PAPER

Generation of a Zoomed Stereo Video Using Two Synchronized Videos with Different Magnifications*

Yusuke HAYASHI[†a)], **Norihiko KAWAI**[†], **Tomokazu SATO**[†], **Miyuki OKUMOTO**[††], ***Members***, ***and*** **Naokazu YOKOYA**[†], ***Fellow***

SUMMARY This paper proposes a novel approach to generate stereo video in which the zoom magnification is not constant. Although this has been achieved mechanically in a conventional way, it is necessary for this approach to develop a mechanically complex system for each stereo camera system. Instead of a mechanical solution, we employ an approach from the software side: by using a pair of zoomed and non-zoomed video, a part of the non-zoomed video image is cut out and super-resolved for generating stereo video without a special hardware. To achieve this, (1) the zoom magnification parameter is automatically determined by using distributions of intensities, and (2) the cutout image is super-resolved by using optically zoomed images as exemplars. The effectiveness of the proposed method is quantitatively and qualitatively validated through experiments.

key words: *stereo video, super-resolution, energy minimization*

1. Introduction

Stereoscopic displays including 3D TV, on which viewers can see stereoscopic vision with or without special glasses, have become popular through the evolution of display devices. In order to make contents of real scenes for stereoscopic displays, typically, two approaches have been employed: One uses two identical video cameras that are arranged in parallel, and the other converts 2D video, which is captured using a single video camera, to 3D [2]. Here, we focus on the situation where the optical zoom magnification is changed while video capturing.

When using two cameras, a straightforward way to handle this situation is to develop a stereo camera system that has two special functions: Synchronization of the optical zoom magnifications and the directions of two cameras' optical axes. However, such a system is mechanically complex, and one problem is that a unique system is required for each kind of stereo camera unit, which results in increasing the development cost. This is especially true for a 3D digital cinema camera system such as that shown in Fig. 1 using 4K ($4,096 \times 2,160$) or 8K ($8,192 \times 4,320$) cameras, because such a camera has a large zoom-adjusting mechanism. On the other hand, automatic 2D/3D conversion often gives unnatural stereoscopic images [2]. Although 2D/3D conversion is often employed in the production of 3D movies, a large amount of manual work is necessary to remove unnaturalness.

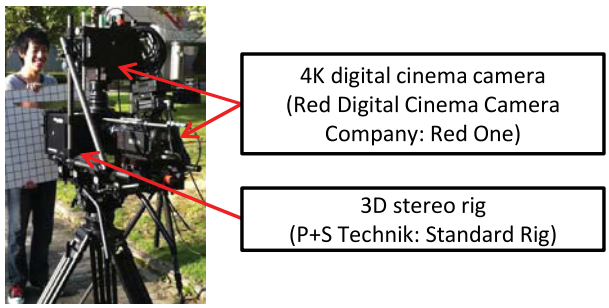


Fig. 1 3D digital cinema camera system consisting of two 4K cameras and a 3D stereo rig.

In this paper, we propose a different approach from conventional ones to make zoomed stereo video for the purpose of generating stereo video contents in offline process. In this study, although we also use two identical video cameras arranged in parallel, a videographer manipulates zoom magnification of only one of the two cameras while keeping that of the other fixed. We then generate a high-resolution zoomed stereo video from a pair of non-zoomed video and optically zoomed video, which has not previously been attempted. Specifically, we first cut out the part of a non-zoomed video image that corresponds to the other optically zoomed image. The cutout part is then super-resolved using the optically zoomed image as the example, so that it has the same resolution. This is effective because there is a high correlation between the target low-resolution image and the high-resolution image captured by two cameras arranged in parallel.

This paper is organized as follows: In Sect. 2, conventional methods for super-resolution relevant to our approach are reviewed. Section 3 presents our proposed method. Experimental results using the proposed and conventional methods are presented and discussed in Sect. 4. Finally, Sect. 5 summarizes this paper.

2. Related Work

Super-resolution methods can be broadly classified into three categories: Reconstruction-based, example-based and

Manuscript received April 20, 2015.
Manuscript publicized June 17, 2015.
[†]The authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The author is with Tokuyama College of Technology, Shunan-shi, 745–8585 Japan.
*This paper is an extended version of the one presented at the 2013 Pacific-Rim Symposium on Image and Video Technology [1].
a) E-mail: hayashi.yusuke.hq7@is.naist.jp
DOI: 10.1587/transinf.2015EDP7150

filter-based methods. In reconstruction-based methods, input images are super-resolved by aligning multiple low-resolution images with sub-pixel accuracy [3]–[7]. These methods have to capture many images of the same scene without moving objects. Since our target scenes often include moving objects, this approach is not suitable for our solution.

On the other hand, in example-based methods, a target low-resolution image is super-resolved using example high-resolution images [8]–[16]. Freeman et al. [9] proposed an example-based learning strategy in which the low-resolution to high-resolution prediction is learned from generic images via a Markov network. The quality of resultant images by example-based methods was improved by introducing some constraints [10], [12]. Sun et al. [10] extended this approach using primal sketch priors to enhance edges. Begin et al. [12] extended this approach by estimating PSF parameters. In addition to constraints, the searching techniques for matching patches were proposed to improve the quality of super-resolved image [14]–[16]. Low-resolution image patches of the leaning dataset are enhanced to increase the accuracy of matching [15]. Hashimoto et al. [16] proposed a method that efficiently searches for correspondences using a binary tree dictionary. However, the learning data set is generally fixed, and thus super-resolution effects are limited. Baker et al. [14] restricted the category of examples in the database by considering that the quality of the resultant image largely depends on the examples.

The filter based methods estimate a high-resolution image from a single low-resolution image by compensating high-frequency components of the input image [17]–[19]. Although these methods are applicable to many scenes, in general, filter-based methods obtain less natural and more blurry results than ones by example-based methods as far as appropriate examples are used.

In the proposed framework where a cutout image from the non-zoomed image is enlarged using a super-resolution technique to the same resolution of the optical-zoomed image, we use the optically zoomed image as an example. By doing this, we can obtain better results than using the database from various images as in the conventional methods because there is a high correlation between the target low-resolution image and the high-resolution image. In addition, we can use the epipolar constraint to improve the speed and precision of matching between the high-resolution and low-resolution images.

3. Generation of Super-Resolved Stereo Video

In this study, we assume that a target scene is captured by using two cameras that are set so that their optical axes and scan lines are parallel (parallel camera configuration). However, even in the case where image sequences are not captured in parallel camera configuration, they can easily be rectified at preprocessing stage. In particular, we calculate a fundamental matrix between a stereo pair of images captured with the same magnification. After that, subsequent

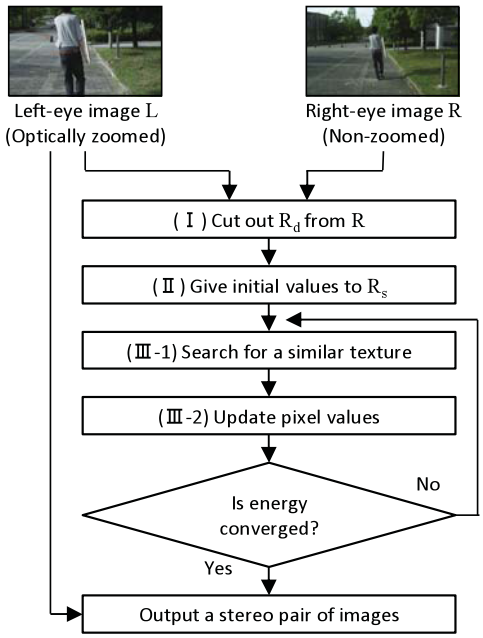


Fig. 2 Flow diagram of the proposed method.

image sequences are rectified using the fundamental matrix for the subsequent processes. In the following, to simplify our explanation, let the optically zoomed left-eye image be denoted by L, and the non-zoomed right-eye image by R.

Figure 2 illustrates the flow diagram of the proposed method. First, a part of the original right image R corresponding to the original left image L is cut out as the cutout right-eye image R_d (Process I). Next, R_d is enlarged to the size of L for giving initial values of the super-resolved image R_s (Process Ⅱ). We then minimize the energy function, which is defined based on the pattern similarity of (R_d, R_s) and the consistency of (L, R_s), to super-resolve the image repeating two processes: search L for a similar texture (Process Ⅲ-1) and update all the pixel values in R_s (Process Ⅲ-2).

When the input image is a color image, the proposed method uses RGB channels in Process I. The RGB channels of the image are then transformed to HSV ones in Process II. After that, the super-resolution process is then applied to V (intensity) channel, and H (hue) and S (saturation) channels are interpolated by bi-cubic interpolation as similar to most of conventional methods for super resolution, which transform RGB channels to intensity and chromatic ones and use the intensity one for super resolution. In the following sections, we detail the methods for cutout and energy minimization.

3.1 Cutout of Non-Zoomed Image

To cut out R_d from the original right image R so that its capturing area corresponds to the shooting range of the original left image L, we estimate transform matrix $\mathbf{M}$ that projects the four corners of the image L onto four points in the image R, as shown in Fig. 3. Here, transform matrix $\mathbf{M}$ is defined

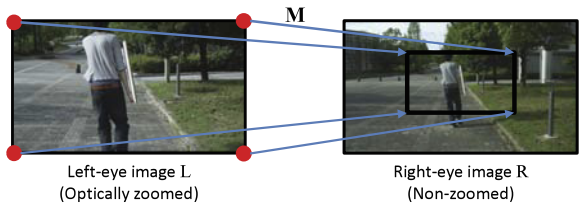


Fig. 3 Projection of four corners using transform matrix **M**.

Fig. 4 Example of graph representing average intensities of scan lines.

as

$$\mathbf{M} = \begin{pmatrix} s & 0 & t_x \\ 0 & s & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{1}$$

Translation parameters t_x and t_y are determined in advance by calibrating the zoom center of the camera. Scaling parameter s (where the zoom magnification is $1/s$) is determined for every frame so that a similarity measure between R_d and L is maximized.

In this study, the similarity measure is defined based on the normalized cross-correlation between two graphs generated using the average intensities of the scanlines in images L and R, as shown in Fig. 4. To successfully generate stereo images that can be fused by human eyes, it is important to align the horizontal lines. Thus, we determine scaling parameter s using horizontal line rather than the vertical line. The graph $h_L(y)$ of L for computing similarity is generated so that the vertical axis is the y coordinate of L, and the horizontal axis is the average pixel value on one horizontal line. For the graph $h_R(y)$, R_d is first cut out from R using the tentative matrix **M** and enlarged to the size of L. The tentative graph $h_R(y)$ is then generated from the enlarged R_d in the same way as for L. The graphs for the R, G and B components are generated, and the sum of the normalized cross-correlations for R, G and B are used as a similarity measure.

For determining the best scaling parameter s using this similarity, in this study, we basically employ exhaustive search. Concretely, in the first frame, we search for the optimal s by discretely shifting s in a given range. In order to prevent non-continuous changing of zoom magnification, in the subsequent frames, a searching range is limited using the value s determined for the previous frame.

Although stereo camera systems usually consist of the same type of two cameras, they often output different color tones because of the individual difference. Therefore, after cutting out the image R_d using determined **M**, we adjust the color tone of R_d so as to be similar with that of the image L for obtaining pixel correspondences between left and right images more correctly in the following energy minimization process and achieving the natural fusion of the generated stereo video. The R, G and B values of the image R_d are linearly transformed so that graphs $h_R(y)$ for RGB fit the graphs $h_L(y)$ for RGB in a least-squares manner.

3.2 Definition of Energy Function

In this section, energy function E that is used for super-resolution of the cutout image R_d is defined using two different kinds of energy terms as given in Eq. (2). E_{ssd} in Eq. (3) represents the pattern dissimilarity between the super-resolved image R_s and the original left image L, and E_{dif} in Eq. (4) represents the intensity difference between the super-resolved image R_s and the cutout image R_d.

$$E = \sum_{\mathbf{x}_i \in R_s} \{\lambda E_{ssd}(\mathbf{x}_i, \mathbf{x}_j) + (1-\lambda) E_{dif}(\mathbf{x}_i, g(\mathbf{x}_i))\}, \tag{2}$$

$$E_{ssd}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\mathbf{p} \in W} \{R_s(\mathbf{x}_i + \mathbf{p}) - L(\mathbf{x}_j + \mathbf{p})\}^2, \tag{3}$$

$$E_{dif}(\mathbf{x}_i, g(\mathbf{x}_i)) = \{R_s(\mathbf{x}_i) - R_d(g(\mathbf{x}_i))\}^2. \tag{4}$$

Here, W is a square window in R_s and L. $\mathbf{p}$ is a shift vector to indicate a pixel in W. λ is a weight for balancing the two terms. $\mathbf{x}_i$ denotes a pixel in R_s, and $\mathbf{x}_j$ is a pixel from L. $R_s(\mathbf{x}_i)$, $R_d(\mathbf{x}_i)$ and $L(\mathbf{x}_i)$ represent the intensities of pixel $\mathbf{x}_i$ in images R_s, R_d and L, respectively. $g(\mathbf{x}_i)$ denotes a pixel position in R_d that corresponds to pixel $\mathbf{x}_i$ in R_s. The relationship is

$$g(\mathbf{x}_i) = \mathbf{M}'\mathbf{x}_i, \tag{5}$$

where matrix $\mathbf{M}'$ is the same as matrix **M** except that the translation parameters are 0. E_{ssd} represents the effect of increasing the resolution of the generated image, and E_{dif} represents the preservation of the texture of the original right-eye image.

3.3 Iterative Energy Minimization

Energy function E is minimized by iterating the following two processes: search for similar patterns in L (Process Ⅲ-1) and update pixel values in R_s (Process Ⅲ-2). This iteration is stopped when the number of iterations reaches a given threshold.

3.3.1 Search Process Ⅲ-1

The whole pixel values in R_s are fixed, and the position $f(\mathbf{x}_i)$ of the texture pattern most similar to $\mathbf{x}_i$ is updated so as to satisfy the following equation.

$$f(\mathbf{x}_i) = \underset{\mathbf{x}_j \in \phi(\mathbf{x}_i)}{\operatorname{argmin}} E_{ssd}(\mathbf{x}_i, \mathbf{x}_j), \tag{6}$$

where $\phi(\mathbf{x}_i)$ is the search region of L that corresponds to the pixel $\mathbf{x}_i$ in the generated image. $\phi(\mathbf{x}_i)$ includes a set of pixels on the epipolar line and several pixels above and below the

epipolar line with consideration of calibration errors in the estimation process.

3.3.2 Update Process III-2

The pixel values $R_s(\mathbf{x}_i)$ in the generated image are updated in parallel so as to minimize energy function E defined in Eq. (2) while keeping all the similar pairs fixed. Energy function E is resolved into the element energy $E(\mathbf{x}_i)$ for each pixel $\mathbf{x}_i$ in R_s.

$$E(\mathbf{x}_i) = \lambda \sum_{\mathbf{p}\in W} \{R_s(\mathbf{x}_i) - L(f(\mathbf{x}_i + \mathbf{p}) - \mathbf{p})\}^2 + (1-\lambda)\{R_s(\mathbf{x}_i) - R_d(g(\mathbf{x}_i))\}^2. \quad (7)$$

Each element energy includes only one parameter and the total energy E consists of the sum of all element energies. Therefore, E can be minimized by minimizing each element energy. $R_s(\mathbf{x}_i)$ that minimizes $E(\mathbf{x}_i)$ can be calculated by differentiating $E(\mathbf{x}_i)$ with respect to $R_s(\mathbf{x}_i)$, and is

$$R_s(\mathbf{x}_i) = \frac{\lambda \sum_{\mathbf{p}\in W} L(f(\mathbf{x}_i+\mathbf{p})-\mathbf{p}) + (1-\lambda) R_d(g(\mathbf{x}_i))}{\lambda N_W + (1-\lambda)}, \quad (8)$$

where N_W denotes the number of pixels in the window.

3.4 Coarse-to-Fine Approach

In order to reduce the computational cost and avoid local minima, we use a coarse-to-fine approach for energy minimization. We first generate an image pyramid. In the coarsest level, the above Processes III-1 and III-2 are iterated until the energy converges. In subsequent levels, the generated texture in the previous level is used for the initial pixel values. Processes III-1 and III-2 are iterated until convergence where the search area for each pixel in R_s in Process III-1 is limited to a range around the memorized corresponding position in the previous level. In addition, in the finest level, we repeat the energy minimization while reducing the size of the window. This enables more detailed textures to be reproduced. This also reduces the effect of the disparity in the corresponding patterns between left and right images.

4. Experiments

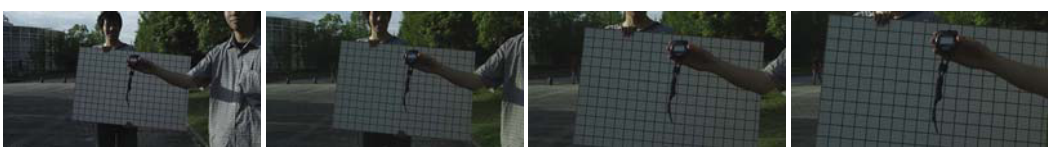
In order to demonstrate the effectiveness of the proposed method, we have performed experiments using stereo images from stereo image datasets [20], [21] and real stereo videos captured using 3D digital cinema camera system as shown in Fig. 1. In the experiments, we empirically determined the window size W for each dataset. We set the number of iterations as five for each dataset by experimentally confirming that the energy almost converges with five iterations in most of the cases.

We first confirmed the validity of the estimation of the zoom scale in Process I with preliminary experiments. Then, the super-resolved results by the proposed method are compared with the results of baseline methods. Finally, we analyze the computational cost of the proposed method.

4.1 Preliminary Experiments: Confirmation of Validity of Zoom Scale Estimation

In this section, we confirm the validity of the scale estimation method described in Sect. 3.1 in simulation. In this experiment, we used two stereo pairs of images (right and left images with $4,096 \times 2,160$ pixels) extracted from two real stereo videos in which the zoom magnifications of the two cameras were the same. To simulate stereo pairs with different zoom magnifications, the right-eye images were resized to half of the original size as non-zoomed images and the left-eye images were cut out and resized as optically zoomed images so that the size of the two images became the same. By doing this, we can simulate the situation where the resolution of right-eye and left-eye cameras is $2,048 \times 1,080$ pixels and the magnification of only the left-eye camera is between 1.0 and 2.0. In this experiment, we cut out each left-eye image to simulate a zoom, changing the magnification from 1.0 to 2.0 with a 0.1 skip and resizing the respective cutout images to $2,048 \times 1,080$ pixels. Figure 5 shows examples of input stereo pairs with different zoom magnifications.

Table 1 shows the estimated zoom magnifications using



(a) Scene 1.

(b) Scene 2.

Fig. 5 Examples of input stereo pairs with different zoom magnifications of two scenes. From left to right: simulated non-zoomed right-eye image, three simulated optically zoomed left-eye images with different magnifications (1.2, 1.6, 2.0).

Table 1 Estimated zoom magnifications.

Ground truth	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Estimated result of scene 1	1.001	1.106	1.204	1.300	1.396	1.497	1.613	1.721	1.831	1.908	1.996
Estimated result of scene 2	1.004	1.104	1.203	1.301	1.406	1.504	1.607	1.709	1.806	1.901	2.000

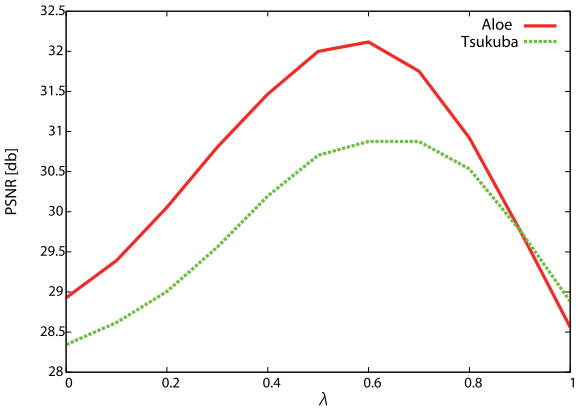


Fig. 6 PSNR with different λ.

the proposed method. The estimated magnifications were almost the same as the ground truth in many cases, and the maximum error ratio between the grand truth and the estimated magnifications was 1.7%.

We confirmed that this level of errors do not affect the stereo fusion by human eyes through a pre-experiment. However, it largely affects the similar pattern search in Process Ⅲ-1. Therefore, in the following experiments, we have set the search region ϕ in Eq. (6) for similar pattern search as the region containing 10 pixels above and below the epipolar line for allowing the estimation errors.

4.2 Experiments Using Stereo Image Datasets

In this experiment, we used two stereo pairs of images (Aloe with 640×554 pixels, and Tsukuba with 384×288 pixels) selected from the stereo image datasets [20], [21]. We reduced the size of each right image to simulate a stereo pair of input images with different zoom magnifications. For the coarse-to-fine approach in this experiment, we did not use an image pyramid because the input image size was small enough, but we reduced the window size W from 7×7 to 5×5 and 3×3 pixels as the energy converges in the original scale.

First, we verified the results using different λ values (from 0.0 to 1.0), which balances the two energy terms. For this simulation, we resized each right image to 1/2 of the original size and super-resolved the images using the proposed method with the known scaling parameter. Figure 6 shows the PSNR values of the generated images for different λ values. From this result, we can confirm that the PSNR becomes higher as λ approaches 0.6.

Next, we compared the results of three methods: the proposed method with $\lambda = 0.6$, example-based super-resolution method [16], and bi-cubic interpolation. We resized each right image to 1/2 and 1/4 of the original size. Figure 7 shows the results that correspond to the two resized ratios using the three methods. By comparing these images, we can confirm that the proposed method successfully reconstructed the high-frequency component. When the resized ratio was 1/4, the difference was especially noticeable. Figure 8 shows the PSNR values of the results of each method. The PSNR values of the proposed method are the highest for both images, when compared with the other methods. However, as shown in Fig. 9, the image generated using our method with $\lambda = 1.0$ includes some incorrect texture patterns. This is because similar patterns do not exist in the optically zoomed high-resolution image due to occlusions.

4.3 Experiments Using Real Stereo Videos

In order to show the practicality of the proposed method, we have conducted the experiment with objective evaluation using real 4K dataset. Here, we captured two stereo videos (Scene A and Scene B) with $4,096 \times 2,160$ pixels using the 3D digital cinema camera system, as shown in Figs. 10 (a), 10 (b), 11 (a) and 11 (b)). We changed the magnification of the left camera and kept the right magnification fixed. In the proposed method, $\lambda = 0.6$, as was suggested by our preliminary experiments (see Fig. 6). For the coarse-to-fine approach, we resized the input image to 1/8, 1/4 and 1/2 of the original size. We set the window size W to be 13×13 pixels. At the finest level, the window size was reduced to 9×9 and 5×5 pixels as the energy converged.

First, we evaluate the quality of each frame of the generated stereo videos by subjectively comparing the result of the proposed method with those by the conventional methods. Figures 10 (c) and 11 (c) show digitally zoomed right-eye image sequences of two scenes using the proposed method. Figures 10 (d)-10 (f) and 11 (d)-11 (f) show the closeups of images generated by bi-cubic interpolation, example-based super-resolution [16] and the proposed method.

From Figs. 10 (f) and 11 (f), we can confirm that the texture generated by the proposed method is clearer than that by both bi-cubic interpolation and the method in [16]. In addition, we can see the difference of ground color with Figs. 10 (b) and 10 (c) is much smaller than that of Figs. 10 (a) and 10 (b). Figure 12 shows histograms for (a) the input cutout and enlarged right-eye image without color adjustment, (b) input left-eye image and (c) generated right-eye image when the estimated zoom magnification is 2.338 in Fig. 10. From the figure, we can also confirm that our method compensated for the color tone of the generated right-eye image.

Next, we objectively evaluate the quality of the gener-

(a) Input images.

(b) Proposed method (Resized ratio: 1/2). (c) Proposed method (Resized ratio: 1/4).

(d) Example-based SR [16] (Resized ratio: 1/2). (e) Example-based SR [16] (Resized ratio: 1/4).

(f) Bi-cubic interpolation (Resized ratio: 1/2). (g) Bi-cubic interpolation (Resized ratio: 1/4).

Fig. 7 Super-resolved results by three methods.

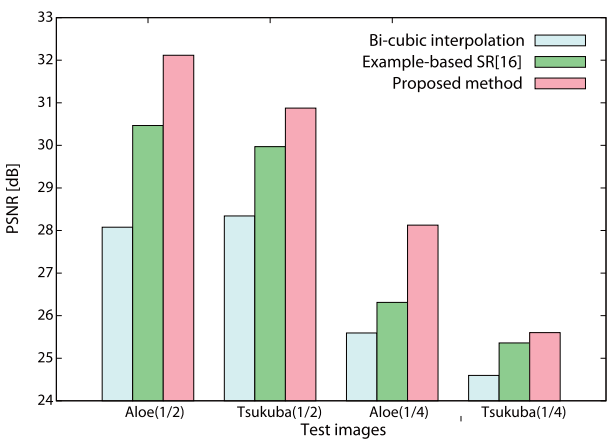


Fig. 8 PSNR of all test images.

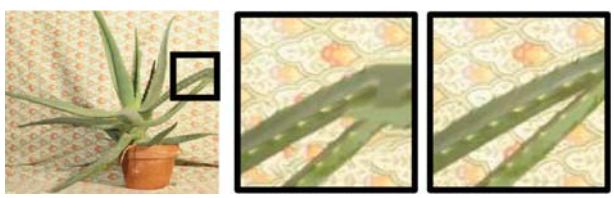


Fig. 9 Example of unnatural texture generated due to occlusions (from left to right: generated image, incorrect texture pattern, ground truth).

ated stereo videos by giving a questionnaire to eight subjects using a head mounted display (SONY HMZ-T1: 1980 × 1080 pixels) for showing stereo videos. In this evaluation, we prepared special videos from six video sequences by horizontally arranging two of three methods generated by above three methods as shown in Fig. 13. In order to evaluate all the combinations of three methods with this layout for six video sequences, we have prepared 18 special videos. The resolution of the special videos is adjusted to that of the HMD. For the questionnaire, the subjects were requested to watch the special videos and responded to the following two questions:

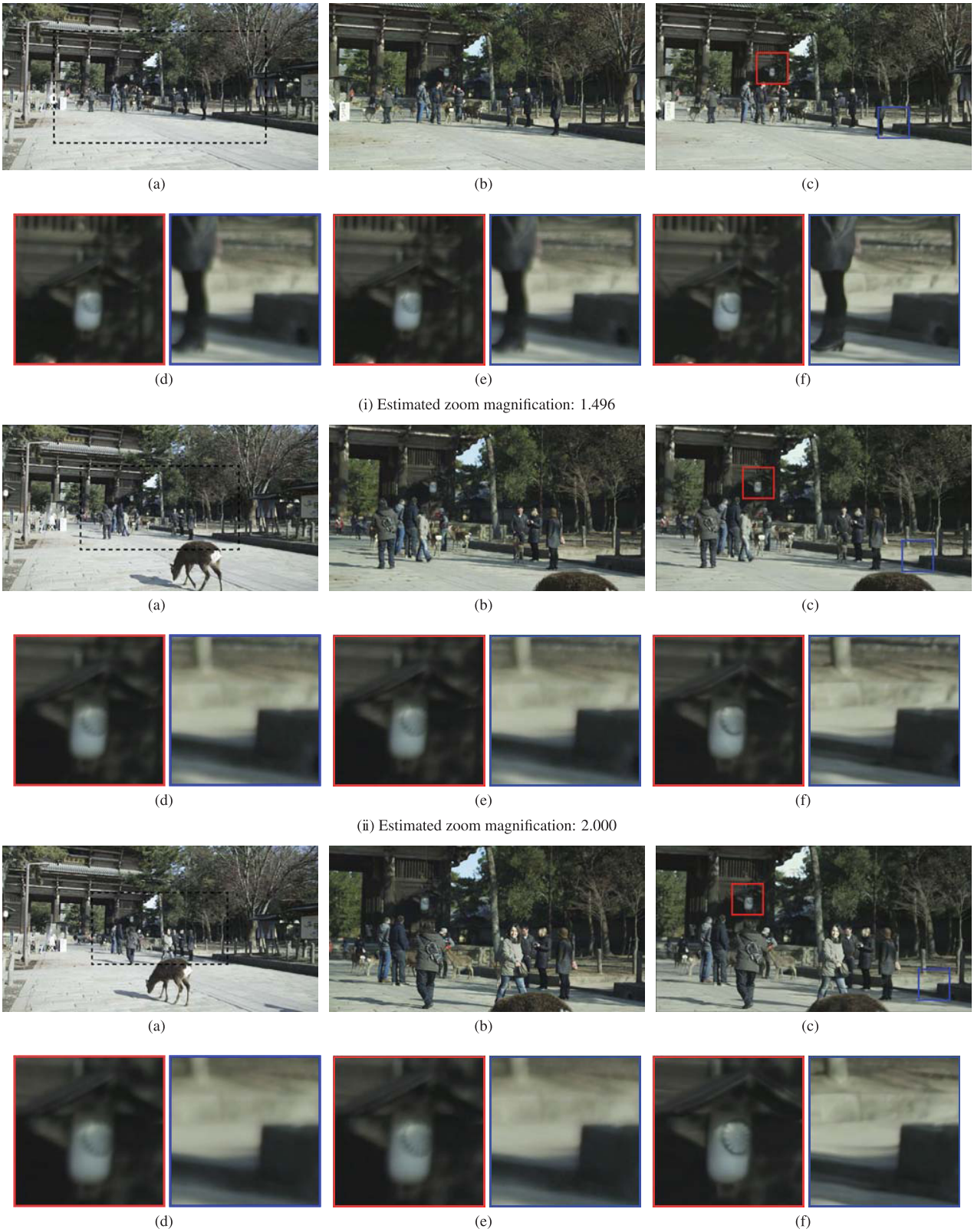
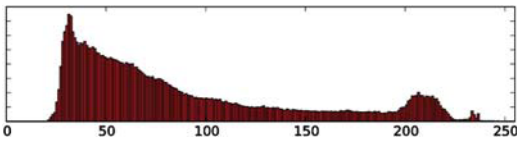
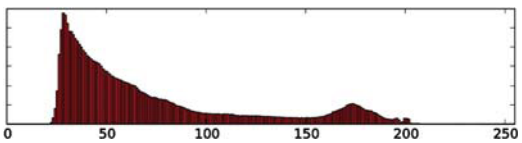


Fig. 10 Input image sequences and generated results for Scene A. (a) input right-eye images (the dotted rectangles show the cutout regions), (b) input left-eye images, (c) digitally zoomed right-eye images using the proposed method, (d)-(f) cutout regions of interest from digitally zoomed right-eye images using three methods (from left to right: bi-cubic interpolation, example-based SR [16], proposed method).

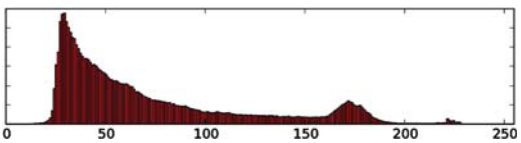
Fig. 11 Input image sequences and generated results for Scene B. (a) input right-eye images (the dotted rectangles show the cutout regions), (b) input left-eye images, (c) digitally zoomed right-eye images using the proposed method, (d)-(f) cutout regions of interest from digitally zoomed right-eye images using three methods (from left to right: bi-cubic interpolation, example-based SR [16], proposed method).



(a) Input cutout and enlarged right-eye image without color adjustment



(b) Input left-eye image



(c) Generated right-eye image

Fig. 12 Example of histograms of R channel for three images when the estimated zoom magnification is 2.338 in Fig. 10 (The vertical axis indicates the number of pixels and the horizontal axis indicates the intensity value).

Left-eye video Right-eye video

Fig. 13 Example of special video in which two results by the different methods are displayed in the upper and lower parts, respectively (In this figure, upper right-eye video is by bi-cubic interpolation, and lower right-eye video is by the proposed method). Subjects watch the fused video by displaying the left image on the left display and the right image on the right display in an HMD.

- Q1: Which stereo video is sharper?
- Q2: Which stereo video is more natural?

The subjects were allowed to answer the questions by selecting one of following the three answers: upper video, lower video, no difference. We randomized the order of special videos presented to the subjects, and allowed them to repeatedly watch the video until they answered the questions. The results of the evaluation are shown in Table 2. Each grid in the tables shows the number of answers in which the method in the row obtained "sharper" or "more natural" compared to the method in the column. From Table 2 (a), we can confirm that the proposed method obtained the larger number of wins than the conventional methods for sharpness, but

Table 2 Experimental results of the questionnaire. Each grid in the tables shows the number of answers in which the method in the row obtained "sharper" or "more natural" compared to the method in the column.

(a) Number of answers in which the method in the row is sharper than that in the column (Q1).

win \ lose	Bi-cubic	Example [16]	Proposed	total
Bi-cubic	-	16	12	28
Example [16]	17	-	12	29
Proposed	23	26	-	49

(b) Number of answers in which the method in the row is more natural than that in the column (Q2).

win \ lose	Bi-cubic	Example [16]	Proposed	total
Bi-cubic	-	18	19	37
Example [16]	13	-	16	29
Proposed	20	15	-	35

subjects sometimes gave better evaluation to the conventional methods than the proposed one. From the result, the proposed method is more effective to generate sharp stereo videos than the conventional ones in many cases. However, we also confirmed that the perception of some people often can compensate for the difference in sharpness when fusing two images even if the sharpness of the right and left images is different. From Table 2 (b), we can confirm that the number of wins for the proposed method is almost the same as bi-cubic interpolation. From the result, we confirmed that the naturalness of the result of the proposed method is not so different from those by the conventional ones although the proposed method sometimes generated unnatural textures as a video because it super-resolves the input image frame by frame. The evaluation results demonstrate the effectiveness of our method for generating zoomed stereo videos for real scenes.

4.4 Analysis of Computational Cost

We analyze the computational cost of the proposed method. For the real stereo videos with 4,096 × 2,160 in Sect. 4.3, approximately five minutes were required to generate each stereo pair of images using a PC (Intel Core i7 3.40GHz of CPU and 8.00 GB of memory). For example, for the frame with estimated zoom magnification 2.000 in scene A as shown in Fig. 10 (ii), it took 30, 1, 205, and 6 seconds for steps (I), (II), (III-1), and (III-2), respectively. From the results, we confirmed that the searching process (Process III-1) occupies the large part of the computational time. In our unoptimized implementation for the searching process, we calculate the sum of squared differences (SSD) between a window centered at a pixel in R_s and windows centered at pixels in the searching area, which is a set of pixels on the epipolar line and 10 pixels above and below it, and the calculation is iterated for all the pixels in R_s. Therefore, the large number of SSD calculation causes the high computational cost.

One of the solutions to reduce the computational cost is to employ the searching algorithm used in the stereo

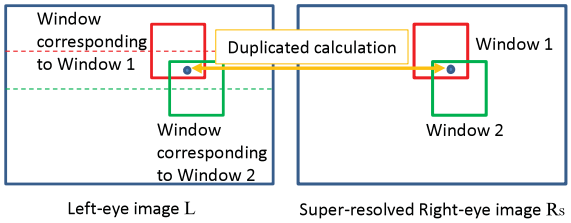


Fig. 14 Duplicated calculation of squared differences.

matching field, which omits the duplicated calculation of the squared differences [22]. As shown in Fig. 14, different pairs of windows between left-eye and super-resolved right-eye images calculate a squared difference between the same pair of pixels. Therefore, when the number of pixels in a window is N_W, the same calculation of a squared difference is iterated N_W times. To avoid this problem, we can first calculate the disparity space image (DSI) [22], which is the squared difference of two images, one of which is shifted a certain pixels. We then calculate the summation of values in the DSI over pixels in a window to obtain the SSD value. By doing this, we can reduce the computational cost for calculating squared differences to $1/N_W$ while the cost for the summation does not change. Suppose that the operation costs for addition, subtraction, and multiplication are same, the total searching cost becomes about one third. Therefore, in the case of this experiment, the time to generate each stereo pair of images is expected to reduce to about two minutes with ignoring the overhead.

Even if the computational cost is reduced by the method mentioned above, the computational cost of the proposed method is still higher than the comparative methods for which it takes less than 1 second. However, considering the applications such as movie generation, we think the proposed method is useful for the practical applications, e.g. like 4K stereo movie generation.

5. Conclusion

We have proposed a novel system for generating a zoomed stereo video from two synchronized videos with different magnifications. In the proposed method, the non-zoomed video image is cut out and super-resolved by energy minimization for generating stereo video without special hardware. In order to achieve this, (1) the zoom magnification parameter is automatically determined by matching distributions of intensities, and (2) the cutout image is super-resolved by using optically zoomed images as exemplars. In experiments using stereo datasets and real videos, we have demonstrated the effectiveness of the proposed method by comparing our results with baseline methods. In the future, we will focus on improving the quality of the generated image by considering occlusions. In addition, we should reduce the computational cost using efficient searching algorithms, e.g. the method in [22] employed in the stereo matching field.

Acknowledgments

This research was partially supported by Grant-in-Aid for Scientific Research (A), No. 23240024 and Challenging Exploratory Research, No. 25540086.

References

[1] Y. Hayashi, N. Kawai, T. Sato, M. Okumoto, and N. Yokoya, "Generation of a super-resolved stereo video using two synchronized videos with different magnifications," Proc. 6th Pacific-Rim Sympo. Image and Video Technology, vol.8333, pp.194–205, 2014.
[2] B. Mendiburu, 3D movie making - stereoscopic digital cinema from script to screen, Focal Press, 2009.
[3] S.C. Park, M.K. Park, and M.G. Kang, "Super-resolution image reconstruction: A technical overview," IEEE Signal Processing Magazine, vol.20, no.3, pp.21–36, May 2003.
[4] A. Iketani, T. Sato, S. Ikeda, M. Kanbara, A. Nakajima, and N. Yokoya, "Super-resolved video mosaicing for documents by extrinsic camera parameter estimation," Proc. Int. Conf. Computer Vision and Graphics, pp.327–336, Sept. 2004.
[5] M. Irani and S. Peleg, "Improving resolution by image registration," Graphical Models and Image Processing, vol.53, no.3, pp.231–239, May 1991.
[6] S. Farsiu, M.D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," IEEE Trans. Image Processing, vol.13, no.10, pp.1327–1344, Oct. 2004.
[7] Z. Lin and H.-Y. Shum, "Fundamental limits of reconstruction-based super resolution algorithms under local translation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.26, no.1, pp.83–97, Jan. 2004.
[8] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.24, no.9, pp.1167–1182, Sept. 2002.
[9] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," IEEE Computer Graphics and Applications, vol.22, pp.56–65, April 2002.
[10] J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum, "Image hallucination with primal sketch priors," IEEE Computer Vision and Pattern Recognition, pp.729–736, June 2003.
[11] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," ACM Trans. Graphics, vol.30, no.2, pp.12:1–12:11, April 2011.
[12] I. Begin and F.P. Ferrie, "Blind super-resolution using a learning-based approach," Proc. Int. Conf. Pattern Recognition, vol.2, pp.85–89, Aug. 2004.
[13] Y.-W. Tai, S. Lin, M.S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," Proc. Int. Conf. Pattern Recognition, pp.2400–2407, June 2010.
[14] S. Baker and T. Kanade, "Hallucinating faces," Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, pp.83–88, 2000.
[15] Z. Xiong, X. Sun, and F. Wu, "Image hallucination with feature enhancement," Proc. Int. Conf. on Pattern Recognition, pp.2074–2081, June 2009.
[16] A. Hashimoto, T. Nakaya, N. Kuroki, T. Hirose, and M. Numa, "Binary tree dictionary for learning-based super-resolution," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J96-D, no.2, pp.357–361, Feb. 2013.
[17] X. Li and M.T. Orchard, "New edge-directed interpolation," IEEE Trans. Image Processing, vol.10, no.10, pp.1521–1527, Oct. 2001.
[18] R. Fattal, "Image upsampling via imposed edge statistics," ACM Trans. Graphics, vol.26, no.3, pp.95:1–95:8, July 2007.
[19] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," IEEE Int. Conf. Computer Vision, pp.349–356, Oct. 2009.

[20] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," Int. Journal of Computer Vision, pp.131–140, 2001.
[21] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.1–8, June 2007.
[22] R. Szeliski and D. Scharstein, "Sampling the disparity space image," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.26, no.3, pp.419–425, March 2004.

Yusuke Hayashi received his B.E. degree from Tokuyama College of Technology in 2013. He received his M.E. degree in information science from the Nara Institute of Science and Technology in 2015 He has been working for AVC Multimedia Software Co., Ltd. from 2015.

Norihiko Kawai received his B.E. degree in informatics and mathematical science from Kyoto University in 2005. He received his M.E. and Ph.D. degrees in information science from the Nara Institute of Science and Technology in 2007 and 2010, respectively. He was a research fellow of the Japan Society for the Promotion of Science and a postdoctoral fellow at the University of California at Berkeley in 2010-2011. He has been an assistant professor at the Nara Institute of Science and Technology since 2011. He is a member of IEEE, IEICE, IPSJ and VRSJ.

Tomokazu Sato received his B.E. degree in computer and system science from the Osaka Prefecture University in 1999. He received his M.E. and Ph.D. degrees in information science from the Nara Institute of Science and Technology in 2001 and 2003, respectively. He was an assistant professor at the Nara Institute of Science and Technology in 2003-2011. He was a visiting researcher at Czech Technical University in Prague in 2010-2011. He has been an associate professor at the Nara Institute of Science and Technology since 2011. He is a member of IEEE, IEICE, IPSJ, VRSJ and ITE.

Miyuki Okumoto received her B.E. degree in computer science from Kyushu Institute of Technology in 1981, and Ph.D. degree in computer science and systems engineering from Yamaguchi University in 2003. Since 1981, she has been with Tokuyama College of Technology. She is presently a professor in the Department of Computer Science and Electronics Engineering. Her research interests are pattern recognition and image processing.

Naokazu Yokoya received his B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University in 1974, 1976, and 1979, respectively. He joined Electrotechnical Laboratory (ETL) of the Ministry of International Trade and Industry in 1979. He was a visiting professor at McGill University in Montreal in 1986-87 and has been a professor at the Nara Institute of Science and Technology since 1992. He has also been a vice president at Nara Institute of Science and Technology since April 2013. He is a fellow of IPSJ, IEICE and VRSJ and a member of IEEE, ACM SIGGRAPH, JSAI, JCSS and ITE.