Image Super-Resolution via Feature-Based Affine Transform

Chih-Chung Hsu [1] and Chia-Wen Lin [2]

[#] *Department of Electrical Engineering, National Tsing Hua University*
Hsinchu, Taiwan

[1] m121754@gmail.com

[3] cwlin@ee.nthu.edu.tw

***Abstract*—State-of-the-art image super-resolution methods usually rely on search in a comprehensive dataset for appropriate high-resolution patch candidates to achieve good visual quality of reconstructed image. Exploiting different scales and orientations in images can effectively enrich a dataset. A large dataset, however, usually leads to high computational complexity and memory requirement, which makes the implementation impractical. This paper proposes a universal framework for enriching the dataset for search-based super-resolution schemes with reasonable computation and memory cost. Toward this end, the proposed method first extracts important features with multiple scales and orientations of patches based on the SIFT (Scale-invariant feature transform) descriptors and then use the extracted features to search in the dataset for the best-match HR patch(es). Once the matched features of patches are found, the found HR patch will be aligned with LR patch using homography estimation. Experimental results demonstrate that the proposed method achieves significant subjective and objective improvement when integrated with several state-of-the-art image super-resolution methods without significantly increasing the cost.**

I. INTRODUCTION

Image/video super-resolution (SR) has become an attractive technique in enhancing the resolution of low-resolution (LR) images because it has many applications such as security in surveillance video, enhancement of aged photos and captured images from low-power devices. For example, a LR image can be taken from a mobile device. If we want to obtain the high-resolution (HR) image from LR image, the image super-resolution can be used to solve such problem.

In general, there are two kinds of the image super-resolution techniques which are multi-frame and single-frame methods. The first-type methods need to obtain multiple input LR images and align these LR images to calculate the missing sub-pixel values within pixels. Since the multi-frame SR methods require accurate alignment, the scaling-up factor of the LR image has its limit practically. On the other hand, single-frame image SR schemes do not suffer from this problem. The single-frame based methods collect a candidate/training pool which may contain a large number of the LR and HR pairs. In the reconstruction stage, each LR patch is replaced with it corresponding HR patch(es) obtained from the candidate/training pool. Therefore, the performance of the single-frame SR methods heavily depends on the comprehensive of the candidate/training pool. This work focuses on enhancing the performance of single frame image SR.

The most representative image SR methods include example-based super-resolution (ES) [2], sparse coding (SC) [4], nonlocal means filter (NLM) [5], and texture synthesis super-resolution (TSS) [6]. All of these methods require a comprehensive dataset as the training set to obtain good visual quality of the reconstructed image. For example, ES requires to collect a comprehensive training set containing HR patches and their LR counterparts and the relationship between LR and HR patches is modeled as Markov Random Fields (MRF). Similarly, TSS uses the texture synthesis technique to hallucinate the HR patch from training set. SC also collects a large number of training samples and learn a set of overcomplete bases. The super-resolved patches can be replaced with a linear combination of the overcomplete basis under an L1 norm constraint.

The basic concept of NLM [5] is to exploit self-similarity in an image. Unlike ES, the pool of candidate patches is obtained from the input LR image itself. Similarly, once the candidate patches for the input LR patch are found, the LR patch is replaced with a linear combination of its corresponding HR candidate patches. Note that, when the candidate poll for NLM is not comprehensive enough, we may not find sufficiently similar candidate patches for an LR patch. As a result patches result in blurring effect due to dissimilar patches [4].

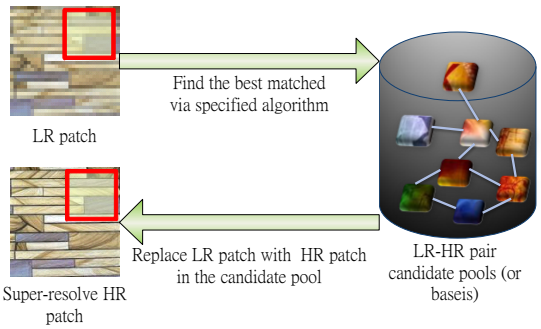
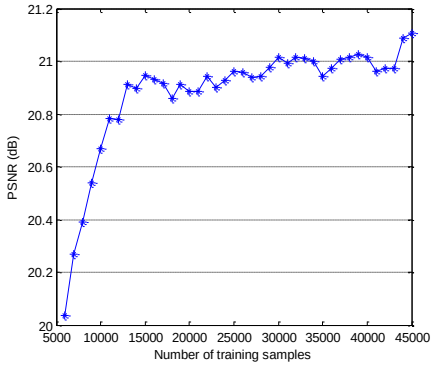


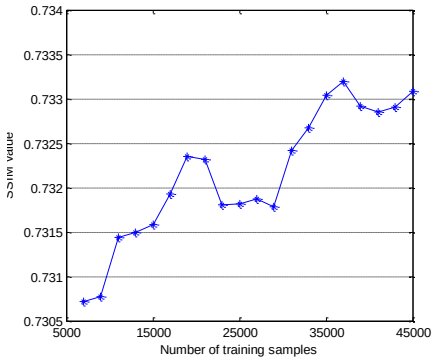
Fig. 1 Illustration of search-based SR methods in [2][4][6][7].

Most of the above state-of-the-art SR schemes can be represented as search-based schemes as illustrated in Fig. 1. The search-based schemes search for each LR input patch in the candidate/training pool a set of best-match HR patches. Consequently, the LR patch is replaced with the HR patch

obtained from the matched HR patches (e.g., a linear combination of the matched HR patches). These search-based SR methods usually rely on search in a comprehensive dataset for appropriate high-resolution patch candidates to achieve good visual quality of reconstructed image. Exploiting different scales and orientations in images can effectively enrich a dataset. A large dataset, however, usually leads to high computational complexity and memory requirement, which makes the implementation impractical. For example, Glasner *el al.*'s work [7] exploits the multi-scale similarity in an image to improve the visual quality of SR. However, if the level of the multi-scale factor is too high, the computational cost will also become very high as well.



(a)



(b)

Fig. 2 An example of objective quality comparison among different sizes of training set under (a) PSNR and (b) SSIM metrics.

Fig. 2 shows that a large number of training samples will result in better visual quality of the reconstructed image using example-based super-resolution [2]. However, such a large dataset results in that the implementation is very difficult due to very high computational cost and memory requirement. Therefore, we introduce the feature extraction process to greatly reduce the number of the training samples so that the visual quality of the reconstructed image will be improved with a reasonable cost. The time complexity comparison among different numbers of training samples is depicted in Fig. 3. As illustrated, the time complexity increases when the number of the training samples increases. Although the time complexity in Fig. 3 is $O(n)$ (or linear time), the time complexity will suddenly increase because the memory requirement of the training samples is greater than the limits of the capacity of physical memory, making the implementation infeasible.

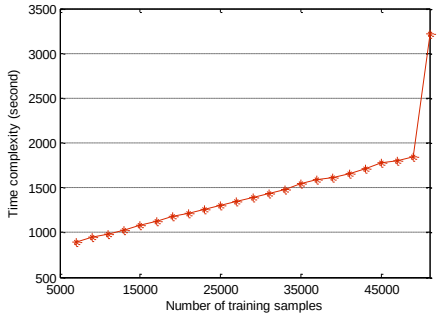


Fig. 3 Time complexity comparison among different number of training samples using example-based super-resolution [2].

This paper proposes a universal framework for enriching the dataset for search-based super-resolution schemes with reasonable computation and memory cost. Toward this end, the proposed method first extracts important features with multiple scales and orientations of patches based on the SIFT (Scale-invariant feature transform) descriptors and then use the extracted features to search in the dataset for the best-match HR patch(es). Once the matched features of patches are found, the found HR patch will be aligned with LR patch using homography estimation. The method proposed in [12] also adopts SIFT to solve the registration problem among LR image sequences. Such conventional multi-frame SR method provides limited performance. However, the proposed method can be integrated with different search-based SR algorithms, which effectively extends the search-space to improve the quality of the reconstructed image with a reasonable time-complexity.

The rest of this paper is organized as follows. Section II presents the proposed general framework of the affine-transform-based patch representation (APR). In Section III, we present the improved image super-resolution methods using the proposed APR framework. Experimental results of the proposed scheme are demonstrated in Section IV. Finally, Section V concludes this paper.

II. Image Super-Resolution via Affine Transform Patch Representation

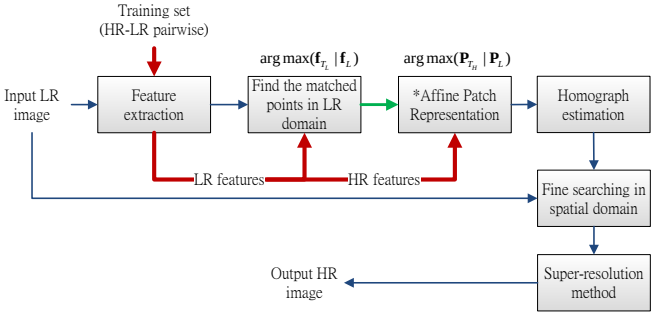


Fig. 4 The flowchart of the proposed APR framework.

As shown in Fig. 4, the proposed APR method is composed of four major steps. In the training stage, the SIFT feature vectors with various scales and orientations are calculated and stored for the training images. In the super-resolution stage, the first step is to calculate the SIFT feature vector of each patch of the input LR image and find the best match in the training set. Second, the homography matrix between the matched pair is established. Then, the matched HR patch is aligned using the calculated homography matrix. Finally, the best match of each input patch is obtained from the aligned patches of the training set in the spatial domain.

The feature extraction step extracts scale- and orientation-invariant features based on the SIFT descriptors. Let the *i*th SIFT feature vector be denoted as $\mathbf{f}_i \in \mathbb{R}^{1\times 128}$, and the patch of an image as $\mathbf{p}$. Finding the most similar patch $\mathbf{p}_j^{T_L}$ in the training set for input LR patch $\mathbf{p}_j^{I_L}$ can be formulated as the following maximum likelihood (ML) problem:

$$\mathbf{p}_j^{T_L *} = \arg\max_{\mathbf{p}_j^{T_L}} p(\mathbf{p}_j^{T_L} \mid \mathbf{p}_i^{I_L}) \tag{1}$$

In general, search-based approaches are used to solve the above problem [2], [5], [7]. However, to achieve good visual quality of a super-resolved image, a large-size training set is usually required, thereby leading to high computational complexity as shown in Fig. 3. To reduce computation while maintaining comparable visual quality, we propose to convert this problem to a feature matching problem by extracting the SIFT features of training set and input LR image so that it is possible to find good candidate patches under different scales and orientations without performing full-search in the spatial domain. Besides, the feature extraction and matching is only performed for visually important patches, whereas the remaining unimportant patches can be super-resolved using a baseline method without sacrificing the visual quality of the reconstructed image.

On the other hand, using the original SIFT features would suppress the edge and corner pixels. Instead, we retain these feature points because the visually important regions in an image usually exist around the edge or corner pixels.

After extracting the SIFT features, the solution to ML problem in (1) can be approximated by maximizing the likelihood between two feature vectors as follows:

$$\mathbf{f}_j^{T_L *} = \arg\max_{\mathbf{f}_j^{T_L}} p(\mathbf{f}_j^{T_L} \mid \mathbf{f}_i^{I_L}) \tag{2}$$

However, since the matched feature vectors ($\mathbf{f}_j^{T_L}$ and $\mathbf{f}_i^{I_L}$) may be obtained from different scales and orientations, the corresponding patch pair ($\mathbf{p}_j^{T_L}$ and $\mathbf{p}_i^{I_L}$) may also be with different scales and orientations. One problem is how to determine these two parameters. Estimating the orientation and scale information directly from the SIFT features is usually not accurate enough, since the values of these parameters have been quantized prior to calculating SIFT features. To obtain an accurate estimate, a homography matrix can be established for estimating the two parameters for the patch pair. Toward this end, given a source patch $\mathbf{p}_s$, we can obtain an encoded patch by

$$\mathbf{p}_e = s(R(\mathbf{p}_s)) + \mathbf{d} \tag{3}$$

where R represents a rotation function, s is a scaling function, $\mathbf{d}$ is the spatial translation. This process can be represented as an affine transformation of coordinates of patches defined below:

$$\begin{cases} x_e = m_1 x_s + m_2 y_s + t_1 \\ y_e = m_3 x_s + m_4 y_s + t_2 \end{cases} \tag{4}$$

where m_1, m_2, m_3, and m_4 are the parameters for controlling the scaling ratio and rotation angle, (t_1, t_2) denotes the translation vector, and (x_e, y_e) and (x_s, y_s) indicate the coordinates of the encoded patch and source patch, respectively. As a result, there are six parameters to be determined. To estimate these parameters, we formulate (4) as the homography estimation problem [9] expressed by

$$\underbrace{\begin{bmatrix} x_e \\ y_e \\ w \end{bmatrix}}_{\mathbf{c}_e} = \begin{bmatrix} m_1 & m_2 & t_1 \\ m_4 & m_5 & t_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix} \simeq \underbrace{\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix}}_{\mathbf{c}_s} \tag{5}$$

where $\mathbf{c}_e$ and $\mathbf{c}_s$ denote the coordinates. This problem can be solved by linear least-squares approximation via rearranging (5) in a simple form as follows:

$$\begin{bmatrix} x_s & y_s & 1 & 0 & 0 & 0 & -x_e x_s & -x_e y_s & -x_e \\ 0 & 0 & 0 & x_s & y_s & 1 & -y_e x_s & -y_e y_s & -y_e \end{bmatrix} \mathbf{h} = \mathbf{A}\mathbf{h} = \mathbf{0} \tag{6}$$

where $\mathbf{h} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{21} & h_{22} & h_{23} & h_{31} & h_{32} & h_{33} \end{bmatrix}^T$.

The form cannot be directly solved by using conventional least-squares solution for $\mathbf{Ah} = \mathbf{b}$. Instead, such problem can be solved using singular-value decomposition (SVD). Let $\mathbf{Ah} = \mathbf{0}$, we have

$$\begin{aligned} f(\mathbf{h}) &= \frac{1}{2}(\mathbf{Ah})^T(\mathbf{Ah}) = \frac{1}{2}\mathbf{h}^T\mathbf{A}^T\mathbf{Ah} \\ \frac{\partial f}{\partial \mathbf{h}} &= 0 = \frac{1}{2}\left(\mathbf{A}^T\mathbf{A} + (\mathbf{A}^T\mathbf{A})^T\right)\mathbf{h} = \mathbf{A}^T\mathbf{Ah} \end{aligned} \tag{7}$$

Consequently, the solution of $\mathbf{h}$ should be equivalent to the eigenvector of $\mathbf{A}^T\mathbf{A}$ that has an eigenvalue of zero. Then, the calculated $\mathbf{h}$ vector is converted to matrix form $\mathbf{H}$. Once the homography matrix for an input patch is obtained, the matrix can be used to align the matched patch from the training set with the input patch. With this correspondence, the pixel values in the encoded patch can be determined using a bilinear function. Consequently, the encoded patch can be obtained using the following homography and warping functions:

$$\mathbf{c}_e^{\text{affine}} = \mathbf{H}^{\text{opt}}\mathbf{c}_s \tag{8}$$

and

$$\mathbf{p}_e(\mathbf{c}_e) = w\left(\mathbf{p}_s\left(\mathbf{c}_e^{\text{affine}}\right)\right) \tag{9}$$

where w is a bilinear interpolation function.

For each feature, we can find the best-match patch and then align it to the input. To increase the matching accuracy, we

perform fine search within a small search window centered around the matched LR patch in the training set as illustrated in Fig. 5. The size of search window **S** is empirically set to be two times larger than the size of input LR patch.

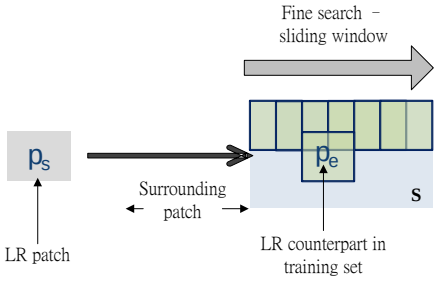


Fig. 5 Illustration of the fine-search process.

Note, some patches may be too smooth to contain not have no feature due to weak edges or smoothness. For these regions, the naïve super-resolution method is used. It is reasonable because these regions contain unimportant information that we would not notice. In other words, the proposed APR method is only used on the important regions.

III. Application to Search-Based Image Super-Resolution

For search-based SR schemes, the visual quality of super-resolved image increases when the number of the candidate/training samples increases. The proposed APR method effectively enriches the candidate/training pool by exploiting the multi-scale and multi-orientation self-similarity within an image. In this section, we shall show that the proposed APR method can be easily integrated with several state-of-the-art search-based SR algorithms to further improve the performance of these SR schemes.

As mentioned previously, ES also adopts the search-and-replacement strategy for super-resolution. Let "SIFT" represent the SIFT features with different scales and orientations and "L1-SIFT" denote the SIFT features with the first-level of scale. To integrate the proposed APR framework with ES method, the SIFT features of images in the training set and the L1-SIFT features of input LR image need to be extracted. Then, the best match between the SIFT and L1-SIFT features can be obtained by finding

$$\mathbf{f}_j^{T_L*} = \arg\min_{\mathbf{f}_j^{T_L}} \| \mathbf{f}_i^{I_L} - \mathbf{f}_j^{T_L} \|_2 \qquad (10)$$

where $\mathbf{f}_i^{I_L}$ denotes the ith feature vector of input LR image $\mathbf{I}_L$ and $\mathbf{f}_j^{T_L}$ denotes the jth feature vector of the candidate pool.

Once the best-match feature vector in the candidate pool is found, the surrounding patches $\mathbf{s}_i^{I_L}$ and $\mathbf{s}_j^{T_L}$ centered at $\mathbf{f}_i^{I_L}$ and $\mathbf{f}_j^{T_L}$ can be retrieved for estimating the corresponding homography matrix. Then, the histogram of the oriented gradients (HoG) is calculated for each pixel within a 3x3 patch in the surrounding patches because the estimation of homography matrix requires four matched pairs to determine the parameters. After finding the four best-match pairs between the surrounding patches $\mathbf{s}_i^{I_L}$ and $\mathbf{s}_j^{T_L}$, the homography matrix can be calculated by

$$\mathbf{H}^{\text{opt}} = \arg\min_{\mathbf{H}} \left\| \mathbf{s}_i^{I_L}(\mathbf{c}_i^{I_L}) - w\left(\mathbf{s}_j^{T_L}\left(\mathbf{H}\mathbf{c}_j^{T_L}\right)\right) \right\|_2 \qquad (11)$$

Note that the homography matrix is estimated with the patch size of 64×64 in the pixel domain. Once the found patch $\mathbf{s}_j^{T_L}$ is rectified using the estimated homography matrix $\mathbf{H}^{\text{opt}}$, fine search is then performed within a small search window in the pixel domain to find the best-match patch $\mathbf{p}_j^{T_L}$ corresponding to the input LR patch $\mathbf{p}_i^{I_L}$ (see Fig. 5).

Once the closest 16 patches $\mathbf{p}_j^{T_L}$ and $j = 1,...,16$ in the training set are found, the MRF is used to solve the spatial coherence problem and super-resolve the input LR patch $\mathbf{p}_i^{I_L}$. Repeating the above process to super-resolve all of the patches which have features. For the patches without the features, the conventional ES method is used to super-resolve those LR patches.

Similarly, the NLM-based SR schemes are also kind of search-based methods. The main difference between ES and NLM is that the training patches for NLM are obtained from the input image itself. Therefore, the L1-SIFT and SIFT features of images in LR image should be extracted simultaneously. For each input patch that can be extracted a SIFT feature vector, we find the corresponding best-match feature vectors that minimizes (10). Then, the homography matrix can be estimated using (11). The advantage is that the best-match patches can be found in an extended candidate pool with multiple scales and orientations, thereby improving the visual quality of the reconstructed image. For those patches that do not have significant SIFT features, the conventional NLM is used instead. After all patches are super-resolved, a deblurring filter is utilized to obtain a sharper HR image.

Although the concept of SC is different from ES, the sparse representation can also be represented as a search-based approach. In general, the solution is to predict the input patch from the linear combination of several candidate patches in the training set [8]. This would lead to the blurring effect in the reconstructed image when the number of the candidates increases. To resolve the problem, sparsity priors can be imposed to solve such problem, aiming to minimize the following cost function with an L1-norm constraint.

To improve the performance of SC, the goal is to find a comprehensive training set to learn a more representative set of overcomplete bases. The proposed APR method can effectively enhance the comprehensiveness of the training set by scaling and rotating the patches in the original training set, while maintaining acceptable computational cost through feature-based matching.

IV. Experimental Results

Our training set contains 26 HR natural images. The scaling-up factor is 3x3. The patch sizes are 3x3 for the ES, SC, and NLM schemes. The level of scale-space used for

SIFT extraction is 5. More simulation results can be found in [13].

The proposed APR framework exploits the multi-scale and multi-orientation self-similarity of patches to enrich the candidate pool. As shown in Fig. 2, when the search pool is enriched, the performance of the search-based super-resolution techniques will be improved.

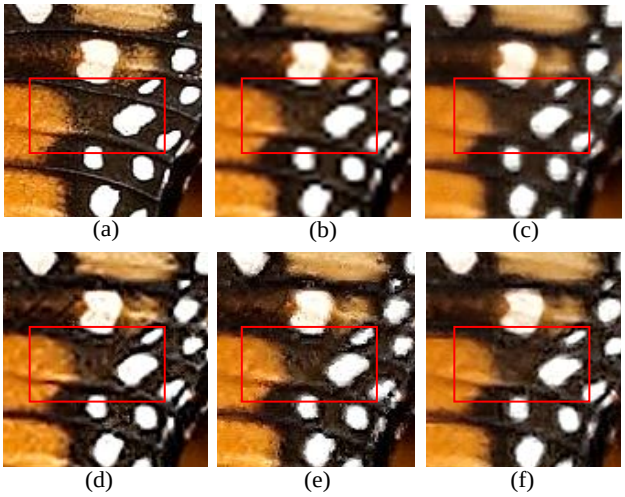


Fig. 6 Comparison of visual quality of reconstructed images: (a) ground-truth, (b) bicubic interpolation, (c) the method in [7] (SSIM: 0.72/MSE:179), (d) SC [4] (SSIM: 0.71/MSE:211), (e) ES [2] (SSIM: 0.72/MSE:198), and (f) ES + APR **(SSIM: 0.76/MSE:161**).

Fig. 6 compare the visual qualities of super-resolved images using the proposed APR method on top of ES (APR + ES) with bicubic interpolation, Glasner *et al.*'s method [7], SC, and ES. As illustrated, the proposed APR framework significantly improves the subjective visual quality of the reconstructed image, especially on edge regions. Besides, the proposed method also introduces fewer noisy artifacts in the smoothing regions compared to the other methods. Table I shows the objective quality comparison among three baseline super-resolution schemes and the proposed APR method on top of the baseline schemes for four test images. Our method significantly improves the objective quality of reconstructed images when it is integrated with the search-based SR methods. For these search-based SR methods, the quality of reconstructed image is dependent on the richness of candidate/training set. By exploiting the multi-scale and multi-orientation self-similarity of an image, the improvement comes from the enriched candidate/training pool which usually makes it easier to find a better match for a patch to be super-revolved so that the artifacts in the reconstructed image can be reduced. According to our experiments, the proposed method leads to 10–20% increase in execution time for different SR methods.

TABLE I
COMPARISON OF OBJECTIVE VISUAL QUALITIES OF RECONSTRUCTED IMAGES USING FOUR STATE-OF-THE-ART SR SCHEMES WITH & WITHOUT THE PROPOSED APR IMPROVEMENT

Methods	*Image* 1	*Image* 2	*Image* 3	*Image* 4
ES [2]	21.3 dB/0.75	19.3 dB/0.71	23.2 dB/0.74	22.8 dB/0.73
ES + APR	22.0 dB/0.79	20.0dB/0.72	23.5 dB/0.81	**23.7 dB/0.79**
Gain	0.7 dB/0.04	0.7 dB/0.01	0.3 dB/0.07	**0.9 dB/0.06**
NLM [5]	22.1 dB/0.73	19.9 dB/0.72	22.6 dB/0.71	23.3 dB/0.77
NLM + APR	22.7 dB/0.74	**20.8 dB/0.81**	23.0 dB/0.76	24.1 dB/0.83
Gain	0.6 dB/0.01	**0.9 dB/0.09**	0.4 dB/0.05	0.8 dB/0.06
SC [4]	22.8 dB/0.75	19.2 dB/0.72	23.3 dB/0.76	21.9 dB/0.78
SC + APR	**23.7 dB/0.76**	19.8 dB/0.75	**23.8 dB/0.81**	22.3 dB/0.82
Gain	**0.9 dB/0.01**	0.6 dB/0.03	**0.5 dB/0.05**	0.4 dB/0.04

Note, the performance of the proposed method relies on extracting gradient features for patch matching across different scales and orientations. Therefore, if an input LR image contains too few gradient features, the improvement with the proposed method would become marginal.

V. CONCLUSION

We have proposed an efficient APR framework to enrich the candidate/training pool of search-based SR schemes by exploiting the multi-scale and multi-orientation self similarity existing in an image. By exploiting the self-similarity of an image, our method significantly enriches the candidate/training pool, making it easier to find a better match for a patch to be super-revolved with the search-based SR schemes. One main contribution of the proposed method is to apply feature-based matching to significantly reduce the high computational cost for searching in a large candidate/training pool. Besides, the proposed APR framework can be easily integrated with state-of-the-art search-based SR algorithms. Experimental results demonstrate that the proposed method effectively improves the visual qualities of super-resolved images both subjectively and objectively.

REFERENCES

[1] B. Baker and T. Kanade, "Limits on superresolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sept. 2002.

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

[3] Q. Shan, Z. Li, J. Jia, and C.-K. Tang, "Fast image/video upsampling," *ACM Trans. Graphics*, vol. 27, no. 5, 2008.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 1-8, 2008.

[5] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol.18, no.1, pp.36-51, Jane 2009.

[6] Y. HaCohen, R. Fattal, and D. Lischinski, "Image upsampling via texture hallucination," *in Proc. IEEE Int. Conf. Comput. Photography*, Cambridge MA USA, pp. 20-30, Mar. 2010.

[7] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," *in Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2009.

[8] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 275-282, Washington DC, 2004.

[9] Dubrofsky and Elan Nathan, "Homography Estimation," MS. Thesis, Department of Computer Science, UBC University.

[10] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346--359, 2008.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91-110, February 2004.

[12] Y. Zhi, Y. Peimin, and L. Sheng, "Super resolution based on scale invariant feature transform," in *Proc. IEEE Int. Conf. Audi., Langu. and Image Process.*, pp.1550-1554, 7-9 July 2008.

[13] *NTHU image super-resolution project.* [*Online*]. *Available:* http://www.ee.nthu.edu.tw/nvlab/SR/index.htm.