IPSJ Transactions on Computer Vision and Applications Vol.5 109–113 (July 2013)

[DOI: 10.2197/ipsjtcva.5.109]

Express Paper

Improving Accuracy of Planar Tracking by Resolving Resolution Inconsistencies

Tomohiro Ushiki[1] Eisuke Ito[1,a)] Takayuki Okatani[1,b)]

Received: March 11, 2013, Accepted: April 24, 2013, Released: July 29, 2013

Abstract: This paper presents a method for improving the accuracy of template-based planar tracking. It has been shown that when the ROI of the input image has a lower resolution than the template, tracking accuracy will deteriorate; then, this can be remedied by blurring the template in response to the motion of the plane. In this study, we show that, conversely, when the template has a lower resolution than the input image, tracking accuracy will deteriorate in a different manner. We then present a method that can simultaneously deal with both cases and thus achieves higher tracking accuracy.

Keywords: image-based tracking, planar tracking, efficient second-order minimization (ESM)

1. Introduction

It is one of the fundamental problems of computer vision to visually track a planar object moving in space. In fact, it is indispensable for tracking features in video images and many applications of augmented reality (AR).

There are two approaches to the problem: the feature-based approach [7], [9], [13] and the template-based approach [1], [2], [3], [4], [5], [6], [10], [11], [12]. The former first extracts primitives such as points, lines, etc. from images, and then determines their geometric relation between the surface texture of the target plane and the input images. This approach tends to be robust, whereas its accuracy and speed are not the best.

The template-based approach is to directly compare the image brightness between the texture of the target plane (i.e., template) and the input images; it determines the pose parameters of the plane by minimizing the sum of the brightness differences. This approach tends to be more accurate owing to the direct comparison of the image brightness at each pixel. It is also fast, since only a small number of iterations are usually necessary to converge by choosing the estimated pose for the last image as an initial value.

The basic assumption behind the template-based approach is that the image brightness at each surface point of the target plane is invariant regardless of how the plane changes its pose in space. However, this assumption of brightness constancy is often invalidated due to several causes such as illumination changes [15], motion blurs [14], etc., and several studies have been conducted to overcome the resulting difficulties so far.

Recently, Ito et al. [8] point out that the decreased resolution of input images, which occurs when the plane moves to a distant place from the camera or when its surface normal has an oblique orientation toward the viewing direction, can also invalidate the assumption, and show a method that can overcome it.

In this study, extending Ito et al.'s study, we consider dealing with the more general cases of resolution inconsistencies between the input images and the template. Ito et al.'s study considers only one half of such inconsistencies. The remaining half is such that the input images have higher resolution than the template. Such cases do often occur, and thus it is important to deal with them. We present a method to be able to deal with the two halves simultaneously in a unified manner.

2. The Template-based Approach

To begin with, we briefly summarize the template-based approach. Let $\mathcal{I}^*(\mathbf{p}^*)$ and $\mathcal{I}(\mathbf{p})$ be the template and the input image, respectively. Here, we consider only gray-scale images. In what follows, we will abuse the notation of image coordinates such as $\mathbf{p}$ and $\mathbf{p}^*$; they will indicate either homogeneous or inhomogeneous coordinates depending on the context.

We write the planar homography that maps a point $\mathbf{p}^*$ in the template to a point $\mathbf{p}$ in the input image as

$$\mathbf{p} \propto \mathbf{H}_0 \mathbf{p}^*. \tag{1}$$

When the brightness constancy assumption is valid, there should exist $\mathbf{H}_0$ such that for any $\mathbf{p}^*$ of the template, it holds that

$$\mathcal{I}(\mathbf{H}_0 \mathbf{p}^*) = \mathcal{I}^*(\mathbf{p}^*). \tag{2}$$

Then, the problem is to obtain such $\mathbf{H}_0$. Considering the presence of image noise, we minimize the sum of squared differences

$$J(\mathbf{x}) = \sum_i \left[\mathcal{I}(\hat{\mathbf{H}}\mathbf{H}(\mathbf{x})\mathbf{p}_i^*) - \mathcal{I}^*(\mathbf{p}_i^*) \right]^2, \tag{3}$$

where $\hat{\mathbf{H}}$ is the latest estimate of $\mathbf{H}_0$ and $\hat{\mathbf{H}}(\mathbf{x})$ is an update we want to determine; $\mathbf{x}$ is an eight-vector that parametrizes the updating homography based on the Lie algebra [3]. When $\hat{\mathbf{H}}$ is close

[1] Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi 980–8579, Japan
a) ito@vision.is.tohoku.ac.jp
b) okatani@vision.is.tohoku.ac.jp

to the true homography and thus $\mathbf{x}$ is small, J can be "linearized" with good accuracy by a low-order polynomial of $\mathbf{x}$, for which it is easy to find the minimizer $\mathbf{x}$. There are several ways of linearization, among which we employ the ESM (efficient second-order minimization) method of Malis et al. [2], [3].

Then, the estimate is updated as $\hat{\mathbf{H}} \leftarrow \hat{\mathbf{H}}\mathbf{H}(\mathbf{x})$, where $\mathbf{x}$ is the minimizer obtained above. This pair of minimization and update is iterated until convergence. Generally, it takes a few dozen iterations for each input image, which can be performed in real time.

3. Overcoming Resolution Inconsistency between Templates and Input Images

3.1 The Case of Decreased Input Image Resolution—Revisiting the Study of Ito et al.

For the subsequent discussion about the case of increased resolution, we summarize here the study of Ito et al. [8]. They considered the case where the warped input image $\mathcal{I}(\mathbf{H}_0\mathbf{p}^*)$ has decreased resolution in the domain of $\mathbf{p}^*$ as compared with the template $\mathcal{I}^*(\mathbf{p}^*)$.

Such decreased resolution is caused by the resolution limit of the imaging system, which can be modeled by a prefilter, which serves as a low-pass filter eliminating the high-frequency component of input images. A standard model of such prefilters is a Gaussian function $f(\mathbf{p}) \propto \exp(-\mathbf{p}^\top\mathbf{p}/(2\sigma_f^2)) = \exp(-(p_x^2 + p_y^2)/(2\sigma_f^2))$.

When the plane has a pose given by $\mathbf{H}_0$ ($\mathbf{p} \propto \mathbf{H}_0\mathbf{p}^*$), the texture of the tracked planar region seen from the camera can be modeled as $\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{p})$. Applying the prefilter $f(\mathbf{p}^*)$ to this, the input image $\mathcal{I}(\mathbf{p})$ can be modeled as

$$\mathcal{I}'(\mathbf{p}) = \mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{p}) * f(\mathbf{p}). \tag{4}$$

Our purpose is to estimate $\mathbf{H}_0$. In the above basic method, its estimate $\mathbf{H}_1$ is determined so that the warped input image $\mathcal{I}(\mathbf{H}_1\mathbf{p}^*)$ is the closest to the template $\mathcal{I}^*(\mathbf{p}^*)$. Abusing notations [*1] for the sake of brevity, the warped input image $\mathcal{I}(\mathbf{H}_1\mathbf{p}^*)$ can be modeled as

$$\mathcal{I}'(\mathbf{H}_1\mathbf{p}^*) = \mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*) * f(\mathbf{H}_1\mathbf{p}^*). \tag{5}$$

It is observed from Eq. (5) that, even if $\mathbf{H}_1$ coincides with the true homography $\mathbf{H}_0$, the right hand side of the equation, which reduces to $\mathcal{I}^*(\mathbf{p}^*) * f(\mathbf{H}_1\mathbf{p}^*)$, will not coincide with the template $\mathcal{I}^*(\mathbf{p}^*)$. Their difference, i.e., $f(\mathbf{H}_1\mathbf{p}^*)$, explains the decreased resolution of input images.

Ito et al. resolved this inconsistency by appropriately lowering the resolution of the template $\mathcal{I}^*(\mathbf{p}^*)$ during tracking, more specifically, by convolving with the template a linear filter simulating the effect of $f(\mathbf{H}_1\mathbf{p}^*)$. Their solution uses two approximations. One is that the pose change within each frame is assumed to be so small that the blurring filter is determined from the estimate $\hat{\mathbf{H}}$ at the last frame; it is fixed during the iterative minimization. The other is that $\hat{\mathbf{H}}$ is approximated by an affine transform $\hat{\mathbf{H}}_A$ for determining the shape of the filter. These make it possible to approximate Eq. (5) with the convolution of a linear filter given by

$$f'(\mathbf{p}^*; \hat{\mathbf{H}}_A) \propto \exp\left(-\frac{1}{2\sigma_f^2}\mathbf{p}^{*\top}\hat{\mathbf{H}}_A^\top\hat{\mathbf{H}}_A\mathbf{p}^*\right). \tag{6}$$

Finally, the objective function is rewritten as follows:

$$J(\mathbf{x}) = \sum_i [\mathcal{I}(\hat{\mathbf{H}}\mathbf{H}(\mathbf{x})\mathbf{p}_i^*) - \mathcal{I}^*(\mathbf{p}_i^*) * f'(\mathbf{p}_i^*; \hat{\mathbf{H}}_A)]^2. \tag{7}$$

It has been experimentally shown [8] that the optimization using this function considerably improves the accuracy and stability of tracking.

3.2 Incorporating the Consideration of Increased Input Image Resolution

When the warped input image is of higher resolution than the template, the brightness constancy assumption is violated similarly. However, this cannot be dealt with by the above method, which considers only the opposite case. As increasing the resolution of the template is unrealistic, we consider artificially decreasing the resolution of input images.

Suppose that we are given a low-resolution template $\mathcal{I}_l^*(\mathbf{p}^*)$. We continue to use $\mathcal{I}^*(\mathbf{p}^*)$ to represent the texture of the target region of the plane, which has higher resolution than $\mathcal{I}_l^*(\mathbf{p}^*)$. In the process of tracking, the input image $\mathcal{I}(\mathbf{p})$ is warped by $\mathbf{H}_1$ to be compared with the corrected template $\mathcal{I}_l^*(\mathbf{p}^*) * f(\mathbf{H}_1\mathbf{p}^*)$. As shown in Eq. (5), the resulting image is modeled as $\mathcal{I}'(\mathbf{H}_1\mathbf{p}^*) = \mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*) * f(\mathbf{H}_1\mathbf{p}^*)$. It is easy to see that even if $\mathbf{H}_1 = \mathbf{H}_0$, it will not coincide with the corrected template because of the difference between $\mathcal{I}^*$ and $\mathcal{I}_l^*$.

A straightforward method to correct this difference is to lower the resolution of the warped input image $\mathcal{I}(\mathbf{H}_1\mathbf{p}^*)$ so that its resolution matches that of the template [*2]. Specifically, incorporating a new linear filter $g(\mathbf{p}^*)$, we apply it to the warped input image as $\mathcal{I}(\mathbf{H}_1\mathbf{p}^*) * g(\mathbf{p}^*)$ and compare against $\mathcal{I}_l^*(\mathbf{p}^*) * f(\mathbf{H}_1\mathbf{p}^*)$. The filtered image is modeled as

$$\mathcal{I}'(\mathbf{H}_1\mathbf{p}^*) * g(\mathbf{p}^*) = (\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*) * g(\mathbf{p}^*)) * f(\mathbf{H}_1\mathbf{p}^*). \tag{8}$$

Thus, it suffices to choose g such that $\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*) * g(\mathbf{p}^*) \approx \mathcal{I}^*(\mathbf{p}^*) * g(\mathbf{p}^*)$ has equal resolution to $\mathcal{I}_l^*(\mathbf{p}^*)$. If the template $\mathcal{I}_l^*(\mathbf{p}^*)$ is of isotropically low-resolution, it will be given as

$$g(\mathbf{p}^*) \propto \exp\left(-\frac{1}{2\sigma_g^2}\mathbf{p}^{*\top}\mathbf{p}^*\right). \tag{9}$$

Unfortunately, there are a few problems with this approach. Firstly, this necessitates warping the input image at every iteration of the minimization. This means that we need to convolve g with the warped input image every iteration, which significantly increases the computational cost (Even though we make maximum use of GPU, filter convolution is computationally expensive). Moreover, as in the case of the decreased resolution, we may consider the within-frame motion of planes to be sufficiently small so as not to affect the image resolution. Thus, it is sufficient to determine the filter solely from the plane pose $\hat{\mathbf{H}}$ estimated at

[*1] Rigorously, $\mathcal{I}'(\mathbf{H}_1\mathbf{p}^*)$ cannot be represented by the convolution of a linear filter because of the nonlinearity of $\mathbf{H}_1$.

[*2] It should be noted that it does not work to vary the size (i.e., pixels) of the template corresponding to its resolution, since using a small sized template will make tracking very unstable.

the last frame.

Therefore, we seek a method that can achieve an equivalent effect by applying some filter to the raw input image. More specifically, we apply a linear filter $h(\mathbf{p})$ to (the ROI of) the input image $\mathcal{I}(\mathbf{p})$ and then warp it by $\mathbf{H}_1$ to compare against the corrected (low-resolution) template $\mathcal{I}_l^*(\mathbf{p}^*) * f'(\mathbf{p}^*; \hat{\mathbf{H}}_A)$. The filtered input image $\mathcal{I}(\mathbf{p}) * h(\mathbf{p})$ can be modeled as

$$\mathcal{I}'(\mathbf{p}) * h(\mathbf{p}) = \mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{p}) * f(\mathbf{p}) * h(\mathbf{p}). \tag{10}$$

Thus, we have only to choose h such that this coincides with the corrected template.

As in the case of the filter f' for blurring the template, we choose for h a two-dimensional Gaussian function

$$h(\mathbf{p}) \propto \exp\left(-\frac{1}{2}\mathbf{p}^\top \mathbf{\Phi}_h^{-1}\mathbf{p}\right), \tag{11}$$

where $\mathbf{\Phi}_h$ is a 2×2 matrix that we want to determine (Note that $\mathbf{p}$ is used here as inhomogeneous coordinates $\mathbf{p} = [p_x, p_y]^\top$). The two filters on the right hand side of Eq. (10) are both Gaussian, and thus it is equivalent to apply the following single Gaussian filter to $\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{p})$:

$$(f \otimes h) \propto \exp\left(-\frac{1}{2}\mathbf{p}^\top(\sigma_f^2\mathbf{I} + \mathbf{\Phi}_h)^{-1}\mathbf{p}\right). \tag{12}$$

By using this, the image obtained by warping the right hand side of Eq. (10) with $\mathbf{H}_1$ is given by

$$\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*) * (f \otimes h)(\mathbf{H}_1\mathbf{p}^*). \tag{13}$$

Adopting similar approximations used in Ref. [8], $(f \otimes h)(\mathbf{H}_1\mathbf{p}^*)$ reduces to

$$(f \otimes h)(\mathbf{H}_1\mathbf{p}^*) \propto \exp\left(-\frac{1}{2}\mathbf{p}^{*\top}\hat{\mathbf{H}}_A^\top(\sigma_f^2\mathbf{I} + \mathbf{\Phi}_h)^{-1}\hat{\mathbf{H}}_A\mathbf{p}^*\right).$$

Similarly, the two filters on the right hand side of Eq. (8) can be merged to the following single Gaussian filter $\mathcal{I}^*(\mathbf{H}_0^{-1}\mathbf{H}_1\mathbf{p}^*)$.

$$(f \otimes g)(\mathbf{p}^*) \propto \exp\left(-\frac{1}{2}\mathbf{p}^{*\top}\left(\sigma_g^2\mathbf{I} + \sigma_f^2(\hat{\mathbf{H}}_A^\top\hat{\mathbf{H}}_A)^{-1}\right)^{-1}\mathbf{p}^*\right).$$

Then, we determine h (i.e., $\mathbf{\Phi}_h$) so that the two combined filters $(f \otimes h)$ and $(f \otimes g)$ coincide with each other. Some calculation leads to

$$h(\mathbf{p}; \hat{\mathbf{H}}_A) \propto \exp\left(-\frac{1}{2\sigma_g^2}\mathbf{p}^\top\hat{\mathbf{H}}_A^{-\top}\hat{\mathbf{H}}_A^{-1}\mathbf{p}\right). \tag{14}$$

Finally, the objective function becomes

$$J(\mathbf{x}) = \sum_i [\mathcal{I}_l(\hat{\mathbf{H}}\mathbf{H}(\mathbf{x})\mathbf{p}_i^*) - \mathcal{I}_l^*(\mathbf{p}_i^*) * f'(\mathbf{p}_i^*; \hat{\mathbf{H}}_A)]^2, \tag{15}$$

where

$$\mathcal{I}_l(\mathbf{p}) = \mathcal{I}(\mathbf{p}) * h(\mathbf{p}; \hat{\mathbf{H}}_A). \tag{16}$$

Note that for each input image, $\hat{\mathbf{H}}_A$ is determined at the beginning of iterative minimization and is fixed during the iterations.

4. Experimental Results

We conducted several experiments to examine the performance of our method. We used a Grasshopper camera of Point Grey Research Inc. and a PC equipped with a GTX580 GPU of nVidia. The input images are 640×480 pixels and we choose the size of templates to be 192×192 pixels. By using the GPU for the non-linear minimization as well as the two convolutions, tracking can be performed in frame rate of 30 Hz.

Figure 1 shows how the two blurring filters $f'(\mathbf{p}^*; \hat{\mathbf{H}}_A)$ of Eq. (6) and $h(\mathbf{p}; \hat{\mathbf{H}}_A)$ of Eq. (11) vary during tracking. It is seen from this that the two filters are complementary with each other, corresponding to the fact that their covariance matrices are the inverse of each other: $(\hat{\mathbf{H}}_A^\top\hat{\mathbf{H}}_A)^{-1} = \hat{\mathbf{H}}_A^{-\top}\hat{\mathbf{H}}_A^{-1}$. Because of this mechanism, the input image and the template will never be blurred in such a way that information is lost.

Figure 2 shows the results of tracking a plane moving between a distant position and a close position to the camera a few times repeatedly. It shows that although there is no clear difference in accuracy between the two methods when both can track the target, Ito et al.'s method failed tracking twice (the red dots are missing), when the plane is closer to the camera. **Figure 3** shows several snapshots of the same tracking results. When the plane is distant, the warped input image (the 2nd row) has the lowest resolution. When it is lower than the template (the leftmost column), Ito et al.'s method applies a blurring filter to the template, which makes the appearances of the warped input and the template similar. As the plane comes closer to the camera, its resolution increases (the second and third columns). Their method managed to deal with this increasing resolution by reducing the amount of the template blurring. However, when the plane comes more closer (the fourth

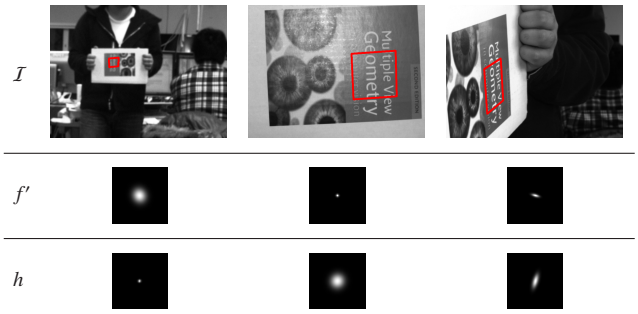

Fig. 1 The two filters $f'(\mathbf{p}^*; \hat{\mathbf{H}}_A)$ and $h(\mathbf{p}; \hat{\mathbf{H}}_A)$ (shown in 51×51 pixel size) work in a complementary way during tracking.

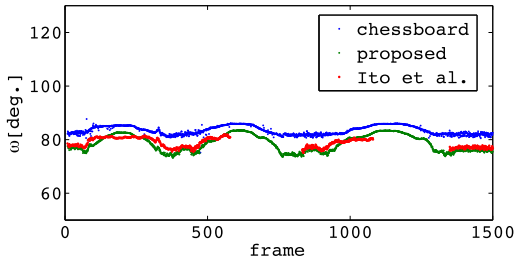

Fig. 2 Temporal variations in the planar poses computed from the results of the proposed method (green dots) and of Ito et al.'s method (red dots). The pose estimated from the images of a chessboard is also shown (blue dots). The vertical axis indicates a component of the three-vector representing the rotational component of planar pose as the angle-axis representation.

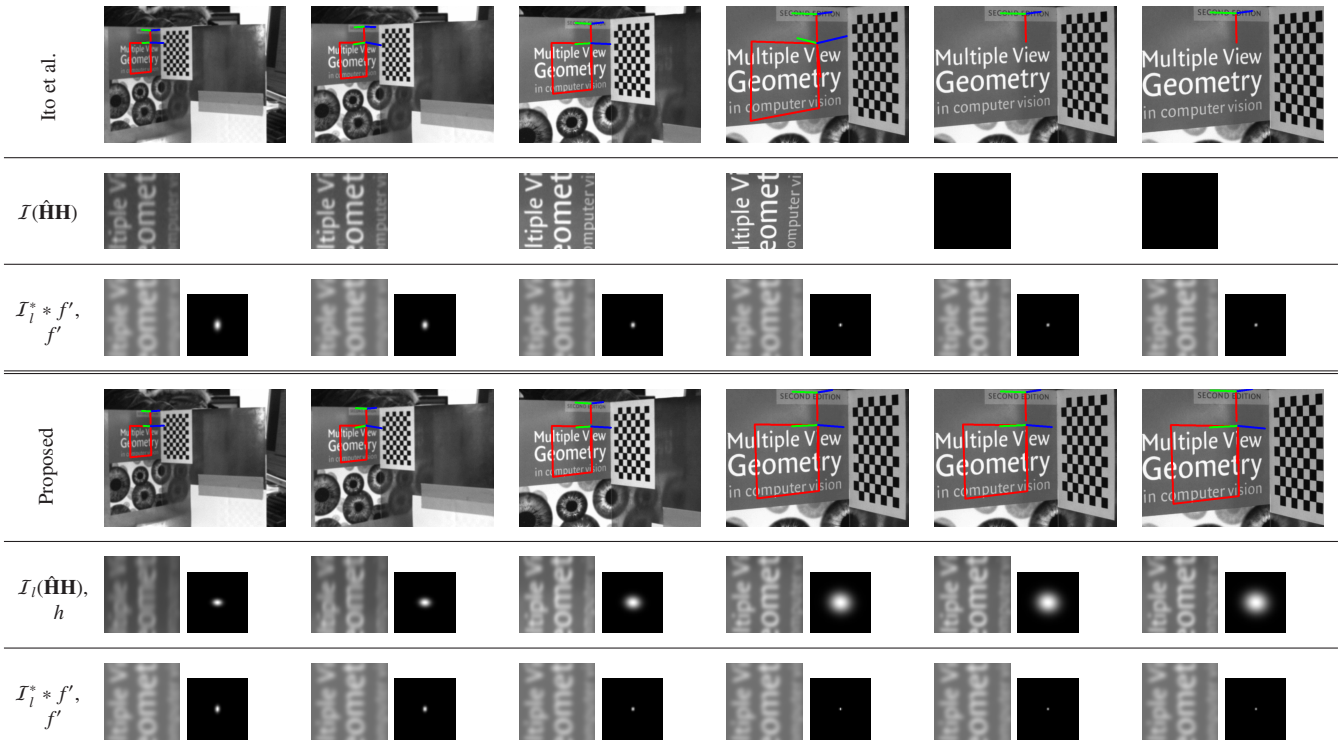

Fig. 3 Snapshots of the tracking results of Fig. 2. From left to right columns, 420, 480, 525, 580, 581, and 625-th frame. Colored axes are overlaid into the input images along with the tracked region. The upper, smaller ones represent the planar poses estimated by the chessboard. The lower ones represent those computed from the estimated homographies.

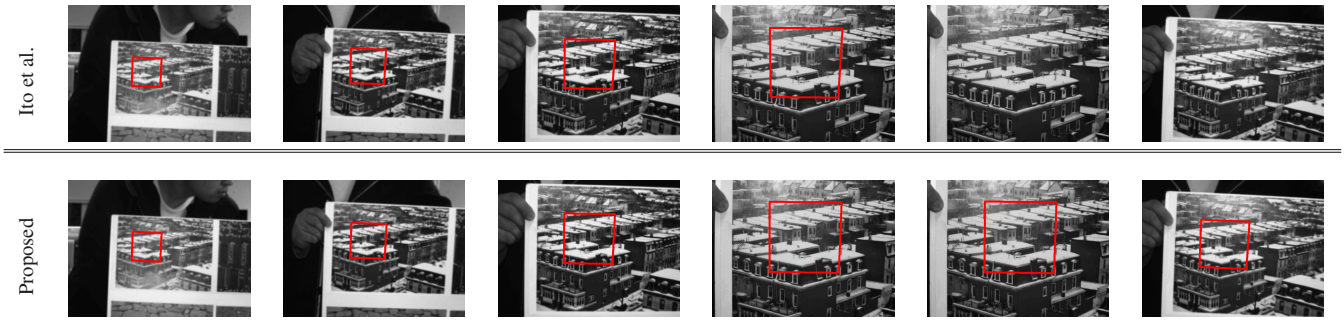

Fig. 4 Snapshots of tracking when using another template. Ito et al.'s method failed tracking from the fifth column, whereas ours can continue tracking.

column), it cannot reduce the blur anymore, resulting in that there is a significant difference between the warped input and the template. Their method could not continue tracking beyond this point (the fifth and sixth columns). On the other hand, the proposed method also applies the blurring filter to the input images, whose size and shape are controlled in a complementary way to the template filter. It can increase the blur of the input image filter whenever that of the template filter is minimized, resulting in that it can continue tracking.

We conducted experiments using a variety of templates, yielding similar results. An example is shown in **Fig. 4**.

5. Summary

We have described a method for planar tracking that can achieve improved accuracy by resolving the resolution inconsistencies between the input images and the template. It extends Ito et al.'s method, which considers only one half of the inconsistencies, to be able to deal with the other half such that the input images are of higher resolution than the template. It can deal with both types of resolution inconsistencies in a unified manner. The experimental results validate the performance of the proposed approach.

References

[1] Baker, S. and Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework, *International Journal of Computer Vision*, Vol.56, pp.221–255 (2004).
[2] Benhimane, S. and Malis, E.: Real-time image-based tracking of planes using Efficient Second-order Minimization, *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp.943–948 (2004).
[3] Benhimane, S. and Malis, E.: Homography-based 2D Visual Tracking and Servoing, *International Journal of Robotics Research*, Vol.26, pp.661–676 (2007).
[4] Dame, A. and Marchand, E.: Accurate real-time tracking using mutual information, *Proc. IEEE Int. Symp. Mixed and Augmented Reality*, pp.47–56 (2010).
[5] Dowson, N. and Bowden, R.: A Unifying Framework for Mutual Information Methods, *Proc. European Conference Computer Vision*, pp.365–378 (2006).

[6] Dowson, N. and Bowden, R.: Mutual Information for Lucas-Kanade Tracking (MILK): An Inverse Compositional Formulation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, pp.180–185 (2008).
[7] Holzer, S., Hinterstoisser, S., Ilic, S. and Navab, N.: Distance transform templates for object detection and pose estimation, *Proc. Computer Vision and Pattern Recognition*, pp.1177–1184 (2009).
[8] Ito, E., Okatani, T. and Deguchi, K.: Accurate and robust planar tracking based on a model of image sampling and reconstruction process, *Proc. ISMAR* (2011).
[9] Lepetit, V. and Fua, P.: Keypoint Recognition Using Randomized Trees, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, pp.1465–1479 (2006).
[10] Lucas, B.D. and Kanade, T.: An iterative image registration technique with an application to stereo vision, *Proc. International Joint Conference on Artificial Intelligence*, pp.674–679 (1981).
[11] Malis, E.: Improving vision-based control using efficient second-order minimization techniques, *Proc. International Conference on Robotics and Automation*, Vol.2, pp.1843–1848 (2004).
[12] Mei, C. and Reid, I.: Modeling and Generating Complex Motion Blur for Real-time Tracking, *Proc. Computer Vision and Pattern Recognition* (2008).
[13] Ozuysal, M., Fua, P. and Lepetit, V.: Fast Keypoint Recognition in Ten Lines of Code, *Proc. Computer Vision and Pattern Recognition*, pp.1–8 (2007).
[14] Park, Y., Lepetit, V. and Woo, W.: ESM-Blur: Handling and Rendering Blur in 3D Tracking and Augmentation, *Proc. International Symposium on Mixed and Augmented Reality* (2009).
[15] Silveira, G. and Malis, E.: Real-time Visual Tracking under Arbitrary Illumination Changes, *Proc. Computer Vision and Pattern Recognition*, Vol.0, pp.1–6 (2007).

(Communicated by *Mitsuru Ambai*)