Spatio-temporal Super-Resolution Using Depth Map

Yusaku Awatsu, Norihiko Kawai, Tomokazu Sato, and Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
http://yokoya.naist.jp/

Abstract. This paper describes a spatio-temporal super-resolution method using depth maps for static scenes. In the proposed method, the depth maps are used as the parameters to determine the corresponding pixels in multiple input images by assuming that intrinsic and extrinsic camera parameters are known. Because the proposed method can determine the corresponding pixels in multiple images by a one-dimensional search for the depth values without the planar assumption that is often used in the literature, spatial resolution can be increased even for complex scenes. In addition, since we can use multiple frames, temporal resolution can be increased even when large parts of the image are occluded in the adjacent frame. In experiments, the validity of the proposed method is demonstrated by generating spatio-temporal super-resolution images for both synthetic and real movies.

Keywords: Super-resolution, Depth map, View interpolation.

1 Introduction

A technology that enables users to virtually experience a remote site is called telepresence [1]. In a telepresence system, it is important to provide users with high spatial and high temporal resolution images in order to make users feel like they are existing at the remote site. Therefore, many methods that increase spatial and temporal resolution have been proposed.

The methods that increase spatial resolution can be generally classified into methods that use one image as input [2,3] and methods that require multiple images as input [4,5,6,7]. The methods using one image are further classified into two types: ones that need a database [2] and ones that do not [3]. The former method increases the spatial resolution of the low resolution image based on previous learning of the correlation between various pairs of low and high resolution images. The latter method increases the spatial resolution by using a local statistic. These methods are effective for limited scenes but largely depend on the database and the scene. The methods using multiple images increase the spatial resolution by corresponding pixels in the multiple images that are taken from different positions. These methods determine pixel values in the super-resolved image by blending the corresponding pixel values [4,5,6] or minimizing

A.-B. Salberg, J.Y. Hardeberg, and R. Jenssen (Eds.): SCIA 2009, LNCS 5575, pp. 696–705, 2009.
© Springer-Verlag Berlin Heidelberg 2009

the difference between the pixel values in an input image and the low resolution image generated from the estimated super-resolved image [7]. Both methods require the correspondence of pixels with sub-pixel accuracy. However, in these methods, the target scene is quite limited because the constraints of objects in the target scene such as planar constraint are often used in order to correspond the points with sub-pixel accuracy.

The temporal super-resolution method increases the temporal resolution by generating interpolated frames between the adjacent frames. Methods have been proposed that generate an interpolated frame by morphing that uses the movement of the points between adjacent frames [8,9]. Generally, the quality of the generated image by morphing largely depends on the number of corresponding points between the adjacent frames. Therefore, especially when many corresponding points do not exist due to occlusions, the methods rarely obtain good results.

The methods that simultaneously increase the spatial and temporal resolution by integrating the images from multiple cameras have been proposed [10,11]. These methods are effective for dynamic scenes but require a high-speed camera that can capture the scene faster than ordinary cameras. Therefore, the methods cannot be applied to a movie taken by an ordinary camera.

In this paper, by paying attention to the fact that determination of dense corresponding points is essential for spatio-temporal super-resolution, we propose the method that determines corresponding points of multiple images with sub-pixel accuracy by one-dimensionally searching for the corresponding points using the depth value of each pixel as a parameter. In this research, each pixel in multiple images is corresponded with high accuracy without the strong constraints for a target scene such as the planar assumption by a one-dimensional search of depth under the condition that intrinsic and extrinsic camera parameters are known. In work similar to our method, the spatial super-resolution method that uses a depth map has already been proposed [12]. However, this method needs stereo-pair images and does not increase the temporal resolution. Our advantages are that: (1) a stereo camera is not needed but only a general camera is needed, (2) the temporal resolution is increased by applying the proposed spatial super-resolution method to a virtual viewpoint located between temporally adjacent viewpoints of input images, and (3) corresponding points are densely determined by considering occlusions based on the estimated depth map.

2 Generation of Spatio-temporal Super-Resolved Images Using Depth Maps

This section describes the proposed method which generates spatio-temporal super-resolved images by corresponding pixels in each frame using depth maps. Here, in this research, a target scene is assumed to be static and camera position and posture of each frame and initial depth maps are given by some other methods like structure from motion and multi-baseline stereo. In the proposed method, the spatial resolution is increased by minimizing the energy function, which is based on the image consistency and the depth smoothness. The

temporal resolution is also increased by the same framework with the spatial super-resolution method.

2.1 Energy Function Based on Image Consistency and Depth Smoothness

Energy function E_f for the target f-th frame is defined by the sum of two different kinds of energy terms:

$$E_f = E_{If} + wE_{Df}, \tag{1}$$

where E_{If} is the energy for the consistency between the pixel values in the super-resolved image of the target f-th frame and those in the input images of each frame, E_{Df} is the energy for the smoothness of the depth map, and w is the weight. In the following, the energies E_{If} and E_{Df} are described in detail.

(1) Energy E_{If} for Consistency
The energy E_{If} is defined based on the plausibility of the super-resolved image of the f-th frame using multiple input images from the a-th frame to the b-th frame ($a \leq f \leq b$) as follows:

$$E_{If} = \frac{\sum_{n=a}^{b} \left|\mathbf{N}(\mathbf{O}_n)(\mathbf{g}_n - \mathbf{m}_{nf})\right|^2}{\sum_{n=a}^{b} |\mathbf{O}_n|^2}. \tag{2}$$

Here, $\mathbf{g}_n = (g_{n1}, \cdots, g_{np})^T$ is a vector notation of pixel values in an input image of the n-th frame and $\mathbf{m}_{nf} = (m_{nf1}, \cdots, m_{nfp})^T$ is a vector notation of pixel values in the image of the n-th frame simulated by the estimated super-resolved image and the depth map of the f-th frame (Fig. 1). $\mathbf{N}(\mathbf{O}_n)$ is a $p \times p$ diagonal matrix whose on-diagonal element is the same as the element of vector $\mathbf{O}_n$. Although E_{If} is basically calculated based on the difference between the input image $\mathbf{g}_n$ and the simulated image $\mathbf{m}_{nf}$, some pixels in the simulated image $\mathbf{m}_{nf}$ do not correspond to pixels in the f-th frame due to occlusions and projection to the outside of the image. Therefore, by using the mask image $\mathbf{O}_n = (O_{n1}, \cdots, O_{np})$ whose element is 0 or 1, the energies of the non-corresponding pixels are not calculated in Eq. (2). Here, the simulated low-resolution image $\mathbf{m}_{nf}$ is generated as follows:

$$\mathbf{m}_{nf} = \mathbf{H}_{fn}(\mathbf{z}_f)\mathbf{s}_f, \tag{3}$$

where $\mathbf{s}_f = (s_{f1}, \cdots, s_{fq})^T$ is a vector notation of pixel values in the super-resolved image and $\mathbf{z}_f = (z_{f1}, \cdots, z_{fq})^T$ is a vector notation of depth values corresponding to the pixels in the super-resolved image $\mathbf{s}_f$. $\mathbf{H}_{fn}(\mathbf{z}_f)$ is the transformation matrix that generates the simulated low-resolution image of n-th frame from the super-resolved image of the f-th frame by using the depth map $\mathbf{z}_f$. $\mathbf{H}_{fn}(\mathbf{z}_f)$ is represented as follows:

$$\mathbf{H}_{fn}(\mathbf{z}_f) = \left[\, \alpha_1 \mathbf{h}_1, \cdots, \alpha_i \mathbf{h}_i, \cdots, \alpha_p \mathbf{h}_p \,\right]^T, \tag{4}$$

where α_i is a normalization factor and $\mathbf{h}_i$ is a q-dimensional vector.

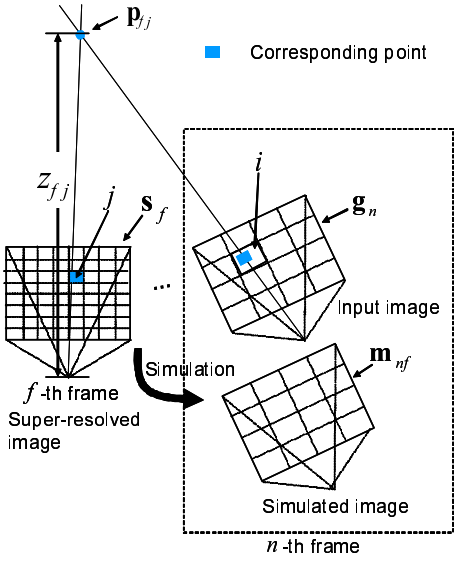


Fig. 1. Relationship between an input image and a super-resolved image

$$\mathbf{h}_i = \left(h_{i1}, \cdots, h_{ij}, \cdots, h_{iq} \right)^T . \tag{5}$$

Here, h_{ij} is a scalar value (1 or 0) that indicates the existence of correspondence between the j-th pixel in the super-resolved image and the i-th pixel in the input image. h_{ij} is calculated based on the estimated depth map as follows:

$$h_{ij} = \begin{cases} 0; d_n(\mathbf{p}_{fj}) \neq i \ \ or \ \ z'_{fj} > z_{ni} + C \\ 1; otherwise, \end{cases} \tag{6}$$

where $\mathbf{p}_{fj}$ indicates the three-dimensional coordinate in the scene corresponding to the j-th pixel in the super-resolved image as shown in Fig. 1 and $d_n(\mathbf{p})$ indicates the index of pixels in the n-th frame onto which $\mathbf{p}$ is projected. As shown in Fig. 2, z'_{fj} is the depth value in the n-th frame converted from the depth value z_{fj} in the f-th frame and z_{ni} is the corresponding depth value in the n-th frame. C is a threshold for determining occlusion.

The normalization factor α_i in Eq. (4) is the number of pixels in the super-resolved image that are projected onto the i-th pixel in the simulated image $\mathbf{m}_{nf}$. α_i is defined as follows using $\mathbf{h}_i$:

$$\alpha_i = \begin{cases} 0 & ; |\mathbf{h}_i| = 0 \\ \frac{1}{|\mathbf{h}_i|^2} & ; |\mathbf{h}_i| > 0. \end{cases} \tag{7}$$

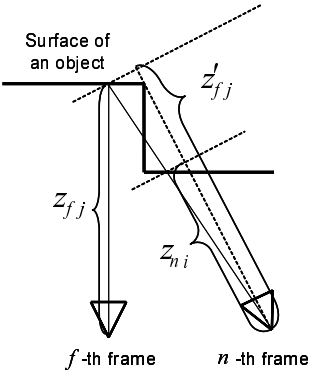


Fig. 2. Difference in depth by occlusion

(2) Energy E_{Df} for smoothness
The energy E_{Df} is defined based on the smoothness of the depth in the target frame as the following equation under the assumption that the depth along x and y direction is smooth in the target scene.

$$E_{Df} = \sum_j ((\frac{\partial^2 z_{fj}}{\partial x^2})^2 + 2(\frac{\partial^2 z_{fj}}{\partial x \partial y})^2 + (\frac{\partial^2 z_{fj}}{\partial y^2})^2), \tag{8}$$

2.2 Spatial Super-Resolution by Depth Optimization

In this research, a super-resolved image is generated by minimizing the energy E_f whose parameters are pixel and depth values in the super-resolved image. As shown in Eq. (2), E_{If} is calculated based on the difference between the input image $\mathbf{g}_n$ and the simulated image $\mathbf{m}_{nf}$. Here, whereas $\mathbf{g}_n$ is invariant, $\mathbf{m}_{nf}$ depends on the pixel values $\mathbf{s}_f$ and the depth values $\mathbf{z}_f$. Because it is difficult to minimize the energy by simultaneously updating the pixel and depth values in this research, the energy E_f is minimized by repeating the following two processes until the energy converges: (i) update of the pixel values $\mathbf{s}_f$ in the super-resolved image keeping the depth values $\mathbf{z}_f$ in the target frame fixed, (ii) update of the depth values $\mathbf{z}_f$ in the target frame keeping the pixel values $\mathbf{s}_f$ in the super-resolved image fixed.

In process (i), the transformation matrix $\mathbf{H}_{fn}(\mathbf{z}_f)$ for the pixel correspondence between the super-resolved image and the input image is invariant because the depth values $\mathbf{z}_f$ in the target frame are fixed. The energy E_{Df} for depth smoothness is also constant. Therefore, in order to minimize the total energy E_f, the pixel values $\mathbf{s}_f$ in the super-resolved image are updated so as to minimize the energy E_{If} for the image consistency. Here, each pixel value s_{fj} in the super-resolved image is updated in a way similar to method [7] as follows:

$$s_{fj} \leftarrow s_{fj} + \frac{\sum_{n=a}^{b}((g_{ni} - m_{nfi})O_{ni})}{\sum_{n=a}^{b} O_{ni}} \tag{9}$$

In process (ii), the depth values $\mathbf{z}_f$ are updated by fixing the pixel values $\mathbf{s}_f$ in the super-resolved image. In this research, because each pixel value in the simulated image $\mathbf{m}_{nf}$ discontinuously changes by the change in the depth $\mathbf{z}_f$, it is difficult to differentiate the energy E_f with respect to depth. Therefore, each depth value is updated by discretely moving the depth within a small range so as to minimize the energy E_f.

2.3 Temporal Super-Resolution by Setting a Virtual Viewpoint

In this research, a temporal interpolated image is generated by applying completely the same framework with the spatial super-resolution to a virtual viewpoint located between temporally adjacent viewpoints of input images. Here, because camera position and posture and a depth map, which are used for spatial super-resolution, are not given for an interpolated frame, it is necessary to set these values.

The position of the interpolated frame is determined by averaging the positions of the adjacent frames. If we want to generate multiple interpolated frames, the positions of adjacent frames are divided internally according to the number of interpolated frames. The posture of the interpolated frame is also determined by interpolating roll, pitch and yaw parameters of adjacent frames. The depth map of the interpolated frame is generated by averaging the depth maps of the adjacent frames.

3 Experiments

In order to demonstrate the effectiveness of the proposed method, spatio-temporal super-resolution images are generated for both synthetic and real movies.

3.1 Spatio-temporal Super-Resolution for a Synthetic Movie

In this experiment, a movie taken in a virtual environment as shown in Fig. 3 was used as input. Here, true camera position and posture of each frame were used as input. As for the initial depth values, Gaussian noise equivalent to an average of one pixel projection error on an image was added to the true depth values and the depth values were used as input. Table 1 shows parameters, and all 31 input frames are used for spatio-temporal super-resolution. In this experiment, a PC (CPU: Xeon 3.4GHz, Memory: 3GB) was used and it took about five minutes to generate one super-resolved image.

Table 1. Parameters in experiment

Input movie	320 240[pixels] 31[frames]
Output movie	640 480[pixels] 61[frames]
Weight w	100
Threshold C	1[m]

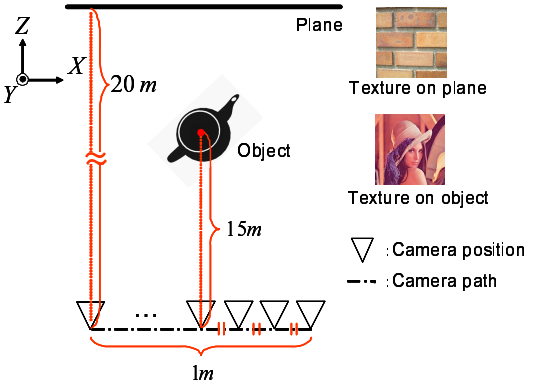
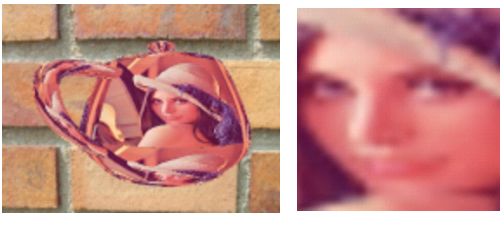
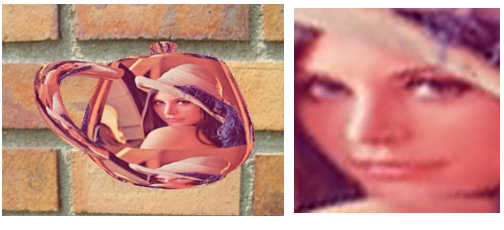


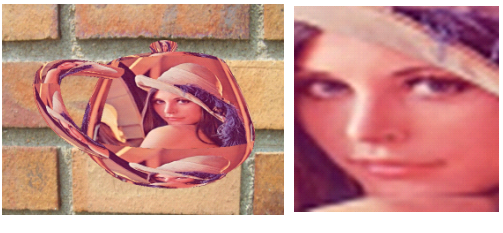
Fig. 3. Experimental environment



(a) Input image (Bilinear interpolation)



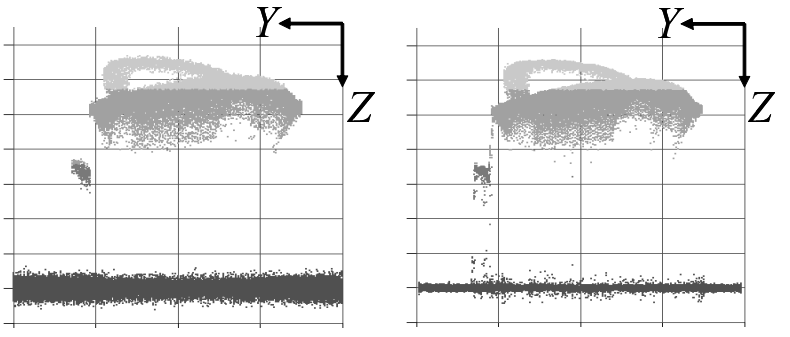
(b) Super-resolved image



(c) Ground truth image

Fig. 4. Comparison of images

Figure 4 shows the enlarged input image by bilinear interpolation (a), the super-resolved image generated by the proposed method (b) and a ground truth image (29-th frame) (c). The right part of each figure is a close-up of the same



(a) Initial depth (YZ plane) (b) Optimized depth (YZ plane)

Fig. 5. Change in depth

region. From Fig. 4, the quality of the image is improved by super-resolution of the proposed method. Figure 5 shows the initial depth values and the depth values after energy minimization. From this figure, the depth values become smooth from the noisy ones.

Next, the spatio-temporal super-resolved images generated by the proposed method were evaluated quantitatively by calculating PSNR (Peak Signal to Noise Ratio) using the ground truth images. Here, as comparison movies, the following two movies were used.

(a) A movie in which the spatial resolution is enlarged by bilinear interpolation and the temporal resolution is the ground truth
(b) A movie in which the interpolation frame is generated by using the adjacent previous frame and the spatial resolution is the ground truth

Figure 6 shows PSNR between the ground truth images and the images by each method. Here, as for movie (b), PSNR only for the interpolated frames is shown because the interpolated frame in movie (b) is the same as the ground truth image. From this figure, the super-resolved images by the proposed method obtained higher PSNR than movie (a). In the interpolated frames, the super-resolved images by the proposed method also obtained higher PSNR than movie (b). However, in the proposed method, the improvement effectiveness of the image quality is small around the first and last frames. This is because there are only a few frames that are taken at spatially close positions from the observed position of the target frame.

3.2 Super-Resolution for a Real Image Sequence

In this experiment, a video movie was taken by Sony HDR-FX1 (1920 × 1080 pixels) from the air and we used a movie that was scaled to 320 × 240 pixels by averaging pixel values as input. As camera position and posture, we used the parameters estimated by structure from motion based on feature point tracking

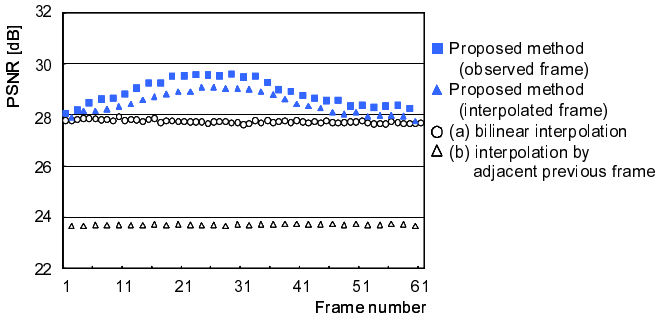
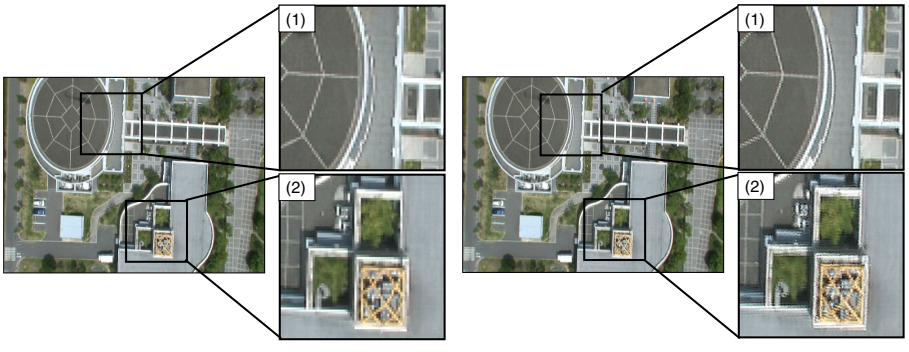


Fig. 6. Comparison of PSNR between the ground truth images and the images by each method



(a) Input image (b) Super-resolved image

Fig. 7. Comparison of input and super-resolved images

[13]. As initial depth maps, we used the interpolated depth map estimated by multi-baseline stereo for interest points [14]. Figure 7 shows the input image of the target frame and the super-resolved image (640 × 480 pixels) generated by using eleven frames around the target frame. From this figure, both the improved part ((1) in this figure) and the degraded part ((2) in this figure) can be observed. We consider that this is because the energy converges to a local minimum because the initial depth values are largely different from the ground truth due to the depth interpolation.

4 Conclusion

In this paper, we have proposed a spatio-temporal super-resolution method by simultaneously determining the corresponding points among many images by using the depth map as a parameter under the condition that camera parameters are given. In an experiment using a simulated video sequence, super-resolved

images were quantitatively evaluated by RMSE using the ground truth image and the effectiveness of the proposed method was demonstrated by comparison with other methods. In addition, a real movie was also super-resolved by the proposed method. In future work, the quality of the super-resolved image should be improved by increasing the accuracy of correspondence of points by optimizing the camera parameters.

Acknowledgments. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 19200016.

References

1. Ikeda, S., Sato, T., Yokoya, N.: Panoramic Movie Generation Using an Omnidirectional Multi-camera System for Telepresence. In: Proc. Scandinavian Conf. on Image Analysis, pp. 1074–1081 (2003)
2. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based Super-Resolution. IEEE Computer Graphics and Applications 22, 56–65 (2002)
3. Hong, M.C., Stathaki, T., Katsaggelos, A.K.: Iterative Regularized Image Restoration Using Local Constraints. In: Proc. IEEE Workshop on Nonlinear Signal and Image Processing, pp. 145–148 (1997)
4. Zhao, W.Y.: Super-Resolving Compressed Video with Large Artifacts. In: Proc. Int. Conf. on Pattern Recognition, vol. 1, pp. 516–519 (2004)
5. Chiang, M.C., Boult, T.E.: Efficient Super-Resolution via Image Warping. Image and Vision Computing, 761–771 (2000)
6. Ben-Ezra, M., Zomet, A., Nayar, S.K.: Jitter Camera: High Resolution Video from a Low Resolution Detector. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 135–142 (2004)
7. Irani, M., Peleg, S.: Improving Resolution by Image Registration. Graphical Models and Image Processing 53(3), 231–239 (1991)
8. Yamazaki, S., Ikeuchi, K., Shingawa, Y.: Determining Plausible Mapping Between Images Without a Priori Knowledge. In: Proc. Asian Conf. on Computer Vision, pp. 408–413 (2004)
9. Chen, S.E., William, L.: View Interpolation for Image Synthesis. In: Proc. Int. Conf. on Computer Graphics and Interactive Techniques, vol. 1, pp. 279–288 (1993)
10. Shechtman, E., Caspi, Y., Irani, M.: Space-Time Super-Resolution. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(4), 531–545 (2005)
11. Imagawa, T., Azuma, T., Sato, T., Yokoya, N.: High-spatio-temporal-resolution image-sequence reconstruction from two image sequences with different resolutions and exposure times. In: ACCV 2007 Satellite Workshop on Multi-dimensional and Multi-view Image Processing, pp. 32–38 (2007)
12. Kimura, K., Nagai, T., Nagayoshi, H., Sako, H.: Simultaneous Estimation of Super-Resolved Image and 3D Information Using Multiple Stereo-Pair Images. In: IEEE Int. Conf. on Image Processing, vol. 5, pp. 417–420 (2007)
13. Sato, T., Kanbara, M., Yokoya, N., Takemura, H.: Camera parameter estimation from a long image sequence by tracking markers and natural features. Systems and Computers in Japan 35, 12–20 (2004)
14. Sato, T., Yokoya, N.: New multi-baseline stereo by counting interest points. In: Proc. Canadian Conf. on Computer and Robot Vision, pp. 96–103 (2005)