UHD Video Super-Resolution using Low-Rank and Sparse Decomposition

Salehe Erfanian Ebadi Valia Guerra Ones Ebroul Izquierdo
Queen Mary University of London, United Kingdom
{s.erfanianebadi, v.guerra, e.izquierdo}@qmul.ac.uk

Abstract

Sparse coding-based algorithms have been successfully applied to the single-image super resolution problem. Conventional multi-image super-resolution (SR) algorithms incorporate auxiliary frames into the model by a registration process using subpixel block matching algorithms that are computationally expensive. This becomes increasingly important as super-resolving UHD video content with existing sparse-based SR approaches become less efficient. In order to fully utilize the spatio-temporal information, we propose a novel multi-frame video SR approach that is aided by a low-rank plus sparse decomposition of the video sequence. We introduce a group of pictures structure where we seek a rank-1 low-rank part that recovers the shared spatio-temporal information among the frames in the group of pictures (GOP). Then we super-resolve the low-rank frame and sparse frames separately. This assumption results in significant time reductions, as well as surpassing state-of-the-art performance both qualitatively and quantitatively.

1. Introduction

The recovery of high-resolution (HR) images and videos from low-resolutions (LR) content is a topic of great interest in digital image processing with applications in many areas such as HDTV [11], medical imaging [20], satellite imaging [23], face recognition [12], immersive content generation, and surveillance [27]. The global super-resolution (SR) problem assumes that the LR image is a noisy, low-pass filtered, and downsampled version of the HR image. This problem is highly ill-posed, as a result of the high-frequency information being lost during the non-invertible low-pass filtering and subsampling. Moreover, the SR problem is practically a one-to-many mapping from the LR to HR space that can have multiple solutions. Finding the correct solution amongst the possible solutions is non-trivial. In SR techniques it is generally assumed that the majority of the high-frequency data is redundant and can be reconstructed accurately from the low-frequency content. The SR methods can be divided into two categories. The multi-image SR (MISR) and single-image SR (SISR) methods. Conventional MISR methods [14], [10], [1] attempt to exploit the explicit redundancy by constraining the problem with additional information, i.e., they normally require multiple low-resolution images of the same scene. However, these models usually require complex subpixel image registration [21] and fusion stages, the accuracy of which directly impacts the quality of the result. The SISR methods attempt to learn the implicit redundancy present in natural data to recover the HR data from the available LR counterpart. These can include but are not limited to the local spatial correlations in images and temporal correlations in videos. A comprehensive survey of recent SISR methods can be found in [25].

A recent thriving family of the SISR methods is sparsity-based techniques that suggest image patches can be well-represented as a sparse linear combination of elements from an appropriately chosen over-complete dictionary [26], [14], [24], [8], [4]. According to this observation, one could seek a sparse representation for each patch of the LR input, and then use the coefficients of this representation to generate the HR output. The learned dictionary should then be able to embed the prior knowledge necessary to constrain the ill-posed problem of SR. Attempts have been made in order to adapt the sparse-based techniques to MISR problem to improve the output quality. Recently, an extension of the model in [26] has been proposed by Kato *et al.* [14], to incorporate multi-frame SR where uni-level HR dictionaries are learned using patches from the HR training images, and the LR dictionary is generated in the testing phase assuming a blur and an estimated translation between the LR target patch and the LR auxiliary patches. In [14], it is assumed the distortion are simple vertical and horizontal translation and the warping operators that are calculated using a sub-pixel block matching algorithm proposed by [21]. A drawback of these approaches is that the registration process is generally computationally expensive. To perform the SR operation on an image, it is necessary to increase the resolution of the LR image to the resolution of the HR image at some stage in the process. Several models based on deep neural networks [3], [17], [19] have achieved this

by upscaling the LR image to the HR space using a single filter, commonly Bicubic interpolation, before reconstruction. That is the SR operation is performed in the HR space, which is sub-optimal and adds computational complexity. Moreover, the previously mentioned methods typically require enormous databases of millions of HR and LR patch pairs, and therefore are computationally expensive.

A reasonable assumption when processing video information is that most of a scene's content is shared by neighboring video frames; except for the scene changes and objects intermittently appearing and then disappearing from the scene. This provides additional redundancy that can be exploited for video super resolution. An SR method that is able to utilize the inherent spatio-temporal information in the video, can potentially demonstrate better performance across a wide range of video SR tasks.

In this paper, we propose a novel multi-frame SR approach for the video SR problem. Our method operates on groups of pictures (GOP) in the LR domain that each contain between 8 to 64 frames. The GOP structures have been used in the literature [28] to accelerate the SR process, using the motion vector, block-size, and prediction residual values that are computed by the video encoder. Here, in each GOP, we calculate a low-rank + block-sparse decomposition [6] in order to separate the static blocks and the dynamic blocks in the video frames, while accounting for the possible camera-induced motion in the background of the scene. We refer to the static blocks that are decomposed in the low-rank component as background, although this may not be the correct nomenclature given the characteristics of this decomposition; similarly we refer to the dynamic (changing) blocks that are decomposed into the sparse component as the foreground. The obtained LR background frame and LR foreground frames are the super-resolved separately with a sparsity-based approach using a compact over-complete dictionary of atoms. Then the HR GOP is reconstructed using the obtained HR background frame and HR foreground frames.

Motivated by [26] the SR part of our algorithm requires only two compact learned dictionaries. Moreover, by super-resolving the background and foreground parts separately, the computation becomes more efficient and scalable, compared with [26], [14], and [10]. The efficiency of our method is two-fold: firstly that the frames in a GOP usually share a significant number of similar blocks, that implicitly enable us to exploit spatio-temporal redundancy in the video. Secondly, we strictly set the rank of the background of each GOP to 1, meaning that we obtain a single image that can be representative of the whole GOP's unchanging pixel structures; this then implies that we only have to super-resolve one background for the whole GOP. The sparse part contains many zero blocks that are super-resolvable by several orders of magnitude faster than its original corresponding frame. Also, the number of operations needed for the matrix decomposition is significantly smaller than that of a block-matching algorithm used in state-of-the-art alternatives. Consequently, these lead to superior performance, both qualitatively and quantitatively, compared to other state-of-the-art alternatives.

The rest of this paper is organized as follows. In Section 2 we describe the fundamentals of sparsity-based SR. Then in Section 3 we introduce our multi-frame video SR method called VSRGOP. The modified approximated RPCA method for SR problem is introduced. Finally, in Section 4 we demonstrate the efficacy of our proposed method by extensive experimental evaluations.

2. Sparse-Based SR

We denote the LR image as Y, and the HR image of the same scene as X. Lowercase y and x denote the low- and high- resolution image patches, respectively. D is used to refer to the dictionary for sparse coding; specifically the D_l and D_h denote the dictionaries for low- and high- resolution image patches, respectively. It has been statistically proven that image patches can be well-represented as a sparse linear combinations of elements, namely atoms of a dictionary taken from a finite and not too big bag [5], [26]. Each vectorized patch $y \in \mathbb{R}^m$ of an LR image Y, can be written as:

$$y = \alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_n D_n, \tag{1}$$

where most of the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ are zero if the atoms $D_1, D_2, \ldots, D_n$ of the dictionary D are properly selected. When $m = n$, D has to be a complete basis to represent any patch. However, when $n > m$ it is possible to find solutions $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ where a considerable number of coefficients α_i are zero. We can conveniently assume a sparse representation for y as each patch is completely determined for a substantially reduced number of parameters that is usually far less than the number of atoms.

To calculate the sparse representation of a patch one needs to determine the appropriate dictionaries D (learning phase), and then estimate the coefficients of the linear combination of the atoms (testing phase). We can find the sparsest α results in the convex Lasso regularized minimization problem below

$$\min_{\alpha} \|D\alpha - y\|_2^2 + \mu\|\alpha\|_1, \tag{2}$$

where μ is a regularization parameter to balance the reconstruction error and sparsity. Different solvers such as Least Angle Regression (LARS), Shooting algorithm, etc., have been used to solve this problem. A systematic way to calculate the dictionary D is solving the following minimization problem

$$\min_{D,Z} \|DZ - X\|_2^2 + \mu\|Z\|_1, \tag{3}$$

where X is the HR training data. The objective function above is non-convex with respect to both D and Z. Z contains the coefficients of the linear combination of the atoms that approximate the training data. The problem above can be solved in an alternating process, by keeping one fixed and solving for the other at a time until convergence. This alternating solution is convex. The selection of training data and the incorporation of structures and characteristics in the dictionaries is application-specific.

2.1. Single-Image SR based on Sparse Coding

Yang *et al.* [26] assume that the degradation from the HR patch x to the LR patch y is nearly linear, where each HR patch and its corresponding LR patch share the same sparse linear coefficients $\alpha = (\alpha_1, \ldots, \alpha_n)$. The high-resolution dictionary D_h and the low-resolution dictionary D_l need to be defined properly. There are then two stages to solving the sparse representation-based SR: the learning phase where the bi-level dictionaries D_h and D_l are constructed, and the testing phase where the vector coefficients α that correspond to each LR patch are calculated.

2.1.1 Learning Phase

We assume that the sparse representation of the HR patches is the same as the sparse representation of the corresponding LR patch; therefore, the set of training samples can be formed by a group of N HR sampled patches X_h and M LR sampled patches Y_l (here $N = M$). The HR and LR vectorized atoms are the columns of the matrices D_h and D_l that solve the following minimization problem

$$\min_{D_h, D_l, Z} \|X_c - D_c Z\|_2^2 + \mu \|Z\|_1, \tag{4}$$

where $X_c = \begin{bmatrix} \frac{1}{\sqrt{N}} X_h \\ \frac{1}{\sqrt{M}} Y_l \end{bmatrix}$ and $D_c = \begin{bmatrix} \frac{1}{\sqrt{N}} D_h \\ \frac{1}{\sqrt{M}} D_l \end{bmatrix}$.

The minimization problem above is non-convex with three variables D_h, D_l and Z. A convex solution would be an alternating process where two variables are kept fixed and the other one is solved until convergence. When D_h and D_l are fixed, the optimization problem is solved by non-negative quadratic linear programming using feature sign (L1QP solver). When Z is fixed, a constrained quadratic programming technique in its dual formulation is used. The details of this solution appears in [15].

2.1.2 Testing Phase

Here, given a LR patch y, the HR desired patch x can be defined as

$$x = D_h \alpha^l, \tag{5}$$

where α^l is the solution of the minimization problem

$$\alpha^l = \arg\min_{\alpha} \|y - D_l \alpha\|_2^2 + \mu \|\alpha\|_1 \tag{6}$$

This problem can be solved using the LARS-Lasso algorithm [7] or the feature-sign search algorithms [15]. To increase perceptual quality of the results a few more steps are required. In order to enforce the compatibility between adjacent patches, the authors in [26] proposed an overlapping strategy that modifies the minimization problem (6) that involves the HR and LR dictionaries. Also, a feature transformation F is used to enforce the high-frequency content of the LR image. Finally, once the HR image has been reconstructed patch by patch using sparse coding, a back-projection algorithm is performed to enforce the global reconstruction constraint to correct for noise in the LR image.

3. VSRGOP: Multi-Frame Video SR

We propose a novel sparse coding-based algorithm for multi-frame SR in videos that is aided by a low-rank and sparse decomposition (LRSD) to fully utilize the spatio-temporal information in the video. To the best of our knowledge only a handful of algorithms based on multi-frame sparse coding-based SR exist in the literature where usually an expensive block-matching algorithm is used. Our algorithm is the first to involve a LRSD step in order to avoid the registration by block-matching. The majority of SR algorithms have been proposed to the SISR problem and do not take into account the temporal information in videos. In [28] the authors proposed to use the motion vectors, block sizes, and prediction residual that is computed by the video encoder in compressed videos to accelerate their algorithm. Low-rank and sparse decomposition (LRSD) methods have been used in many applications such as background subtraction [6], [9], robust subspace clustering, etc.; however, these LRSD models are not suitable for the problem at hand. To adapt the LRSD to the SR problem, we propose a novel modified approximated RPCA model where the low-rank component L is a rank-1 matrix and the sparse matrix S has a tree-regularized block structure.

As discussed before, the main limitation of using the sparse coding-based algorithms for video SR is the high computational cost associated with the super-resolving frames individually. Here, we propose a novel approach that alleviates the high computational cost. Our method obtains greater visual quality while achieving significant reduction of the number of floating point operations.

We propose to super-resolve the LR video in GOPs of F frames with $F = [8, 16, 24, 32, 64]$; we decompose each GOP into a low-rank component L that contains mostly the static unchanging parts of the scene and a sparse component S that contains dynamic pixels, changes in the scene, and possible noise. Then each obtained L and S image

for the frames in the GOP are upsampled separately using the sparse coding method described in the previous section. Notice that since we perform the SR on a low-rank component that is obtained by decomposing a GOP, we implicitly incorporate temporal information into our SR approach. Another advantage of this method is that, since the sparse component S is expected to contain very few non-zero blocks of pixels, the upsampling for each sparse image can be performed with several orders of magnitude faster than that of a non-sparse image. Therefore, the spatio-temporal information in the GOP are fully exploited without having to calculate any block matching, complex registration, or relying on motion vectors calculated by the video encoder. Then the shared information between the images in the GOP that is contained in the matrix L is upsampled only once – again providing time savings – as opposed to having to perform the upsampling for each frame individually. This is supported by empirical evidence that we will explain later. The LRSD provides a robust motion compensation possibility for the cases where camera-induced motion is present in the video sequence. The assumption of low-rankness and sparsity itself gives a good cue for being able to describe the global motion in the scene as transformations between the low-rank images in adjacent frames. We find that in videos containing camera-induced motion, our method performs better than the state-of-the-art alternatives.

3.1. LRSD for SR Problem

Given a set of frames in a GOP of N frames $I = \{I_1, I_2, \dots, I_n\}$, we can form the matrix $A \in \mathbb{R}^{m \times n}$ by stacking the frames in I as columns in the matrix A. The problem of finding a low-rank matrix L and a sparse matrix S such that $A = L + S$ has been extensively studied in the literature [2], [29], [18], [9], [6]. In [6], the authors propose a modified approximated RPCA where they solve a 3-term decomposition problem. We are interested in decomposing the matrix A into 2 terms L and S as

$$\min_{rank(L) \leq r, S, \tau} \|A \circ \tau - L - S\|_F^2 + \lambda \psi(S) \qquad (7)$$

where we have strictly set $rank(L) \leq r \leq rank(A)$. $\|\cdot\|_F$ is the Frobenius norm of a matrix defined as $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$; λ is a scalar that controls the amount of data in S. We find that setting it to $\lambda = 1/\sqrt{max(m,n)}$ works well for our experimental data. τ stands for some transformation describing the global motion induced by camera motion (e.g. 2D affine transformations, or 3D projective transformations).

The matrix S contains noise and sparse components. Similar to [6] we use a tree-structured sparse component since it better describes the spatial connectivity of the pixels in the sparse matrix. The scene in a frame can be described using a tree structure by subdivision where each child node is a subset of its parent node and the nodes of the same depth level do not overlap. Denote $\mathcal{G}$ as a set of groups from the power set of the index set $\{1, \dots, m\}$, with each group $G \in \mathcal{G}$ containing a subset of these indices. The aforementioned tree-structured groups used in this paper are formally defined as follows: A set of groups $\mathcal{G}$ is said to be *tree-structured* in $\{1, \dots, m\}$ if $\mathcal{G} = \{\dots, G_1^i, G_2^i, \dots, G_{b_i}^i, \dots\}$ where $i = 0, 1, 2, \dots, d$, d is the depth of the tree, $b_0 = 1$ and $G_1^0 = \{1, 2, \dots, m\}$, $b_d = m$ and correspondingly $\{G_j^d\}_{j=1}^m$ are singleton groups. Let G_j^i be the parent node of a node $G_{j'}^{i+1}$ in the tree, we have $G_{j'}^{i+1} \subseteq G_j^i$. We also have $G_j^i \cap G_k^i = \emptyset, \forall i = 1, \dots, d, j \neq k, 1 \leq j, k \leq b_i$. Similar group structures are also considered in [6], [13]. With the above notation, a general tree-structured sparsity-inducing norm can be written as

$$\psi(S) = \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|S_{G_j^i}\|_{2,1}, \qquad (8)$$

where $S_{G_j^i}$ is a vector with entries equal to those of S for the indices in G_j^i and 0 otherwise. w_j^i are positive weights for groups G_j^i chosen as $w_j^i = 1/max(A_{G_j^i})$ to enforce illumination invariance in the regularization scheme across patches. The regularizer $\psi(\cdot)$ on S is chosen to be $\|\cdot\|_{2,1}$. $\ell_{2,1}$-norm is a group sparsity inducing norm that acts in a tree-structured which involves a hierarchical partition of the m variables in S into groups.

The optimization problem (7) is solved via an alternating minimization strategy described in [6]. First an initialization of τ is found, by pre-aligning all the frames in the GOP to the middle frame. Then τ is linearized via the robust multiresolution method proposed in [16], [18]. Then the function is minimized for L and S separately until convergence as

$$L^t = \underset{rank(L) \leq r}{\arg\min} \|A \circ \tau - L - S^{t-1}\|_F^2 \qquad (9)$$

$$S^t = \underset{S}{\arg\min} \|A \circ \tau - L^t - S\|_F^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{b_i} w_j^i \|S_{G_j^i}\|_{2,1} \qquad (10)$$

Both these subproblems have non-convex constraints. Their global solutions L^t and S^t exist. In particular, the two subproblems can be solved by updating L^t via singular value hard thresholding of $A - S^{t-1}$ [30], and updating S^t via our structured-sparsity inducing norms with a soft-thresholding with λ. The penalty term in (10) assures the structured-sparsity of S w.r.t. the defined tree-structured groups.

Using the LRSD method we propose the VSRGOP algorithm shorthand for *Video Super Resolution using Groups of Pictures*. The parameters that we need to set for this algorithm are: number of atoms of the dictionaries, patch

Algorithm 1 VSRGOP Algorithm

Input: LR frames of the GOP
Output: HR frames of the GOP
Learning phase: Construct of the bilateral dictionaries D_h and D_l following the strategy by [26]. (This phase can be performed in advance and use D_h and D_l as inputs of the algorithm.)
Testing phase:

1) Estimate the LRSD of matrix A, while estimating the camera motion as $A \circ \tau \approx L + S$, where $rank(L) = 1$, S is block-sparse, and τ is the transformation parameter.
2) Construct a HR version of the frame corresponding to background frame using the SISR algorithm described in Section 2.1.
3) For all the frames in the GOP $(1, 2, \ldots, N)$ construct a HR version of the frames corresponding to the columns S using the SISR algorithm.
4) Reconstruct the SR version of the GOP with the HR background and HR foreground frames, applying the inverse transformation.

size, number of frames in GOP, the overlap size of patches, regularization parameter, and scale factor. Algorithm 1 describes VSRGOP steps in detail. Following the strategy in [26], in steps 2 and 3 of Algorithm 1 we use a high-pass filtering in order to extract local features that correspond to the high-frequency content. Also, a back-projection step is performed as part of both these steps. Where the back-projection is used in our tests we refer to it as VSRGOP + BP. In step 4 the HR backgorund and HR foreground frames are simply added to create the SR video.

4. Experiments

In this section we show a comparative study of the performance of the proposed algorithm for video and single image SR. We first demonstrate the SR results obtained by applying our method on video sequences from our test databases. Then we show that our method can be successfully applied to the SISR problem despite being a video SR algorithm by nature. Finally we move on to discuss how various influential factors for the proposed algorithm affect the global reconstruction, as well as the computational complexity. For video super-resolution we use the following datasets: **BBC**[1], **Ultra Video Group (UVG)**[2], and **SJTU**[3] [22]. These three datasets comprise of 27 videos of 10 seconds each at 60fps. For our tests we use all the frames in the videos. Since by default we choose GOP size of 8 frames, we report average results for an 8-frame GOP where applicable. For single image super-resolution we use the publicly available **Set5**[4] and **Set14**[5] datasets. Our algorithm is implemented in MATLAB and run on a Core i7-4770 CPU @3.40GHz (single core) and 32GB of RAM. We compare our method against state-of-the-art in sparse coding SR methods, namely Kato *et al.* [14], Yang *et al.* [26], and a state-of-the-art deep learning approach by Dong *et al.* [3], as well as the baseline Bicubic interpolation. We set the parameters of our algorithm for these experiments as: The dictionaries D_h and D_l are learned using 100,000 patches extracted from 57 HR natural images. The number of atoms in the dictionary is set to 512. Scale factors 2 and 4 are used. Patch size is 10, regularization parameter μ 0.15, and tolerance 0.05.

Following previous works, for our video SR experiments, we only consider the luminance channel in YCbCr color space, as humans are more sensitive to luminance changes. The chroma components of the original video are interpolated using plain Bicubic interpolation. The evaluations for the Kato *et al.* [14], and the Yang *et al.* [26] models are calculated based on the MATLAB code and models provided by their respective authors. We have provided supplementary material for all our tests, that includes the qualitative results. Please find the supplementary material available online here `https://goo.gl/SKkG9V`. Code for our algorithm will be publicly available online upon acceptance of the paper.

4.1. Qualitative Evaluation

We later demonstrate that our method is able to obtain high image quality metric values, however, the final judge for the image quality is the human viewer. It has been observed that although some methods generate visually appealing images, their Peak Signal-to-Noise (PSNR) values could be subjectively lower. Hence, the PSNR alone is not a reliable criterion for visual image quality.

To make a visual comparison between our model with other sparse-based methods, we super-resolve a GOP of 8 frames (the first 8 frames of a video) from all our test videos. We then compare the middle frame of the GOP with the corresponding SR image obtained by other algorithms. You can see the results of super-resolving a GOP of 8 frames from 1080p to 4K UHD with an upscaling factor 2 in Figure 1. Our algorithm is able to handle camera-induced motion in the background of the sequence well.

In Figure 2 we demonstrate a comparison between our method and four other methods. Here, a sequence has been super-resolved from 480×270 to 1080p with an upscaling

[1]The BBC has produced and made available the BBC video sequences for use under the Creative Commons Attribution-NonCommercial 3.0 licence.

[2]These sequences and all intellectual property rights therein remain the property of Digiturk. These videos may be used according to Creative Commons Attribution-NonCommercial 3.0 Unported `http://creativecommons.org/licenses/by-nc/3.0/deed.en_US`. The dataset can be obtained from: `http://ultravideo.cs.tut.fi/`

[3]SJTU 4K Video Sequences: `http://medialab.sjtu.edu.cn/web4k/index.html`

[4]`http://www.ifp.illinois.edu/~dingliu2/iccv15/html/SRdemoFrame_set5.html`

[5]`http://www.ifp.illinois.edu/~dingliu2/iccv15/html/SRdemoFrame_set14.html`

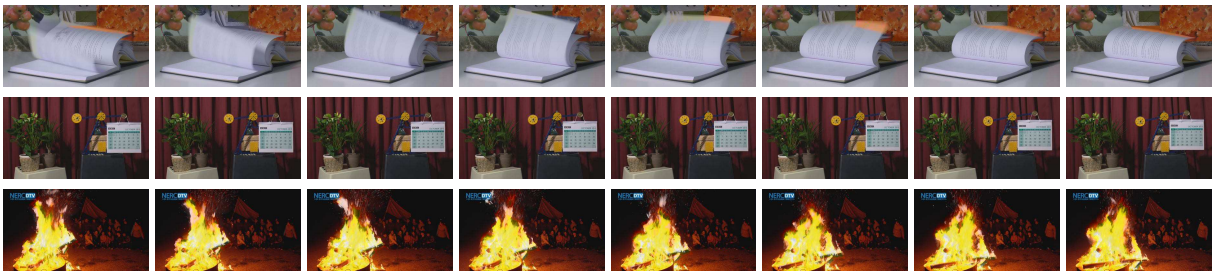

Figure 1: A GOP of 8 frames in Book, CalendarAndPlants, and CampfireParty sequences up-sampled with upscaling factor 3 (1080p to 4K UHD) with the VSRGOP + BP. Please refer to the supplementary material (available online `https://goo.gl/SKkG9V`) for full-size images.

factor 4. A cropped region of the image is shown that contains edges of printed fonts, as well as smooth texture and shading. While VSRGOP obtains better results than Bicubic and Yang [26], our method plus the Back-Projection (VSRGOP + BP) obtains higher visual reconstruction as well as better PSNR. The results in Kato + BP [14] tend to have grid-like and jagged artifacts.

Figure 3 shows more results for super-resolving sequences from 480×270 resolution to 1080p. In general our method is able to produce better texture, edge, and smooth-shaded region definitions for all the test videos; yet at the same time, the PSNR values of our results are the highest among competitors. While Bicubic interpolation produces overly smooth and watercolor-like images, our VSRGOP + BP is able to recreate both high-frequency and low-frequency components in the images. Kato + BP [14] is able to hallucinate the high-frequency content very well, however, it fails to produce visually appealing results on smoother regions. Moreover, the ringing and jagged artifacts produced by Kato + BP can be seen in the first three examples (HoneyBee, Jockey, and ParkAndBuildings sequences).

Visually our VSRGOP + BP method produces better results in general. The obtained PSNR values for our multi-frame algorithm demonstrate superior performance as well. The advantage of using bilateral dictionaries compared with the unilateral dictionaries suggested by [14] is corroborated with our empirical results. Moreover, the visual results show that our multi-frame strategy outperforms the single-image algorithm in [26] and the multi-frame algorithm in [14]. As we will discuss later, the advantage of our method not only limited to higher qualitative performance, but also it achieves this with significant reduction of computational cost.

4.2. Quantitative Evaluation

In this section we analyze the proposed method's performance with PSNR image quality metric. Also, we compare the time consumption of our algorithm against state-of-the-art sparse-based SR methods.

Table 1 shows the mean PSNR values for super-resolving all the frames in each of our test sequences individually. On average our algorithm outperforms contenders for the SR problem. Our method provides between 0.77dB to 3.72dB improvement over its sparse-based predecessor, and between 0.52dB to 0.81dB improvement over the state-of-the-art sparse-based SR method. In Table 2 we show an average time consumption comparison between our method and its predecessor sparse-based method [26] and state-of-the-art sparse-based method [14], for processing a 600-frame sequence. Our method is between $1.3\times$ to $1.6\times$ faster than its sparse-based predecessor and $271.1\times$ to $424.6\times$ faster than the state-of-the-art sparse-based SR method.

Recently, deep learning algorithms have had a great success in the SR problem. We have selected the best published method SRCNN [3] with the 9-5-5 architecture trained on ImageNet dataset, and report its results in Table 3. Here an upscaling factor 4 is used. Our method outperforms SRCNN by 2.18dB. However, the advantage of deep learning based methods is that they can be used in real-time processing. Although for applications such as medical imaging, where exact reconstruction is vitally important our method offers to be a better alternative.

5. Conclusions

In this paper we introduced a new sparsity-based video super-resolution method, that exploits the spatio-temporal information of the video sequence by a low-rank and sparse decomposition algorithm. Our method builds upon sparse representations in terms of coupled dictionaries jointly trained from high- and low-resolution image patch pairs. Our low-rank and sparse decomposition provides significant reductions in computation cost, while increasing the visual and quantitative quality of the reconstruction results by exploiting the spatio-temporal information that can be shared among adjacent frames of a video. Extensive ex-

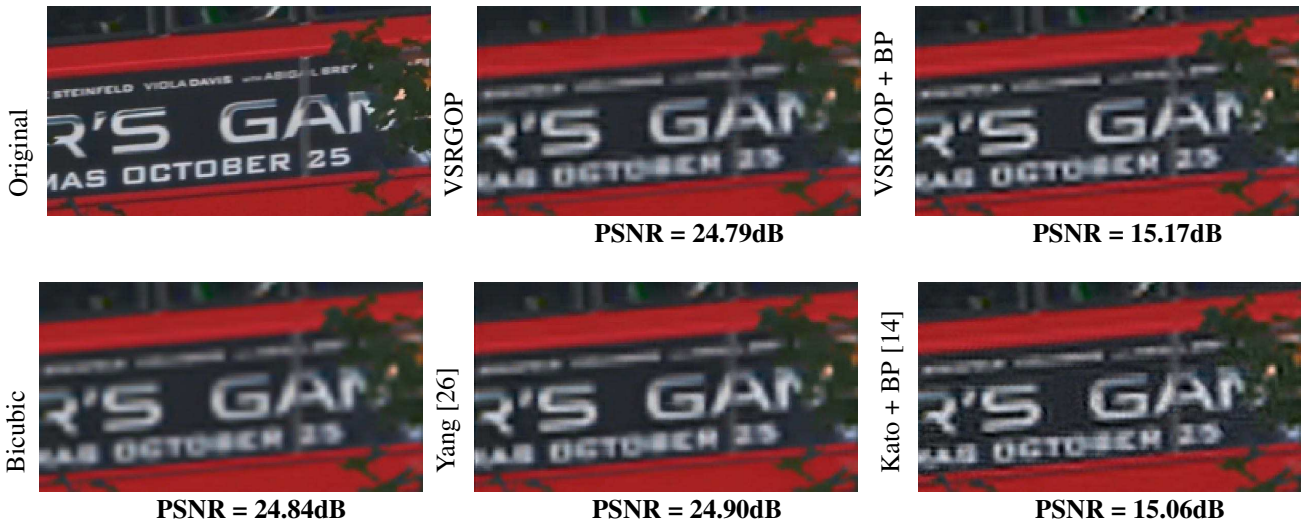

Figure 2: Qualitative comparison for up-sampling the frame 2 of Vehicles sequence from 480×270 to 1080p using different methods. Please refer to the supplementary material (available online `https://goo.gl/SKkG9V`) for full images. For each sequence a crop of the image, as well as its respective full-image PSNR is shown.

Table 1: Mean PSNR for up-sampling from 1080p to 4K UHD with upscaling factor 2, and from 480×270 to 1080p with upscaling factor 4 for all the frames in the sequences of 3 datasets. Our method provides between 0.77dB to 3.72dB improvement over its sparse-based predecessor, and between 0.52dB to 0.81dB improvement over the state-of-the-art sparse-based SR method.

1080p to 4K UHD

	VSRGOP	VSRGOP + BP	Kato [14]	Kato + BP [14]	Yang [26]	Bicubic
mean	37.41	**39.95**	31.61	39.43	36.23	39.29

480×270 to 1080p

	VSRGOP	VSRGOP + BP	Kato [14]	Kato + BP [14]	Yang [26]	Bicubic
mean	31.54	**32.32**	25.16	31.51	31.55	31.72

Table 2: Average time consumption comparison between our method and its predecessor sparse-based method [26] and state-of-the-art sparse-based method [14], for processing 1 frame. Our method is between 1.3× to 1.6× faster than its sparse-based predecessor and 271.1× to 424.6× faster than the state-of-the-art sparse-based SR method.

1080p to 4K UHD

	VSRGOP + BP	Yang [26]	Kato + BP [14]
time (h:mm:ss.s)	**0:08:20.9**	0:10:32.4	58:57:5.5

480×270 to 1080p

	VSRGOP + BP	Yang [26]	Kato + BP [14]
time (h:mm:ss.s)	**0:01:32.9**	0:02:30.6	6:59:43.0

perimental evaluation on 3 video datasets indicate the efficacy and effectiveness of the proposed algorithm in video super-resolution for HD and UHD content. Furthermore, we demonstrated the efficacy of our method for the single-image super-resolution problem, and showed that it can be successfully applied to single images, yet at the same time providing better reconstruction quality as well as less computation time. In future, we will investigate techniques to obtain real-time performances with our VSRGOP + BP method.

Table 3: Comparison with state-of-the-art Super-Resolution method with a Deep Learning approach SRCNN 9-5-5 [3] trained on ImageNet dataset, using an upscaling factor 4.

	SRCNN [19]	VSRGOP + BP
Bosphorus	37.53	**45.21**
ReadySetGo	33.69	**38.38**
Beauty	**39.48**	35.55
YachtRide	33.17	**42.11**
ShakeNDry	36.68	**39.48**
HoneyBee	**40.51**	38.23
Jockey	**41.55**	38.98
mean	37.52	**39.70**

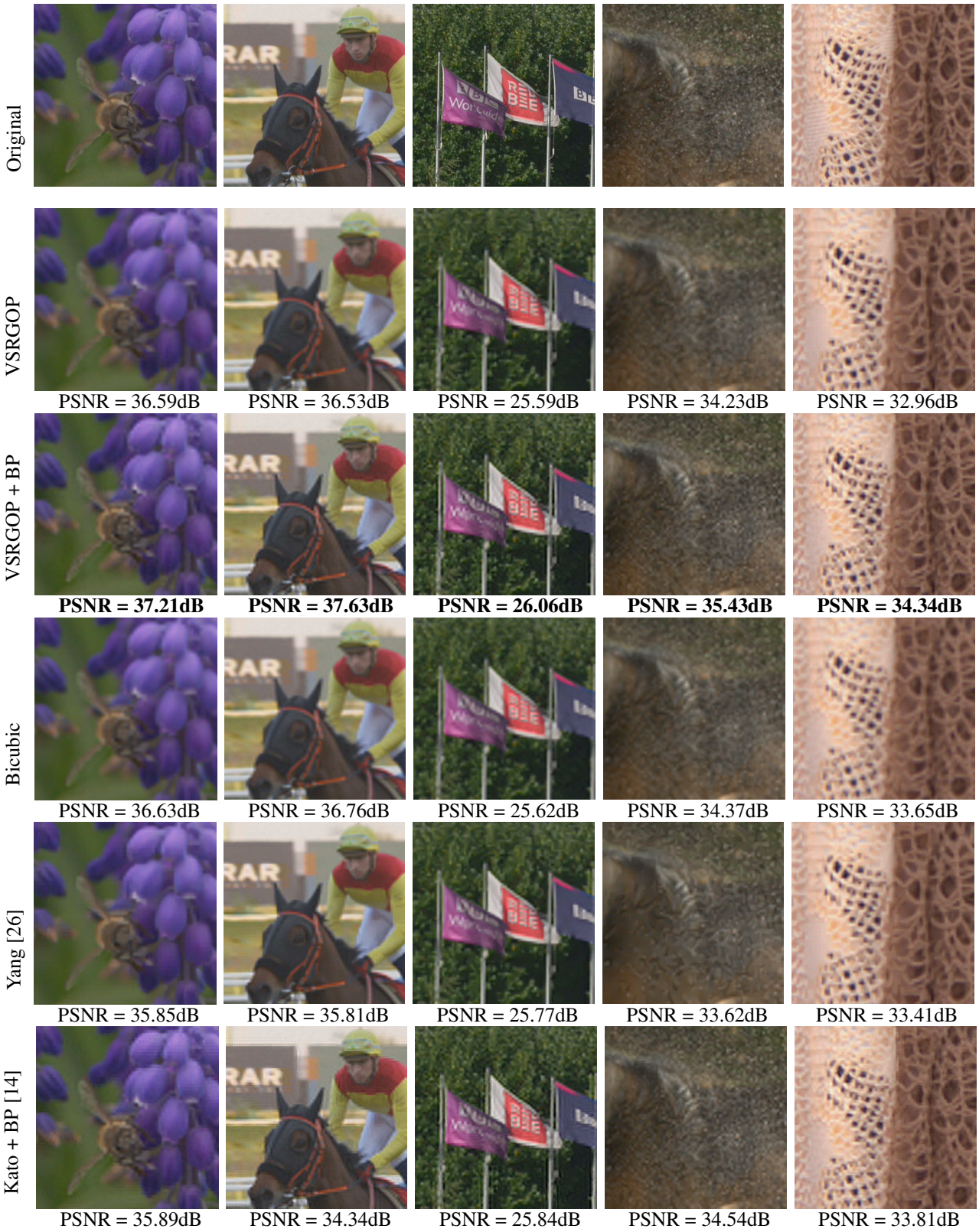

Figure 3: Qualitative comparison for up-sampling sequences from 480×270 to 1080p using different methods. Please refer to the supplementary material (available online `https://goo.gl/SKkG9V`) for full images. For each sequence a crop of the image, as well as its respective full-image PSNR is shown.

References

[1] S. Borman and R. L. Stevenson. Super-resolution from image sequences-a review. In *Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on*, pages 374–378. IEEE, 1998.

[2] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[4] W. Dong, L. Zhang, G. Shi, and X. Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[5] D. L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[6] S. E. Ebadi and E. Izquierdo. Foreground segmentation via dynamic tree-structured sparse rpca. In *European Conference on Computer Vision*, pages 314–329. Springer, 2016.

[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[8] M. Elad. Sparse and redundant representations: From theory to applications in signal and image processing. 2010.

[9] S. Erfanian Ebadi, V. Guerra Ones, and E. Izquierdo. Efficient background subtraction with low-rank and sparse matrix decomposition. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4863–4867, Sept 2015.

[10] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[11] T. Goto, T. Fukuoka, F. Nagashima, S. Hirano, and M. Sakurai. Super-resolution system for 4k-hdtv. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4453–4458. IEEE, 2014.

[12] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE transactions on image processing*, 12(5):597–606, 2003.

[13] K. Jia, T.-H. Chan, and Y. Ma. Robust and practical face recognition via structured sparsity. In *Computer Vision–ECCV 2012*, pages 331–344. Springer, 2012.

[14] T. Kato, H. Hino, and N. Murata. Multi-frame image super resolution based on sparse coding. *Neural Networks*, 66:64–78, 2015.

[15] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19:801, 2007.

[16] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of visual communication and image representation*, 6(4):348–365, 1995.

[17] C. Osendorfer, H. Soyer, and P. Van Der Smagt. Image super-resolution with fast approximate convolutional sparse coding. In *International Conference on Neural Information Processing*, pages 250–257. Springer, 2014.

[18] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012.

[19] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

[20] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. ORegan, and D. Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 9–16. Springer, 2013.

[21] M. Shimizu and M. Okutomi. Multi-parameter simultaneous estimation on area-based matching. *International Journal of Computer Vision*, 67(3):327–342, 2006.

[22] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia. The SJTU 4K video sequence dataset. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 34–35. IEEE, 2013.

[23] M. Thornton, P. M. Atkinson, and D. Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3):473–491, 2006.

[24] Z. Wang, J. Yang, H. Zhang, Z. Wang, Y. Yang, D. Liu, and T. S. Huang. *Sparse Coding and its Applications in Computer Vision*. World Scientific, 2015.

[25] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014.

[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[27] L. Zhang, H. Zhang, H. Shen, and P. Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010.

[28] Z. Zhang and V. Sze. Fast: Free adaptive super-resolution via transfer for compressed videos. *arXiv preprint arXiv:1603.08968*, 2016.

[29] T. Zhou and D. Tao. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In *International conference on machine learning*. Omnipress, 2011.

[30] T. Zhou and D. Tao. GoDec: Randomized low-rank and sparse matrix decomposition in noisy case. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 33–40. ACM, June 2011.