SPATIO-TEMPORAL RESOLUTION ENHANCEMENT OF VIDEO SEQUENCE BASED ON SUPER-RESOLUTION RECONSTRUCTION

Miki Haseyama, Daisuke Izumi and Makoto Takizawa

School of Information Science and Technology, Hokkaido University
N-14 W-9, 060-0814, Sapporo, Japan

ABSTRACT

This paper proposes a new spatio-temporal resolution enhancement method of video sequences based on super-resolution reconstruction. The proposed method derives a new observation model based on feature point correspondence between successive frames. The observation model is defined by including the motion estimation function in computing the warping matrix. Also, a new constraint is introduced to the optimization formula for estimating the parameters of the observation model, in order to achieve effective resolution-enhancement. By the newly obtained matrix and the new constraint introduction make accurate high-resolution and high frame rate video sequences. Simulation results are shown to confirm the high performance of the proposed method.

Index Terms— Video processing, super-resolution, frame rate up-conversion, image enlargement, spatio-temporal resolution enhancement

1. INTRODUCTION

In most electronic imaging applications, video with high resolution and high-frame rate are desired and often required. In recent work, several methods have been proposed for getting high-resolution and high-frame rate video sequences from low resolution sequences.

For frame-rate up-conversion, interpolation of video frames methods have been proposed [1, 2, 3]. They estimate correspondence of pixels between two contiguous frames by using the motion vectors. According to the correspondence, the intensities of the interpolated frames are obtained. However, since this method assumes that neighboring pixels have similar motion vectors each other, the interpolation results sometimes contain artifacts, especially near boundaries between regions whose motions are different. Actually, in their interpolation results, over smoothing appears in the boundaries. Further, to obtain high resolution images or sequences from observed multiple low-resolution ones, super-resolution image reconstruction[4] has been proposed. Its basic premise for increasing the spatial resolution in Super-resolution techniques is the availability of multiple low resolution images captured from the same scene. However, to obtain an edge-preserving high resolution image, we have to define appropriate functions in the algorithm.

This paper proposes a new spatio-temporal resolution enhancement method of video sequences based on super-resolution reconstruction. The proposed method derives a new observation model based on feature point correspondence between successive frames. The observation model is defined by including the motion estimation function in computing the warping matrix. Also, a new constraint is introduced to the optimization formula for estimating the parameters of the observation model, in order to achieve effective resolution-enhancement. By the newly obtained matrix and the new constraint introduction make accurate high-resolution and high frame rate video sequences. Simulation results are shown to confirm the high performance of the proposed method.

This paper is organized as follows. Firstly, in Section 2, super-resolution image construction the image morphing and the global motion estimation, which are used for the proposed method, are simply introduced. Second, in Section 3, the proposed method is explained in detail. Finally, experimental results are shown to verify the high performance of the proposed method in Section 4.

2. PREPARATION – SUPER-RESOLUTION IMAGE RECONSTRUCTION

In this section, some notations of the super-resolution processing are defined, which are used for derivation of the proposed method. The super-resolution is a resolution enhancement approach, which is to obtain an high resolution image or sequence from observed multiple low-resolution images. Now, a low-resolution sequence, which has \( M \) frames of size of \( N_1 \times N_2 \) pixels, is given and \( l \)-th frame is denoted in lexicographical notation as the vector \( Y_l = [y_l(1), y_l(2), \ldots, y_l(N_1N_2)]^T \) \( (l = 1, \ldots, M) \). Consider the desired high resolution image of size \( N_1q \times N_2q \) pixels, which is written in lexicographical notation as the
vector \( \mathbf{X} = [x(1)x(2) \cdots x(N_1N_2q^2)]^T \). The observation model for the super-resolution reconstruction is represented as follows:

\[
Y_l = D_lB_lF_l \mathbf{X} + V_l \quad \text{for} \quad 1 \leq l \leq M, \tag{1}
\]

where \( D_l \) represents a downsampling matrix of size \( N_1N_2 \times N_1N_2q^2 \), \( B_l \) is a blur matrix of size \( N_1N_2q^2 \times N_1N_2q^2 \), \( F_l \) is a warp matrix of size \( N_1N_2q^2 \times N_1N_2q^2 \), and \( V_l \) is a noise vector. Using the observation model defined in Eq. (1), the estimate of \( \hat{\mathbf{X}} \), denoted by \( \hat{\mathbf{X}} \), can be given by

\[
\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \left( \sum_{l=1}^{M} ||D_lB_lF_l \mathbf{X} - Y_l||^2 + \alpha||K \mathbf{X}||^2 \right), \tag{2}
\]

where \( \alpha \) is a regularization parameter, and \( K \) is a matrix of size \( q^2N_1N_2 \times q^2N_1N_2 \) and works to be a high-pass filter. From Eq. (2), we can get the high resolution image \( \hat{\mathbf{X}} \). By repeatedly applying the above to the low resolution sequence, which has more than \( M \) frames, with sliding window length \( M \) frames, we can get the high resolution sequence. However, if the simple application, that is to say each frame of the high resolution sequence is independently generated each other, we cannot expect high performance. It is because moving object might not be estimated smoothly. Therefore, in order to perform high resolution sequence reconstruction, a new observation model, in which camera motion estimation and object moving estimation are embedded, is proposed in the following section. By this modification, the proposed method can achieve spatio-temporal resolution enhancement simultaneously.

3. SPATIO-TEMPORAL HIGH RESOLUTION RECONSTRUCTION OF VIDEO SEQUENCE BASED ON SUPER RESOLUTION

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

\[
Y_l = D_lB_l \left( F_l^c F_l^b \right) \mathbf{X} + V_l. \tag{3}
\]

There is the difference between the above equation and Eq. (1) on the warp matrix. The warp matrix in Eq. (3) is \( F_l^c \) \( F_l^b \), where \( F_l^c \) presents a camera motion matrix, and \( F_l^b \) represents a motion matrix of the moving object in the sequence. The details to obtain spatio-temporal resolution enhancement results are shown below.

Subsection 3.1, in order for spatio-temporal resolution enhancement, firstly derives the motion estimation technique in the interpolation frames by using the known low resolution frames. Next, subsection 3.2 proposes the technique of simultaneously performing high-resolution and high frame up-conversion with using the motion estimated by 3.1.

3.1. Motion Estimation in Interpolated Frame

This subsection explains the technique of performing motion estimation in the interpolated frames with the low resolution frames. That is to explain how to compute the matrices \( F_l^c \) and \( F_l^b \) by the low resolution frames.

In order to accurately estimate the motion of the camera and the moving objects in the video sequence, we adopt Scale Invariant Feature Transform (SIFT) [5]. By using SIFT, the motion can be estimated with high precision, even when lighting conditions change. Let us explain the procedures below.

Procedure1: 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 [5] in the frames \( Y_l \) and \( Y_{l+1} \), respectively, where \( N_l \) and \( N_{l+1} \) are the total numbers of the feature points. The best correspondence feature point with \( p_l(j) \) 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 corresponding point pairs 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 vector of \( p_{l+1}(k) \) and the feature vector of the best matched feature point in the frame \( Y_l \), and \( D_{second} \) is the distance between the feature vector of \( p_{l+1}(k) \) and the second best matched feature vector. This matching strategy shown in Eq. (4) is commonly used, such as shown in [6].

Procedure2: Definition of trajectory vector

Procedure 1 is applied to \( Y_l \) and \( Y_{l+1} \) in \( l = 1, \cdots, M - 1 \). From the remained feature points, we select the points which can be tracked completely from \( l = 1 \) to \( l = M - 1 \). For example, if a feature point in \( Y_l \), whose coordinate is \( (x_l, y_l) \), is corresponding to a feature point in \( Y_{l+1} \), whose coordinate is \( (x_{l+1}, y_{l+1}) \), where \( l = 1, \cdots, M - 1 \); these pairs are selected, and their trajectory vector \([?]\) is defined by

\[
t_p = (x_1, y_1, x_2, y_2, \cdots, x_M, y_M)^T. \tag{5}
\]

Procedure3: Clustering of feature points

All the trajectory vectors obtained in Procedure 2 are clustered, where the number of clusters is determined by AIC. From the clustering result, the cluster \( C^3 \), to which most trajectory vectors belong, is selected.

Procedure4: Calculation of trajectory vector

By using the feature point correspondence between
two frames $Y_l$ and $Y_{l+1}$, which remains in Procedure1 and Procedure2, the transformation matrix $H_l (l = 1, \cdots, M)$ (3 × 3 matrix) is computed as follows. The feature points in $Y_l$ tracked by trajectory vectors belong to the cluster $C^i$ have the coordinates $(x_{i,1}^l, y_{i,1}^l)$ (i = 1, ⋯, φ; φ is the total number of the feature points belonging to $C^i$). For the feature points in $(x_{i,1}^l, y_{i,1}^l)$ the following projection transformation matrix $H_l$ is obtained below.

$$P_{l+1}^i = H_l P_l^i,$$

where $P_l^i = (p_{i,1}^l, \cdots, p_{i,\phi}^l)$ and $P_{l+1}^i = (x_{i,1}^l, y_{i,1}^l, 1)^t$. However, since the solution to Eq. (6) generally is not necessarily obtained, we compute $H_l$ by a least-squares method using singular value decomposition. By using $H_l$, the trajectory vector for each pixel is recalculated.

Procedure5: Computation of $F_l^c$ and $F_l^b$

By applying the third order spline interpolation to the trajectory vectors obtained in Procedure4, $F_l^c$ can be obtained, because the obtained motion is dominant motion of the video sequence, that is from camera motion. However, since the actual motion includes the motion of the moving objects, only $F_l^c$ cannot express it, because there are moving objects in the video sequence; therefore, it is detected by

$$\frac{1}{W^2} \sum_{w_1=1}^{W} \sum_{w_2=1}^{W} \left( |Y_l(x + w_1, y + w_2) - Y_{l+1}(x + v_2, y + v_2)| \right) > T_i$$

(7)

Where $Y_l(x, y)$ is the intensity in the coordinate $(x, y)$ of $Y_l$, and $T_i$ is a threshold. Further, the motion vector $v = (v_x, v_y, 0)^t$ of the pixel $s_l = (x, y, 1)^t$ of the frame $Y_l$ is calculated by $w = (w_1, w_2, 0)^t$ below.

$$v = H_l (s_l + w) - (s_l + w)$$

(8)

As a result, Eq. (7) is equivalent to block matching using the absolute error in the frame, and we evaluate an error about the block of neighborhood $W \times W$. In the pixels detected by Eq. (7), $F_l^b$ is computed by the above motion estimation method, and if no pixels satisfy Eq. (7), $F_l^b = I$.

Based on the motion estimation results obtained above, the interpolation frames are computed by the third order spline interpolation. When the given frames are super-resolutionized, the interpolation is not performed and each original pixel is used as it is.

3.2. Estimation of Super resolution and frame rate up-conversion Video Sequence

By using the motion estimation results obtained in the previous subsection, in order to accurately achieve spatio-temporal resolution enhancement, the super-resolution in 2 is applied to video sequence as follows.

$$\hat{X} = \arg \min_X \left( \sum_{l=1}^{M} ||D_l B_l (F_l^c F_l^b) X - Y_l||^2 + \alpha ||K X||^2 + \beta ||X - X_+||^2 + \lambda ||X - X_-||^2 \right)$$

where $X_+$ and $X_-$ are two known frames nearest to $X$, and express the high-resolution frame by super-resolution processing. Using the above equation, which includes the new terms in Eq. (2), we can keep smoothness between the interpolation frame and its nearest known frame. Further, $\beta$ and $\lambda$ express normalization parameters. By the above equation, high-resolution and high frame rate are achieved simultaneously.

4. EXPERIMENTAL RESULTS

The performance of the proposed method is verified in this section. We use a test video sequence City of size 704 × 576 pixels, 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 subsample it to 352 × 288 pixels and 30 fps. Then we apply the proposed method to the low sequence and generate the resolution enhanced video sequences at the original resolution, that is of size 704 × 576 pixels and 60 fps.

For subjective evaluation, Fig. 1 is shown as follows: (a) is an original frame of City, (b) is the resolution enhancement result by the proposed method and, (c) is the result by combination use of the bicubic interpolation and [1]. Further, their enlarged portions, which include building windows, are shown in Fig. 2 (a)–(c). From these figures, we can see that the result by the proposed method keeps sharper edges than (c), especially in the window frames.

The proposed method is also applied to the other frames and the PSNR, which is defined below, is plotted in Fig. 3.

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

(9)

where $\text{MAX}$ is the maximum possible intensity, and $\text{MSE}$ is the mean square error of the original image and the enhanced result. From Fig. 3, we can see that the proposed method has on average 0.85 higher SNR than [1].

5. CONCLUSIONS

We have proposed the method of performing super-resolution and frame-rate up-conversion simultaneously by modifying super-resolution processing. In the proposed method, the new observation model has been utilized and the suitable optimization scheme has been adopted for supporting spatio-temporal resolution enhancement. Also, the simulation results were shown to confirm the high performance of the proposed method.
Fig. 1. Spatio-temporal resolution enhancement results. (a) A frame of an original sequence “City” of size $704 \times 576$ pixels and 60fps. (b) Resolution enhancement result by the proposed method. (c) Resolution enhancement by the method in [1].

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

6. REFERENCES


