JOINT IMAGE REGISTRATION AND EXAMPLE-BASED SUPER-RESOLUTION ALGORITHM

Hyo-Song Kim, Jeyong Shin, and Rae-Hong Park

Department of Electronic Engineering, School of Engineering, Sogang University
35 Baekbeom-ro, Mapo-gu, Seoul, 121-742, Korea

ABSTRACT

Supper-resolution (SR) methods are classified into two different methods: image registration (IR)-based methods and example-based methods. The proposed joint SR method is focused on estimating high-resolution (HR) video sequences from low-resolution (LR) ones by combining the two different methods. The IR-based SR method collects information from adjacent frames to reconstruct HR images in the video sequence. Example-based SR methods give good textures and strong edges in the result HR video. In this paper, IR-based and example-based SR methods are fused based on the gradient features. The proposed joint SR method gives smaller peak signal to noise ratio than the example-based method, however it shows better reconstruction results on high-level features such as characters in images. Experimental result of the proposed joint SR method shows less noise and higher contrast than the example-based method.

KEYWORDS

Super-Resolution, Image Registration, Motion Estimation, Motion Compensation, Example-Based Learning, Sparse Coding, Neigborhood Regression.

1. INTRODUCTION

Supper-resolution (SR) methods are traditionally classified into two different classes: image registration (IR)-based and example-based methods [1]. IR is the task of finding the motion between two or more frames of the same scene. It is noted that the motion estimated by IR does not necessarily describe the true motion of either the camera or the object in a scene. The most common way to allow a reliable implementation of the IR is to estimate the motion using only two-dimensional translations, under the assumption of small motion within the scene. In [1] more elaborated methods are presented including methods using rotation and scaling, basically intended for scanned documents.

IR-based SR methods can be classified into two main approaches: featured-based methods and block-based methods. The first approach fits a motion model by using a sparse set of point correspondences, while the second one uses the information of the entire pixels within search area. Feature-based SR methods can cause false matches due to their regression nature and require a large number of points in order to achieve a high level of accuracy. On the contrary, the block-based methods can obtain the motion information with a good accuracy for producing a high-resolution (HR) image. While the feature-based methods use a small number of points, the block-based methods utilize all the overlapping blocks of the adjacent images. A lot of block-based methods estimate image motion by minimizing a cost function between two motion-corrected images [2], however suffer from a very high computational cost.

Although SR methods which use motion information between adjacent frames have been successful on some of the video sequences, they are not suitable for the video sequences with stationary objects and background, which do not contain information to fill in the sub-pixels. Moreover, estimating a true motion vector in spatially high-frequency regions and handling occluded regions is not a simple task. Thus, SR techniques on a single image are still important even though the goal is to perform SR on video sequences. Most single image SR methods use example-based approach, which learns from example images to form a dictionary and apply it to each image patch. Xiong *et al*. [3] used the soft information and decision on the one-to-many correspondence of dictionaries to solve the dimensionality gap between low-resolution (LR) and HR spaces. Timofte *et al*. improved the conventional sparse coding method by introducing anchored neighborhood regression (ANR) method [4]–[6]. Zhu *et al*. [7] used optical flow on image patches to form deformable patches, in order to make the learned dictionary more expressive.

2. Related Work

2.1. IR-Based SR Methods

IR-based SR methods [1]–[2] generally concentrate on spatial domain and typically consist of two processes: IR and HR reconstruction. IR is the process that calculates the motion parameters based on a specific motion model. HR reconstruction incorporates the estimated motion parameters into inverse estimation. The aliasing effect among LR images may reduce the accuracy of registration. Accurate registration is a challenging task because motion parameters are calculated from a number of aliased LR images. To deal with the problem, various SR algorithms have been proposed to reduce the registration error on the final estimated HR image.

2.2. Dictionary-Based SR Methods

A large number of dictionary-based SR methods [3]–[8] have been developed for the last decade. They use dictionary of image patches or patch-based atoms that are trained to represent natural image patches efficiently, which brings advantages in both time complexity and accuracy of the SR results.

2.2.1. Neighbor Embedding Methods

Neighbor embedding methods are generally used in patch-based methods. They assume that the LR input patches can be approximated by a weighted sum of HR trained patches. LR and HR patches are learned simultaneously in the training phase. A typical method, locally linear embedding [3], can be written as

$$\mathbf{x} = \sum_{i=1}^{K} w_i \mathbf{x}'_i \, , \tag{1}$$

where $\mathbf{x}$ denotes the result HR patch and $\mathbf{x}'_i$ represents the i th candidate HR patch, which is the pair of the i th nearest neighbor LR patch of the input LR patch. K is the number of neighbors in consideration and w_i denotes the weight of the i th candidate HR patch.

2.2.2. Sparse Coding Methods

Unlike neighbor embedding methods, sparse coding methods try to find efficient representations, dictionary atoms, of the image patches. Zeyde *et al.* [8] built an efficient and improved method to train a sparse dictionary. They built dictionaries for both LR patches and their corresponding HR patches by using joint optimization on these two patches. After LR and HR dictionaries are trained, a sparse representation $\boldsymbol{\alpha}^*$ of an input LR image patch can be calculated as

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \left\| D_l \boldsymbol{\alpha} - \mathbf{y} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_1 , \qquad (2)$$

where D_l denotes the LR dictionary, $\boldsymbol{\alpha}$ is a sparse representation of LR patch, $\mathbf{y}$ represents the input LR patch, and λ denotes the regularization parameter which controls the significance of the sparsity constraint (l_1-norm in the second term) over the modelling error (l_2-norm in the first term).

2.2.3 ANR Method

ANR method [4] reformulates the dictionary of Zeyde *et al.* [8] to pre-compute regression matrices used to calculate the result HR patches. Instead of using LR and HR dictionaries directly, it considers the nearest neighbors of a specific j th dictionary atom $\mathbf{d}_l^j$, which is the j th column vector of the LR dictionary D_l. Using only the nearest neighbors of $\mathbf{d}_l^j$, a local neighborhood LR dictionary N_l^j is calculated along with its corresponding local neighborhood HR dictionary N_h^j. Instead of l_1 -norm, l_2 -norm is used for the sparsity constraint to calculate a sparse representation $\boldsymbol{\beta}^*$, which can be written as

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \left\| N_l \boldsymbol{\beta} - \mathbf{y} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_2 , \qquad (3)$$

where the superscript j is omitted for simplicity. (3) can be solved in a closed-form solution by ridge regression [9], which is written as

$$\boldsymbol{\beta}^* = (N_l^T N_l + \lambda I)^{-1} N_l^T \mathbf{y} , \qquad (4)$$

and then followed by

$$\mathbf{x} = N_h \boldsymbol{\beta}^* = P^j \mathbf{y} , \qquad (5)$$

where $\mathbf{x}$ denotes the SR result and P^j is the projection matrix which projects an input LR patch directly onto $\mathbf{x}$. Note that P^j can be pre-calculated, which improves the algorithm a lot in terms of the time complexity. In summary, ANR learns off-line regressors for fast SR, while improving qualities using the neighborhood concept.

2.2.4. Adjusted ANR Method

Adjusted ANR (A+) method [5] is an improved ANR algorithm, which is based on the following observation. First, the dictionary atoms are sparsely sampled in the space, whereas the training pool of image patch samples obtained in off-line training is practically near-infinite. Second, the local manifold around an atom is spanned better by dense training samples than by the dictionary atoms.

Based on the observation, A+ method reformulates the optimization problem (3), which can be written as

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \left\|S_l \boldsymbol{\delta} - \mathbf{y}\right\|_2^2 + \lambda \left\|\boldsymbol{\delta}\right\|_2 , \tag{6}$$

where S_l denotes a LR dictionary, which contains K *neighboring training samples* (possibly pre-calculated per atom in off-line training). S_l replaces N_l that contains *neighboring atoms* in (3). As A+ uses the same baseline of ANR method, (6) can also be solved by ridge regression. Thus, the closed-form solution can be obtained as

$$\mathbf{x} = S_h \boldsymbol{\delta}^* = \hat{P}^j \mathbf{y} , \tag{7}$$

where S_h denotes HR dictionary corresponding to S_l and $\hat{P}^j$ represents the projection matrix obtained from S_l. Note that S_h projects an input LR patch $\mathbf{y}$ directly onto HR patch $\mathbf{x}$. Using such strategy, A+ method chooses better neighborhood for local dictionaries, which drastically improves the SR result.

3. PROPOSED JOINT SR METHOD

The proposed IR-based SR method generates image grid by referencing neighboring frames. The overall process of the IR-method SR is illustrated in Figure 1. The resolution of image grid is *N* times higher than that of the original image in both width and height. *N* is related to the sub-pixel accuracy for motion search. If *N* is set to 4, quarter-pel accuracy motion estimation is performed on the neighboring reference frames. All pixels in image grid are reconstructed using the information of neighboring reference frames. The reconstructed image grid has full HR to reconstruct an HR image using down-sampling. The down-sampling process uses a super sampling anti-aliasing (SSAA) method, which performs patterned sampling method such as grid, random, Poisson, jitter, and rotated grid. Rotated grid is well known as good for removing edges [10], therefore we use it for image sampling.

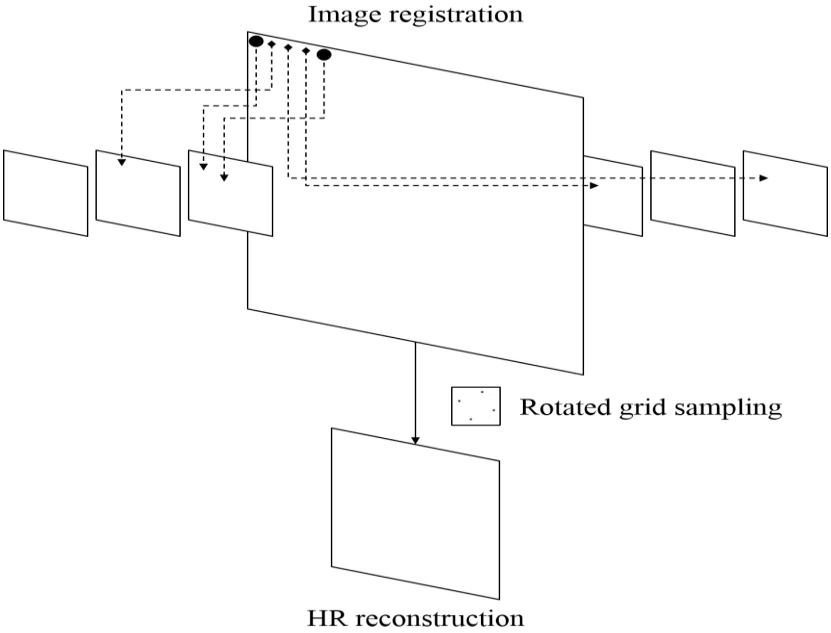
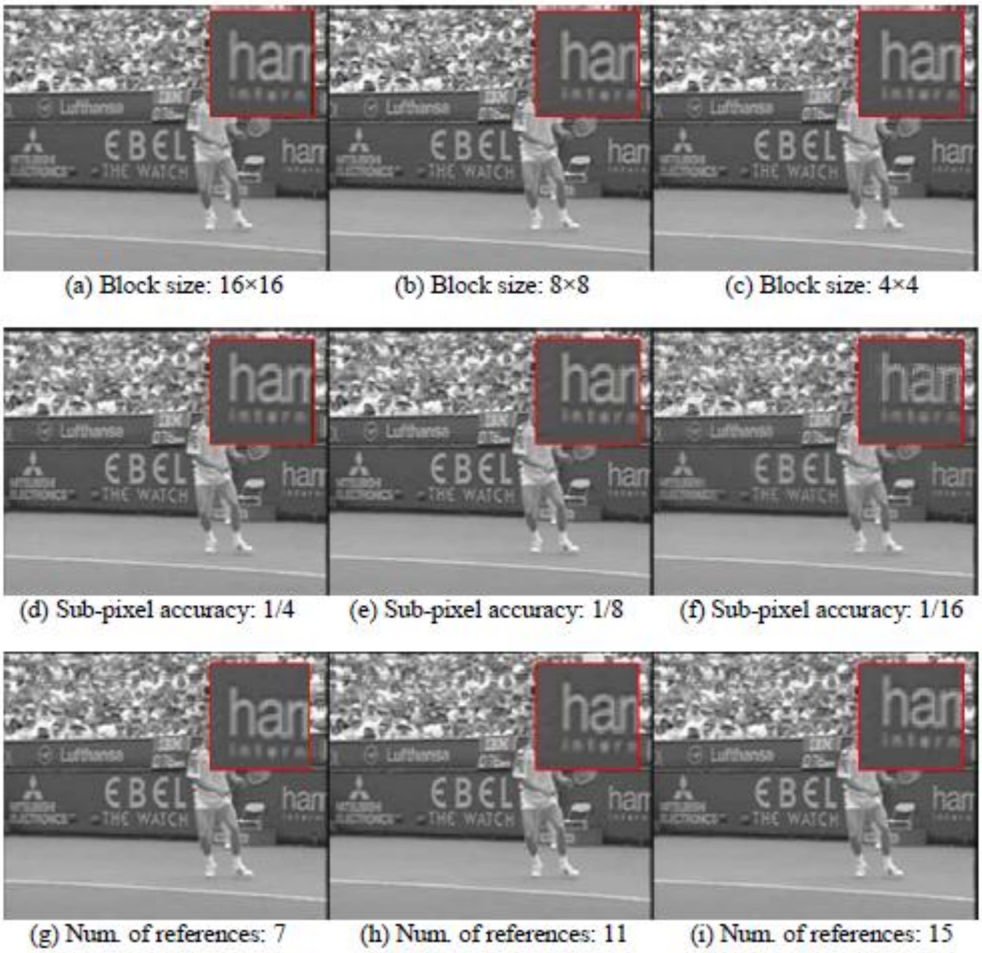


Figure 1. An example of the proposed IR-based SR method. The sub-pixel accuracy is set to four. Circles denote the integer-pels and diamonds represent the sub-pels.

Figure 2 shows experimental results of the proposed IR-based SR method according to different parameter setting. We adjust three parameters, which are block sizes, sub-pixel accuracy, and the number of reference frames. Default parameter setting is as follows: block size = 32×32, sub-pixel accuracy = 1/2, and the number of reference frames = 3. As shown in Figures 2(a)-2(c), small block size reduces block artifact in object boundaries, however, computational burden is increased. Figures 2(d)-2(f) show that high level of sub-pixel accuracy increases the quality of the reconstructed HR image, however, a lot of holes are produced (see Figure 2(f)) because default parameter setting for the number of references is too small. To solve this problem, a large number of references are used to obtain lots of temporal information (observed from Figures 2(g)-2(i)), however, irrelevant information can be presented by scene change. Figure 2(i) shows the quality loss using irrelevant reference frames in case of fast motion. In consideration of both image quality and computational complexity, we use the following parameter setting: block size = 8×8, sub-pixel accuracy = 1/8, and the number of reference frames = 11.



(a) Block size: 16×16 (b) Block size: 8×8 (c) Block size: 4×4

(d) Sub-pixel accuracy: 1/4 (e) Sub-pixel accuracy: 1/8 (f) Sub-pixel accuracy: 1/16

(g) Num. of references: 7 (h) Num. of references: 11 (i) Num. of references: 15

Figure 2. Experimental results of the proposed IR-based SR method according to different parameter setting for the *Stefan* image sequence.

Adjusted ANR method [5] is used for example-based SR. After reconstructing an HR image using both IR- and example-based methods, the proposed method combines the two HR images [11] by using a gradient-based weight function. The final reconstructed image is calculated by

$$F_J(i,j) = \omega^{IR}(i,j) F_{IR}(i,j) + \omega^{EX} F_{EX}(i,j) \tag{8}$$

where F_J represents the final reconstructed image, F_{IR} denotes the HR image reconstructed by the IR method, and F_{EX} is the HR image reconstructed by the example-based method. The weight functions are defined as

$$\omega^{IR} = \frac{|g \circ F_{IR}(i,j)|}{|g \circ F_{IR}(i,j)| + |g \circ F_{EX}(i,j)|} \tag{9}$$

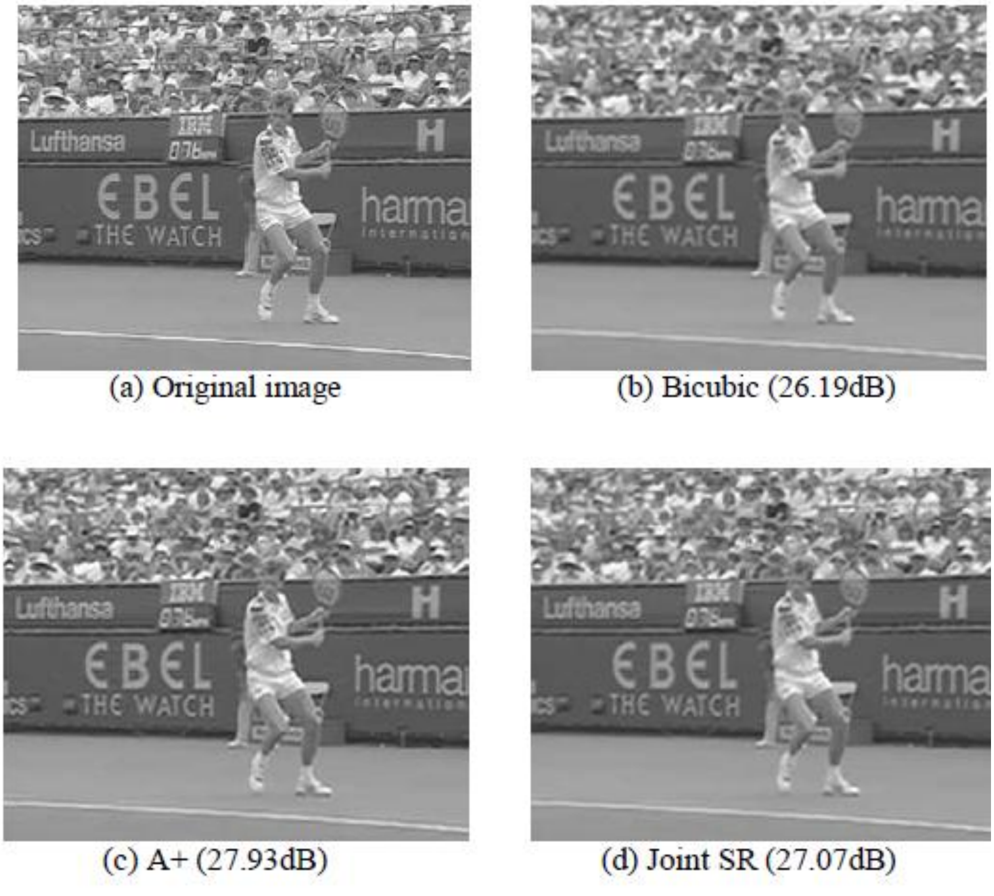
$$\omega^{EX} = \frac{|g \circ F_{EX}(i,j)|}{|g \circ F_{IR}(i,j)| + |g \circ F_{EX}(i,j)|} \tag{10}$$

where g represents gradient operation and $\circ$ denotes convolution operation. The joint SR method enhances the reconstructed HR image quality by weighting two images reconstructed by different approaches.

4. Experimental Results and Discussions

The simulated image sequence is a common intermediate format (CIF) video, of which spatial resolution equals 352 × 288 at 30 frames per second. The test sequence named *Stefan* contains dynamic scenes playing tennis. One can observe cluttered background located at the top of the scene caused by crowd, fast motions of the tennis players, large camera motion tracking of the players, and characters of various sizes on the walls.

In this paper, the CIF videos are down-sampled to quarter CIF (QCIF) size, which equals 176 × 144. SR is performed on the created QCIF videos to recover CIF videos. For example-based SR, the number of the atoms in the dictionary κ is set to 16. For registration-based SR, sub-pixel accuracy N is set to 8, with the search range=32, and the number of reference frames=11.



(a) Original image

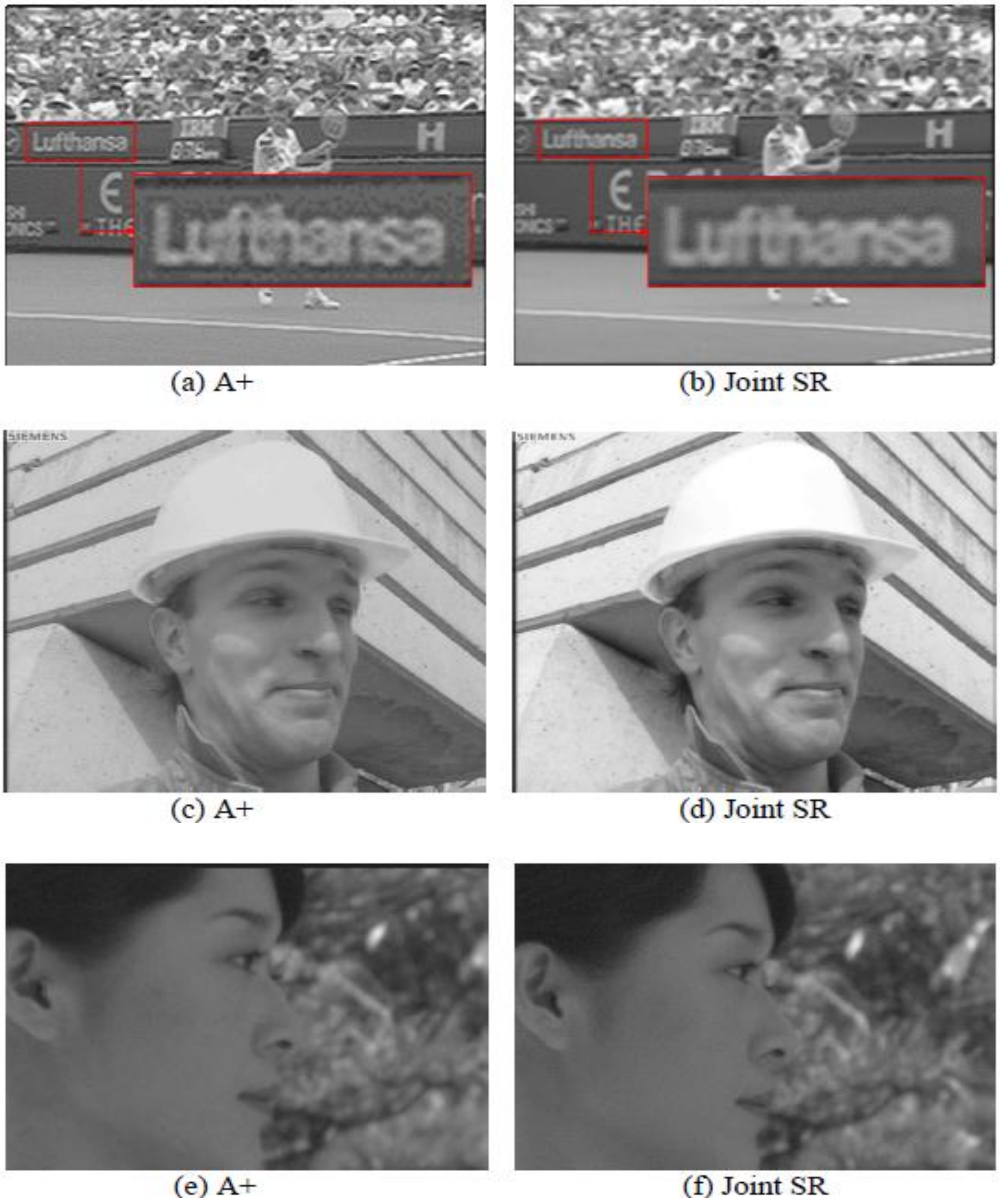
(b) Bicubic (26.19dB)

(c) A+ (27.93dB)

(d) Joint SR (27.07dB)

Figure 3. Experimental results and performance comparison in terms of the PSNR for the LR image sequence (*Stefan*).

Figure 3 shows the SR result on the *Stefan* sequence with their cropped/zoomed images at the corners. For a reference of the comparison, SR result using bi-cubic interpolation is shown in Figure 3(b). Example-based SR result is shown in Figure 3(c). It can be observed that example-based SR gives fine details on the complex textures such as crowd. Also, the result by example-based SR shows the best result on the strong edges (see shoulder of the tennis player, horizontal lines across the image, and so on). However, English characters located at the bottom left are not clearly readable, since there is not sufficient information to recover the characters in a single LR image. The IR-based SR method collects information from a number of adjacent frames to build an HR image, even though there is only little evidence in a single frame. Figure 3(d) shows the SR result by the proposed joint SR method which takes advantages from both IR- and example-based approaches although the peak signal to noise ratio (PSNR) is somewhat decreased. However, it shows better SR quality on characters (shown in Figures 4(a) and 4(b)).



(a) A+ (b) Joint SR

(c) A+ (d) Joint SR

(e) A+ (f) Joint SR

Figure 4. Quality comparison of the proposed joint SR with the A+ for three different image sequences.

Figures 4(a)-4(b) compare the proposed joint SR with A+ method for the *Stefan* sequence, whereas Figures 4(c)-4(d) for the *Foreman* sequence. The result of the proposed joint SR method shows less noise and higher contrast than A+ method. Figures 4(e) and 4(f) show the experimental results performed on the full HD-size (1920×1080) *Kimono* image sequence. The proposed joint SR method up-scales the *Kimono* sequence from full HD to UHD 4K (3840×2160) resolution. In contrast with the *Stefan* sequence (LR sequence), the experimental results of the Kimono sequence (HR sequence) shows little difference in qualitative comparison.

5. CONCLUSION

The proposed SR method is focused on reconstructing HR video sequences from LR video sequences. The proposed IR-based SR method successfully collects information from adjacent frames to reconstruct English characters in the video sequences. However, the example-based SR method gives better textures and strong edges in the result HR video. In this paper, IR- and example-based SR methods are fused based on the gradient features. The proposed joint SR method gives smaller PSNRs than the example-based method, however it shows better reconstruction results on high-level features. Future work will focus on optimizing the joint SR method using convolutional neural network to reduce the time complexity of the algorithm.

6. ACKNOWLEDGMENTS

This work was supported in part by the Brain Korea 21 Plus.

REFERENCES

[1] Y. Tian and K.-H. Yap, “Joint image registration and super-resolution from low-resolution images with zooming motion,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1224–1234, Jul. 2013.

[2] C. Liu and D. Sun, “On Bayesian adaptive video super resolution,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[3] Z. Xiong, D. Xu, X. Sun, and F. Wu, “Example-based super-resolution with soft information and decision,” *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1458–1465, May 2013.

[4] R. Timofte, V. D. Smet, and L. V. Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in *Proc. IEEE Int. Conf. Computer Vision*, pp. 1920–1927, Sydney, Australia, Dec. 2013.

[5] R. Timofte, V. D. Smet, and L. V. Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in *Proc. Asian Conf. Computer Vision*, pp. 1–15, Singapore, Nov. 2014.

[6] R. Timofte and L. V. Gool, “Adaptive and weighted collaborative representations for image classification,” *Pattern Recognition Letters*, vol. 43, pp. 127–135, Jul. 2014.

[7] Y. Zhu, Y. Zhang, and A. L. Yuille, “Single image super-resolution using deformable patches,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 2917–2924, Columbus, OH, USA, Jun. 2014.

[8] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in *Lecture Notes in Computer Science: Curves and Surfaces*, J.-D. Boissonnat and P. Chenin, Eds., Springer, pp. 711–730, 2012.

[9] Y. Zhu, Y. Zhang, and A. L. Yuille, “Single image super-resolution using deformable patches,” in *Proc. IEEE Computer Vision and Pattern Recognition*, pp. 2917–2924, Columbus, OH, USA, Jun. 2014.

[10] R. Barringer and T. A. Moller, “A4: Asynchronous adaptive anti-aliasing using shared memory,” *ACM Trans. Graphics*, vol. 32, no. 4, pp. 100:1–100:10, Jul. 2013.

[11] K. Lee and C. Lee, “High quality spatially registered vertical temporal filtering for deinterlacing,” *IEEE Trans. Consumer Electronics*, vol. 59, no. 1, pp. 182–189, Feb. 2013.