LETTER

Super-Resolution Reconstruction for Spatio-Temporal Resolution Enhancement of Video Sequences

Miki HASEYAMA[†a)], *Member*, Daisuke IZUMI[†], *and* Makoto TAKIZAWA[†], *Student Members*

SUMMARY A method for spatio-temporal resolution enhancement of video sequences based on super-resolution reconstruction is proposed. A new observation model is defined for accurate resolution enhancement, which enables subpixel motion in intermediate frames to be obtained. A modified optimization formula for obtaining a high-resolution sequence is also adopted.

key words: *video processing, super-resolution, frame rate up-conversion, image enlargement, spatio-temporal resolution enhancement*

1. Introduction

Videos with high resolution and high frame rate are desired and often required for most electronic imaging applications, and several methods for achieving high resolution and high frame rate have been proposed.

For frame rate up-conversion, video frame interpolation methods have been proposed [1]–[3]. In these methods, correspondence of pixels between two contiguous frames is estimated by using motion vectors, and intensities of the interpolated frames are obtained according to the estimation. However, since these methods assume that neighboring pixels have similar motion, the interpolation results sometimes contain artifacts, especially near boundaries between regions for which motions are different. Actually, in the interpolation results, over-smoothing or wrong moving objects appear at the boundaries.

To obtain high-resolution images from observed multiple low-resolution images, super-resolution image reconstruction has been proposed. Its basic premise for increasing spatial resolution is the availability of multiple low-resolution images captured from the same scene [4]. Most of the super-resolution image reconstruction methods consist of the three functions, registration, interpolation and restoration. It is well known that estimation of motion information as registration is important for the success of super-resolution image reconstruction.

Therefore, if high-resolution and high-frame rate video sequences are simultaneously desired, a simple combination of the above methods cannot provide sufficient quality because of motion estimation error. Another problem is that a simple combination does not consider the smoothness of the motion between two adjacent frames. It causes natural movement cannot be reconstructed.

In this letter, a new method for spatio-temporal resolution enhancement of video sequences based on super-resolution reconstruction is proposed. In the proposed method, a new observation model is defined and a cost function for estimation of spatio-temporal high-resolution sequences is newly realized. Owning to these two new features, the proposed method can achieve accurate spatio-temporal resolution enhancement.

2. Traditional Super-Resolution Image Reconstruction

Super-resolution is a resolution enhancement approach for obtaining a high-resolution image or sequence from observed multiple low-resolution images. A low-resolution sequence, which has M frames of $N_1 \times N_2$ pixels in size, is given and l-th frame is denoted in lexicographical notation as the vector $\boldsymbol{Y}_l = [y_l(1)y_l(2)\cdots y_l(N_1N_2)]^T$ $(l = 1, \cdots, M)$. Consider the desired high-resolution image of $N_1q_1 \times N_2q_2$ $(q_1 \geq 1,\ q_2 \geq 1)$ pixels in size, which is written in lexicographical notation as the vector $\boldsymbol{X} = [x(1)x(2)\cdots x(N_1q_1N_2q_2)]^T$. The observation model for super-resolution reconstruction is represented as follows:

$$\boldsymbol{Y}_l = \boldsymbol{D}_l\boldsymbol{B}_l\boldsymbol{F}_l\boldsymbol{X} + \boldsymbol{V}_l \qquad \text{for} \quad 1 \leq l \leq M, \tag{1}$$

where $\boldsymbol{D}_l$ represents a downsampling matrix of size $N_1N_2 \times N_1q_1N_2q_2$, $\boldsymbol{B}_l$ is a blur matrix of size $N_1q_1N_2q_2 \times N_1q_1N_2q_2$, $\boldsymbol{F}_l$ is a warp matrix of size $N_1q_1N_2q_2 \times N_1q_1N_2q_2$, and $\boldsymbol{V}_l$ is a noise vector. Using Eq. (1), the estimate of $\boldsymbol{X}$, denoted by $\hat{\boldsymbol{X}}$, can be given by

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} \left(\sum_{l=1}^{M} \|\boldsymbol{D}_l\boldsymbol{B}_l\boldsymbol{F}_l\boldsymbol{X} - \boldsymbol{Y}_l\|^2 + \alpha\|\boldsymbol{K}\boldsymbol{X}\|^2 \right), \tag{2}$$

where α is a regularization parameter, $\boldsymbol{K}$ is a matrix of size $N_1q_1N_2q_2 \times N_1q_1N_2q_2$ and works as a high-pass filter, and $\|\cdot\|^2$ presents an l_2 norm. From Eq. (2), we can obtain the high-resolution image $\hat{X}$.

By repeatedly applying the above to the low-resolution sequence, which has more than M frames, with sliding window length of M frames, we can obtain a high-resolution sequence. However, since this simple application only generates each frame of the high-resolution sequence independently of its adjacent frames, motion estimation error remains, and smoothness of the motion as an inherent characteristic of the video sequence is not considered. Therefore, we cannot expect high performance by this simple application.

Manuscript received March 15, 2012.
Manuscript revised May 18, 2012.

[†]The authors are with the School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: miki@ist.hokudai.ac.jp

DOI: 10.1587/transinf.E95.D.2355

3. Super-Resolution Reconstruction for Spatio-Temporal Resolution Enhancement of Video Sequences

We carried out the following procedures in order to achieve high-resolution sequence reconstruction:

(i) **New observation model**
The traditional observation model in Eq. (1) is modified to be suitable for spatio-temporal resolution enhancement. Its warp matrix is computed basis of estimated on the motion. In order to accurately estimate the motion, we assume that it consists of both camera motion and moving object motion and separately estimate these two motions.

(ii) **Cost function suitable for spatio-temporal resolution enhancement**
A new constraint is introduced to the optimization of cost function for obtaining a high-resolution sequence. The optimization formula works for keeping smooth motion between contiguous frames.

The above procedures enable spatio-temporal resolution enhancement to be achieved. The procedures are explained in detail below.

3.1 Modified Observation Model and Motion Estimation in Intermediate Frames

In the proposed method, we modify the observation model shown in Eq. (1) to be suitable for spatio-temporal super-resolution reconstruction as follows:

$$\boldsymbol{Y}_l = \boldsymbol{D}_l \boldsymbol{B}_l \left(\boldsymbol{F}_l^c \boldsymbol{F}_l^b \right) \boldsymbol{X} + \boldsymbol{V}_l. \tag{3}$$

The difference between the above equation and Eq. (1) is the warp matrix. The warp matrix in Eq. (1) $\boldsymbol{F}_l$ corresponds to $\boldsymbol{F}_l^c \boldsymbol{F}_l^b$ in Eq. (3), where $\boldsymbol{F}_l^c$ presents a camera motion matrix, and $\boldsymbol{F}_l^b$ represents a motion matrix of the moving object in the sequence.

Let us explain how to compute the matrices $\boldsymbol{F}_l^c$ and $\boldsymbol{F}_l^b$ by low-resolution frames. Scale Invariant Feature Transform (SIFT) [5] is adopted for motion estimation because it is well known that motion can be estimated with high precision by using SIFT even when lighting conditions change. The procedures are as follows.

Procedure 1: Feature point extraction
The feature points $p_l(j)$ and $p_{l+1}(k)$ ($j = 1, 2, \cdots, N_l$; $k = 1, 2, \cdots, N_{l+1}$) are detected by SIFT in the frames Y_l and Y_{l+1}, respectively, where N_l and N_{l+1} are the total numbers of feature points. The best matched feature point of $p_l(j)$ is selected among $p_{l+1}(k)$ ($k = 1, 2, \cdots, N_{l+1}$) according to the criterion in [5], which is defined as the distance between their feature vectors. In our method, the matching points satisfying the following equation remain to be processed in Procedure 2:

$$\frac{D_{first}}{D_{second}} < T_d. \tag{4}$$

In the above equation, T_d is a predefined threshold, D_{first} is the distance between the feature vectors of $p_l(k)$ and its best matched feature point in the frame Y_{l+1}, and D_{second} is the distance between the feature vectors of $p_l(k)$ and its second-best matched feature point in the frame Y_{l+1}. The matching strategy in Eq. (4) is commonly used, as shown in [6].
This procedure is repeatedly applied to each pair of frames, Y_l and Y_{l+1} in $l = 1, \cdots, M-1$.

Procedure 2: Calculation of trajectory vectors
We search all of the remaining feature points in **Procedure 1** for sets of points that can be completely tracked from $l = 1$ to $l = M-1$. Each set of completely tracked points has a trajectory vector. If its feature point in Y_l, whose coordinate is (x_l, y_l), corresponds to a feature point in Y_{l+1}, whose coordinate is (x_{l+1}, y_{l+1}), where $l = 1, \cdots, M-1$; its trajectory vector [7] is expressed as

$$\boldsymbol{t} = \begin{bmatrix} \boldsymbol{p}_1 \\ \boldsymbol{p}_2 \\ \vdots \\ \boldsymbol{p}_M \end{bmatrix}, \tag{5}$$

where $\boldsymbol{p}_l = [x_l \; y_l]^T$. All of the trajectory vectors are clustered by a clustering method such as that described in [7]. From the clustering results, the cluster C^1 to which most trajectory vectors belong is selected.

Procedure 3: Calculation of transformation matrix $\boldsymbol{H}^l$
The trajectory vectors belonging to cluster C^1 are written as

$$\boldsymbol{t}_k^{C^1} = \begin{bmatrix} \boldsymbol{p}_{1,k}^{C^1} \\ \boldsymbol{p}_{2,k}^{C^1} \\ \vdots \\ \boldsymbol{p}_{M,k}^{C^1} \end{bmatrix}, \qquad k = 1, \cdots, \phi, \tag{6}$$

where ϕ is the total number of trajectory vectors belonging to C^1, and $\boldsymbol{p}_{l,k}^{C^1} = [x_{l,k} \; y_{l,k}]^T$, where $(x_{l,k}, y_{l,k})$ is the coordinate of the feature point in Y_l, which is tracked by the k-th trajectory vector. The transformation matrix $\boldsymbol{H}_l$ ($l = 1, \ldots, M-1$) is obtained by minimization of the following criterion:

$$\sum_{k=1}^{\phi} \left\| \boldsymbol{p}_{l+1,k}^{C^1} - \boldsymbol{H}^l \begin{bmatrix} \boldsymbol{p}_{l,k}^{C^1} \\ 1 \end{bmatrix} \right\|^2, \tag{7}$$

where

$$\boldsymbol{H}^l = \begin{bmatrix} h_{1,1}^l & h_{1,2}^l & h_{1,3}^l \\ h_{2,1}^l & h_{2,2}^l & h_{2,3}^l \end{bmatrix}. \tag{8}$$

The transformation matrices H_l ($l = 1, \ldots, M$), which are obtained by minimization of Eq. (7), represent the dominant motion in the given video sequence, which is generally the camera motion.

Procedure 4: Computation of $\boldsymbol{F}_l^c$ and $\boldsymbol{F}_l^b$
According to the computed elements of $\boldsymbol{H}^l$, that is h^l's, the matrix $\boldsymbol{F}_l^c$ can be obtained. However, the video sequence also includes the motion of moving objects, and

it cannot be expressed only by $\boldsymbol{F}_l^c$; therefore, we have to prepare another matrix, $\boldsymbol{F}_l^b$, to express it. In order to compute $\boldsymbol{F}_l^b$, we assume that the pixels in the moving objects satisfy the following equation:

$$\frac{1}{(2W+1)^2}\sum_{w_1=-W}^{W}\sum_{w_2=-W}^{W}(|Y_l(x+w_1,y+w_2) - Y_{l+1}(x',y')|) > Th, \quad (9)$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \boldsymbol{H}^l \begin{bmatrix} x+w_1 \\ y+w_2 \\ 1 \end{bmatrix}. \quad (10)$$

Here, $Y_l(x,y)$ is the intensity of the pixel in the coordinate (x,y) of the l-th frame, $W \times W$ is the size of the block to be watched for the pixel detection, and Th is a predefined threshold. For the pixels detected by Eq. (9), $\boldsymbol{F}_l^b$ is computed by using the motion estimated by the block matching, and then $\boldsymbol{F}_l^c = \boldsymbol{I}$. If the pixels do not satisfy Eq. (9), $\boldsymbol{F}_l^b = \boldsymbol{I}$, and $\boldsymbol{F}_l^c$ is used, which has already been computed in **Procedure 3**.

By applying cubic spline interpolation to the motion estimated above, the motion in intermediate frames can be computed.

3.2 Estimation of Spatio-Temporal Resolution Enhancement Sequence

In order to accurately achieve spatio-temporal resolution enhancement, a suitable optimization scheme instead of Eq. (2) is required. Therefore, we introduce a new constraint to the cost function as follows:

$$\hat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}}\left(\sum_{l=1}^{M}\|\boldsymbol{D}_l\boldsymbol{B}_l\left(\boldsymbol{F}_l^c\boldsymbol{F}_l^b\right)\boldsymbol{X} - \boldsymbol{Y}_l\|^2 + \alpha\|\boldsymbol{K}\boldsymbol{X}\|^2 + \beta\|\boldsymbol{X} - \boldsymbol{X}_+\|^2 + \lambda\|\boldsymbol{X} - \boldsymbol{X}_-\|^2\right), \quad (11)$$

where $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ are the two given frames adjacent to $\boldsymbol{X}$, and β and λ are normalization parameters.

The new constraints, the third and fourth terms in the right side of Eq. (11), which are different from Eq. (2), work so that the motion between a given frame and its neighbor intermediate frames can be smoothly reconstructed.

4. Experimental Results

The performance of the proposed method is verified in this section. We used the test video sequence *City* of 704 × 576 pixels in size, 8 bit/pixel and 60 fps, and its total number of frames is 300. In order to obtain its low frame rate and low-resolution video sequence, we subsampled it to 352 × 288 pixels and 30 fps. Then we applied the proposed method to the low-resolution sequence and generated resolution enhanced video sequences at the original resolution, that is, of 704 × 576 pixels in size and 60 fps. When applying the proposed method to the test sequences, the parameters were set as $T_d = 0.01$ and $Th = 5.0$, which were simply determined by [8].

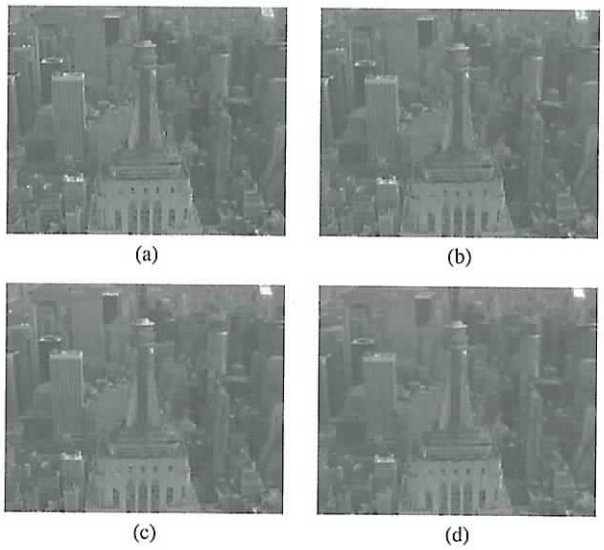


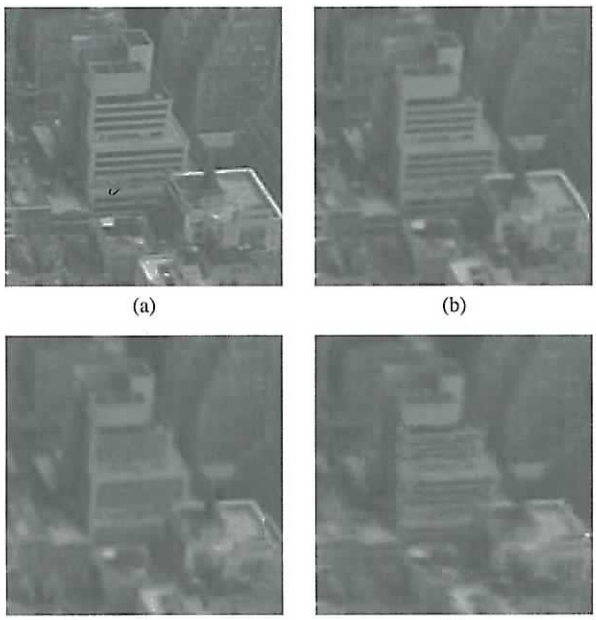
Fig. 1 Spatio-temporal resolution enhancement results. (a) The 127th frame of an original sequence "*City*". (b) Resolution enhancement result by the proposed method. Its PSNR is 33.2 dB. (c) Resolution enhancement by the method in [1]. Its PSNR is 32.6 dB. (d) Resolution enhancement by the method in [9]. Its PSNR is 32.2 dB.

For subjective evaluation, Fig. 1 is shown as follows: (a) is the 127th frame of *City*, (b) is the resolution enhancement result obtained by the proposed method, (c) is the result obtained by combined use of bicubic interpolation for enlargement and [1] for frame rate up-conversion, and (d) is the result obtained by [9] for up-conversion. Their enlarged portions, which include building windows, are shown in Fig. 2 (a)–(d). From these figures, we can see that the image obtained by the proposed method has sharper edges than those in (c) and (d), especially in texture of the buildings.

The proposed method was also applied to other frames of *City*, and the PSNR, which is defined below, is plotted in Fig. 3 (a).

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \quad (12)$$

where MAX is the maximum possible intensity, 255 in the experiments, and MSE is the mean square error of the original frame and the enhanced result. The average PSNR of the frames is also shown in Table 1 (a). The performance was also verified by using another test video sequence, *Coastguard*, of 352 × 288 pixels in size, 8 bit/pixel and 30 fps, and with a total of 150 frames, where the low frame rate and low-resolution video sequence is 176×144 pixels in size and 15 fps. The PSNR is plotted in Fig. 3 (b), and the average PSNR is shown in Table 1 (b). Based on the results shown in Fig. 3, we can see that the proposed method provides better performance in most of the frames than those in the other two methods for both two sequences. From Table 1, it can



(a) (b) (c) (d)

Fig. 2 Enlarged figures of Fig. 1 (a), (b), (c) and (d).

Table 1 Average PSNR of the spatio-temporal resolution enhancement results of each sequence.

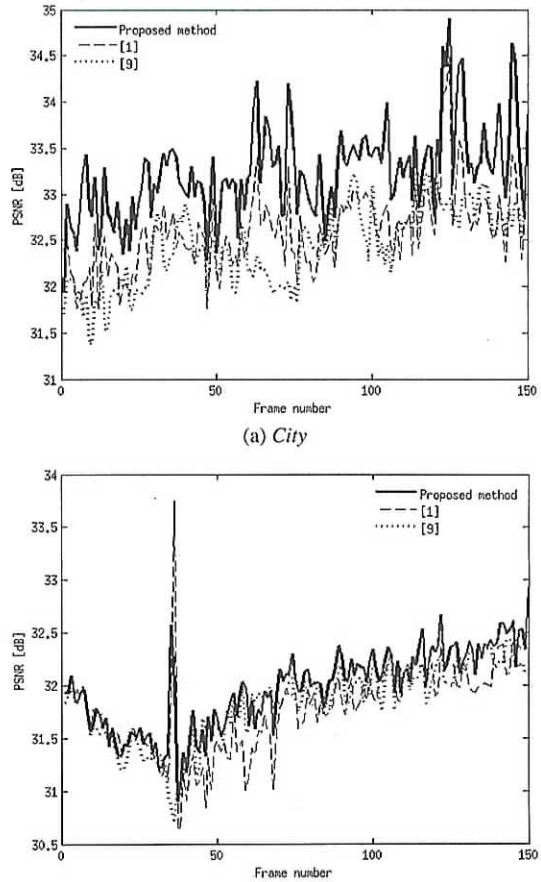
	Proposed Method	Method in [1]	Method in [9]
(a) "City"	33.2 dB	32.6 dB	32.4 dB
(b) "Coastguard"	32.0 dB	31.8 dB	31.8 dB

be verified that the average PSNR of the proposed method is higher than [1] and [9]. Further, note that the motion in the video sequence "*City*" is mainly the camera motion, and "*Coastguard*" includes both the camera motion and the object movement. Since the proposed method provides good performance in both their results, we can see that the camera motion estimation and the object movement estimation, which are embedded in the proposed method, are effective in resolution enhancement.

5. Conclusions

A new method for spatio-temporal resolution enhancement of video sequences based on super-resolution reconstruction has been proposed.

A new observation model, in which camera motion estimation and object movement estimation are embedded, has been realized. In addition a new constraint has been introduced into the cost function for obtaining a high-resolution sequence, the effect of which is to ensure smooth motion between contiguous frames. Finally, by experiments using actual video sequences, we have been able to verify the high performance of the proposed method.



(a) *City*

(b) *Coastguard*

Fig. 3 PSNR Comparison of the proposed method with [1] and [9].

References

[1] R. Castagno, P. Haavisto, and G. Ramponi, "A method for motion adaptive frame rate up-conversion," IEEE Trans. Circuits Syst. Video Technol., vol.6, no.5, pp.436–446, 1996.

[2] T.-Y. Kuo, J. Kim, and C.-C.J. Kuo, "Motion-compensated frame interpolation scheme for H.263 codec," Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '99), vol.4, pp.491–494, May-June 1999.

[3] H.A. Karim, M. Bister, and M.U. Siddiqi, "Multiresolution motion estimation for low-rate video frame interpolation," EURASIP Journal on Applied Signal Processing, vol.2004, no.11, pp.1708–1720, 2004.

[4] S.C. Park, M.K. Park, and M.G. Kang, "Super-resolution image reconstruction: A technical overview," IEEE Signal Process. Mag., vol.20, no.3, pp.21–36, 2003.

[5] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol.6, no.5, pp.436–446, 1996.

[7] Y. Sugaya and K. Kanatani, "Multi-stage optimization for multi-body motion segmentation," IEICE Trans. Inf. & Syst., vol.E87-D, no.7, pp.1935–1942, July 2004.

[8] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern., vol.9, no.1, pp.62–66, 1979.

[9] C. Wang, L. Zhang, Y. He, and Y.-P. Tan, "Frame rate up-conversion using trilateral filtering," IEEE Trans. Circuits Syst. Video Technol., vol.20, no.6, pp.886–893, 2010.