IMAGE RESOLUTION ENHANCEMENT BASED ON NOVEL VIEW SYNTHESIS

Yusuke Hayashi, Norihiko Kawai, Tomokazu Sato and Naokazu Yokoya

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayamacho, Ikoma, Nara, Japan

ABSTRACT

This paper proposes an example-based method to increase the resolution of a low-resolution image. In the proposed method, we generate example images by a novel view synthesis technique using 3D geometry reconstruction and camera pose estimation from a video or images capturing the same scene. We then increase the resolution by minimizing an energy function by searching for the optimal example from the generated example images. The proposed method has less limitations on camera positions and geometry of the target scene than those in conventional methods. Experiments demonstrate the effectiveness of the proposed method by qualitatively comparing the results of the proposed and conventional methods.

***Index Terms*—** Super-resolution, 3D reconstruction, energy minimization, novel view synthesis

1. INTRODUCTION

With the increase of the resolution of display devices, there exist emerging demands for techniques of high-resolution video/image generation from low-resolution one. To meet the demands, super-resolution techniques have been widely investigated. Super-resolution methods can be broadly classified into three categories: filter-based, reconstruction-based and example-based methods. The filter based methods estimate a high-resolution image from a single low-resolution image by compensating high-frequency components of the input image [1, 2]. In reconstruction-based methods, input images are super-resolved by aligning multiple low-resolution images with sub-pixel accuracy [3, 4, 5]. On the other hand, in example-based methods, a target low-resolution image is super-resolved using example high-resolution images [6, 7, 8, 9]. It is necessary for the example-based methods to use suitable example images because the effect of resolution enhancement depends on examples [7].

This study considers the situation where we take a video while walking, and the captured video often includes the scenes in which some objects are captured from far and near. In such a video, a pair of low-resolution and high-resolution textures of the same objects can easily be found. In a different scenario, some photos with high-resolution for the same scene may be found on the Internet. In these situations, where appropriate examples are available, we think example-based methods are more effective than other methods. Therefore, this study focuses on the example-based methods.

As one of the state-of-the-art methods that try to find suitable examples, an example-based super-resolution method has been proposed that retrieves web images capturing the same scene with a target image [10]. However, in the method, the example images are aligned to the target low-resolution image using a single homography matrix. Therefore, it is difficult to obtain good results when the capturing positions of example images are different from that of the target image, and the target scene is not planar.

This paper proposes a method to increase the resolution of a low-resolution image using example images generated by a novel view synthesis technique using 3D geometry reconstruction and camera pose estimation from a video or web images capturing the same scene. Specifically, after reconstructing 3D geometry of a target scene and estimating camera poses from input target and reference images, we first warp the input images to the viewpoint of the target image using the estimated 3D geometry and camera poses to compensate for the difference in appearance and make rough correspondences between images. The warped images used as examples are then narrowed down using the camera poses for effectively increasing the resolution. Finally, we increase the resolution of the low-resolution image by minimizing an energy function defined based on the spatial frequency and texture similarity between the target and selected example images, which makes accurate correspondences. The novel view synthesis and energy minimization enable to have less limitations on camera poses and geometry of the target scene than those in the conventional method.

2. PROPOSED METHOD

In our method, the resolution of target image I_t is increased using K reference images that capture the same scene as the target one. Figure 1 illustrates the flow diagram of the proposed method. First, 3D geometry of a target scene is reconstructed using input images including target and reference ones (1). Next, the reference images are warped to the view-

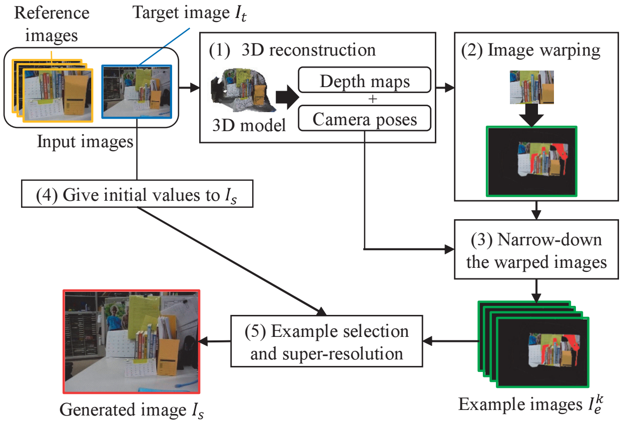


Fig. 1. Flow diagram of the proposed method

point of the target image using the reconstructed 3D geometry, and the warped images are used as example candidates $\{I_e^k | k = 1, \cdots, K\}$ for increasing the resolution (2). The candidate images are then narrowed down using camera poses (3). After that, target image I_t is enlarged by bi-cubic interpolation to the size of the target resolution for giving initial values to generated image I_s (4). The final result of I_s is obtained by energy minimization using the example images (5).

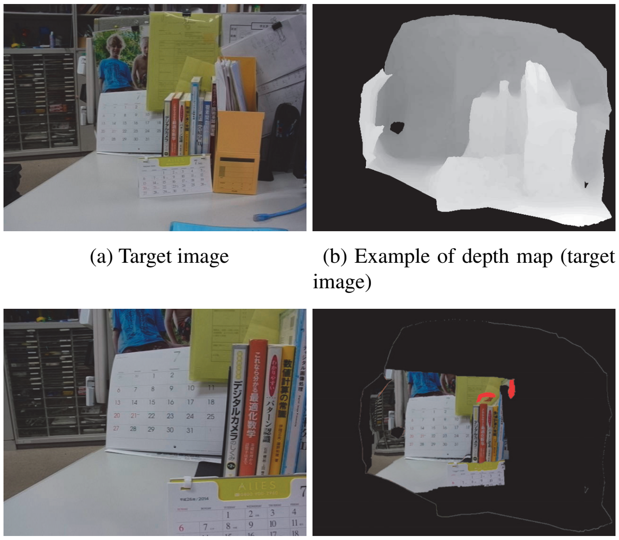
2.1. 3D reconstruction and image warping

In the proposed method, we estimate camera poses of input images including both target and reference ones and reconstruct 3D geometry of the scene by applying Structure from Motion (SfM) [11, 12] and Multi View Stereo (MVS) [13] to input images. We then generate a depth map with a target resolution for each input image from the reconstructed 3D geometry as shown in Figs. 2(a) and (b). Next, the reference images (Fig. 2(c)) are warped to the viewpoint of the target image by projecting pixel values of the reference images using the depth map of the target image and the estimated camera poses. In warping images, we check the consistency of depths between the target and reference images, and the regions of inconsistency caused by occlusions or estimation errors of 3D model are set as unusable ones as shown in the red regions in Fig. 2(d).

2.2. Narrowing down of reference images

Generally, the warped image that is generated from the input image captured near objects includes many higher frequency components than that captured far from them as shown in Fig. 3. Based on this fact, we narrow down the warped reference images so that we can use example images with high-frequency components in the energy minimization process.

Specifically, we determine the pixels in the warped images corresponding to a pixel in I_s. We then select top T warped images of the smallest depth values of the corresponding pixels. By this way, we independently select T example images for each pixel in I_s.

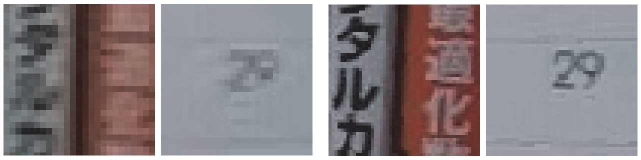

(a) Target image

(b) Example of depth map (target image)

(c) Example of reference

(d) Warped image

Fig. 2. Intermediate images generated by the proposed method.


(a) Long distance

(b) Short distance

Fig. 3. Comparison of high-frequency textures in warped images according to the distances from the cameras to the object.

2.3. Resolution enhancement by energy minimization

The resolution-enhanced image I_s of target image I_t is generated by minimizing an energy function using the example images and the original target image. It should be noted that we do not have pixel values in the warped images for the pixels of which depth values do not exist in the depth map of the target image. For these pixels, we just leave initial values generated by bi-cubic interpolation. In this process, the input image is transformed from RGB to YCbCr. After that, the resolution enhancement process is applied to Y (intensity) channel, and Cb and Cr (chromatic) 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, we first give the definition of the energy function and then describe the minimization process.

2.3.1. *Definition of energy function*

Energy function E is defined using two different kinds of energy terms as follows:

$$E = \sum_{\mathbf{x}_i \in I_s} \{E_{sr}(\mathbf{x}_i, \mathbf{x}_j, k) + \beta E_{data}(\mathbf{x}_i)\}, \tag{1}$$

where E_{sr} represents the pattern dissimilarity between generated image I_s and example image I_e^k, and this term gives the effect of increasing the resolution of generated image I_s using the texture including high-frequency components in example image I_e^k. E_{data} represents the intensity difference between generated image I_s and original target image I_t, and the term gives the effect of preserving the structure of target image I_t onto generated image I_s. β is a weight for balancing the two terms. E_{sr} and E_{data} are defined as follows, respectively:

$$E_{sr}(\mathbf{x}_i, \mathbf{x}_j, k) = \omega_{(\mathbf{x}_j, k)} \sum_{\mathbf{p} \in W} \{I_s(\mathbf{x}_i + \mathbf{p}) - I_e^k(\mathbf{x}_j + \mathbf{p})\}^2, \tag{2}$$

$$E_{data}(\mathbf{x}_i) = \{I_s(\mathbf{x}_i) - I_t(\mathbf{D}\mathbf{x}_i)\}^2. \tag{3}$$

Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ denote pixels in I_s and I_e^k, respectively. $I_s(\mathbf{x}_i)$, $I_e^k(\mathbf{x}_j)$ and $I_t(\mathbf{x}_i)$ represent the intensities of pixels in images I_s, I_e^k and I_t, respectively. $\mathbf{p}$ is a shift vector to indicate a pixel in a square window W. $\mathbf{D}$ transfers a pixel position $\mathbf{x}_i$ in I_s to the corresponding pixel in I_t. $\omega_{(\mathbf{x}_j, k)}$ is a reciprocal of the total of the power spectrum values that is larger than a threshold obtained by Fourier transforming the window region centered at $\mathbf{x}_j$ in example image I_e^k. This term enables to select an example image with high-frequency components from the selected T images, which often include motion blurs and defocuses even after the example images are narrowed down using the camera poses.

2.3.2. *Iterative energy minimization*

Energy function E is minimized by iterating the following two processes: (i) search for similar textures in the selected example images and (ii) update pixel values in I_s.

In the process (i), we determine two parameters, pixel position $\mathbf{x}_j$ and example image index k, by searching T example images for the position $\mathbf{x}_j$ around which the pattern is most similar to that around $\mathbf{x}_i$ so that we minimize Eq. (2). The searching region is a certain range of $L \times L$ pixels around the coordinate $\mathbf{x}_i$ in T example images because the example images are roughly aligned by image warping. The searching compensates for the misalignment caused by geometric errors. The selected example image index and pixel corresponding to $\mathbf{x}_i$ are represented by $n(\mathbf{x}_i)$ and $f(\mathbf{x}_i)$.

In the process (ii), pixel values $I_s(\mathbf{x}_i)$ in the generated image are updated in parallel so as to minimize the energy function E while keeping all the similar texture pairs fixed. Energy function E is resolved into element energy $E(\mathbf{x}_i)$ for each pixel $\mathbf{x}_i$ in I_s:

$$\begin{aligned} E(\mathbf{x}_i) \;=\; & \sum_{\mathbf{p} \in W} \omega_{(f(\mathbf{t}), n(\mathbf{t}))} \{I_s(\mathbf{x}_i) - I_e^{n(\mathbf{t})}(f(\mathbf{t}) - \mathbf{p})\}^2 \\ & + \beta\{I_s(\mathbf{x}_i) - I_t(\mathbf{D}\mathbf{x}_i)\}^2, \end{aligned} \tag{4}$$

$$\mathbf{t} = \mathbf{x}_i + \mathbf{p}. \tag{5}$$

E can be minimized by minimizing each element energy $E(\mathbf{x}_i)$ because energy E consists of the sum of all element energies. $I_s(\mathbf{x}_i)$ that minimizes $E(\mathbf{x}_i)$ can be calculated by differentiating $E(\mathbf{x}_i)$ with respect to $I_s(\mathbf{x}_i)$, and is

$$I_s(\mathbf{x}_i) = \frac{\sum_{\mathbf{p} \in W} \omega_{(f(\mathbf{t}), n(\mathbf{t}))} I_e^{n(\mathbf{t})}(f(\mathbf{t}) - \mathbf{p}) + \beta I_t(\mathbf{D}\mathbf{x}_i)}{\sum_{\mathbf{p} \in W} \omega_{(f(\mathbf{t}), n(\mathbf{t}))} + \beta}. \tag{6}$$

3. EXPERIMENTS AND RESULTS

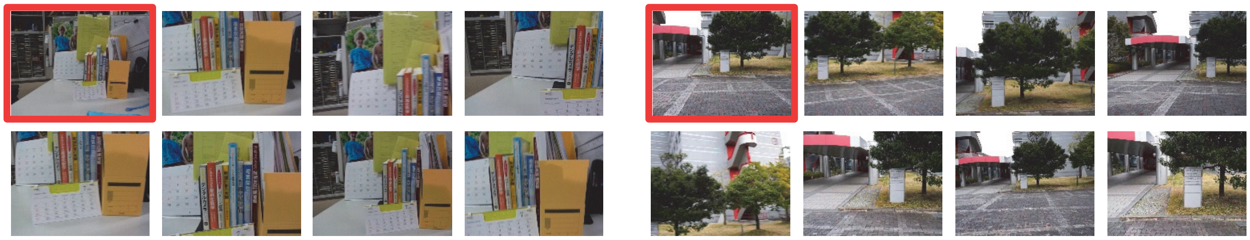
In this section, we evaluate the performance of the proposed method by subjectively comparing the results of the proposed method with those by the conventional methods. In the experiment, we captured two videos consisting of 60 frames with 640×480 pixels while moving a camera in indoor and outdoor environments as shown in Fig. 4. We selected one image from the input images as the target image, and the target resolution is set to 1280×960 (magnification factor is 2). We experimentally determined the parameters of the proposed method as: $W = 25 \times 25$, $\beta = 0.0005$, $L = 10$, and $T = 5$.

Figure 5 shows the results by bi-cubic interpolation (a), example-based method [9] (b) and the proposed method (c). From these results, we can confirm that the high frequency components are successfully generated by the proposed method, and the results of the proposed method are much clearer than those of both bi-cubic interpolation and conventional example-based method [9]. However, in our method, we cannot enhance the resolution of the regions that have no depth values because of the limited area of reconstructed geometry. Therefore, the unnatural change in the resolution appears on the boundary of the enhanced and the other regions as shown in Fig. 6.

4. CONCLUSION

In this paper, we have proposed a method to increase the resolution of a low-resolution image using example images generated by a novel view synthesis technique using 3D geometry. Our contribution is to have less limitations on camera positions and geometry of the target scene than those in the conventional methods. Our experimental result has demonstrated that our proposed method successfully generate high-resolution images. In future work, we should attempt to enhance the resolution of the regions that have no depth values.

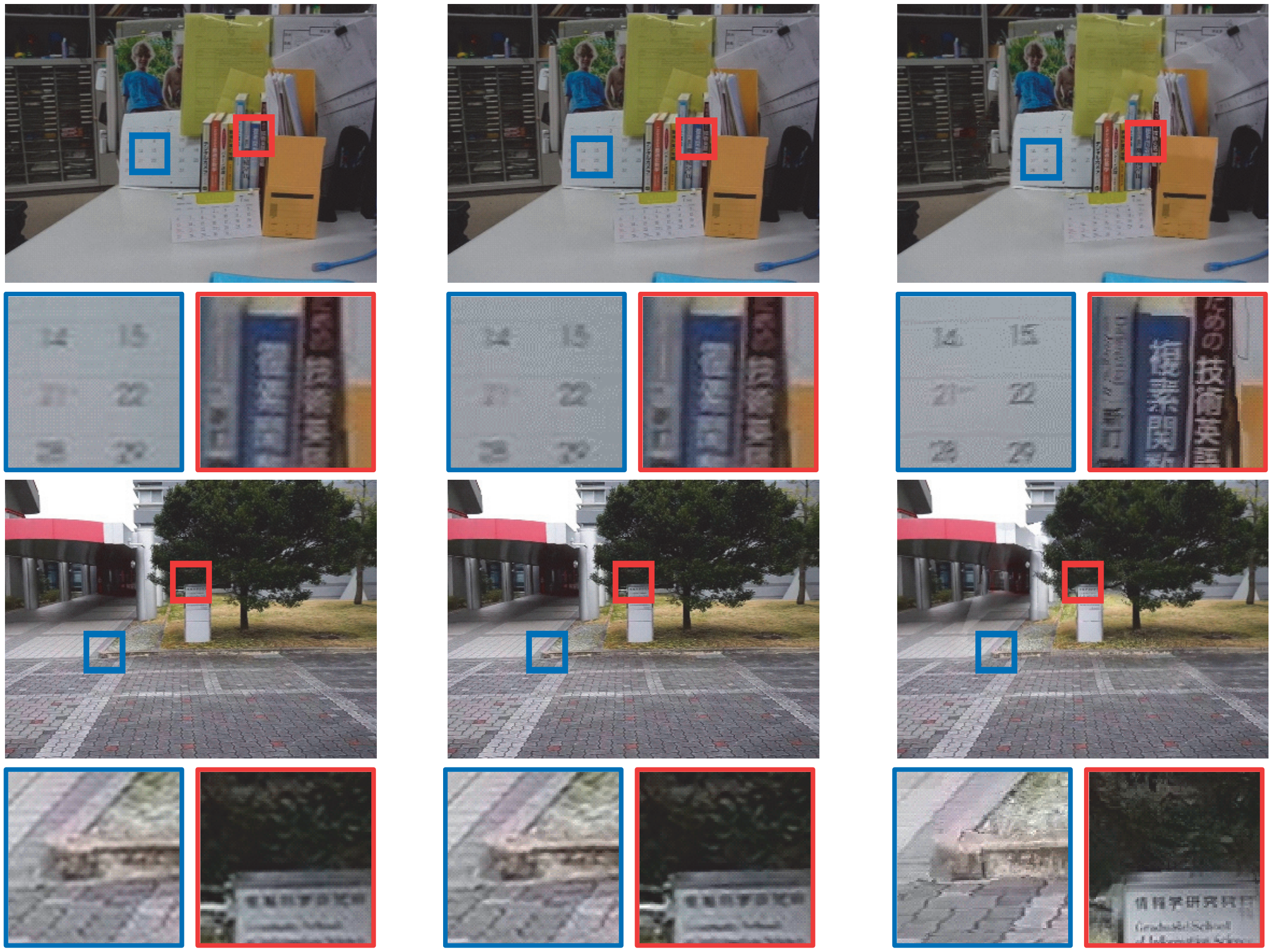
Acknowledgment This research was partially supported by JSPS KAKENHI Nos. 23240024 and 25540086.



(a) Scene 1

(b) Scene 2

Fig. 4. Examples of input images including a target image (upper left) and reference images.



(a) Bi-cubic interpolation

(b) Example based method [9]

(c) Proposed method

Fig. 5. Experimental results

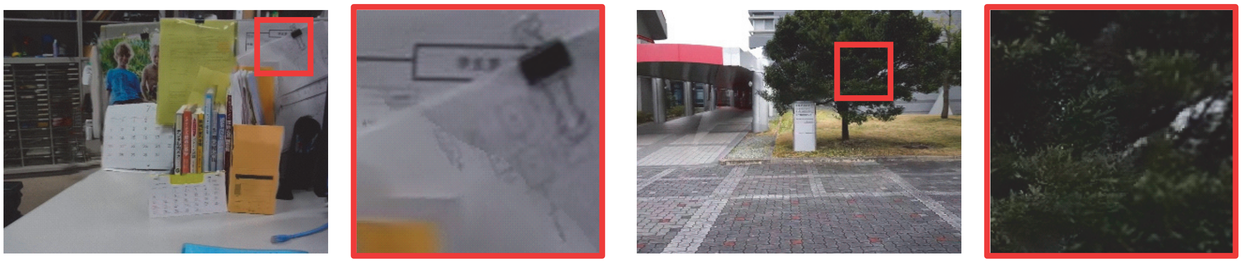


Fig. 6. Example of unnatural change in resolution caused by the missing of depth values in the target image

5. REFERENCES

[1] X. Li and M.T. Orchard, "New edge-directed interpolation," *IEEE Trans. on Image Processing*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.

[2] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Trans. on Graphics*, vol. 26, no. 3, pp. 95:1–95:8, July 2007.

[3] M. Irani and S. Peleg, "Improving resolution by image registration," *Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, Apr. 1991.

[4] 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.

[5] S. Farsiu, M.D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[6] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-based super-resolution," *IEEE Trans. on Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, Mar. 2002.

[7] S. Baker and T. Kanade, "Hallucinating faces," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Mar. 2000, pp. 83–88.

[8] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, Sep. 2002.

[9] A. Hashimoto, T. Nakaya, N. Kuroki, T. Hirose, and M. Numa, "Binary tree dictionary for learning-based super-resolution," *IEICE Trans. on Information and Systems (Japanese Edition)*, vol. J96-D, no. 2, pp. 357–361, Feb. 2013.

[10] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving web images," *IEEE Trans. on Image Processing*, vol. 22, no. 12, pp. 4865–4878, Dec. 2013.

[11] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. IEEE Int. Conf. on 3D Vision*, June 2013, pp. 127–134.

[12] C. Wu, S. Agarwal, B. Curless, and S.M. Seitz, "Multicore bundle adjustment," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, June 2011, pp. 3057–3064.

[13] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, June 2011, pp. 3121–3128.