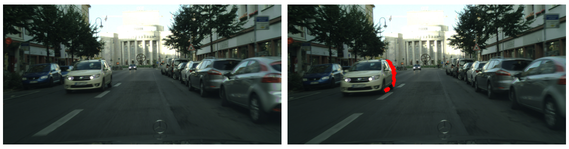
A Kalman filter is applied to the translation and rotation estimation according to the model; consequently, object motion can be estimated while suppressing the effect of noise. Furthermore, it is possible to predict the motion of the object by finding the parameters of the next frame based on the Kalman filter. From these predicted parameters, the 3D point $\mathbf{X}_{t+1}^j$ is predicted as follows:

$$\mathbf{X}_{t+1}^j = \hat{\mathbf{R}}_{t+1}^j \mathbf{X}_t^j + \hat{\mathbf{T}}_{t+1}^j \quad (10)$$

Using these predicted shapes in the next frame, a future image can be rendered.

3.3 Image Inpainting by Deep Image Prior

The future image, which is synthesized using the method mentioned above is constructed based on the



(a) input image (b) predicted image

Figure 3: Example of a missing area caused by an occlusion in predicted images.

observation image of the current time. Therefore, the appropriate image cannot be synthesized when several regions in the future images cannot be observed at the current time due to occlusions. For example, Fig. 3 shows a scene where the vehicle fitted camera is moving forward, and the vehicle in the left part of the image is also moving forward. In this case, the region occluded by the front vehicle in the current frame cannot be rendered and is not in the predicted image, as shown in the red area. Therefore, to synthesize a natural predicted image, it is necessary to interpolate such missing regions correctly.

A method to interpolate such a missing image is called inpainting, and various methods have been proposed(Yu et al., 2018). In our system, the missing region is interpolated based on a deep image prior(Ulyanov et al., 2018). The deep image prior is one of the image synthesis methods of CNN. The image synthesis in the deep image prior is based on the network architecture combined with some image data rather than training images. Therefore, the method does not need training on CNN. Thus, an image can be interpolated without training the network using the deep image prior.

Let us consider the case where the network architecture f synthesizes an image $f(\mathbf{z}, \mathbf{W})$ similar to an image $\mathbf{x}$ based on input noise $\mathbf{z}$ and parameters $\mathbf{W}$ of the network. Furthermore, the image $\mathbf{x}$ includes a missing region and an image mask $\mathbf{m}$ replacing the region to 0 as $\mathbf{x} \odot \mathbf{m}$. In this case, the deep image prior optimizes the parameters $\mathbf{W}$ by minimizing E for interpolating the missing region as follows:

$$E = ||(f(\mathbf{W}, \mathbf{z}) - \mathbf{x}) \odot \mathbf{m}||^2 \qquad (11)$$

From the estimated $\mathbf{W}$, image$f(\mathbf{W}, \mathbf{z})$ is synthesized without any missing regions. Furthermore, by replacing the missing part of the generated image, the interpolated image can be synthesized. A natural predicted image can be synthesized without any missing regions using this method.

In the general case, the image is estimated from the randomly generated noise image $\mathbf{z}$. However, in this research, we use the observation image $\mathbf{x}_t$ at the current time instead of the noise. In this case, the image is synthesized based on the evaluation equation as follows:

Figure 4: CityScapes Dataset.

$$E = ||(f(\mathbf{W}, \mathbf{x}_t) - \mathbf{x}) \odot \mathbf{m}||^2 \qquad (12)$$

As the image of the current time and future time have a strong correlation with each other, it can be expected that the image can be synthesized more effectively and efficiently from the current image. Additionally, our research targets sequential video frames, and it is considered that the images at continuous times are similar. Therefore, the image generation network $\mathbf{W}$ is also the same in each frame. Thus, the estimated parameters in the previous frame are used as the initial parameters of the network. As a result, more efficient parameter estimation can be achieved; hence this initialization can reduce the image estimation time.

4 EXPERIMENTAL RESULTS

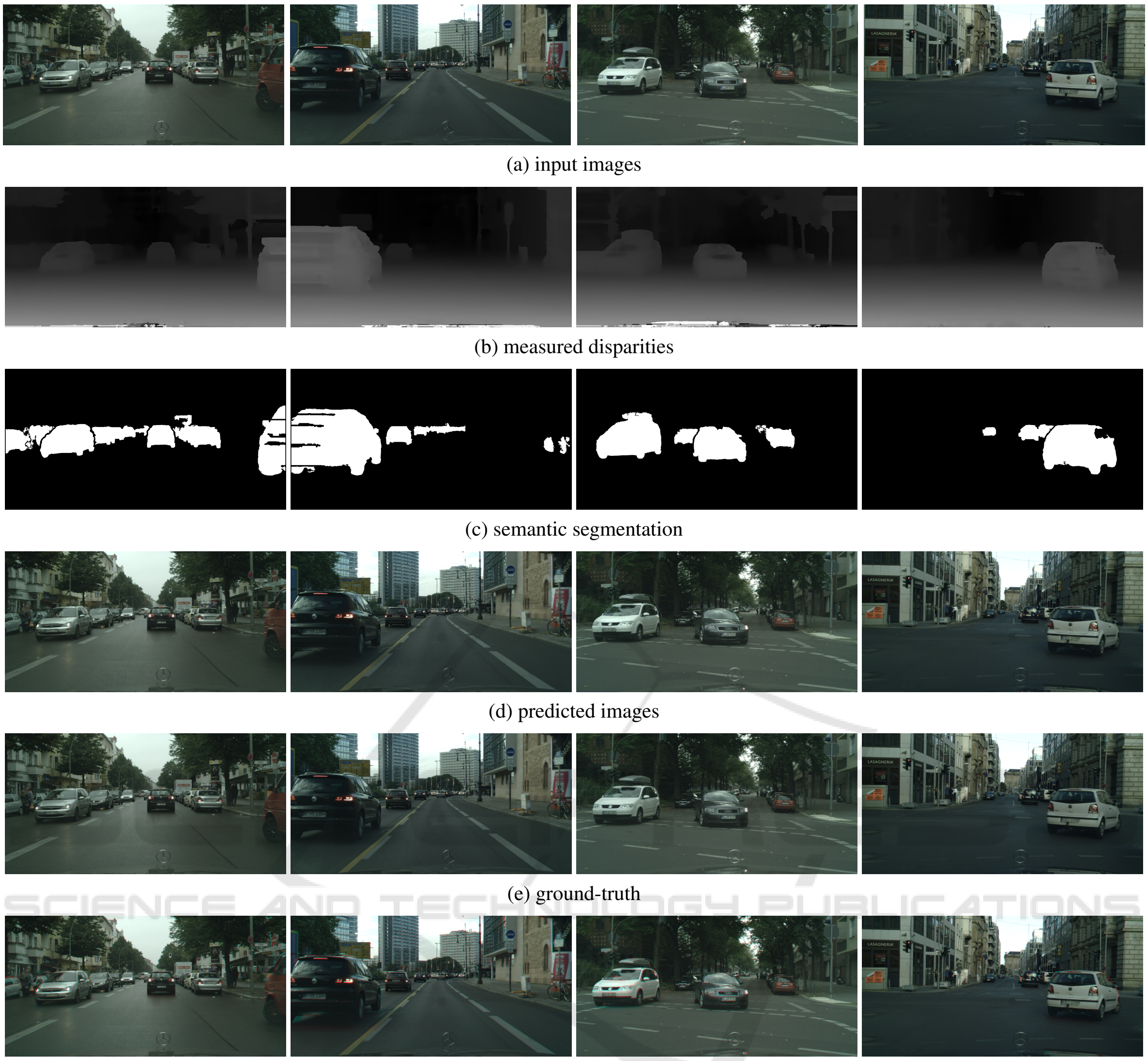
4.1 Environment

In this section, we show several experimental results of our proposed method. In these experiments, we used CityScapes Dataset(Cordts et al., 2016). The dataset includes many videos taken by the stereo cameras in several cities. The framerate of the videos is 17 fps. An example of the image included in the data set is shown in Fig.4.

Moreover, the dataset includes image labels for semantic segmentation. In this dataset, 2975 sequences were used for training the CNN for semantic segmentation, and the others were used as test data.

4.2 Image Prediction Results

The predicted images of the proposed method are shown in Fig. 5. In the figures, (a) represents an image at the current time, (b) represents the disparities obtained from SGM, (c) represents the results of semantic segmentation to dynamic and static objects, and (d) represents the predicted images in the next frame using our method. The ground truth of the predicted images is shown in (e). Furthermore, (f) shows an image where the predicted image and ground-truth image are multiplexed. In these multiplexed images, the red component of each pixel is from the ground-truth. The green and blue components are from the predicted images. The locations,



(a) input images

(b) measured disparities

(c) semantic segmentation

(d) predicted images

(e) ground-truth

(f) multiplexed images of the predicted images and ground-truth

Figure 5: Experimental results on a straight road: (a) input image, (b) measured disparities, (c) semantic segmentation results, (d) predicted images, (e) ground-truth and (f) comparison between ground-truth and predicted images are shown.

where the color shifts occur, show the difference between the predicted images and ground-truth.

In these results, the predicted results are similar to the ground-truth. Moreover, a significant color shift does not occur in the multiplexed image, which indicates that our proposed method can predict future images accurately. Notably, the third image include multiple objects that move independently. For example, when our vehicle turns to the right, the right vehicle in the video turns to the right and left vehicle moves forward in the figure. Despite these complex movements in the scene, the synthesized image is similar to the ground-truth. These results indicate that the method can predict the future condition of each object independently, and it is practical to synthesize the correctly predicted images.

Figure 6 shows longer span prediction results. In this figure, the prediction results based on the current input image are shown in the right column in Fig.5, from one frame (59 ms) ahead to four frames (256 ms) ahead, are shown in each column. Remarkably, natural images can be predicted when a four frames ahead image is predicted. Although some differences are found in (iii) the multiplexed image, (ii) the predicted images are relatively accurate. The results indicate that our proposed method is reliable, and it can predict future images.

Figure 7 shows another longer span prediction results. In this figure, the same frame images were estimated from the different input images. Each col-

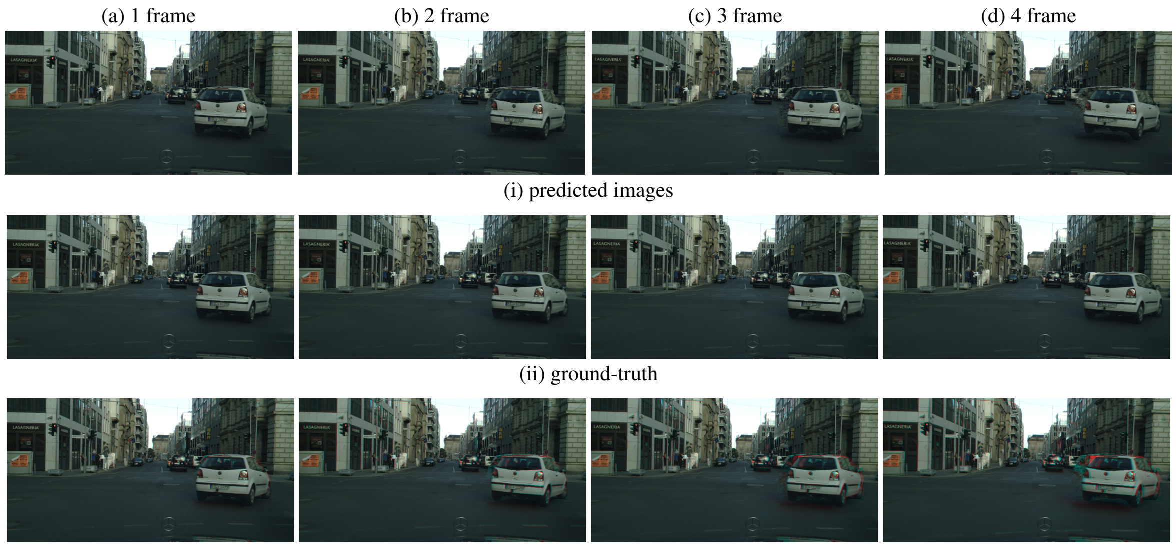

(a) 1 frame
(b) 2 frame
(c) 3 frame
(d) 4 frame
(i) predicted images
(ii) ground-truth
(iii) multiplexed images

Figure 6: Long span prediction results: Top row shows predicted images of our proposed method. Second row shows the ground truth, and third row shows multiplexed images of the predicted images and ground truth.

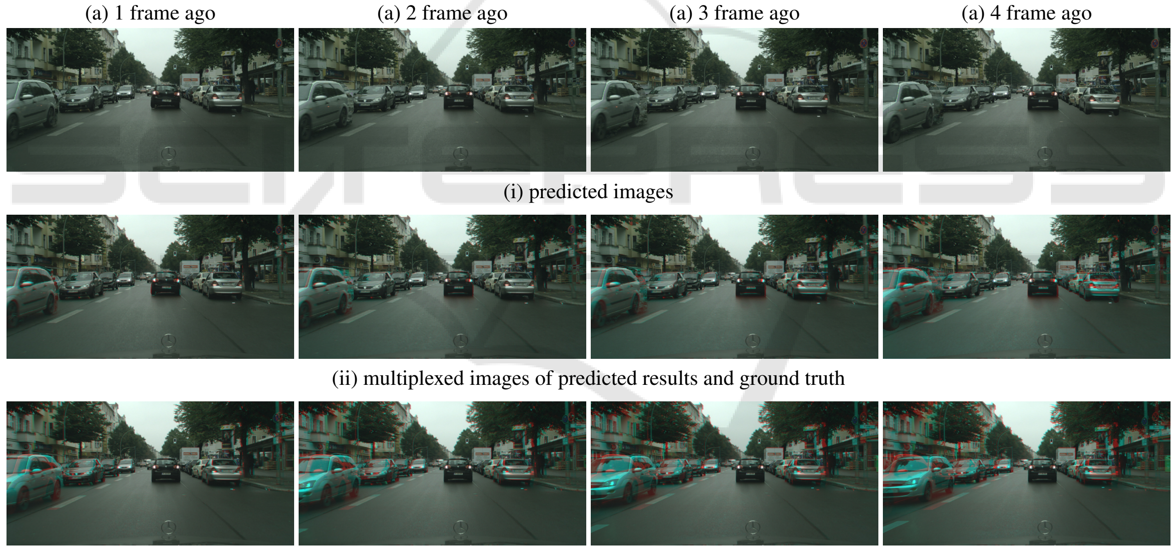

(a) 1 frame ago
(a) 2 frame ago
(a) 3 frame ago
(a) 4 frame ago
(i) predicted images
(ii) multiplexed images of predicted results and ground truth
(iii) multiplexed images of input images and ground truth

Figure 7: Comparison between long span prediction and short span prediction: Each column shows predicted results and multiplex image from1∼4 frames ago images.

umn in the first and second row shows predicted results and multiplexed images from 1∼4 ago images, respectively. The third row shows multiplexed images of input images and ground-truth. In this result, (iii) row shows input images are much different from the ground-truth since the change of the whole image is very large. In contrast, the difference between predicted images and ground-truth is tiny. The fact indicates that our proposed method can predict the change of the image even if the whole image changes drastically.

4.3 Evaluation

We last show the evaluation results of our image prediction. In this evaluation, we chose 20 sequence, which includes 600 images, from the CityScapes dataset, and we synthesized predicted images for all chosen images.

We first evaluated the effectiveness of image inpainting by deep image prior. In this experiment, predicted images are estimated from 1 frame ago images and RMSE between the ground-truth and predicted

Table 1: RMSEs from the ground-truth for original input image, predicted images without inpainting and predicted images with inpainting.

	original	without inpainting	predicted image
RMSE	16.04	11.57	10.77

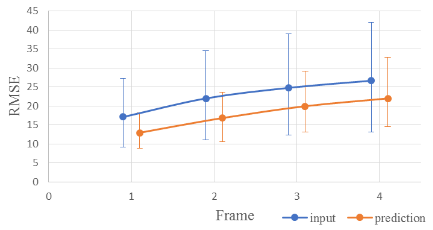


Figure 8: RMSE between ground-truth and predicted/input images.

images were computed. Table 1 shows computed RMSEs for original input images, predicted images without image inpainting, and predicted images with inpainting. The table shows that the RMSE decreases with object motion prediction and image inpainting. The results indicate that the image inpainting technique provides better results for video prediction.

We next synthesized predicted images from 1∼4 ago images, respectively. We computed RMSEs between ground-truth and the predicted images. Figure 8 shows average RMSE for each result. In this figure, error bars for each point show the minimum and the maximum error. For comparison, RMSE between the input images and ground-truth are shown by an orange line. In this result, RMSEs of predicted results always are lower than ones of input images. The results show that our proposed method can predict future images for various images. Note that the minimum errors for input images and predicted images are mostly the same since the data includes mostly static sequences. This fact indicates that our proposed method can predict future images for static sequences as well as dynamic sequences.

5 CONCLUSION

In this paper, we proposed a future image prediction method from a stereo image in a driving scene. In this method, 3D shapes in the scene are reconstructed using the stereo method, and the reconstructed shapes are separated into multiple objects by semantic image segmentation. In the motion estimation of each separated object, the Kalman filter is used, and the filter predicts the future condition of the objects. From the conditions of the predicted objects, future images are rendered. Furthermore, the deep image prior is applied to the predicted images to interpolate the missing areas of the images caused by occlusion, and we finally predict the future natural images without missing areas. Several experimental results of a public dataset show that our proposed method can predict future images even when the input scene includes multiple objects moving independently.

REFERENCES

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

Finn, C., Goodfellow, I., and Levine, S. (2016). Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 64—-72.

Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814 vol. 2.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *arxiv*.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Sharma, S., Ansari, J. A., Murthy, J. K., and Krishna, K. M. (2018). Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3508–3515. IEEE.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.

Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). Generating videos with scene dynamics. In *In Advances In Neural Information Processing Systems (NIPS)*, pages 613—621.

Vondrick, C. and Torralba, A. (2017). Generating the future with adversarial transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. (2018). Generative image inpainting with contextual attention. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5505–5514.