SELF-SIMILARITY MODELING FOR INTERPOLATION AND EXTRAPOLATION OF MULTI-VIEWPOINT IMAGE SETS

Takeshi Naemura and Hiroshi Harashima

Department of Electrical Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
E-mail : naemura@harashima.t.u-tokyo.ac.jp

ABSTRACT

For the 3-D image communication, a flexible interpolation technique of a Multi-Viewpoint Image Set (MVIS) is required. If we could recover the 3-D structure from an MVIS, interpolative images would be obtained. This approach, however, can not be easily applied to arbitrary objects because of the difficulty in stereo-matching. In this paper, the self-similarity modeling which has been studied in the field of data compression is chosen as a modeling scheme which is scalable to any resolution, and the interpolation is virtually achieved by enhancing the resolution of the viewpoint-axis of an MVIS. This new approach is free from stereo-matching and is shown to achieve interpolative images with an SNR of about 35 dB.

1. INTRODUCTION

For the 3-D image communication, the possibilities and the problems of a Multi-Viewpoint Image Set (MVIS) as a 3-D image format are discussed. Especially, an interpolation technique is focused and examined. Because of the difficulty in stereo-matching, structure recovery method seems not so robust for 3-D image communication system. So, the new approach for interpolation of an MVIS should be free from stereo-matching.

The self-similarity of images has been applied to data compression methods. In these methods, the original image is approximated by a fractal image which is scalable to any resolution. On the other hand, interpolation can be achieved by enhancing the resolution of the viewpoint-axis, that is, virtually increasing the number of viewpoints of an MVIS. Toward this end, the scalability of the self-similarity modeling is exploited to the viewpoint-axis of an MVIS for the purpose of interpolation, in this paper.

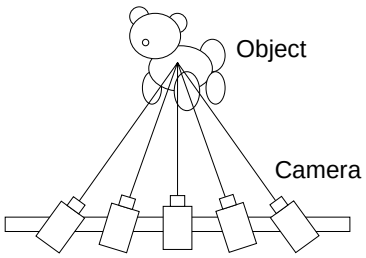


Figure 1. An example of camera configuration

2. MULTI-VIEWPOINT IMAGE SETS

An MVIS is a set of 2-D images which results when several cameras look at the same object from slightly different viewpoints. Fig.1 shows an example of such an input system that gives five viewpoints image set. Other configurations are also possible; for example, the light axes of cameras may be set parallel, or the cameras may be placed around the object.

As a 3-D image format, an MVIS has two merits. One is the compatibility with the 2-D TV format. This means that an input system of an MVIS which can be built with 2-D cameras is more easily applicable to any object than other input systems like holographic system which use lasers. Another is the generality for any 3-D TVs, that is, most kinds of 3-D TVs including holographic methods can display an MVIS. But it has also a demerit. It is just a set of 2-D images, so we have to solve the problem of stereo-matching, when we need to recover the structure.

2.1. Interpolation techniques of an MVIS

If the number of viewpoints is large enough, we can produce some autostereoscopic visual effect. But as the number increases, the amount of data expands and the input system becomes more complex. Considering the spatial correlation between 2-D images of an MVIS, it seems not to be necessary for the input system to have so many cameras. So, an interpolation technique which can synthesize many-viewpoint images from fewer-viewpoint images is needed.

Most of previous works on an MVIS have concentrated on data compression methods which employ the concept of disparity compensation [1]. This method may be useful for data compression when the number of viewpoints is fixed. The number, however, should be variable for the flexible 3-D communication system. On the other hand, structure recovery method for data compression has been also examined [2]. Though this method can synthesize interpolative images, it requires the accuracy of camera arrangement and ignores the lightening conditions, because of the difficulty in stereo-matching.

The new approach presented in this paper for interpolation of an MVIS should be free from stereo-matching, and applicable to several camera configurations and lightening conditions.

2.2. Geometrical structure of an MVIS

An MVIS has geometrical structure which reflects the real world as illustrated in Fig.2. The MVIS on the left side

of Fig.2 has N viewpoints and is sorted by viewpoint locations. This MVIS has 3 axes (horizontal, vertical, viewpoint). When the vertical coordinate is fixed, we can generate a 2-D image called epipolar-image (EPI) whose axes are horizontal and viewpoint only. The right side rectangular of Fig.2 is one of them. In an EPI, crossings of areas mean occlusions, and gradations in shades have some information of lightening conditions in the real world. By considering the camera configuration, this structure of EPIs can be used for structure recovery [2].

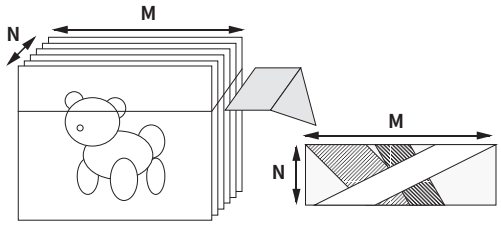


Figure 2. Geometrical structure of an MVIS

This structure of an EPI almost appears to be available for interpolation of an MVIS, but its analysis has the difficulty in stereo-matching. In this paper, the self-similarity of this structure is exploited to interpolate an MVIS without analyzing the structure, that is, without recognizing the real world.

3. SELF-SIMILARITY MODELING

Application of the fractal theory to iterated transformations has been studied and examined for 2-D image coding [3]. Fractal coding schemes exploit the self-similarity within the original image to obtain a fractal approximation. The fractal approximation as a reconstructed image has a unique feature : the scalability to any resolution. To take advantage of this feature, the self-similarity modeling is applied to EPIs of an MVIS for the purpose of interpolation.

3.1. Modeling Process

The basic idea of modeling process is as follows :

1. The original image f is divided into range blocks $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N$ so that $f = \mathbf{R}_1 \cup \mathbf{R}_2 \ldots \cup \mathbf{R}_N$ and $\mathbf{R}_i \cap \mathbf{R}_j = 0$ when i is not equal to j. The range blocks cover the whole image and do not overlap.
2. The image is also divided into larger blocks called domain blocks $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_M$, which may overlap.
3. For each range block $\mathbf{R}_i$, a matching domain block $\mathbf{D}_j$ is searched among all the domain blocks so that the $\mathbf{D}_j$ transformed with a contractive transformation T_i is similar to $\mathbf{R}_i$, that is $\mathbf{R}_i \approx T_i\mathbf{D}_j$.

The range and domain blocks are parallelograms of several inclinations in EPIs as described in **3.2**.

Transformation T_i is a composition of a spatial contraction φ, a contrast scaling α_i, and an isometry ι_i, of the form:

$$T_i\mathbf{D}_j = \alpha_i \ \{\iota_i \ \varphi \ (\mathbf{D}_j - \mu_{Dj})\} + \mu_{Ri}.$$

where μ_{Dj} and μ_{Ri} are the average values of brightness level of $\mathbf{D}_j$ and $\mathbf{R}_i$, respectively.

In this paper, the root-mean-square (RMS) distortion is used to determine the closeness between a range block and a domain block. The searching of a matching domain block consists in finding such a transformation T_i and domain block $\mathbf{D}_j$ that minimizes the error function :

$$e^2 = \frac{1}{XYZ} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} \sum_{z=0}^{Z-1} \{\mathbf{r}_i(x,y,z) - \alpha_i \mathbf{d}_j(x,y,z)\}^2 \quad (1)$$

where

$$\mathbf{r}_i = \mathbf{R}_i - \mu_{Ri}, \ \mathbf{d}_j = \iota_i \ \varphi \ (\mathbf{D}_j - \mu_{Dj}).$$

Spatial contraction φ The domain block $\mathbf{D}_j$ is spatially contracted to the size of range blocks and the inclination of $\mathbf{D}_j$ is converted to that of $\mathbf{R}_i$. In this paper, when the size of a range block is $rx \times ry \times rz$ ($horizontal \times vertical \times viewpoint$), the size of a domain block is $2rx \times 2ry \times rz$ as described in **3.3**. So, the pixel values of the contracted domain block are the average values of four neighboring pixels in the domain block.

Contrast scaling α_i The contrast scaling α_i is used to let the contrast of $\mathbf{D}_j$ reflect that of $\mathbf{R}_i$. α_i must be contractive, that is, $|\alpha_i| \leq 1$.

To find the minimum of the error function(1), the partial derivatives of e^2 must be set to zero. Thus the optimized α_i is obtained as follows:

$$\alpha_{opt} = \frac{C_{rd} \ (0,0)}{\sigma_d^2}$$

where C_{rd} denotes the covariance of $\mathbf{r}_i$ and $\mathbf{d}_j$, σ_d^2 the variance of $\mathbf{d}_j$ [4].

Isometry transformation ι_i The rotation and flip operations those are called isometries shuffle pixels within a block, in a deterministic way. When the contrast scaling α_i is optimized, the error function (1) becomes:

$$e^2 = \sigma_r^2 - \alpha_{opt}^2 \ \sigma_d^2 = \sigma_r^2 - \{\frac{C_{rd} \ (0,0)}{\sigma_d}\}^2.$$

The variances σ_r and σ_d are independent from the isometry ι_i. Thus among all the isometries, one that maximizes $C_{rd} \ (0,0)$ is selected as ι_i to minimize the error funciton.

In this paper, just two isometries are chosen in order to save calculation time by considering the structure of an MVIS. They are :

1. Identity, and
2. Orthogonal reflection about mid-horizontal axis and mid-viewpoint axis.

3.2. Partitioning technique

If rz is equal to $Z/2$, that is, an EPI is divided at the middle point of the viewpoint-axis, the extrapolative images will be synthesized near the center of the viewing zone by enhancing the resolution of the viewpoint-axis. In this case, the block distortion in an EPI will degrade the extrapolative

images. So, in this paper, rz is equal to Z, that is, an EPI is not divided along the viewpoint-axis.

If data compression is desired, range blocks should not overlap. For the purpose of interpolation, both range and domain blocks are allowed to overlap to take advantage of the self-similarity. Fig.3 shows how an EPI is divided.

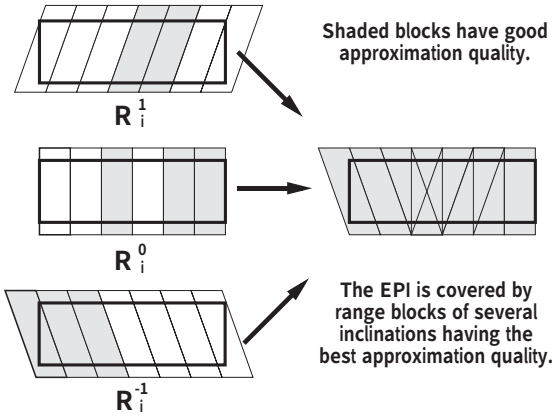


Figure 3. Partitioning of an EPI

An EPI is divided into parallelogram range blocks $\mathbf{R}_i^k$ whose inclination is k and is also divided into parallelogram domain blocks $\mathbf{D}_j^l$ whose inclination is l. For each inclination k, range blocks $\mathbf{R}_i^k (i = 1, \ldots, N)$ cover the whole image and do not overlap. But the self-similarity of an EPI seems not to be exploited efficiently when the inclination k is fixed. So, in this paper, range blocks of several different inclinations are selected so that they cover the whole image even if they overlap. The selection process is as follows:

1. For each dividing inclination k, the modeling process explained in **3.1** is executed for the original MVIS using domain blocks of several inclinations.
2. For each pixel in an EPI, there are several range blocks which include that pixel and have different inclination k. Among such range blocks, the best one, that is, such a range block that has the best approximation quality is selected.

After this selection, the EPI is covered by range blocks of several inclination having the best approximation quality.

3.3. Synthesizing Process

Starting from an arbitrary MVIS of the same size as the original MVIS ($X \times Y \times Z$), each range block $\mathbf{R}_i^k (rx \times ry \times Z)$ is computed from the corresponding matching domain block $\mathbf{D}_j^l (2rx \times 2ry \times Z)$. Computing all the range blocks once is called an iteration. After several iterations, the reconstructed MVIS will be very close to the original MVIS.

Assuming that the similarity between a range block and the corresponding matching domain block is saved in any resolution, interpolation is virtually achieved. For instance, $3Z$-viewpoint image set will be synthesized when the size of the starting MVIS, $\mathbf{R}_i^k$ and $\mathbf{D}_j^l$ are set to $X \times Y \times 3Z$, $rx \times ry \times 3Z$ and $2rx \times 2ry \times 3Z$, and the inclination k and l are set to $k/3$ and $l/3$, respectively. Fig.4 shows an example of modeling and synthesizing processes.

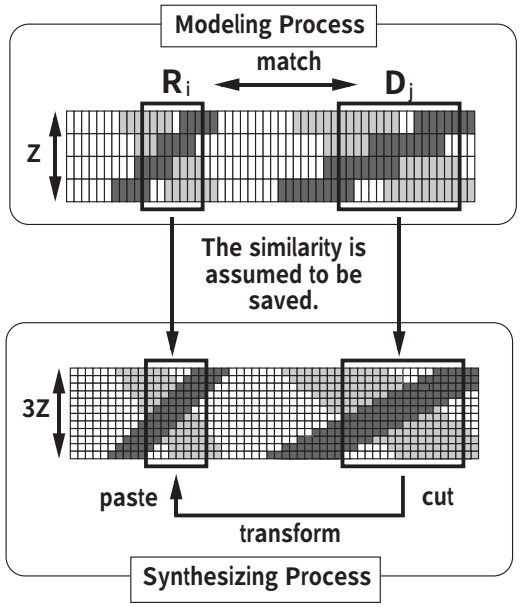


Figure 4. Modeling and Synthesizing Processes

4. EXPERIMENTAL RESULTS

The original MVIS used in an experiment here is 8 bits/pel and $176 \times 144 \times 18 viewpoints$. The experimental process is as follows:

1. From just six images (frame No. 2, 5, 8, 11, 14, 17) of the original images, the self-similarity of the original MVIS is estimated. The size of a range block is $8 \times 1 \times 6$. But such a range block that the approximation quality is too bad is divided into two range blocks ($4 \times 1 \times 6$).
2. Using the result of step 1, an MVIS having 18 viewpoints is synthesized.
3. By comparing the synthesized image with the original images, the ability of this method is evaluated.

By exploiting the result of step 1, we can synthesize an MVIS having arbitrary number of viewpoints, because EPIs are not divided along the viewpoint-axis. The number, however, must be 18 in order to compare with the original MVIS.

In this case, two images (frame No. 1, 18) are extrapolated and ten images (frame No. 3, 4, 6, 7, 9, 10, 12, 13, 15, 16) are interpolated. Fig.5 and 6 show an interpolative image (frame No.10) and an extrapolative image (frame No.18), respectively. The edge busyness is perceived, especially in the extrapolative image. To avoid this demerit, the range block should be extended along the vertical-axis, that is, it should not be a plain block like $8 \times 1 \times 6$ but a hexahedron block like $4 \times 2 \times 6$.

Fig.7 shows evaluation of the ability of our method by comparing reconstructed images with the original images. Interpolative images have an SNR of about 35 dB, and extrapolative images have that of about 30 dB.

Fig.8 shows an EPI which contains the left eye and nose of the dog. The 18-viewpoint image set (c) is synthesized by

Figure 5. the interpolative image (frame No. 10)

Figure 6. the extrapolative image (frame No. 18)

exploiting the self-similarity of subsampled 6-viewpoint image set (b), and is compared with the original 18-viewpoint image set (a).

5. CONCLUSIONS

For the purpose of interpolation of an MVIS, the scalability of the self-similarity modeling is exploited to the viewpoint-axis. This new method is shown to achieve interpolative images with an SNR of about 35 dB.

For the 3-D image communication, we do not require such images that are viewed from outside of the viewing zone. Within the viewing zone, this method, though it does not recover the 3-D structure, is applicable for several camera configurations and lightening conditions because it just exploits the self-similarity of an MVIS.

REFERENCES

[1] M.E.Lukacs : "Predictive Coding of Multi-Viewpoint Image Sets", *ICASSP '86*, pp. 521-524, 1986.

[2] T.Fujii, *et al.* : "Data Compression for an Autostereoscopic 3-D image", *PCS '93*, 13.21, 1993.

[3] A.E.Jacquin : "Image Coding Based on a Fractal Theory of Iterated Contractive Image Transformations", *IEEE Trans. Image Process.*, **1**, 1, pp. 18-30, 1992.

[4] H.Kambara, *et al.* : "Fast Coding Algorithm for Iterated Transformation Theory-Based Coding", *IEICE Spring Conference '93*, SA-3-2, 1993.

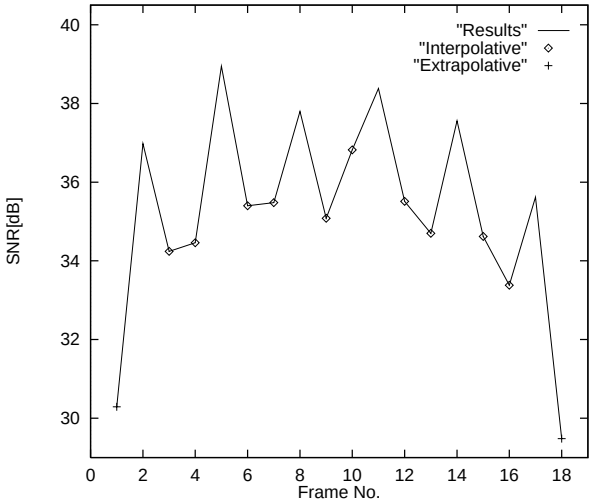
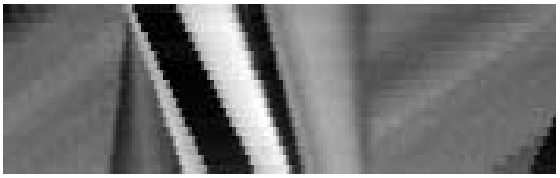
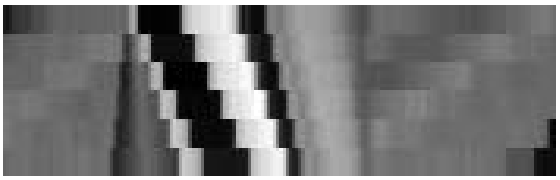


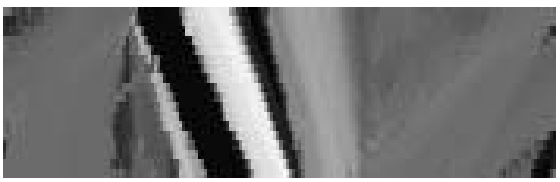
Figure 7. SNR vs. Frame



(a) an EPI of original 18-viewpoint image set



(b) an EPI of subsampled 6-viewpoint image set



(c) an EPI of synthesized 18-viewpoint image set

Figure 8. Epipolar images