Adaptive Reference Image Selection for Temporal Object Removal from Frontal In-vehicle Camera Image Sequences

Toru Kotsuka[1], Daisuke Deguchi[2], Ichiro Ide[1] and Hiroshi Murase[1]
[1]*Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*
[2]*Information & Communication Headquarters, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, Japan*

Keywords: In-vehicle Camera, Temporal Object Removal, Adaptive Reference Image Selection.

Abstract: In recent years, image inpainting is widely used to remove undesired objects from an image. Especially, the removal of temporal objects, such as pedestrians and vehicles, in street-view databases such as Google Street View has many applications in Intelligent Transportation Systems (ITS). To remove temporal objects, Uchiyama et al. proposed a method that combined multiple image sequences captured along the same route. However, when spatial alignment inside an image group does not work well, the quality of the output image of this method is often affected. For example, large temporal objects existing in only one image create regions that do not correspond to other images in the group, and the image created from aligned images becomes distorted. One solution to this problem is to select adaptively the reference image containing only small temporal objects for spatial alignment. Therefore, this paper proposes a method to remove temporal objects by integration of multiple image sequences with an adaptive reference image selection mechanism.

1 INTRODUCTION

In recent years, image inpainting is widely used to remove undesired objects from an image. Especially, there is a strong need for removal of temporal objects (ex. pedestrians and vehicles) from street-view databases such as Google Street View[1] so that they can be used for Intelligent Transportation Systems (ITS) technologies such as geo-localization of vehicles (Matsumoto et al., 2000).

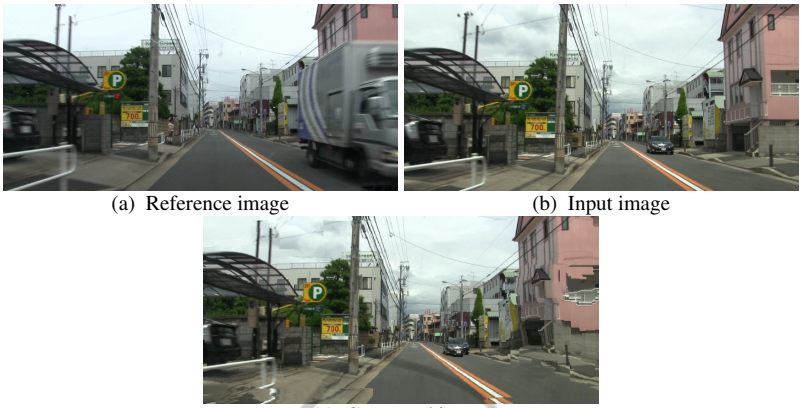
Methods to remove temporal objects in an image can be categorized into three approaches; (1) using a single image, (2) using a single image sequence, and (3) using multiple image sequences. The first approach synthesizes the background scene of a target region selected manually (Bertalmio et al., 2000). It requires only one image as an input, but it is impossible to restore the *true* background scene.

The second approach integrates frames captured as one image sequence (Kawai et al., 2014). Using the difference of appearances between frames, this method can remove temporal objects automatically and restore most of the background scene. However, some temporal objects, for example, parked vehicles, cannot be removed since they are observed as static objects in the image sequence. The third approach integrates multiple image sequences captured along the same route (Uchiyama et al., 2010). By using multiple image sequences, this approach can remove temporal objects even if they are observed as static objects in a certain image sequence. Therefore a method that removes temporal objects using multiple image sequences is the most suitable for constructing a street-view database.

The method proposed by Uchiyama et al. uses frame alignment between image sequences and spatial alignment inside an image group as a preprocessing step to integrate multiple image sequences. Through frame alignment, all images captured at the same location are grouped. Then, spatial alignment is performed by aligning all images with a reference image selected from each image group. Finally, an image without temporal objects is generated by fusion of images in each image group. However, this method can result in poor output image quality due to failure of spatial alignment inside an image group. For example, when a temporal object region existing in only one image cannot be aligned with other images. Generally, in such cases, it is necessary to estimate correspondences from the surroundings of the temporal object. But, if the temporal object in the reference image is large, this estimation will not work correctly.

[1]http://www.google.co.jp/help/maps/streetview/

Kotsuka T., Deguchi D., Ide I. and Murase H..
Adaptive Reference Image Selection for Temporal Object Removal from Frontal In-vehicle Camera Image Sequences.
DOI: 10.5220/0005357102330239
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 233-239
ISBN: 978-989-758-089-5



(a) Reference image (b) Input image

(c) Generated image

Figure 1: The generated image is distorted depending on the reference image. If there are temporal objects in the reference image, the pixel correspondence fails.

(a) (b) (c) (d) (e) (f)

Figure 2: Examples of temporal objects.

This causes image deformation due to inappropriate spacial alignment, as shown in Figure 1.

One way to solve this problem is to select the reference image for spatial alignment adaptively such that it contains only small temporal objects. Thus, this paper proposes a method to remove temporal objects by fusion of multiple image sequences with an adaptive reference image selection mechanism. The aim of this method is to reduce deterioration of image quality that is one of the biggest problems in the state-of-the-art methods for temporal object removal.

In the following, Section 2 explains the details of the proposed method. Next, Section 3 describes the experiments. The results of the experiments are discussed in Section 4. Finally, we conclude this paper in Section 5.

2 TEMPORAL OBJECTS REMOVAL

2.1 Definition of Temporal Objects

We define temporal objects as follows:

- Objects that exist at a certain time, but not at all times.

In other words, temporal objects do not exist constantly in the same location. Figure 2 shows examples of temporal objects in typical road scenes, such as pedestrians (a, b), a bicycle (c), and a moving vehicle (d). Parked vehicles (e, f) are also temporal objects since they do not exist in every image sequence. On the other hand, a vehicle existing in all image sequences (every time) is not treated as a temporal object.

2.2 Strategy for Adaptive Reference Image Selection

The proposed method makes the following assumption for selecting a reference image:

- If we obtain multiple images at the same location, only background pixels can be aligned correctly.

Therefore, a reference image can be selected as the image which has the maximum number of pixel correspondences in the image group. This can be represented by the following objective function $G(i)$, and the i-th image in an image group is selected as a reference image that maximizes $G(i)$.

$$
\begin{aligned}
G(i) &= \sum_{j=1, j\neq i}^{n} \sum_{\boldsymbol{x}} g_1(\boldsymbol{x}, i, j), \qquad (1)\\
g_1(\boldsymbol{x}, i, j) &= \begin{cases} 1 & \text{if } ||g_2(\boldsymbol{x}, i, j)|| \leq l \\ 0 & \text{otherwise} \end{cases}, \\
g_2(\boldsymbol{x}, i, j) &= v_{i\to j}(\boldsymbol{x}) + v_{j\to i}(\boldsymbol{x} + v_{i\to j}(\boldsymbol{x})),
\end{aligned}
$$

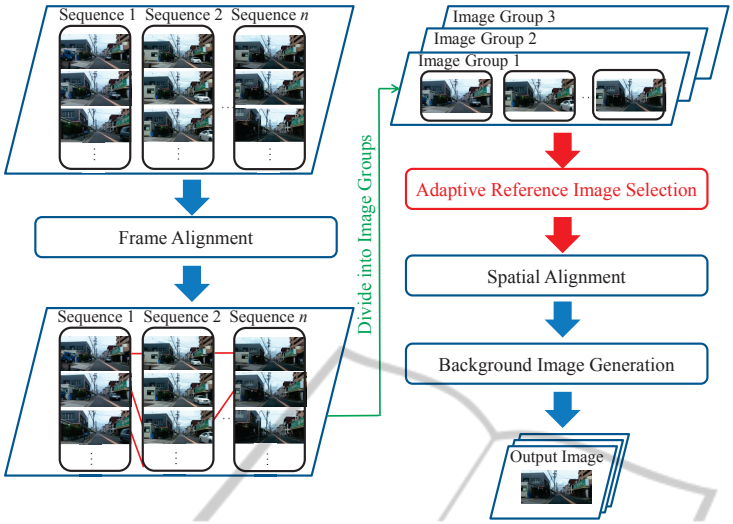


Figure 3: Process flow of the proposed method.

where $\boldsymbol{x}$ is a pixel in the image, $v_{i\to j}(\boldsymbol{x})$ is the displacement between pixel $\boldsymbol{x}$ in image i and the pixel in image j which is matched to pixel $\boldsymbol{x}$. The number of pixel correspondences are calculated, such that the matching error is less than $l = 4$ pixels. The image which has the maximum number of pixel correspondences is selected as the reference image in the image group.

2.3 Algorithm

According to the definition in Section 2.1, we assume that parts of the images captured at the same location at different times do not include temporal objects. Thus we combine parts of the images considered as background to remove temporal objects.

The proposed method removes temporal objects using a majority voting scheme. Temporal object removal is performed by integration of frontal in-vehicle camera image sequences captured along the same route at different times. Figure 3 shows the framework of the proposed method. The proposed method consists of four phases; (1)frame alignment between image sequences, (2)adaptive reference image selection, (3)spatial alignment inside an image group, and (4)background image generation.

Frame alignment between image sequences is the process to compensate for different vehicle speeds between image sequences. After frame alignment, the aligned images are grouped and the following processes are performed for every result.

Spatial alignment inside each image group is the process to compensate for the difference in appearance depending on the vehicle position. In this process, a reference image is adaptively selected.

Background image generation is the process that removes temporal objects by combining images in the image group. From the assumption in Section 2.2, we remove temporal objects by majority voting.

The following sections describe each process in detail.

2.3.1 Frame Alignment Between Image Sequences

Frame alignment between image sequences is performed by DP matching (Dynamic Time Warping) using measurement of the positional relationship between two camera locations based on epipolar geometry (Kyutoku et al., 2012). This method is known for effective matching of frontal camera image sequences captured along a common route and is robust to occlusions. The epipole is calculated in all image pairs between the reference image sequence and input image sequences. The nearer the captured location is between two cameras, the longer the distance is between the image center and the epipole. This property is used as a measure for DP matching of the image sequences and image groups captured along the same route.

2.3.2 Adaptive Reference Image Selection

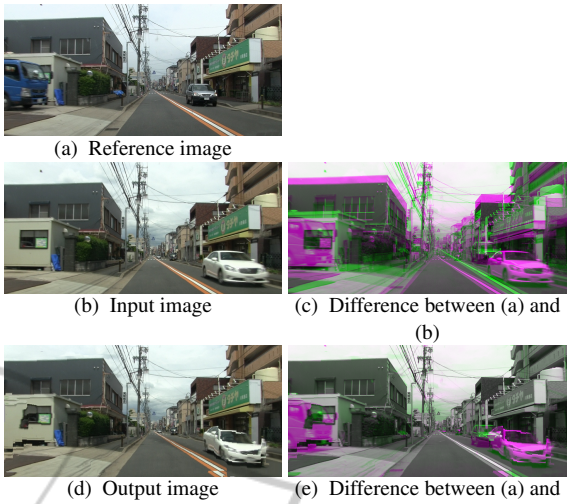
Appearances of images in an image group differ depending on the vehicle position. Before overcoming this, adaptive reference image selection is performed.

A reference image is selected for each image group by the method outlined in Section 2.2 using SIFT flow (Liu et al., 2011). SIFT flow calculates SIFT features at each pixel and associates all pixels

between images correctly using belief propagation. By using SIFT features, the robustness to rotation, scale, and illumination changes of the standard SIFT algorithm are achieved, and the technique is more effective than the optical flow method.

2.3.3 Spatial Alignment Inside an Image Group

After adaptive reference image selection, spatial alignment inside image groups is performed relative to the reference image. In each case, pixels in the input image are rearranged based on their corresponding positions in the reference image. In detail, all images in an image group are warped to the reference image according to the flow field obtained by the SIFT flow algorithm. Figure 4 shows the result of this process. The colored regions show the difference between two images in Figures 4(c) and (e). Figure 4(d) is the output image which is rearranged to make the image more similar to the reference image shown in Figure 4(a).



(a) Reference image
(b) Input image
(c) Difference between (a) and (b)
(d) Output image
(e) Difference between (a) and (d)

Figure 4: Spatial alignment using SIFT flow. The colored region in (c) and (e) shows the difference between two images.

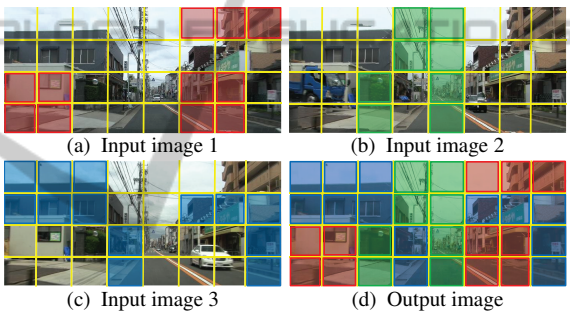
2.3.4 Background Image Generation

Temporal object removal to create a background image is performed by fusion of images in each image group. First, each image is divided into W patches, and vectors consisting of the pixel values of each patch are created. In this paper, we formulate the problem of background image generation as optimal patch selection in each image group. Here, we introduce an index vector $\mathbf{n}$ consisting of the selected image indices in the image group. The index vector $\mathbf{n}$ is selected by minimizing the following objective function $F(\mathbf{n})$,

$$F(\mathbf{n}) = \sum_{w=1}^{W} [(1-\lambda) f_w(n_w) + \lambda g_w(n_w)], \qquad (2)$$

where n_w is the w-th patch of the selected image and f_w is the penalty term for temporal objects and calculated by using the vector median filter (Astola et al., 1990). The vector median filter is a filter which outputs the central vector of a field of input vectors. The central vector is defined as the vector which minimizes the sum of the distance between itself and other vectors, and f_w is the sum of the distance between the central vector and other vectors. Since we assume that for a majority of timings, the input patches do not contain temporal objects, the central vector is considered to be the background.

Each vector median filter match selection is performed independently, so illumination conditions of the neighborhood patches are not preserved. To solve this problem, a penalty g_w to represent the discontinuity of neighborhood patches is employed. If the index



(a) Input image 1
(b) Input image 2
(c) Input image 3
(d) Output image

Figure 5: Result of fusion of an image group. Output (d) is generated by fusion of images (a), (b) and (c).

$n_{w+\alpha}$ in the neighborhood patch is different from n_w in a certain patch, g_w adds the distance between the neighborhood patch's vector $n_{w+\alpha}$ and the one in n_w. λ is the weight between f_w and g_w.

Figure 5 shows the selection of the patch by $F(\mathbf{n})$ and image fusion. In fact, each patch is overlapped and localized, and the integration of overlapped patches is done by alpha-blending.

3 EXPERIMENT

3.1 Dataset

We prepared a dataset composed of five frontal in-vehicle camera image sequences captured along the same route. The route was a straight main road and contained some temporal objects. Each image sequence had an image resolution of 720×340 pixels,

recorded at 24 fps, and contained 1,000 images. Every one image out of 25 images was used for evaluation.

3.2 Evaluation Experiments

We performed the following two experiments to evaluate the effectiveness of the proposed method. First is an experiment for evaluating the accuracy of the adaptive reference image selection. Second is an experiment for evaluating an accuracy of removal of temporal objects. In addition to the proposed method, we prepared a comparative method without adaptive reference image selection, where all reference images are selected from a specific image sequence.

3.2.1 Accuracy of Adaptive Reference Image Selection

We evaluated the accuracy of the adaptive reference image selection. First, we prepared the ground truth, which consisted of the most suitable reference images containing the fewest temporal objects in each image group. Second, we counted the pixels of the temporal objects in each reference image selected by the proposed and the comparative methods. We compared the number of temporal object pixels between the ground truth and the proposed and the comparative methods.

3.2.2 Accuracy of Temporal Objects Removal

We evaluated the accuracy of temporal object removal by comparing the number of pixels corresponding to temporal objects in the proposed and the comparative methods.

4 RESULTS AND DISCUSSIONS

4.1 Adaptive Reference Image Selection

Table 1 shows the average difference of the number of temporal object pixels in each reference image. Figure 6 shows the time transition of the number of temporal object pixels compared with the ground truth. The result of the proposed method was closer to the ground truth than that of the comparative method. These show how adaptive reference image selection helps to choose an appropriate reference image with fewer temporal object regions as a starting point for temporal object removal. However, the proposed method could not choose the best reference image in some image groups due to failure of SIFT flow calculation. In future works, we need to investigate more accurate spatial alignment methods.

Table 1: The average number of pixels of temporal objects in each reference image (pixels).

Ground truth	Proposed method	Comparative method
693	2,106	4,906

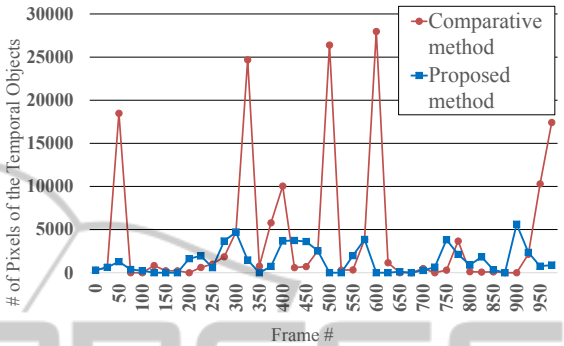


Figure 6: Time transition of the number of temporal object pixels in reference images selected by the proposed and the comparative methods compared with the ground truth.

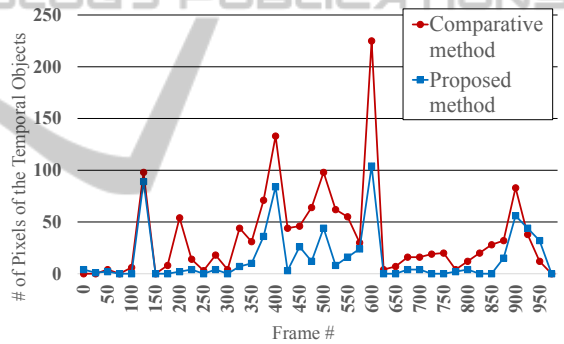


Figure 7: Time transition of the number of temporal object pixels in the result of the proposed and the comparative methods.

4.2 Temporal Object Removal

The average number of pixels for temporal objects in each output image was 16 pixels in the proposed method and 36 pixels in the comparative method. Figure 7 shows the time transition of the number of pixels corresponding to the temporal objects remaining in each output image. Both the proposed and the comparative methods could remove almost all temporal objects. This shows the effectiveness of temporal object removal using multiple image sequences. Figure 8 shows examples of the output image for both methods. As seen in Figure 8(a), the image quality de-

(a) Comparative method (b) Proposed method

Figure 8: Examples of output images.

teriorated in the comparative method. Because adaptive reference image selection was not performed, there were many regions where the alignment did not work well. Conversely, since the proposed method employs adaptive reference image selection, the image alignment failed in fewer areas. This is why the proposed method gives better output image quality than the comparative method. As seen in Figure 8(b), the proposed method removed temporal objects and reduced the distortion of visual image quality when compared to the comparative method.

Therefore, we confirmed that adaptive reference image selection is one of solution to prevent image distortion when integrating multiple image sequences for the removal of temporal objects.

5 CONCLUSION

In this paper, we introduced the concept of adaptive reference image selection, and proposed a method to remove temporal objects by fusion of multiple image sequences. We confirmed that adaptive reference image selection is one solution to prevent image distortion when integrating multiple image sequences for the removal of temporal objects.

For future work, we will apply the proposed method to image sequences with more crowded scenes. In this situation, the assumption stated in Section 2.3 may not be satisfied. To solve this problem, use of additional input image sequences and between-frames information (Kawai et al., 2014) may be effective. Furthermore, we would like to develop a spatial alignment method more robust to occlusions than the standard SIFT flow method.

ACKNOWLEDGEMENTS

Parts of this research were supported by a Grant-in-Aid for Young Scientists from MEXT, a Grant-In-Aid for Scientific Research from MEXT, and a CREST project from JST, and Nagoya University COI. The authors would like to thank Mr. David Robert Wong for his useful comments.

REFERENCES

Astola, J., Haavisto, P., and Neuvo, Y. (1990). Vector median filters. *Proc. IEEE*, 78(4):678–689.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. (2000). Image inpainting. In *Proc. 27th Int. Conf. and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH2000)*, pages 417–424.

Kawai, N., Inoue, N., Sato, T., Okura, F., Nakashima, Y., and Yokoya, N. (2014). Background estimation for a single omnidirectional image sequence captured with a moving camera. *IPSJ Trans. on Computer Vision and Applications*, 6:68–72.

Kyutoku, H., Deguchi, D., Takahashi, T., Mekada, Y., Ide, I., and Murase, H. (2012). Substraction-based forward obstacle detection using illumination insensitive feature for driving support. In *Proc. ECCV2012 Workshop on Computer Vision in Vehicle Technology (CVVT2012): From Earth to Mars*, pages 515–525.

Liu, C., Yuen, J., and Torralba, A. (2011). SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(5):978–994.

Matsumoto, Y., Sakai, K., Inaba, M., and Inoue, H. (2000). View-based approach to robot navigation. In *Proc. 2000 IEEE/RSJ Int. Conf. on Intelligent Robots and System (IROS2000)*, pages 1702–1708.

Uchiyama, H., Deguchi, D., Takahashi, T., Ide, I., and Murase, H. (2010). Removal of moving objects from a street-view image by fusing multiple image sequences. In *Proc. 20th IAPR Int. Conf. on Pattern Recognition (ICPR2010)*, pages 3456–3459.