Sparse Representation-Based Multiple Frame Video Super-Resolution

Qiqin Dai, Seunghwan Yoo, Armin Kappeler, and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*— In this paper, we propose two multiple-frame super-resolution (SR) algorithms based on dictionary learning (DL) and motion estimation. First, we adopt the use of video bilevel DL, which has been used for single-frame SR. It is extended to multiple frames by using motion estimation with sub-pixel accuracy. We propose a batch and a temporally recursive multi-frame SR algorithm, which improves over single-frame SR. Finally, we propose a novel DL algorithm utilizing consecutive video frames, rather than still images or individual video frames, which further improves the performance of the video SR algorithms. Extensive experimental comparisons with the state-of-the-art SR algorithms verify the effectiveness of our proposed multiple-frame video SR approach.**

***Index Terms*— Video super-resolution, dictionary learning, sparse coding, optical flow, motion estimation.**

I. Introduction

VIDEO super-resolution, namely estimating the high-resolution (HR) frames from low-resolution (LR) input sequences, has become one of the fundamental problems in image and video processing and has been extensively studied for decades. With the popularity of high-definition display devices, such as High-definition television (HDTV), or even Ultra-high-definition television (UHDTV), on the market, there is an avid demand for transferring LR videos into HR videos so that they are displayed on high resolution TV screens.

Figure 1 shows the degradation model relating the HR sequence to the LR sequence which is the input to the SR algorithms. The HR frames I_k^h are of size $LN_1 \times LN_2$ and the degraded LR frames $\tilde{I}_k^l$ are of size $N_1 \times N2$, where L represents the down-sampling factor. The original multiple HR frames are related through warping based on the motion fields. The HR frames are smoothed with a blur kernel, down-sampled and contaminated by additive Gaussian noise to generate the observed LR frames. The degradation model of the k^{th} frame is therefore given by

$$\tilde{I}_k^l = DBI_k^h + \epsilon_k, \tag{1}$$

Manuscript received March 14, 2016; revised September 21, 2016 and October 22, 2016; accepted November 17, 2016. Date of publication November 21, 2016; date of current version December 12, 2016. This work was supported by the Samsung DMC Research and Development Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter Tay.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: qiqindai2012@u.northwestern.edu; seunghwanyoo2013@u.northwestern.edu; arminkappeler2011@u.northwestern.edu; aggk@eecs.northwestern.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2631339

where I_k^h and $\tilde{I}_k^l$ are the HR and LR frames, respectively, written in lexicographical notation as vectors, B represents the blur matrix, D is the down-sampling matrix and ϵ_k represents the Gaussian noise vector. Although Equation (1) provides the relationship between the k^{th} HR and LR frames, we can find the relationship between any two frames $\tilde{I}_i^l$ and I_k^h via the motion model. In that sense, Equation (1) can be extended to

$$\tilde{I}_k^l = DBC(d_{i,k})I_i^h + \epsilon_{i,k}, \tag{2}$$

where $C(d_{i,k})$ is the warping matrix generated by the motion vectors $d_{i,k}$, mapping frame i into frame k, and $\epsilon_{i,k}$ captures both the mis-registration error and the Gaussian noise. For $i = k$, Equation (2) turns into Equation (1), since $C(d_{i,k})$ becomes the identity operator. Equation (2) provides the additional observations for the LR frame $\tilde{I}_k^l$, for various values of $i \neq k$. The objective of the multiple frame SR algorithm is to operate on the observed multiple LR frames $\tilde{I}_k^l$ provided by Equation (2) for various values of i and obtain an estimate of the HR frame I_k^h.

SR techniques have been extensively studied in the literature. Detailed literature reviews of this topic can be found in [1]–[4]. With one class of approaches multiple observations are used in increasing the resolution of one image, as described in Equation (2). Such multiple observations can be due to global sub-pixel motion between the camera and the scene or due to the dynamics of the scene, i.e., the sub-pixel motion of individual objects in the scene. In the former case either multiple still cameras or one still camera which changes its position are used. The motion vectors $d_{i,k}$ in this case are constant for the whole frame but they typically represent more complicated motions than simple translation, such as rotation (e.g., [5]). In the latter case, one video camera is typically used, which might move as well resulting in global shifts amongst frames, but the additional information about the frame to be super-resolved is provided by the motion of objects, as is depicted in the neighboring frames. This is the case of video SR considered in this paper, in which case the motion vectors $d_{i,k}$ in Equation (2) depend on the pixel location. In designing video SR algorithms, the degradation matrix B is either considered known or is estimated from the data, along with the motion vectors $d_{i,k}$, the original HR frames, and the noise level, either simultaneously [6]–[9], or sequentially [10], [11]. Recently, Hung *et al.* [12] proposed a method based on codebooks derived from key-frames and achieved good SR performance on compressed videos. Zhou *et al.* [13] proposed to retrieve high-frequency details from complementary multi-frames by non-uniform interpolation, depending on registered

Fig. 1. Degradation and SR model. The original HR video frames are related to each other by motion fields. The HR sequence is then degraded to generate the observed sequence according to Equation (1). Our proposed dictionary based video SR algorithm estimates the HR sequence, as well as the motion field.

LR frames with sub-pixel accuracy. They further improved the SR performance in [14] when the number of LR inputs is small by taking advantage of nonlocal self-similarity to fit local surfaces. Liu and Sun [6], [15] proposed to estimate the blur kernel, noise level, motion field and HR frames jointly by Maximum-a-Posteriori (MAP) inference. Ma *et al.* [16] presented an algorithm that extended the same idea to handle motion blur. Liao *et al.* [17] proposed to apply a traditional multi-frame SR method [18] to obtain SR drafts with different motion estimation parameters, and then to combine the SR drafts through a deep convolutional neural network (CNN).

Another class of SR approaches is represented by single frame SR, where a single observation is used to increase the resolution of one frame. Due to the limited LR information, example-based or learning approaches, such as dictionary learning (DL) approaches [19]–[24], showed recently promising single frame SR performance. These methods learn the non-linear mapping from an LR frame to the corresponding HR frame through an HR/LR training data set in the training phase and apply the learned non-linear mapping to an LR observation in the testing phase. DL approaches have also been utilized for deblurring [25] and denoising of images and image sequences [26], [27]. For the SR of a still image using dictionary techniques, typically only one observation of an LR image is utilized. The mapping from an LR to an HR image, as depicted by Equation (1) is learned during training and is captured by the structures of two coupled, LR and HR, dictionaries. No explicit use of the degradation matrix B is made during the sparse coding based reconstruction of the HR frame. Some methods [19], [20], [22], [23] might include a back-projection step, thus using matrix B, as a final refinement step. However, based on our knowledge, the first work reported in the literature on the application of DL to video SR is the work in [21]. According to it, block-based motion estimation is performed among input LR keyframes and DL is only applied for single-frame SR when the motion compensation error is larger than a threshold. The approach reported in [21] however does not provide sub-pixel precision in motion estimation and does not utilize any of the advanced DL techniques. Later the work in [28] utilized the semi-coupled DL technique [24] to super-resolve each LR frame individually and performed a weighted fusion of the super-resolved HR frames by nonlocal similarity match [29]. However, the nonlocal similarity match is also block-based and do not fully exploit the sub-pixel shift information. Also the initial HR frames estimation by the semi-coupled DL and patch similarity match are performed sequentially, so the reconstruction error by the semi-coupled DL SR will not be minimized in the later SR steps.

In this paper, we propose an approach for video SR, according to which multiple LR observations of an HR video frame are utilized according to Equation (2) for both designing coupled dictionaries connecting the sparse representation of LR and HR image frames, as well as for reconstructing an HR frame. We borrow two ideas from single frame SR, namely, bilevel coupled dictionary [19], [20], [22], [23] and multiple-dictionaries [24], [30], to be explained later. We incorporate them into a multiple frame SR framework, according to which the non-redundant information contained in LR frames which are typically related by sub-pixel shifts among them is utilized to generate an HR frame. We propose a multiple dictionary multiple frame video SR algorithm utilizing sub-pixel accurate motion estimation. With our proposed SR approach, the estimated optical flow is utilized to obtain multiple frame high accuracy registration and an HR frame is reconstructed from multiple LR frames. The moving parts in a scene can be super-resolved by the sub-pixel shift information while for the stationary parts, the SNR improves due to the multiple observation of the same scene. As far as registration error is concerned, we address it by adapting the weight parameter that enforces the similarity of multiple LR observations, so that our proposed algorithm has the ability to move between single frame bilevel coupled dictionary [22], [23] SR approach and multi-frame SR approach, and perform at least as good as any of these two approaches.

The multiple frame SR performance is further improved by training dictionaries from consecutive video frames. Most dictionary learning techniques [19]–[21], [24]–[27], [31], [32] use still images or individual video frames to train the dictionaries. However, this causes an inconsistency in multiple frame SR

since we are super-resolving videos while the dictionaries are trained from still images. The proposed training from videos incorporates temporal information into the dictionaries, and makes the training and testing phases consistent. Although as a result the training phase becomes more complicated, the testing phase remains the same.

Because our proposed SR method is a learning method, we do not explicitly model and estimate the blur kernel (matrix B in Equation (2)) in the sparse coding reconstruction of the HR frame in the SR testing phase. Clearly, in the training phase, the HR and LR patch pairs carry the blur information which will be incorporated into the resulting trained HR and LR dictionary pairs. To handle the potential mismatch of the blur kernel in training and testing phase, an idea similar to the one in [25] can be applied. Multiple blurred and downsampled versions of the same HR video will be used to train LH/HR dictionary pairs (assume there are N such pairs). All such dictionary pairs will then be used to reconstruct N HR videos from one LR observation during testing. A decision criterion can be adopted to decide which reconstruction is the preferred one. For example, from the N HR reconstructions N LR observations can be generated using the N different blur kernels. All these N LR generated observations will be compared against the actual observation and the one with the smallest error (say the k^{th} one) will determined which HR reconstruction (the k^{th} one) will be chosen. This way a blur identification is indirectly performed.

Based on the results reported in the literature [1], [2], the quality of the multiple-frame SR critically depends on the accuracy of the motion estimates. The two important characteristics of the motion field are that 1) it should have sub-pixel accuracy and 2) it should be dense. There is a plethora of techniques in the literature for estimating a dense motion field [33], [34]. Optical flow techniques assume that the optical flow is preserved over time. This information is utilized to form the optical flow equation connecting spatial and temporal gradients. More recent optical flow algorithms [35], [36] use a variational coarse-to-fine framework to handle large displacements.

In-depth and comprehensive experiments demonstrate that our proposed SR framework outperforms the state-of-the-art super resolution frameworks, such as, NE+NNLS [37], NE+LLE [38], ANR [39], SR-CNN [40], Enhancer [41] and Bayesian [6] on UHD (4K) sequences.

Our main contributions lie in the following three aspects:

- We extended the bilevel coupled dictionary learning based single frame SR method [22], [23] from a single dictionary to multiple dictionaries (Section II-A).
- We extended the bilevel coupled dictionary learning based single frame SR method [22], [23] from a single frame to multiple frames by developing two approaches: a batch approach and a recursive approach (Section II-B II-C).
- We proposed and developed an approach for training the dictionaries from consecutive video frames instead from individual still images (Section III).

This paper is an extension of our previous work [42]. The extension and additional contributions lie in the following aspects:

- We proposed a recursive multiple frame video super-resolution algorithm in Section II-C and the corresponding algorithm for training dictionaries from videos in Section III-B.
- We utilized an adaptive weight parameter which depends on the mis-registration error (Equation 9).
- We introduced an iteration between motion estimation and HR frame estimation for both the batch approach and recursive approach.
- We illustrated a detailed algorithm for training dictionaries from videos.
- We introduced multiple SR steps for large upscale factors.
- We included more comprehensive experimental results.

The rest of the paper is organized as follows. Section II presents our proposed dictionary based multiple-frame SR framework. Section III illustrates a novel dictionary training strategy, that of training from videos. Section IV provides experimental results, and finally conclusions are drawn in Section V.

II. Dictionary Based Multiple-Frame Super-Resolution Approach

Given the LR image sequence $\{I_1^l, \ldots, I_k^l, \ldots\}$, the goal of SR is to estimate the HR sequence $\{I_1^h, \ldots, I_k^h, \ldots\}$. Since each frame is primarily correlated with its neighbors and to also reduce computation, when we are super-resolving the k^{th} frame I_k^h, only the adjacent $(M+N)$ frames $I_{k-M}^l, \ldots, I_{k+N}^l$ are used. Clearly when $N = 0$, causal processing is performed.

In this section, we introduce two approaches to find the sparse representation of an LR patch y_k by incorporating motion information from the neighboring frames, namely, the batch approach and the temporally recursive approach. The core idea of these two approaches originates from the fact that image registration through motion compensation provides multiple observations of the same scene, enabling the SR algorithm to take advantage of the details lost in the k^{th} frame but present in past or future frames. For simplicity the super-resolution framework will be derived for gray-scale images; however, it can be easily extended to handle color image.

A. Multiple Bilevel Dictionary Learning

The first task we address is the coupled learning of high and low resolution dictionaries over a large database of training HR images. Each HR image I_j^h in the training database is degraded by blur and noise and down-sampled, according to Equation (1), resulting in the corresponding LR image $\tilde{I}_j^l$. Each LR image $\tilde{I}_j^l$ is up-sampled using bicubic interpolation to become I_j^l, so that I_j^h and I_j^l have the same size. In the remaining part of the paper, when dealing with LR frames, we refer to I_j^l, which is the bicubically interpolated LR frame $\tilde{I}_j^l$. I_j^h and I_j^l are then divided into patches of size $W \times W$; the corresponding i^{th} patches out of L total patches are lexicographically ordered to form vectors x^i and y^i, respectively. In the dictionary learning phase, we aim at finding HR and LR dictionaries D^h and D^l such that the sparse representation of any HR patch over D^h is identical to that of the corresponding

LR patch over D^l. In order to do so, Yang *et al.* [22], [23] formulated the following bilevel optimization problem

$$\min_{D^h, D^l} \sum_{i=1}^{L} \left\| x^i - D^h z^i \right\|_2^2$$
$$\text{s.t. } z^i = \arg\min_{\alpha^i} \left\| F y^i - F D^l \alpha^i \right\|_2^2 + \lambda \left\| \alpha^i \right\|_1$$
$$\left\| D^h(:,k) \right\|_2 \le 1, \left\| D^l(:,k) \right\|_2 \le 1, \quad \forall k, \tag{3}$$

where α^i contains the sparse representation of the i^{th} HR/LR patch, $\|.\|_2$ and $\|.\|_1$ represent the l_2 and the l_1 vector norms, respectively, λ is the regularization parameter which controls the sparsity of the sparse coefficient a^i, F is a linear operator which extracts features of the LR patches, and $\left\| D^h(:,k) \right\|$ and $\left\| D^l(:,k) \right\|$ indicate the k^{th} column of matrices D^h and D^l, respectively.

In the testing phase, given an observed LR patch y, we first solve the following LASSO problem

$$z = \arg\min_{\alpha} \left\| F y - F D^l \alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1, \tag{4}$$

and then the sparse coefficient z is applied to the HR dictionary D^h to obtain the HR patch x corresponding to y, that is,

$$x = D^h z. \tag{5}$$

With the bilevel dictionary learning technique, in the training phase, when updating the sparse coefficient z^i in the so referred to as the lower level, the optimization is consistent with the optimization in the testing phase (Equation (4)), thus guaranteeing good reconstruction accuracy. Improved SR results have been reported with this bilevel formulation in [22] and [23] compared to the previous formulation in [19] and [20]. Because of the diverse structures and textures in images of different styles, using a general coupled dictionary is often not good enough to super-resolve all variations in image patches. Considering the fact that image patches, according to their appearance, can be classified into different categories (such as textures, flat regions, edges, etc.), we train a coupled dictionary for each such category. The heuristic clustering strategy in [24] is integrated in our framework. More specifically, K-Means clustering is performed on sampled LR training patches y after applying the feature filter F. Let y_c^i be the i^{th} LR training patch belonging to cluster c, which has in total L_c patches, and x_c^i its corresponding HR training patch. The coupled dictionary $(D_c^l \; D_c^h)$ is then trained on $\left\{x_c^i, y_c^i\right\}_{i=1}^{L_c}$ based on Equation (3).

After learning the C coupled dictionaries $\{(D_1^l \; D_1^h), \ldots, (D_C^l \; D_C^h)\}$, during the testing phase, for a sample LR patch y, the most appropriate dictionary c^* is determined via

$$c^* = \arg\min_{c=1\ldots C} \| O_c - F y \|_2^2, \tag{6}$$

where O_c is the centroid of the columns of the c^{th} LR dictionary. Here, the Euclidean distance between the centroid and the LR patch is used as the similarity measure. The best dictionary pair $(D_{c^*}^l \; D_{c^*}^h)$ is then used to find the HR version of y (denoted by x) by solving Equation (4).

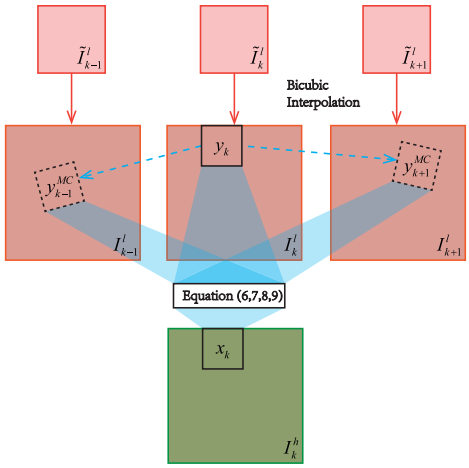

Fig. 2. Batch approach (the figure is depicted for the case $M = N = 1$).

B. A Batch Multiple Frame Video Super-Resolution Algorithm

A dictionary based batch multiple-frame video SR algorithm is shown in Fig. 2 (when $M = N = 1$). The three consecutive LR frames are shown in pink while the HR frame corresponding to the middle LR frame is depicted in green. We want to fill in the patch x_k which is the HR version of the patch y_k in the k^{th} frame, by combining information from patch y_k, the motion compensated patches $y_{k-M}^{MC}, \ldots, y_{k-1}^{MC}, y_{k+1}^{MC}, \ldots, y_{k+N}^{MC}$ and the pre-trained multiple coupled dictionaries $(D_c^l \; D_c^h)$.

With this approach, we alternate optimizing for the motion field and the HR frames I_k^h. In the first iteration, the motion field is estimated based on the LR input frames $\left\{I_{k+j}^l\right\}_{j=-M}^{j=N}$. The optical flow method in [36] is applied to obtain the motion field with sub-pixel accuracy. Then the motion compensated versions of y_k are computed according to the motion field in the past and future frames, denoted by $\left\{y_{k+j}^{MC}\right\}_{j=-M}^{j=N, j\neq 0}$. To super-resolve y_k in the k^{th} frame, the most appropriate LR dictionary indexed by c^*, out of the C possible choices, is found via Equation (6). Then the best dictionary pair $(D_{c^*}^l \; D_{c^*}^h)$ is picked to find the HR version of y_k according to

$$\min_{\substack{\alpha_k, \alpha_{k+j}^{MC}, \\ j=-M,\ldots,N, j\neq 0}} \left\| F y_k - F D_{c^*}^l \alpha_k \right\|_2^2 + \sum_{j=-M, j\neq 0}^{N} \left\| F y_{k+j}^{MC} - F D_{c^*}^l \alpha_{k+j}^{MC} \right\|_2^2 + \lambda \left(\|\alpha_k\|_1 + \sum_{j=-M, j\neq 0}^{N} \left\| \alpha_{k+j}^{MC} \right\|_1 \right) + \sum_{j=-M, j\neq 0}^{N} \gamma_j \left\| D_{c^*}^h \alpha_k - D_{c^*}^h \alpha_{k+j}^{MC} \right\|_2^2 \tag{7}$$

$$x_k = D_{c^*}^h \alpha_k, \tag{8}$$

where α_{k+j}^{MC} is the sparse representation of y_{k+j}^{MC}. The first two terms in Equation (7) ensure the fidelity to the LR

observations (similar to Equation (4)). The middle two terms are l_1 regularizers promoting the sparse representation of the LR patches by the LR dictionaries and the last term enforces the similarity of the reconstructed HR patches from past and future frames to the current frame. Only α_k is used in Equation (8) to reconstruct the HR patch in the current frame. The regularization parameter λ is chosen experimentally, while the choice of the γ_j's is described below.

After the reconstruction of the HR frames $\left\{I^h_{k+j}\right\}^{j=N}_{j=-M}$ in the first iteration, we update the motion field based on these HR frames by applying the optical flow algorithm in [36], since it typically results in a higher accuracy motion field than the one resulting by using the LR frames $\left\{I^l_{k+j}\right\}^{j=N}_{j=-M}$. We can alternate updating the motion field and the HR frames $\left\{I^h_{k+j}\right\}^{j=N}_{j=-M}$ until convergence.

An important point to be taken into account is that the desired accuracy on motion estimation will not be reached if images have a lot of aliasing. Notice that the mis-registration error between y_k and y^{MC}_{k+j}, i.e., $e(k,k+j) = \left\|y_k - y^{MC}_{k+j}\right\|_2$, is proportional to $\left\|D^h_{c^*}\alpha_k - D^h_{c^*}\alpha^{MC}_{k+j}\right\|_2$. Therefore, γ_j in Equation (7) should be small when $e(k,k+j)$ is relatively large, and vice versa, in other words they are inversely proportional. The exponential function of the mis-registration is applied here to formalize this relationship, as in [43],

$$\gamma_j = \beta_1 \cdot \exp(-\beta_2 \cdot e(k,k+j)^2), \tag{9}$$

where β_1 and β_2 are adjusted experimentally. If the registration error is large, γ_j will become small and the proposed method in Equation (7) degenerates to a single frame super-resolution method, since we weakly enforce the similarity of the reconstructed HR patches in the temporal domain.

C. A Recursive Multiple Frame Video Super-Resolution Algorithm

In this section, we propose a novel temporally recursive algorithm for dictionary-based multiple-frame video SR. By using information from already super-resolved frames in the past, the recursive method provides efficient computation, reduced storage, high quality super-resolution results and no delay in processing.

As depicted in Figure (3), with the recursive approach, unlike the batch approach, only past frames are used in order to super-resolve y_k. This way the algorithm is temporally causal therefore there is no delay by waiting for future LR frames prior to super-resolving the current one. Because neighboring frames exhibit redundant information, using HR information from previously super-resolved frames can improve the quality of the current SR frame.

Given an LR patch y_k in the k^{th} frame, the most suitable LR dictionary indexed by c^* is first found via Equation (6). Like the iteration estimation process of the HR frames and motion field in the batch approach (Section II-B), in the first iteration, the motion field is estimated by the optical flow method in [36] with sub-pixel accuracy based on the

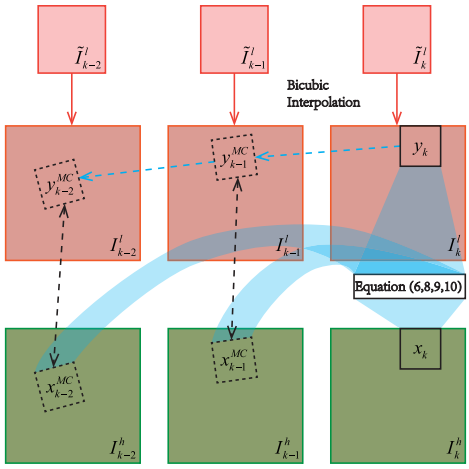

Fig. 3. The recursive approach, (the figure is depicted for the case $N = 2$).

LR frames $\left\{I^l_{k-j}\right\}^{j=N}_{j=0}$. Motion compensated versions of y_k $\left(\left\{y^{MC}_{k-j}\right\}^{j=N}_{j=1}\right)$ are then found according to the motion field. Subsequently, their corresponding HR patches $\left(\left\{x^{MC}_{k-j}\right\}^{j=N}_{j=1}\right)$ are determined by the motion field as well and substituted into the following temporally recursive model

$$\min_{\alpha_k} \left\|Fy_k - FD^l_{c^*}\alpha_k\right\|^2_2 + \lambda\left\|\alpha_k\right\|_1 + \sum_{j=1}^{N}\gamma_j\left\|D^h_{c^*}\alpha_k - x^{MC}_{k-j}\right\|^2_2 \tag{10}$$

The first term in the above equation ensures the fidelity to the data, i.e., the current LR observations, while the second term promotes the sparsity of the solution α_k. The last term enforces the similarity between the reconstructed HR patches of the current frame $(D^h_{c^*}\alpha_k)$ and the previous reconstructed HR patches $\left(\left\{x^{MC}_{k-j}\right\}^{j=N}_{j=1}\right)$. Also γ_j is selected adaptively according to Equation (9). Similarly to the batch approach, the corresponding HR patch x_k is obtained according to Equation (8). The reconstruction error will not propagate to future frames due to this adaptive weight. Assume that frame I^h_k has large reconstruction error in a certain region. Its motion compensated patches to frame $(k+1)$ will have large registration error, in which case γ_j will be small and Equation (10) will degenerate to a single frame super-resolution method. The reconstructed frame I^h_{k+1} will have smaller reconstruction error and will provide helpful HR information to reconstruct frame $(k+2)$, and so on.

Similarly to the batch approach, after the reconstruction of the HR frame I^h_k in the first iteration, a more accurate motion field can be estimated based on the HR frames $\left\{I^h_{k-j}\right\}^{j=N}_{j=0}$ by applying the optical flow algorithm in [36]. The motion field and the HR frames $\left\{I^h_{k-j}\right\}^{j=N}_{j=0}$ are updated in an alternate fashion until convergence.

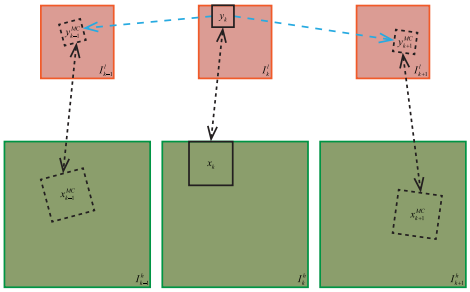

Fig. 4. Batch approach: training from consecutive frames when $M = N = 1$.

Unlike the batch approach, with the use of motion compensated HR patches $\left(\left\{x_{k-j}^{MC}\right\}_{j=1}^{j=N}\right)$ from the super-resolved previous HR frames, only the coefficients α_k of the patches in the current frame are estimated, which significantly reduces both storage and computation.

III. Training Dictionaries From Videos

Typically, for the dictionary learning process, all training patches are sampled from still images or individual video frames. This causes some inconsistency in training and testing, since clearly we are trying to super-resolve videos while the dictionaries are trained from still images. Also the optimizations for training (Equation (3)) and testing (Equation (7) or Equation (10)) are not consistent.

We therefore propose two new dictionary training algorithms based on consecutive video frames and motion estimation for both the batch and recursive approaches. Both algorithms are applied to each cluster (Section II-A) separately, so in the following equation we omit the dependency on a particular cluster for simplifying the notation.

A. Video Training for the Batch Approach

As shown in Figure (4), during training, a number of consecutive video frames from the training videos are used. In the k^{th} training video sequence of total L_s video sequences, the original HR frames, $\left\{I_{k+j}^{h}\right\}_{j=-M}^{j=N}$, are degraded to obtain the LR frames $\left\{I_{k+j}^{l}\right\}_{j=-M}^{j=N}$. Motion estimation is then performed utilizing the $(M + N + 1)$ frames to find the corresponding patches $\left\{y_{k+j}^{MC}\right\}_{j=-M}^{j=N,j\neq 0}$ $\left(\left\{x_{k+j}^{MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ of y_k (x_k) in the past and future frames. Let L_p be the number of sampled patches in each scene. The coupled dictionary $(D^l \ D^h)$ for the batch multiple-frame video SR approach is then trained on $\left\{\left\{x_k^i, y_k^i, \left\{y_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$ according to the bilevel dictionary learning in Equation (11), as shown at the bottom of this page, above.

The objective function in Equation (11) is highly nonlinear and nonconvex. Similarly to [22] and [23], we alternate optimizations over D^h, D^l and $\left(z_k^i \ \left\{z_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ while keeping the remaining of the terms fixed. When D^h and D^l are fixed, the optimization over $\left(z_k^i \ \left\{z_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ becomes a standard LASSO problem as reformulated in Equation (12), as shown at the bottom of this page.

$$\min_{D^h,D^l} \sum_{k=1}^{L_s}\sum_{i=1}^{L_p}\left\|x_k^i - D^h z_k^i\right\|_2^2$$

$$\text{s.t.} \quad \underset{j=-M,\ldots,N,j\neq 0}{z_k^i, z_{k+j}^{i\ MC}} = \underset{\substack{\alpha_k^i,\alpha_{k+j}^{i\ MC},\\ j=-M,\ldots,N,j\neq 0}}{\arg\min} \left\|Fy_k^i - FD^l\alpha_k^i\right\|_2^2 + \sum_{j=-M,j\neq 0}^{N}\left\|Fy_{k+j}^{i\ MC} - FD^l\alpha_{k+j}^{i\ MC}\right\|_2^2$$
$$+ \lambda\left(\left\|\alpha_k^i\right\|_1 + \sum_{j=-M,j\neq 0}^{N}\left\|\alpha_{k+j}^{i\ MC}\right\|_1\right) + \sum_{j=-M,j\neq 0}^{N}\gamma_j\left\|D^h\alpha_k^i - D^h\alpha_{k+j}^{i\ MC}\right\|_2^2$$
$$\left\|D^h(:,k)\right\|_2 \le 1, \left\|D^l(:,k)\right\|_2 \le 1, \quad \forall k \tag{11}$$

$$\min_{z_{k-M}^{i\ MC},\ldots,z_k^i,\ldots,z_{k+N}^{i\ MC}} \left\|\begin{bmatrix} FD^l & & & & \\ & \ddots & & & \\ & & FD^l & & \\ & & & \ddots & \\ -\gamma_{-M}D^h & & \gamma_{-M}D^h & & FD^l \\ & \ddots & \vdots & \ddots & \\ & & \gamma_N D^h & & -\gamma_N D^h \end{bmatrix}\begin{bmatrix} z_{k-M}^{i\ MC} \\ \vdots \\ z_k^i \\ \vdots \\ z_{k+N}^{i\ MC} \end{bmatrix} - \begin{bmatrix} Fy_{k-M}^{i\ MC} \\ \vdots \\ Fy_k^i \\ \vdots \\ Fy_{k+N}^{i\ MC} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}\right\|_2^2 + \lambda\left\|\begin{bmatrix} z_{k-M}^{i\ MC} \\ \vdots \\ z_k^i \\ \vdots \\ z_{k+N}^{i\ MC} \end{bmatrix}\right\|_1 \tag{12}$$

When $\left(z_k^i \ \left\{z_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ and D^h are fixed, the optimization over D^l is reduced to

$$\min_{D^l} \sum_{k=1}^{L_s}\sum_{i=1}^{L_p}\left(\left\|Fy_k^i - FD^l z_k^i\right\|_2^2 + \sum_{j=-M,j\neq 0}^{N}\left\|y_{k+j}^{i\ MC} - FD^l z_{k+j}^{i\ MC}\right\|_2^2\right)$$
$$\text{s.t. } \left\|D^l(:,k)\right\|_2 \le 1, \quad \forall k, \tag{13}$$

which is a quadratically constrained quadratic program (QCQP) [44] that can be efficiently optimized using conjugate gradient descent [45]. The l_2 norm constraint can be satisfied by simply projecting each column onto the unit ball at each iteration according to Equation (14), that is,

$$D^l(:,k) = \frac{D^l(:,k)}{\max\left(1, \left\|D^l(:,k)\right\|_2\right)}. \tag{14}$$

Finally, when we fix $\left(z_k^i \ \left\{z_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ and D^l, by collecting terms containing D^h in both upper and lower levels, the optimization over D^h becomes

$$\min_{D^h} \sum_{k=1}^{L_s}\sum_{i=1}^{L_p}\left(\left\|x_k^i - D^h z_k^i\right\|_2^2 + \sum_{j=-M,j\neq 0}^{N}\gamma_j\left\|D^h(z_k^i - z_{k+j}^{i\ MC})\right\|_2^2\right)$$
$$\text{s.t. } \left\|D^h(:,k)\right\|_2 \le 1, \quad \forall k, \tag{15}$$

which is also a QCQP [44] and can be optimized by conjugate gradient descent [45]. The projection to the unit ball becomes

$$D^h(:,k) = \frac{D^h(:,k)}{\max\left(1, \left\|D^h(:,k)\right\|_2\right)}. \tag{16}$$

Algorithm 1 summarizes the complete procedure of our coupled dictionary learning algorithm for sequential video training.

Notice that the lower level optimization of Equation (11) in the training phase is consistent with the optimization in the testing phase of multiple-frame sequential SR in Equation (7). Therefore the training and testing phases are consistent and the accuracy in sequentially reconstructing one frame from multiple frames is guaranteed.

To train multiple dictionaries, Algorithm 1 is applied to each cluster separately. Feature filter F is applied on the LR patch y_k^i to cluster each training patch set $\left\{x_k^i, y_k^i, \left\{y_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right\}$.

B. Video Training for the Recursive Approach

Similarly to Section III-A, a number of consecutive video frames are used in the training phase, as depicted in Figure (5). The original HR frames $\left\{I_{k-j}^h\right\}_{j=0}^{j=N}$, are degraded to obtain the

Algorithm 1 Coupled Dictionary Learning: Training From Video for Batch Approach

input: training patch sets $\left\{\left\{x_k^i, y_k^i, \left\{y_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$

1: Initialization: initialize $D^{l(0)}$ and $D^{h(0)}$ by Equation (3) based on $\left\{\left\{x_k^i, y_k^i\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$, $n=0$
2: **repeat**
3: Update $\left(z_k^i, \left\{z_{k+j}^{i\ MC}\right\}_{j=-M}^{j=N,j\neq 0}\right)$ according to Equation (12);
4: Update $D^{l(n+1)}$ from $D^{l(n)}$ according to Equation (13);
5: Project the columns of $D^{l(n+1)}$ onto the unit ball according to Equation (14);
6: Update $D^{h(n+1)}$ from $D^{h(n)}$ according to Equation (15);
7: Project the columns of $D^{h(n+1)}$ onto the unit ball according to Equation (16);
8: n=n+1;
9: **until** convergence

output: coupled dictionaries $D^{l(n)}$ and $D^{h(n)}$.

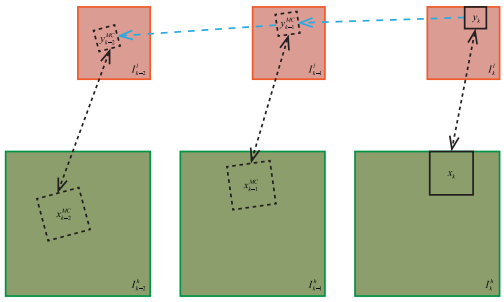

Fig. 5. Recursive approach: training from consecutive frames when $N = 2$.

LR frames $\left\{I_{k-j}^l\right\}_{j=0}^{j=N}$. The backwards corresponding patches $\left\{y_{k-j}^{i\ MC}\right\}_{j=1}^{j=N}$ $\left(\left\{x_{k-j}^{i\ MC}\right\}_{j=1}^{j=N}\right)$ to y_k^i $\left(x_k^i\right)$ are obtained by motion estimation, performed on the LR frames. Let y_k^i be the i^{th} LR training patch, x_k^i the corresponding HR training patch to (y_k^i) and $x_{k-j}^{i\ MC}$ the motion compensated patch of x_k^i in the $(k-j)^{th}$ HR frame. We then train the coupled dictionary $(D^l\ D^h)$ for the recursive multiple-frame approach based on $\left\{\left\{y_k^i, \left\{x_{k-j}^{i\ MC}\right\}_{j=1}^{j=N}, x_k^i\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$ by the following bilevel optimization

$$\min_{D^h, D^l} \sum_{k=1}^{L_s}\sum_{i=1}^{L_p}\left\|x_k^i - D^h z_k^i\right\|_2^2$$
$$\text{s.t. } z_k^i = \arg\min_{\alpha_k^i}\left\|Fy_k^i - FD^l z_k^i\right\|_2^2 + \lambda\left\|z_k^i\right\|_1 + \sum_{j=1}^{N}\gamma_j\left\|x_{k-i}^{i\ MC} - D^h z_k^i\right\|_2^2$$
$$\left\|D^h(:,k)\right\|_2 \le 1, \ \left\|D^l(:,k)\right\|_2 \le 1, \quad \forall k. \tag{17}$$

Algorithm 2 Coupled Dictionary Learning: Train From Video for Recursive Approach

input: training patch sets $\left\{\left\{y_k^i, \left\{x_{k-j}^{i\ MC}\right\}_{j=1}^{j=N}, x_k^i\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$

Initialization: initialize $D^{l(0)}$ and $D^{h(0)}$ by Equation (3) based on $\left\{\left\{y_k^i, x_k^i\right\}_{i=1}^{i=L_p}\right\}_{k=1}^{k=L_s}$, $n=0$

repeat

  Update z_k^i according to Equation (18);

  Update $D^{l(n+1)}$ from $D^{l(n)}$ according to Equation (19);

  Project the columns of $D^{l(n+1)}$ onto the unit ball according to Equation (14);

  Update $D^{h(n+1)}$ from $D^{h(n)}$ according to Equation (20);

  Project the columns of $D^{h(n+1)}$ onto the unit ball according to Equation (16);

  n=n+1;

until convergence

output: coupled dictionaries $D^{l(n)}$ and $D^{h(n)}$.

The optimization strategy from Section III-A can be applied here by alternating optimization over D^l, z_k^i, and D^l. When D^h and D^l are fixed, optimizing over z_k^i is a standard LASSO problem

$$\min_{z_k^i} \left\| \begin{bmatrix} Fy_k^i \\ \gamma_1 x_{k-1}^{i\ MC} \\ \vdots \\ \gamma_N x_{k-N}^{i\ MC} \end{bmatrix} - \begin{bmatrix} FD^l \\ \gamma_1 D^h \\ \vdots \\ \gamma_N D^h \end{bmatrix} z_k^i \right\|_2^2 + \lambda \left\| z_k^i \right\|_1. \quad (18)$$

In the next step, by fixing D^h and z_k^i, the optimization over D^l is reduced to

$$\min_{D^l} \sum_{k=1}^{L_s} \sum_{i=1}^{L_p} \left\| Fy_k^i - FD^l z_k^i \right\|_2^2 \quad \text{s.t.} \left\| D^l(:,k) \right\|_2 \le 1, \quad \forall k, \quad (19)$$

which can be carried out by conjugate gradient descent [45] followed by projection onto the unit ball (Equation (14)).

Finally, the optimization over D^h is carried out by fixing D^l and z_k^i, and solving the following QCQP problem

$$\min_{D^h} \sum_{k=1}^{L_s} \sum_{i=1}^{L_p} \left(\left\| x_k^i - D^h z_k^i \right\|_2^2 + \sum_{j=1}^{N} \gamma_j \left\| x_{k-j}^{i\ MC} - D^h z_k^i \right\|_2^2 \right) \quad \text{s.t.} \left\| D^h(:,k) \right\|_2 \le 1, \quad \forall k, \quad (20)$$

and then projecting onto the unit ball (Equation (14)).

The iterative procedure of the coupled dictionary learning algorithm for recursive video training is summarized in Algorithm 2.

Algorithm 2 can be applied on each cluster separately to train multiple dictionaries. Feature filter F on the LR patch y_k^i is utilized to cluster each training patch set $\left\{y_k^i, \left\{x_{k-j}^{i\ MC}\right\}_{j=1}^{j=N}, x_k^i\right\}$.

IV. Experimental Results

Our two proposed algorithms extend the bilevel dictionary learning [22], [23] in two aspects: from single dictionary to multiple dictionaries and from single frame to multiple frames. We first show that each extension is beneficial by comparing the SR performances of single dictionary single frame SR (Bilevel), multiple dictionaries single frame SR (MDSF), single dictionary multiple frames SR (SDMF-B for the batch approach, SDMF-R for the recursive approach), multiple dictionary multiple frames SR (MDMF-B for the batch approach, MDMF-R for the recursive approach) and MDMF-B/MDMF-R with the proposed video training (MDMF-B-VT/MDMF-R-VT). We also compare the performance of the proposed algorithm with state-of-the-art video SR algorithms, such as Enhancer [41], Bayesian [6], Bayesian-MB [16] and DraftCNN [17].

A. Implementation Details

We performed an extensive set of experiments utilizing frames of a 4K video database [46]. There is a high demand of upscaling videos of low resolutions to 4K resolution (2160×3840) these days due the proliferation of 4K monitors. Upscaling of 1080P (1080×1920) or 540P (540×960) resolution to 4K videos is a representative example used in this paper, resulting in an upscale factor of 2 and 4, respectively. In detail, for upscale factor 2, there are in total 57 scenes in the 4K video database [46]. LR (1080×1920) frames result from the degradation of the original HR (2160×3840) frames by the Matlab function "imresize", which is experimentally found to represent a Gaussian blur kernel with variance approximately equal to 0.4, thus specifying the B matrix in Equation (1). 50 scenes are used for training and 7 for testing. In the training phase of these experiments, 800,000 patch sets are sampled from the center frame and the motion compensated neighboring frames for training the dictionary from videos, while the same 800,000 patches in center frames are used for training the dictionary from images. The patch size is 5×5 and no feature filter F is applied to the LR patches. The reason for not doing so is that we verified experimentally that by using for example four high-pass filters, as was done in [22] and [23], does not provide any sizeable advantage. In addition, four high-pass filter will increase the dimension of the LR dictionary atoms by a factor of four, thus increasing considerably the required computation. λ is chosen to be 0.02 by a parameter traversing experiment, as shown in Figure (6). β_1 and β_2 are chosen to be equal to 0.2 and $\frac{1}{3 \times max(e(k,k+j))}$ according to the convexity criteria in [43], respectively. Every dictionary for the SDSF, SDMF-B, and SDMF-R approaches has 512 atoms and the dictionary for the MDSF, MDMF-B and MDMF-R approaches has 8 subdictionaries with 512 atoms each. For the multiple-dictionary methods in the testing phase, we solve the LASSO problem with only one sub-dictionary

TABLE I
PSNR VALUES (IN dB) OF THE SR FRAME FOR VARIOUS METHODS AND TEST SCENES (BEST RESULTS ARE SHOWN IN BOLD)

	Bicubic	Bilevel	SDMF-B	SDMF-R	MDSF	MDMF-B	MDMF-R	MDMF-B-VT	MDMF-R-VT
Scene 2	45.27	46.12	46.81	46.41	46.79	47.66	46.86	**48.14**	47.41
Scene 8	38.18	39.94	40.08	40.32	40.34	40.59	40.60	40.98	**41.05**
Scene 18	41.43	43.04	43.41	43.69	43.37	43.92	44.19	44.32	**44.46**
Scene 25	44.40	46.69	47.52	47.68	47.37	48.45	47.83	**49.19**	48.59
Scene 33	40.22	42.95	43.08	43.55	43.27	43.68	44.05	**44.49**	44.48
Scene 45	42.43	43.72	44.07	44.18	44.05	44.49	44.28	44.60	**44.62**
Scene 48	33.90	36.10	36.20	35.66	36.55	36.67	36.07	**36.91**	36.64

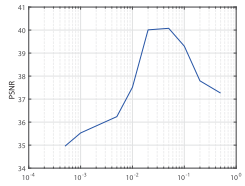

Fig. 6. λ is traversed to find its optimal value. For each tested λ, we perform the multiple frame SR according to Equation (7) and compute its corresponding PSNR value.

and the computation for assigning patches to each cluster (Equation (6)) is negligible, therefore the comparison is fair.

In the testing phase of upscale factor 2, 6 consecutive video frames are super-resolved by each method. 5×5 patches are extracted with overlap of 4 pixels between adjacent patches. The multiple estimates of the same pixel from different overlapping patches are averaged to obtain the final result. For those multiple-frame batch SR methods, the current LR frame, together with one LR backward and one LR forward frames (i.e., $M = N = 1$), are utilized to estimate the current frame. For those multiple-frame recursive SR methods, the current LR frame, together with two LR/HR super-resolved backward frames (i.e., $N = 2$) are used to estimate the current frame. We tested a number of optical flow estimation algorithm [34]. Based on their comparison we are using the method in [36] in all reported experiments.

For an upscale scale factor of 4, similarly to [23], we found experimentally that utilizing the trained coupled dictionaries for an upscale factor of 2 and upscaling the frames twice with an upscale factor of 2 in each step provides better SR results than training and testing directly with an upscale factor of 4.

For color video frames, we apply our video SR algorithm to the luminance channel only, since humans are more sensitive to illumination changes. The color layers (Cb, Cr) are upscaled using bicubic interpolation. The results of the various methods are evaluated in terms of PSNR (peak signal-to-noise ratio) and SSIM [47] on the luminance channel.

B. Effect of the Proposed Extensions

Our two proposed methods are based on the bilevel dictionary learning [22], [23], which is a single dictionary single frame SR method. Since our methods extend it to use multiple dictionaries and multiple frames, we perform a controlled experiment for each extension here to show that the proposed extensions are effective. All multiple-frame SR methods utilize

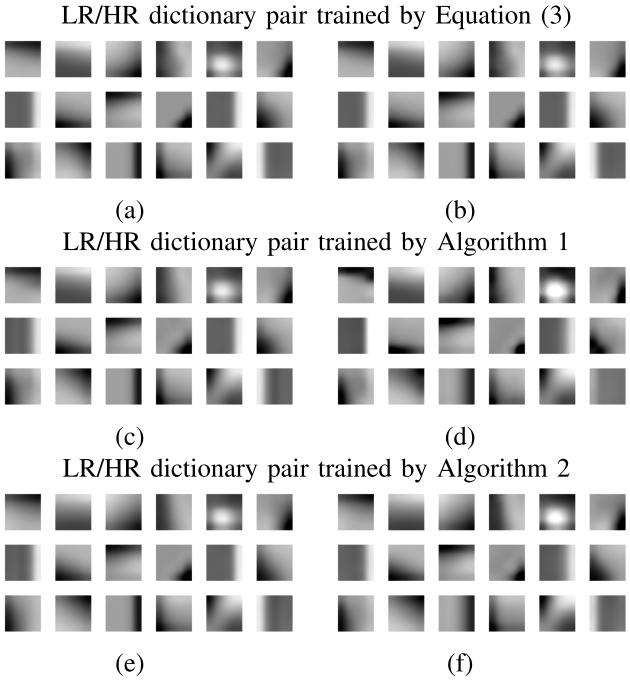

Fig. 7. Atoms of LR/HR dictionary pairs trained by three different dictionary learning algorithms. Each 5×5 atom is upscaled by a factor of 6 by bicubic interpolation for better visualization. (a) D^l. (b) D^h. (c) D^l. (d) D^h. (e) D^l. (f) D^h.

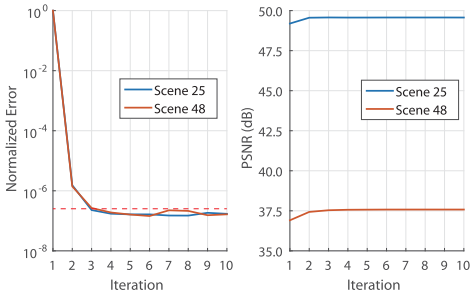

Fig. 8. The iteration process of the batch approach. The normalized error $\left\|I^h_{p+1} - I^h_p\right\|^2_F / \left\|I^h_{p+1}\right\|^2_F$ of Scenes 25 and Scene 48 as a function of iteration is shown in the left image and the corresponding PSNR values are shown in the right image.

one iteration in updating the motion field and HR frames, since the effect of iteratively updating motion fields and HR frames will be discussed in Section IV-C.

Table I shows the peak signal-to-noise ratio (PSNR) of the SR frames in dB (the dB values are averaged over 6

TABLE II
PSNR VALUES (IN dB) OF THE SR FRAME FOR VARIOUS METHODS, ITERATIONS AND SCENES

	Bicubic	MDSF	MDMF-B-VT			MDMF-R-VT		
			Iter=1	Iter=2	Convergence	Iter=1	Iter=2	Convergence
Scene 2	45.27	46.79	48.14	48.24	**48.26**	47.41	47.98	48.11
Scene 8	38.18	40.34	40.98	41.43	41.62	41.05	41.53	**41.72**
Scene 18	41.43	43.37	44.32	44.66	44.73	44.46	44.85	**44.94**
Scene 25	44.40	47.37	49.19	49.56	49.57	48.59	49.57	**49.68**
Scene 33	40.22	43.27	44.49	45.19	45.33	44.48	45.30	**45.44**
Scene 45	42.43	44.05	44.60	44.73	44.76	44.62	44.86	**44.91**
Scene 48	33.90	36.55	36.91	37.43	**37.56**	36.64	37.16	37.31

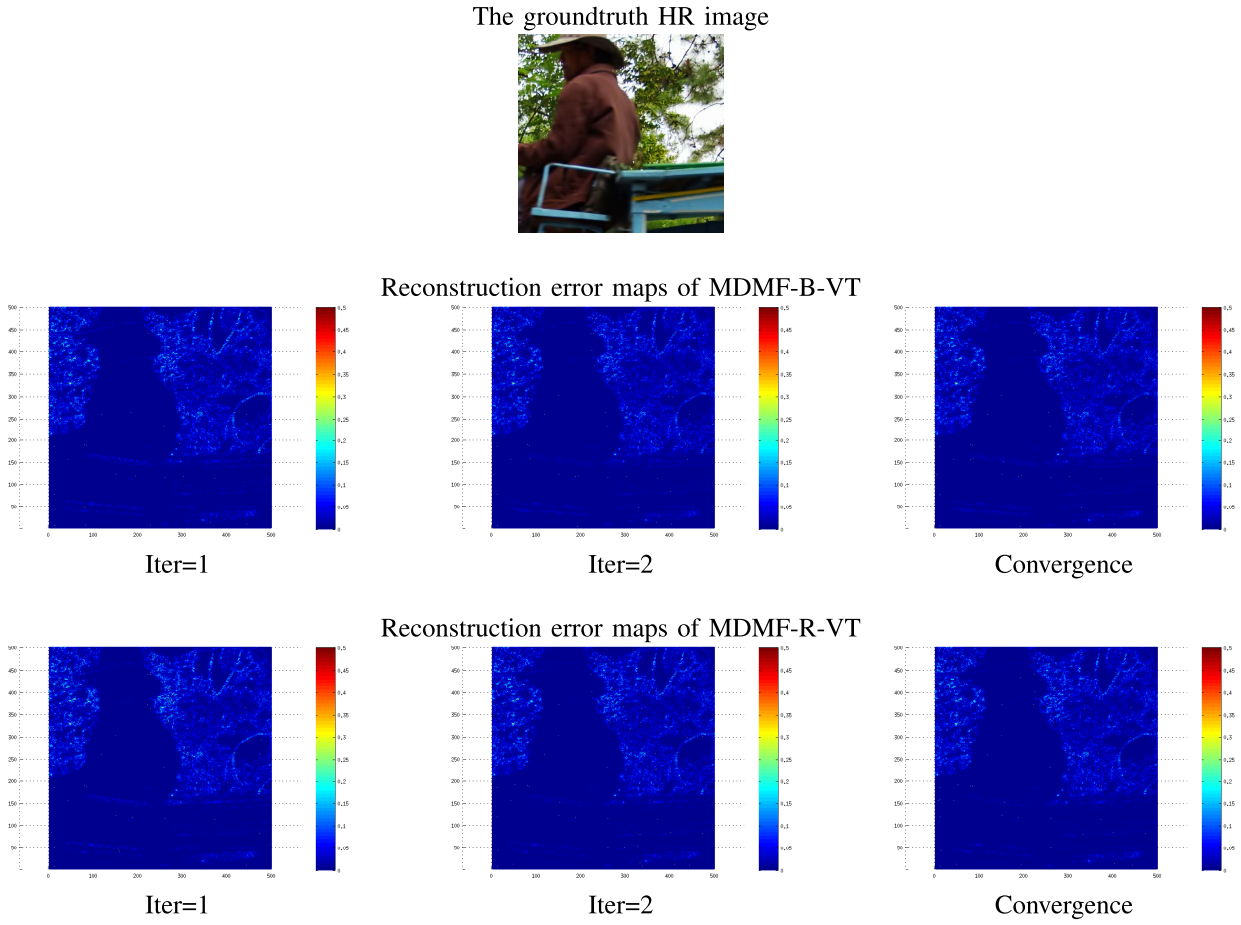

Fig. 9. Reconstruction error maps of a cropped region in scene 48, with different iteration number.

testing frames) for various algorithms and test sequences. The best results are shown in bold. From these experiments it is concluded that DL based multiple-frame SR methods (SDMF-B, SDMF-R, MDMF-B, MDMF-R) outperform single frame SR (Bilevel, MDSF). We can also see that multiple-dictionary SR methods perform better than single-dictionary SR methods, by comparing results of SDMF-B with MDMF-B and SDMF-R with MDMF-R. Finally, the proposed training dictionaries from video algorithms (Algorithm 1 and Algorithm 2), MDMF-B-VT and MDMF-R-VT, further improve the SR results over MDMF-B and MDMF-R.

We show in Figure (7) LR and HR dictionary atoms resulting from the various dictionary training approaches we have considered. 18 atoms from the D^l dictionary and the corresponding atoms in the D^h dictionary trained according to Equation (3) are shown respectively in Figure (7a) and (7b). The same 18 LR/HR atom pairs resulting from Algorithm 1 and Algorithm 2 are shown respectively in Figures (7c), (7d) and (7e), (7f). Notice that dictionaries D^l and D^h trained from Equation (3) is the initializations of D^l and D^h for Algorithms 1 and 2. As shown in Figure (7), sharper HR atoms result in general from our proposed training Algorithms 1 and 2 (compare Figure (7b), (7d) and (7f)).

TABLE III

PSNR VALUES (IN dB, TOP) AND SSIM VALUES (BOTTOM) OF EXPERIMENTAL RESULTS COMPARING OUR PROPOSED METHODS WITH THE STATE-OF-THE-ART METHODS FOR UPSCALE FACTOR 2 (BEST RESULTS ARE SHOWN IN BOLD)

Video	Bicubic	Bilevel [22], [23]	NE+NNLS [37]	NE+LLE [38]	ANR [39]
Scene 2	44.87 0.9830	46.88 0.9879	46.85 0.9851	46.53 0.9834	46.91 0.9857
Scene 8	38.05 0.9738	39.95 0.9842	40.04 0.9824	41.08 0.9817	40.27 0.9832
Scene 18	40.81 0.9738	43.31 0.9849	43.18 0.9820	43.28 0.9816	43.45 0.9833
Scene 25	44.31 0.9917	47.44 0.9961	46.69 0.9938	47.38 0.9936	47.85 0.9952
Scene 33	39.42 0.9786	42.81 0.9904	42.70 0.9879	43.37 0.9889	43.59 0.9902
Scene 45	42.23 0.9718	44.11 0.9810	43.62 0.9772	43.89 0.9776	44.11 0.9791
Scene 48	33.81 0.9668	36.05 0.9808	35.78 0.9774	36.24 0.9785	36.39 0.9799
Average	40.50 0.9771	42.94 0.9865	42.69 0.9837	43.11 0.9836	43.22 0.9851
Video	SR-CNN [40]	Enhancer [41]	Bayesian [6]	MDMF-B-VT	MDMF-R-VT
Scene 2	47.41 0.9859	46.10 0.9854	46.23 0.9874	**48.26** **0.9882**	48.11 **0.9882**
Scene 8	41.08 0.9852	39.42 0.9823	39.73 0.9828	41.62 **0.9884**	**41.65** 0.9882
Scene 18	43.94 0.9844	42.93 0.9844	43.16 0.9842	44.74 0.9877	**45.04** **0.9884**
Scene 25	48.29 0.9955	46.17 0.9938	46.36 0.9954	**49.57** **0.9970**	49.07 0.9967
Scene 33	43.83 0.9907	43.05 0.9908	42.65 0.9900	**45.33** 0.9937	45.11 **0.9938**
Scene 45	44.32 0.9797	43.11 0.9764	43.59 0.9790	44.76 0.9812	**44.84** **0.9823**
Scene 48	37.48 0.9826	35.24 0.9751	35.27 0.9770	**37.57** **0.9846**	36.89 0.9821
Average	43.76 0.9863	42.29 0.9840	42.43 0.9851	**44.55** **0.9887**	44.39 0.9885

TABLE IV

PSNR VALUES (IN dB, TOP) AND SSIM VALUES (BOTTOM) OF EXPERIMENTAL RESULTS COMPARING OUR PROPOSED METHODS WITH THE STATE-OF-THE-ART METHODS FOR UPSCALE FACTOR 4 (BEST RESULTS ARE SHOWN IN BOLD)

Video	Bicubic	Bilevel [22], [23]	NE+NNLS [37]	NE+LLE [38]	ANR [39]
Scene 2	39.58 0.9648	40.50 0.9662	41.32 0.9691	41.12 0.9675	41.32 0.9691
Scene 8	32.13 0.9013	32.46 0.9099	33.00 0.9145	32.95 0.9187	32.81 0.9107
Scene 18	35.65 0.9122	36.37 0.9209	36.76 0.9243	36.82 0.9249	36.76 0.9243
Scene 25	36.10 0.9515	37.02 0.9546	37.90 0.9622	37.78 0.9607	37.49 0.9587
Scene 33	32.15 0.8899	33.44 0.9140	33.79 0.9157	33.94 0.9188	34.00 0.9206
Scene 45	36.13 0.9101	36.71 0.9155	37.12 0.9193	37.27 0.9211	37.35 0.9226
Scene 48	27.25 0.8514	28.03 0.8730	28.04 0.8710	28.20 0.8757	28.26 0.8780
Average	34.14 0.9116	34.93 0.9220	35.42 0.9252	35.44 0.9268	35.43 0.9263
Video	SR-CNN [40]	Enhancer [41]	Bayesian [6]	MDMF-B-VT	MDMF-R-VT
Scene 2	43.17 0.9703	40.62 0.9695	39.18 0.9660	**43.48** 0.9737	42.90 **0.9740**
Scene 8	33.40 0.9198	32.09 0.9121	31.73 0.8972	**33.48** **0.9266**	33.42 0.9250
Scene 18	37.50 0.9280	36.44 0.9308	35.70 0.9183	**37.68** 0.9331	37.65 **0.9341**
Scene 25	38.35 0.9633	37.44 0.9621	35.34 0.9473	**39.03** **0.9702**	38.75 0.9687
Scene 33	34.57 0.9230	34.67 0.9304	32.14 0.8945	**34.92** 0.9363	34.86 **0.9374**
Scene 45	37.90 0.9253	37.15 0.9267	35.76 0.9083	**38.42** **0.9340**	38.10 0.9316
Scene 48	28.73 0.8883	27.75 0.8679	26.76 0.8393	**28.75** **0.8921**	28.49 0.8842
Average	36.23 0.9311	35.17 0.9285	33.80 0.9101	**36.54** **0.9380**	36.31 0.9364

In conclusion, our proposed multiple frames SR, utilizing multiple dictionaries and training dictionaries from videos are effective individually and their benefits in SR are cummulative, as the proposed MDMF-B-VT and MDMF-R-VT algorithms provide best SR results.

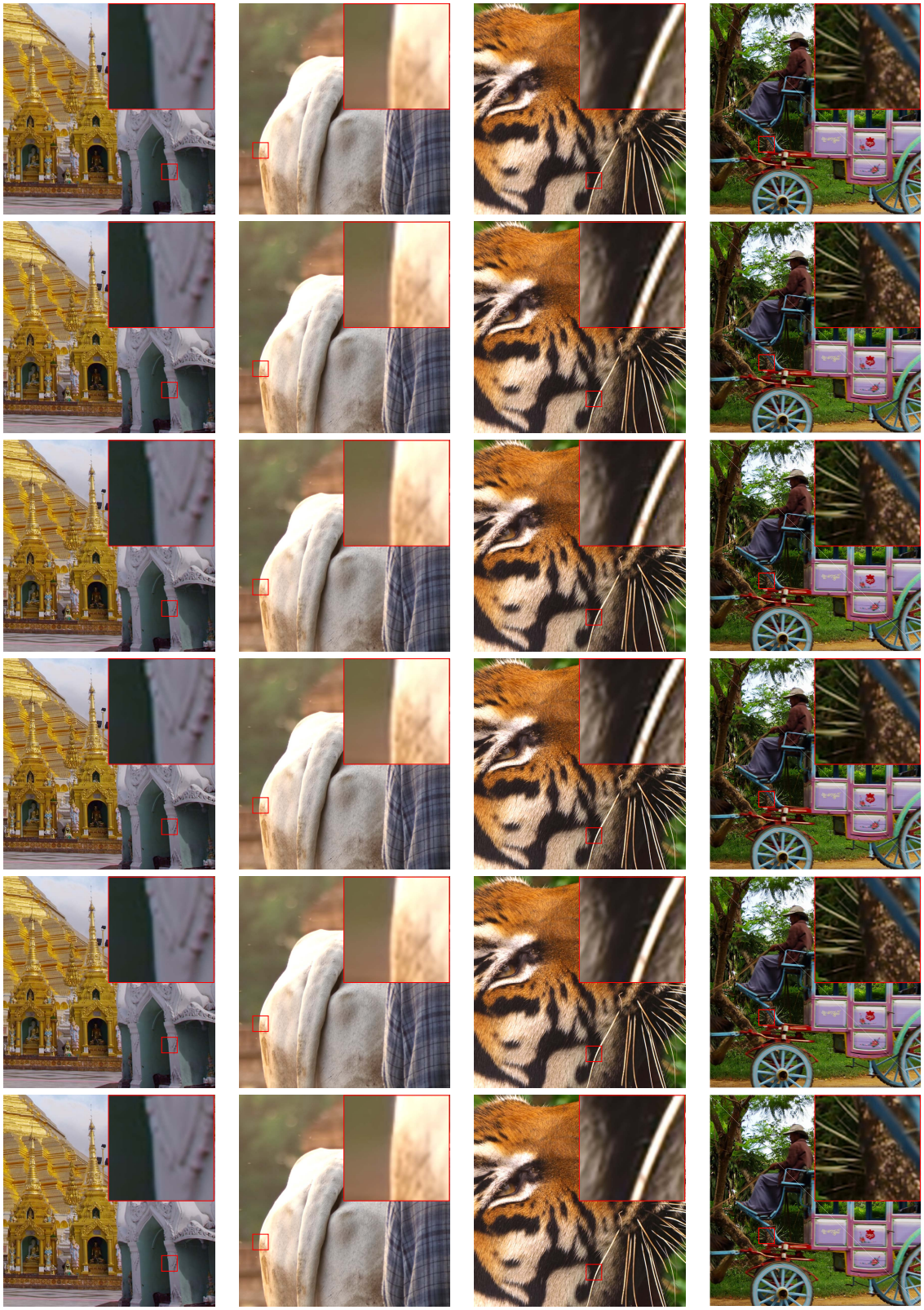

Fig. 10. Visual Comparison of SR results. Column left to right is scene 8, scene 25, scene 33 and scene 48, respectively. Row top to bottom is Bicubic, Bilevel [22], [23], Enhancer [41], Bayesian [6], proposed MDMF-B-VT and proposed MDMF-R-VT, respectively. Our proposed algorithms can generate natural-looking frames without noticeable visual artifacts. Because the testing frames have high resolution, results are better viewed in zoomed PDF.

C. Effect of Iteration

The proposed batch approach (Section II-B) and recursive approach (Section III-B) by alternating optimizations update the motion fields and reconstructed HR frames I^h. To demonstrate the convergence of the iteration process, we calculate the normalized error $\left\| I^h_{p+1} - I^h_p \right\|^2_F / \left\| I^h_{p+1} \right\|^2_F$

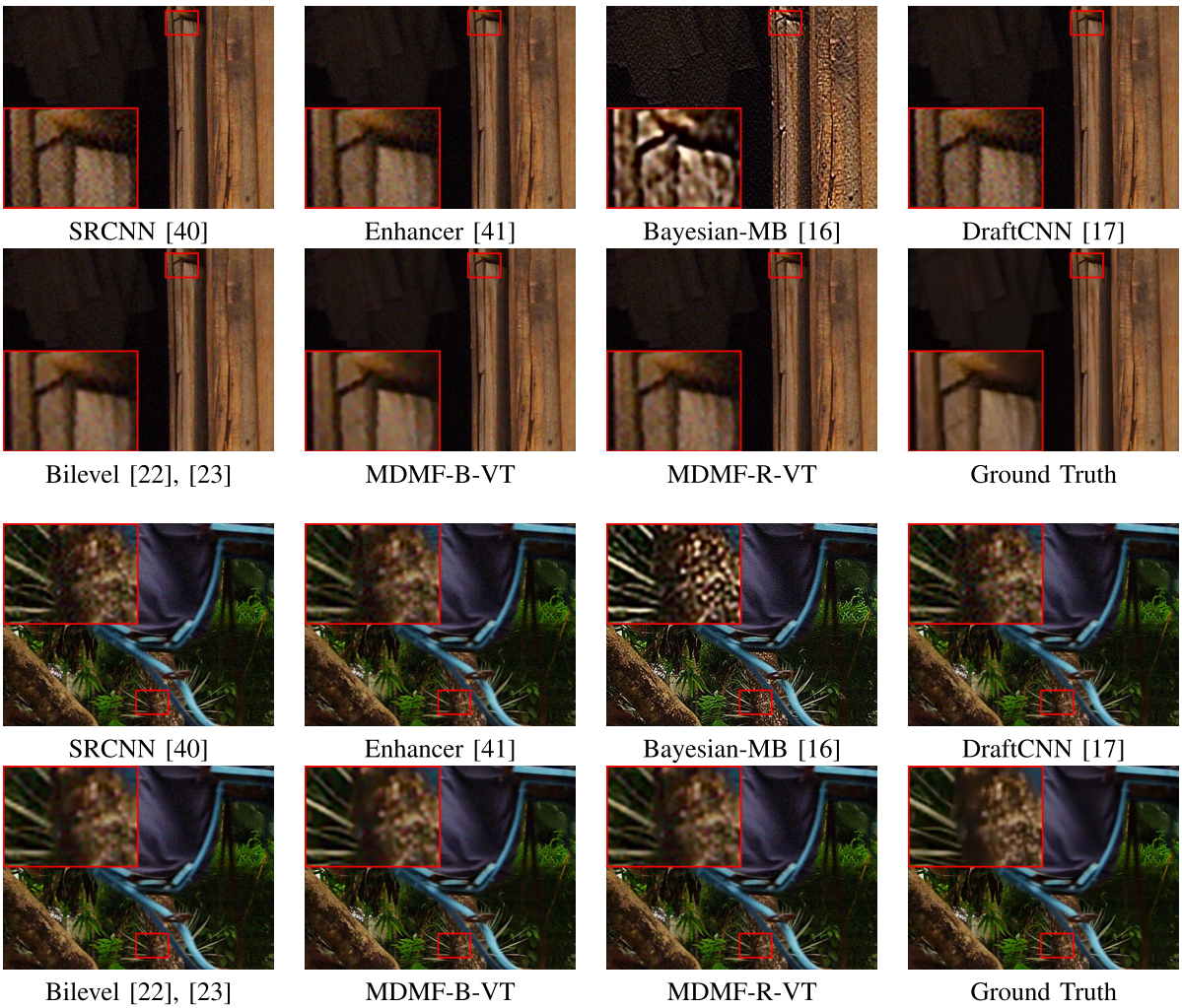

Fig. 11. Visual Comparison of SR results of different SR methods when Gaussian noise variance equals to 0.001. Our proposed algorithms suppress the noise and generate the closest HR frames to the Ground Truth HR frames. Readers are suggested to zoom in to see the details.

(I_p^h is the reconstructed HR frame at the p^{th} iteration) at each iteration. This normalized error is shown in Figure (8) (left) for the batch approach (MDMF-B-VT) for two of the experiments, and the corresponding PSNR is in Figure (8) (right). As shown in Figure (8), the iteration process converges fast. Similar results are also observed with the recursive approach. In all experiments, we terminate the iteration when the normalized error is below the threshold of 5×10^{-7}.

We visualize the reconstruction error maps of a cropped region of the 6^{th} frame in scene 48 in Figure (9), which has a global panning motion of the background with the local motion of the foreground. From the heat maps, the reconstruction error in the background texture region decreases as the iteration progresses, also the error in the handle in the foreground almost disappears at the final result.

More interestingly, as shown in Table II, we observe that although the batch SR algorithm outperforms the recursive SR algorithm at iteration 1, their performance is comparable in the final iteration, illustrating that the batch approach is more robust to errors in motion estimation and that both approaches have similar performance when motion estimation is precise.

D. Comparison With State-of-the-Art Results

In the previous Sections IV-B and IV-C, we show that our extensions of single frame bilevel SR methods [22], [23] are effective and the iterative updates of the motion field and HR frames improve the SR performance. Here we compare our proposed methods, MDMF-B-VT and MDMF-R-VT, with other state-of-the-art methods, including Bayesian [6] and a commercial software Enhancer [41], and six single frame SR methods including Bicubic, Bilevel [22], [23], NE+NNLS [37], NE+LLE [38], ANR [39] and SR-CNN [40]. Two more state-of-the-art methods [16], [17] will be compared in Section IV-E with smaller spatial resolution because their implementation is extremely slow on 4K resolution.

According to Table III and Table IV, our proposed approaches (MDMF-B-VT and MDMF-R-VT) provide the best SR performance compared to all other methods for both upscale factors of 2 and 4, demonstrating the effectiveness of the proposed algorithms. Although the Bayesian SR method [6] evaluates the blur kernel, noise level and super-resolved frames simultaneously, it requires the motion compensation of 30 consecutive frames in the backward and

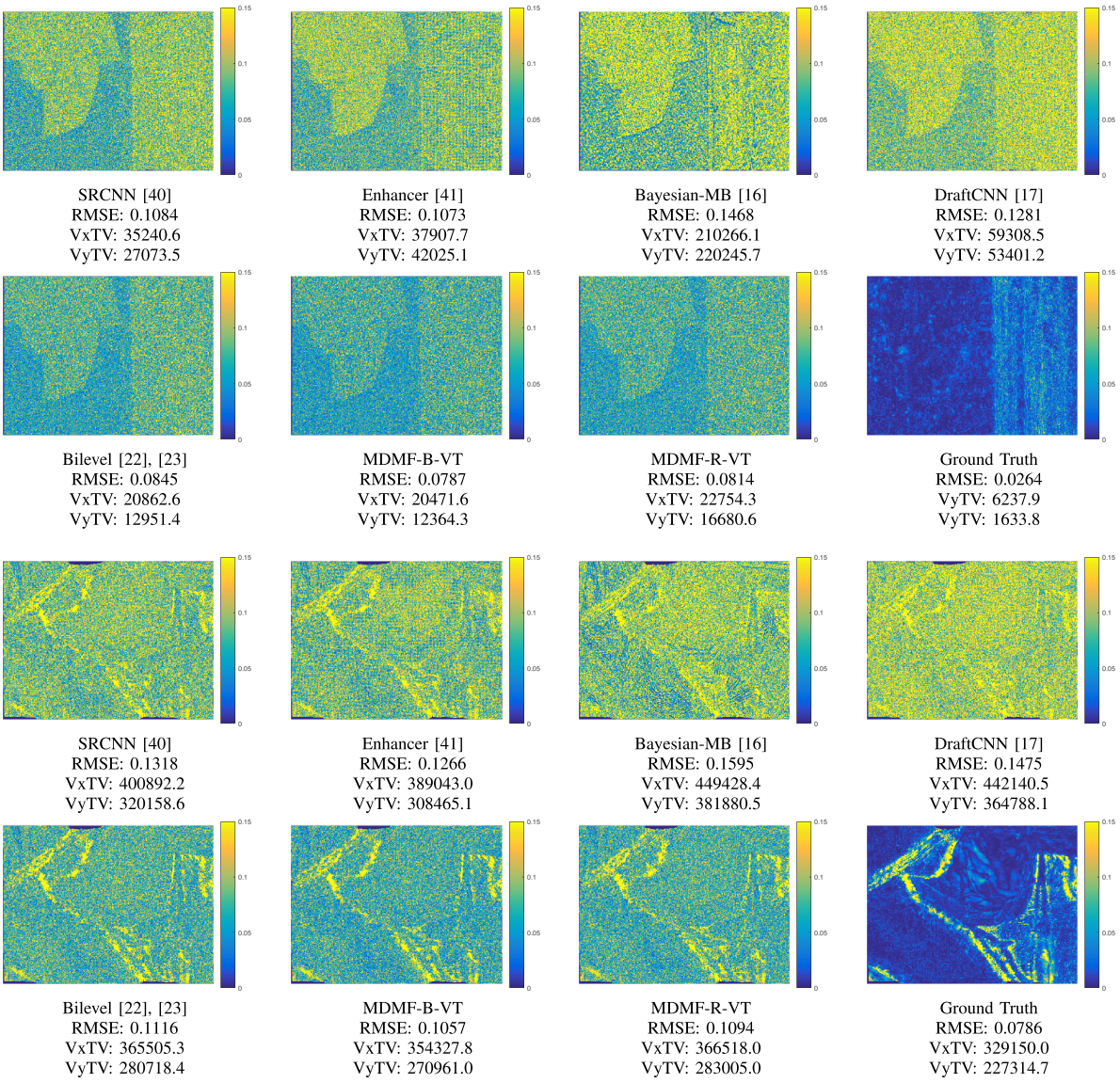

Fig. 12. Visual Comparison of the motion compensation error of the SR results by different SR methods when Gaussian noise variance equals to 0.001. Our proposed algorithms have the smallest motion compensation error from both the error head map and RMSE metric, illustrating the advantages in temporal smoothness of the super-resolved frames.

forward directions, which is computationally infeasible with 4K videos because of the memory and computational limitations. When we drop the consecutive frames from 30 to 3, the SR performance of [6] is not as good as ours. In Figure (10), we compare the visual quality of our upscaled images with the result produced by several recent state-of-the-art SR methods. We notice that all these SR methods produce sharper images than bicubic interpolation, however artifacts are introduced. Next we notice that our proposed method has fewer artifacts and shaper edges compared to all other methods.

E. Robustness to Noise

In this section, we evaluate the noise robustness of different SR algorithms by adding Gaussian noise to the LR input frames. The center regions (480×640) of the original 4K frames are utilized as the HR ground truth, in order to compare with two more state-of-the-art video SR methods, Bayesian-MB [16] and DraftCNN [17]. The LR input frames (240×320) are obtained by spatially downsampling the HR frames by a factor of 2 and adding white Gaussian noise with variance 0.001. Different SR methods are applied to increase the spatial resolution by a factor of 2. We also show the experimental results with no additional Gaussian noise (noise variance 0).

As shown in Table V, the SR performance of all methods is reduced when noise is added, as expected. The HR dictionaries for the dictionary learning based methods, Bilevel [22], [23], MDMF-B-VT and MDMF-R-VT, are trained with noise free HR frames, so the reconstructed HR frames naturally contain less noise. The sparse coding problem in SR testing phase is also proven to be robust to noise [48], so better SR performance is obtained by the dictionary learning based methods (Bilevel [22], [23], MDMF-B-VT and MDMF-R-VT).

TABLE V

PSNR VALUES (IN dB, TOP) AND SSIM VALUES (BOTTOM) OF EXPERIMENTAL RESULTS COMPARING OUR PROPOSED METHODS WITH THE STATE-OF-THE-ART METHODS (BEST RESULTS ARE SHOWN IN BOLD) UNDER DIFFERENT NOISE CONDITIONS

Video	Bicubic		Bilevel [22], [23]		NE+NNLS [37]		NE+LLE [38]		ANR [39]		SR-CNN [40]	
Noise Variance	0	0.001	0	0.001	0	0.001	0	0.001	0	0.001	0	0.001
Scene 2	45.53	32.68	47.78	37.51	46.57	33.31	46.66	32.85	47.16	32.91	46.67	36.13
	0.9806	0.7085	0.9875	0.8920	0.9843	0.7377	0.9833	0.7144	0.9854	0.7180	0.9835	0.8499
Scene 8	35.01	31.60	36.80	33.49	37.11	32.01	37.22	31.70	37.17	31.72	38.29	32.75
	0.9424	0.7462	0.9645	0.8708	0.9640	0.7728	0.9643	0.7550	0.9655	0.7578	0.9686	0.8401
Scene 18	41.65	32.38	44.72	36.68	44.22	33.18	44.53	32.74	44.89	32.80	44.65	35.48
	0.9780	0.7168	0.9902	0.8949	0.9876	0.7530	0.9874	0.7311	0.9892	0.7345	0.9873	0.8546
Scene 25	43.94	32.47	47.60	37.21	46.05	33.31	47.26	32.93	47.87	32.97	46.87	35.96
	0.9917	0.7246	0.9971	0.9090	0.9952	0.7628	0.9947	0.7416	0.9966	0.7447	0.9935	0.8698
Scene 33	35.94	31.97	39.65	34.62	40.46	32.85	41.21	32.58	41.36	32.60	40.88	33.27
	0.9606	0.7941	0.9839	0.9054	0.9809	0.8237	0.9830	0.8093	0.9842	0.8113	0.9834	0.8792
Scene 45	44.60	33.26	46.99	37.55	46.24	33.87	46.63	33.50	47.08	33.55	46.57	36.44
	0.9850	0.7584	0.9912	0.9092	0.9889	0.7808	0.9885	0.7613	0.9904	0.7647	0.9887	0.8768
Scene 48	34.96	31.62	36.88	33.47	36.74	31.98	37.25	31.84	37.38	31.89	**38.39**	32.77
	0.9660	0.7921	0.9788	0.9059	0.9764	0.8165	0.9782	0.8024	0.9796	0.8048	0.9809	0.8806
Average	40.23	32.28	42.92	35.79	42.48	32.93	42.97	32.59	43.27	32.63	43.19	34.69
	0.9720	0.7487	0.9847	0.8982	0.9825	0.7782	0.9828	0.7593	0.9844	0.7622	0.9837	0.8644
Computation time	-		15.0 s		75.9 s		13.1 s		1.5 s		2.6 s	

Video	Enhancer [41]		Bayesian [6]		Bayesian-MB [16]		DraftCNN [17]		MDMF-B-VT		MDMF-R-VT	
Noise Variance	0	0.001	0	0.001	0	0.001	0	0.001	0	0.001	0	0.001
Scene 2	46.91	33.95	45.34	30.26	43.95	32.56	**47.94**	32.88	**47.94**	**40.55**	47.91	40.13
	0.9849	0.7597	0.9821	0.5983	0.9836	0.7044	**0.9880**	0.7150	0.9877	**0.9539**	0.9875	0.9414
Scene 8	36.68	32.23	36.25	29.08	36.72	31.35	37.43	31.70	**38.80**	**35.61**	38.59	35.10
	0.9632	0.7875	0.9645	0.6528	0.9632	0.7448	0.9671	0.7556	0.9733	**0.9314**	**0.9734**	0.9209
Scene 18	45.41	33.77	42.69	29.74	42.17	32.37	45.80	32.80	**46.68**	39.47	46.39	38.95
	0.9909	0.7743	0.9879	0.5981	0.9859	0.7188	0.9919	0.7321	**0.9929**	0.9581	0.9926	0.9457
Scene 25	46.53	33.92	44.38	29.73	42.96	32.47	46.41	32.88	49.40	**39.99**	**50.13**	39.85
	0.9952	0.7828	0.9953	0.6112	0.9926	0.7284	0.9962	0.7395	0.9978	**0.9690**	**0.9981**	0.9584
Scene 33	39.53	32.97	38.29	29.40	38.26	31.73	39.03	32.08	**42.79**	**36.39**	41.19	34.77
	0.9832	0.8312	0.9792	0.7108	0.9782	0.7903	0.9846	0.7972	**0.9896**	**0.9488**	0.9870	0.9293
Scene 45	46.08	34.58	44.26	31.28	44.76	33.23	47.35	33.45	47.61	**40.10**	**47.76**	39.67
	0.9884	0.8075	0.9868	0.6645	0.9878	0.7531	0.9913	0.7603	**0.9918**	**0.9564**	**0.9918**	0.9480
Scene 48	35.95	32.18	35.67	29.23	35.39	30.95	36.34	31.52	38.10	**34.44**	37.35	32.93
	0.9692	0.8278	0.9755	0.7141	0.9637	0.7779	0.9745	0.7960	**0.9820**	**0.9416**	0.9784	0.9170
Average	42.44	33.37	40.98	29.82	40.60	32.09	42.90	32.47	**44.47**	38.08	44.19	37.34
	0.9821	0.7958	0.9816	0.6500	0.9793	0.7454	0.9848	0.7565	**0.9879**	0.9513	0.9870	0.9372
Computation time	1.7 s		114.2 s		76.0 s		2165.7 s		159.3 s		115.5 s	

By comparing the SR results of Bilevel [22], [23] with MDMF-B-VT and MDMF-R-VT, we found out that better SR performance is obtained by utilizing multiple LR noisy input frames. The proposed MDMF-B-VT consistently outperforms MDMF-R-VT, since it estimates the sparse coefficients of 3 noisy LR patches simultaneously.

The average computation time for all SR algorithms to super-resolve 1 frame is also shown in Table V. All experiments except Enhancer and Bayesian-MB are performed on a Linux workstation with an Intel Xeon E5-2630 processor with 2.4GHz and 64 GB RAM. The Enhancer and the Bayesian-MB algorithm were only available for the Windows operating system and were tested on a different workstation with Intel i7-6820 processor with 2.70GHz and 16 GB RAM. Notice that our proposed methods MDMF-B-VT and MDMF-R-VT can be sped up by a factor of 4 approximately if we only apply 1 iteration instead of 4 iterations. For MDMF-B-VT, the motion estimation takes 21.5s and the sparse coefficients inference of Equation (7) takes 14.4s on average for one iteration. For MDMF-R-VT, the motion estimation takes 22.1s and the sparse coefficients inference of Equation (10) takes 9.1s on average for one iteration. So our proposed methods can be further sped up by utilizing faster motion estimation methods and sparse coefficients inference algorithms.

We visually compare the SR results of our proposed methods with several other state-of-the-art SR methods, when white Gaussian noise with variance 0.001 is added to the LR input frames. We notice that the dictionary learning based methods, Bilevel [22], [23], MDMF-B-VT and MDMF-R-VT, outperform others in suppressing the noise. The proposed MDMF-B-VT algorithm provides the sharpest HR frame with few artifacts.

The temporal continuity between adjacent super-resolved HR frames is compared in Figure (12) by visualizing the motion compensation error of two adjacent super-resolved HR frames by different SR algorithms. The optical flow estimation method in [36] is applied to estimate the motion field between two adjacent super-resolved HR frames, and the second frame is warped to the first one according to the computed motion field. The difference between the first frame and the warped second frame is visualized to compare the temporal smoothness of different SR algorithms. The main idea behind this is that if two adjacent super-resolved frames are temporally smooth, then an accurate motion field can be estimated and

the resulting motion compensated difference will be small. In quantifying this difference we compute the RMSE (Root-Mean-Square Error). The smoothness of the motion field is of course also indicative of the temporal continuity between adjacent frames. One can imagine situations where the RMSE of the displaced frame difference is small but the motion field exhibits large variations. We therefore also compute the Total Variation (TV) of the estimated motion field vectors, in both the horizontal (VxTV) and vertical (VyTV) directions. In comparing the temporal smoothness of video frames, both the RMSE of the displaced frame difference and the TV of the motion field should be taken into account; the smaller such measures the higher the temporal smoothness. As shown in Figure (12), our proposed MDMF-B-VT method produces the smallest RMSE on the motion compensation error, as well as the smallest TV on the motion vector, demonstrating that it better explores the spatio-temporal correlation of consecutive frames. Notice that our proposed MDMF-R-VT method produces the second smallest RMSE on the motion compensation error while have larger TV on the motion vector compared to Bilevel [22], [23], so its temporal smoothness is similar to Bilevel [22], [23]. However, its SR performance is still 2.3 dB better than Bilevel [22], [23] on average according to Table V. It is also interesting to point out that according to Table V, the single frame SR method SRCNN [40] outperforms the multiple frame SR method Enhancer [41] in terms of the averaged single frame PSNR and SSIM metrics, while Enhancer [41] has smaller motion compensation error of adjacent frames according to Figure (12), illustrating that multiple frame SR methods provide an advantage in terms of the temporal smoothness of the super-resolved HR frames.

V. Conclusion

In this paper we presented two novel video SR frameworks, the batch approach and the recursive approach, based on dictionary learning and motion estimation. According to them, the HR patches are estimated from multiple corresponding LR patches or previously super-resolved HR patches in multiple frames, making the dictionary-based reconstruction algorithm more accurate. The dictionary training algorithms that utilize multiple frames of the training videos further improved the SR performance by making the training and testing phases consistent. We performed experiments with 4K videos and showed that our methods outperform the state-of-the-art algorithms, based either on quantitative analysis or visual comparison.

References

[1] S. Borman and R. L. Stevenson, "Super-resolution from image sequences—A review," in *Proc. IEEE Comput. Soc. Circuits Syst., Midwest Symp.*, Aug. 1998, pp. 374–378.

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

[3] A. K. Katsaggelos, R. Molina, and J. Mateos, "Super resolution of images and video," *Synth. Lect. Image, Video, Multimedia Process.*, vol. 1, no. 1, pp. 1–134, 2007.

[4] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014.

[5] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Adv. Comput. Vis. Image Process.*, vol. 1, no. 2, pp. 317–339, 1984.

[6] C. Liu and D. Sun, "A Bayesian approach to adaptive video super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 209–216.

[7] C. A. Segall, R. Molina, and A. K. Katsaggelos, "High-resolution images from low-resolution compressed video," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 37–48, May 2003.

[8] C. A. Segall, A. K. Katsaggelos, R. Molina, and J. Mateos, "Bayesian resolution enhancement of compressed video," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 898–911, Jul. 2004.

[9] S. P. Belekos, N. P. Galatsanos, and A. K. Katsaggelos, "Maximum a posteriori video super-resolution using a new multichannel image prior," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1451–1464, Jun. 2010.

[10] Q. Shan, Z. Li, J. Jia, and C.-K. Tang, "Fast image/video upsampling," *ACM Trans. Graph.*, vol. 27, no. 5, Dec. 2008, Art. no. 153.

[11] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1958–1975, Sep. 2009.

[12] E. M. Hung, R. L. D. Queiroz, F. Brandi, K. F. D. Oliveira, and D. Mukherjee, "Video super-resolution using codebooks derived from key-frames," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1321–1331, Sep. 2012.

[13] F. Zhou, W. Yang, and Q. Liao, "Interpolation-based image super-resolution using multisurface fitting," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3312–3318, Jul. 2012.

[14] F. Zhou, S.-T. Xia, and Q. Liao, "Nonlocal pixel selection for multisurface fitting-based super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2013–2017, Dec. 2014.

[15] C. Liu and D. Sun, "On Bayesian adaptive video super resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 346–360, Feb. 2014.

[16] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5224–5232.

[17] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 531–539.

[18] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1187–1193, Aug. 2001.

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

[20] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[21] B. C. Song, S.-C. Jeong, and Y. Choi, "Video super-resolution algorithm using bi-directional overlapped block motion compensation and on-the-fly dictionary training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 274–285, Mar. 2011.

[22] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang, "Bilevel sparse coding for coupled feature spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2360–2367.

[23] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[24] S. Wang, D. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2216–2223.

[25] R. Nakagaki and A. K. Katsaggelos, "A VQ-based blind image restoration algorithm," *IEEE Trans. Image Process.*, vol. 12, no. 9, pp. 1044–1053, Sep. 2003.

[26] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–35, Jan. 2009.

[27] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[28] M. Liang, J. Du, and L. Li, "Learning-based video superresolution reconstruction using spatiotemporal nonlocal similarity," *Math. Problems Eng.*, vol. 2015, 2015, Art. no. 687074.

[29] X. Gao, Q. Wang, X. Li, D. Tao, and K. Zhang, "Zernike-moment-based image super resolution," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2738–2747, Oct. 2011.

[30] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.
[31] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.*, 2012, pp. 711–730.
[32] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3501–3508.
[33] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, p. 185–203, Aug. 1981.
[34] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.
[35] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2432–2439.
[36] M. Drulea and S. Nedevschi, "Total variation regularization of local-global optical flow," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 318–323.
[37] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. 23rd Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 135.1–135.10.
[38] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2004, pp. I–I.
[39] R. Timofte, V. De, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1920–1927.
[40] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
[41] Infognition. (2010). *Video Enhancer*. [Online]. Available: http://www.infognition.com/videoenhancer/
[42] Q. Dai, S. Yoo, A. Kappeler, and A. K. Katsaggelos, "Dictionary-based multiple frame video super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 83–87.
[43] M. K. Park, A. K. Katsaggelos, and M. G. Kang, "Regularized high-resolution image reconstruction considering inaccurate motion information," *Opt. Eng.*, vol. 46, no. 11, p. 117004, Nov. 2007.
[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[45] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2006, pp. 801–808.
[46] H. Inc. (2014). *Harmonic 4k Footage*. [Online]. Available: http://www.harmonicinc.com/resources/videos/4k-video-clip-center
[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[48] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 457–464.

Qiqin Dai received the B.S. degree in automation from Zhejiang University, China, in 2012. He is currently pursuing the Ph.D. degree with the Image and Video Processing Laboratory, Northwestern University, Evanston, IL, USA. His research interests include digital image processing, computer vision, computational photography, and high dynamic range imaging.

Seunghwan Yoo received the B.E. and M.S. degrees in electrical engineering from Sogang University, Seoul, South Korea, in 2005, and 2007, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with Northwestern University, Evanston, IL, USA. He joined the Image and Video Processing Laboratory in 2013. His research interests include image and video super-resolution and deconvolution.

Armin Kappeler received the B.S. degree in electrical engineering from the Hochschule für Technik Rapperswil, Rapperswil-Jona, Switzerland, in 2004, and the M.S. and Ph.D. degrees in 2012 and 2016, respectively. He joined the Image and Video Processing Laboratory, Northwestern University, Evanston, IL, USA, in 2010. He is currently with Yahoo Inc., Sunnyvale, CA, USA, where he is involved in an image classification algorithm using deep neural networks. His research focuses on deep neural networks for image and video restoration and classification.

Aggelos K. Katsaggelos (F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotle University of Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, in 1981 and 1985, respectively. He has an appointment with the Argonne National Laboratory. In 1985, he joined the Department of Electrical Engineering and Computer Science, Northwestern University, where he is currently a Professor. He has authored extensively in the areas of multimedia signal processing and communications (over 230 journal papers, 500 conference papers, and 40 book chapters). He has co-authored the *Rate-Distortion Based Video Compression* (Kluwer, 1997), the *Super-Resolution for Images and Video* (Claypool, 2007), the *Joint Source-Channel Video Transmission* (Claypool, 2007), and the *Machine Learning Refined* (Cambridge University Press, 2016). He holds 25 international patents. He has supervised 54 Ph.D. theses. He was a BOG Member of the IEEE Signal Processing Society (1999–2001) and a member of the Publication Board of the IEEE Proceedings (2003–2007). He is a member of the Award Board of the IEEE Signal Processing Society and a fellow of the SPIE (2009). He is also a member of the Academic Staff, NorthShore University Health System, and a Faculty with the Department of Linguistics. He was a recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), the IEEE Signal Processing Society Best Paper Award (2001), the IEEE ICME Paper Award (2006), the IEEE ICIP Paper Award (2007), an ISPA Paper Award (2009), and a EUSIPCO Paper Award (2013). He was the Ameritech Chair of Information Technology and the AT&T Chair. He holds the Joseph Cummings Chair. He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008). He was the Editor-in- Chief of the IEEE *Signal Processing Magazine* (1997–2002).