A Nuclear-Norm Model for Multi-Frame Super-Resolution Reconstruction from Video Clips

Rui Zhao and Raymond HF Chan

Abstract We propose a variational approach to obtain super-resolution images from multiple low-resolution frames extracted from video clips. First the displacement between the low-resolution frames and the reference frame is computed by an optical flow algorithm. Then a low-rank model is used to construct the reference frame in high resolution by incorporating the information of the low-resolution frames. The model has two terms: a 2-norm data fidelity term and a nuclear-norm regularization term. Alternating direction method of multipliers is used to solve the model. Comparison of our methods with other models on synthetic and real video clips shows that our resulting images are more accurate with less artifacts. It also provides much finer and discernable details.

Keywords Image processing · Super-resolution · Low-rank approximation

1 Introduction

Super-resolution (SR) image reconstruction from multiple low-resolution (LR) frames has many applications, such as in remote sensing, surveillance, and medical imaging. After the pioneering work of Tsai and Huang [28], SR image reconstruction has become more and more popular in image processing community, see, for example, [3, 8, 10, 12, 19, 25–27]. SR image reconstruction problems can be classified into two categories: single-frame super-resolution (SFSR) problems and multi-frame super-resolution (MFSR) problems. In this paper, we mainly focus on the multi-frame case, especially the MFSR problems from low-resolution video sequences. Below, we first review some existing work related to MFSR problems.

R. Zhao
Department of Mathematics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
e-mail: rzhao@math.cuhk.edu.hk

R.H.F. Chan (✉)
Department of Mathematics, City University of Hong Kong, KLN, Hong Kong
e-mail: rchan.sci@cityu.edu.hk

D. A. Bini et al. (eds.), *Structured Matrices in Numerical Linear Algebra*,
Springer INdAM Series 30, https://doi.org/10.1007/978-3-030-04088-8_16

Bose and Boo [3] considered the case where the multiple LR image frames were shifted with affine transformations. They modeled the original high-resolution (HR) image as a stationary Markov–Gaussian random field. Then they made use of the maximum a posteriori scheme to solve their model. However, the affine transformation assumption may not be satisfied in practice, for example, when there are complex motions or illumination changes. Another approach for SR image reconstruction is the one known as patch-based or learning-based. Bishop et al. [2] used a set of learned image patches which capture the information between the middle and high spatial frequency bands. They assumed a priori distribution over such patches and made use of the previous enhanced frame to provide part of the training set. The disadvantage of this patch-based method is that it is usually time consuming and sensitive to the off-line training set. Liu and Sun [18] applied Bayesian approach to estimate simultaneously the underlying motion, the blurring kernel, the noise level, and the HR image. Within each iteration, they estimated the motion, the blurring kernel, and the HR image alternatively by maximizing a posteriori, respectively. Based on this work, Ma et al. [20] tackled motion blur in their paper. An expectation-maximization (EM) framework is applied to the Bayesian approach to guide the estimation of motion blur. These methods used optical flow to model the motion between different frames. However, they are sensitive to the accuracy of flow estimation. The results may fail when the noise is heavy.

In [6], Chan et al. applied wavelet analysis to HR image reconstruction. They decomposed the image from previous iteration into wavelet frequency domain and applied wavelet thresholding to denoise the resulting images. Based on this model, Chan et al. [7] later developed an iterative MFSR approach by using tight-frame wavelet filters. However, because of the number of framelets involved in analyzing the LR images, the algorithm can be extremely time consuming.

Optimization models are one of the most important image processing models. Following the classical ROF model [24], Farsiu et al. [11] proposed a total variation-l_1 model where they used the l_1 norm for the super-resolution data fidelity term. However, it is known that TV regularization enforces a piecewise solution. Therefore, their method will produce some artifacts. Li et al. [16] used l_1 norm of the geometric tight-framelet coefficients as the regularizer and adaptively mimicking l_1 and l_2 norms as the data fidelity term. They also assumed affine motions between different frames. The results are therefore not good when complex motions or illumination changes are involved.

Chen and Qi [9] recently proposed a single-frame HR image reconstruction method via low rank regularization. Jin et al. [14] designed a patch-based low rank matrix completion algorithm from the sparse representation of LR images. The main idea of these two papers is based on the assumption that each LR image is downsampled from a blurred and shifted HR image. However, these works assumed that the original HR image, when considered as a matrix, has a low rank property, which is not convincing in general.

In this paper, we show that the low rank property can in fact be constructed under MFSR framework. The idea is to consider each LR image as a downsampled instance of a *different* blurred and shifted HR image. Then when all these different HR images are properly aligned, they should give a low rank matrix; therefore, we

can use a low-rank prior to obtain a better solution. Many existing works assume that the shift between two consecutive LR frames is small, see, e.g., [1, 11, 22, 30, 31]. In this paper, we allow illumination changes and more complex motions other than affine transformation. They are handled by an optical flow model proposed in [13]. Once the motions are determined, we reconstruct the high-resolution image by minimizing a functional which consists of two terms: the 2-norm data fidelity term to suppress Gaussian noise and a nuclear-norm regularizer to enforce the low-rank prior. Tests on seven synthetic and real video clips show that our resulting images are more accurate with less artifacts. It can also provide much finer and discernable details.

The rest of the paper is organized as follows: Section 2 gives a brief review of a classical model on modeling LR images from HR images. Our model will be based on this model. Section 3 provides the details of our low-rank model, including image registration by optical flow and the solution of our optimization problem by alternating direction method. Section 4 gives experimental results on the test videos. Conclusions are given in Sect. 5.

To simplify our discussion, we now give the notation that we will be using in the rest of the paper. For any integer $m \in \mathbb{Z}$, I_m is the $m \times m$ identity matrix. For any integer $l \in \mathbb{Z}$ and positive integer $n \in \mathbb{Z}^+$, there exists a unique $0 \leq \widetilde{l} < n$ such that $\widetilde{l} \equiv l \mod n$. Let $N_n(l)$ denote the $n \times n$ matrix

$$N_n(l) = \begin{bmatrix} 0 & I_{n-\widetilde{l}} \\ I_{\widetilde{l}} & 0 \end{bmatrix}. \tag{1}$$

For a vector $\mathbf{f} \in \mathbb{R}^n$, $N_n(l)\mathbf{f}$ is the vector with entries of $\mathbf{f}$ cyclic-shifted by l.

Define the downsampling matrix D_i and the upsampling matrix D_i^T as

$$D_i(n) = I_n \otimes \mathbf{e}_i^T \text{ and } D_i^T(n) = I_n \otimes \mathbf{e}_i, \ i = 0, 1, \tag{2}$$

where $\mathbf{e}_0 = [1, 0]^T$, $\mathbf{e}_1 = [0, 1]^T$, and $\otimes$ is the Kronecker product. For $0 \leq \epsilon \leq 1$, define $T_n(\epsilon)$ to be the $n \times n$ circulant matrix

$$T_n(\epsilon) = \begin{bmatrix} 1-\epsilon & \epsilon & \cdots & 0 \\ 0 & 1-\epsilon & \ddots & \vdots \\ \vdots & \ddots & \ddots & \epsilon \\ \epsilon & \cdots & 0 & 1-\epsilon \end{bmatrix}. \tag{3}$$

This matrix performs the effect of linear interpolation shifted by ϵ.

For a matrix $X_{m\times n}$, the *nuclear norm* $\|\cdot\|_*$ of $X_{m\times n}$ is given by

$$\|X_{m\times n}\|_* = \sum_{i=1}^{r} |\sigma_i|,$$

where $\sigma_i, i = 1, 2, \cdots, r$ are *singular values* of $X_{m\times n}$.

2 Low-Resolution Model with Shifts

Consider a LR sensor array recording a video of an object. Then it gives multiple LR images of the object. Unless the object or the sensor array is completely motionless during the recording, the LR images will contain multiple information of the object at different shifted locations (either because of the motion of the object or of the sensor array itself). Our problem is to improve the resolution of one of the LR images (called the reference image) by incorporating information from the other LR images.

Let the sensor array consist of $m \times n$ sensing elements, where the width and the height of each sensing element are L_x and L_y, respectively. Then, the sensor array will produce an $m \times n$ discrete image with mn pixels, where each of these LR pixels is of size $L_x \times L_y$. Let r be the upsampling factor, i.e., we would like to construct an image of resolution $rm \times rn$ of the same scene. Then the size of the HR pixels will be $L_x/r \times L_y/r$. Figure 1a shows an example. The big rectangles with solid edges are the LR pixels and the small rectangles with dashed edges are the HR pixels.

Let $\{g_i \in \mathbb{R}^{m\times n}, 1 \le i \le p\}$ be the sequence of LR images produced by the sensor array at different time points, where p is the number of frames. For simplicity we let g_0 be the reference LR image which can be chosen to be any one of the LR images g_i. The displacement of g_i from the reference image g_0 is denoted by $(\epsilon_i^x L_x, \epsilon_i^y L_y)$, see the solid rectangle in Fig. 1a labeled as g_i. For ease of notation, we will represent the 2D images g_i, $0 \le i \le p$, by vectors $\mathbf{g}_i \in \mathbb{R}^{mn}$ obtained by stacking the columns of g_i. We use $\mathbf{f} \in \mathbb{R}^{r^2mn}$ to denote the HR reconstruction of g_0 that we are seeking.

We model the relationship between $\mathbf{f}$ and $\mathbf{g}_0$ by averaging, see [3, 8]. Figure 1b illustrates that the intensity value of the LR pixel is the weighted average of the

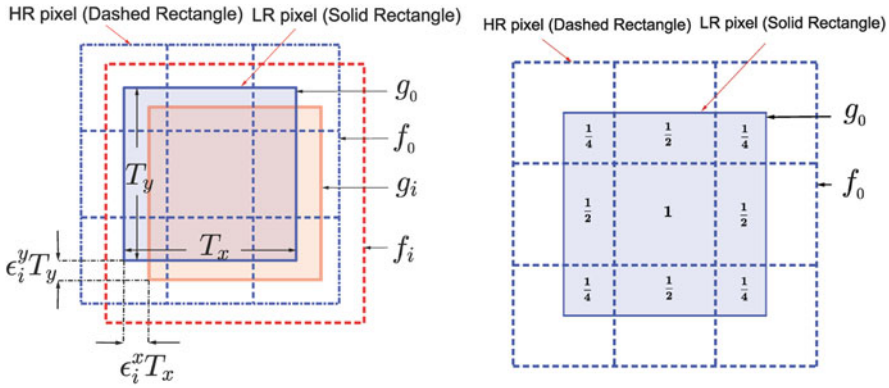


Fig. 1 LR images with displacements. (**a**) Displacements between LR images. (**b**) The averaging process

intensity values of the HR pixels overlapping with it. The weight is precisely the area of overlapping. Thus, the process from $\mathbf{f}$ to each of the LR images g_i can be modeled by [8]

$$\mathbf{g}_i = DKA_i\mathbf{f} + \mathbf{n}_i,\ i = 1, 2, \cdots, p, \tag{4}$$

where $D = D_0(n) \otimes D_0(m) \in \mathbb{R}^{mn \times r^2mn}$ is the downsampling matrix defined by (2); $K \in \mathbb{R}^{r^2mn \times r^2mn}$ is the average operator mentioned above; $A_i \in \mathbb{R}^{r^2mn \times r^2mn}$ is the warping matrix which measures the displacement between g_i and g_0; and $\mathbf{n}_i$ is the additive unknown noise. In this paper, we assume for simplicity that the noise is Gaussian. Other noise models can be handled by choosing suitable data fidelity terms.

The warping matrix A_i, $1 \leq i \leq p$, is to align the LR pixels in $\mathbf{g}_i$ at exactly the middle of the corresponding HR pixels in $\mathbf{f}$, exactly like the $\mathbf{g}_0$ is w.r.t $\mathbf{f}_0$ in Fig. 1b. Once this alignment is done, the average operator K, which is just a blurring operator, can be written out easily. In fact, the 2D kernel (i.e., the point spread function) of K is given by vv^T, where $v = [1/2, 1, \ldots, 1, 1/2]^T$ with $(r-1)$ ones in the middle, see [3]. The A_i are more difficult to obtain. In the most ideal case where the motions are only translation of less than one HR pixel length and width, A_i can be modeled by $A_i = T_n(\epsilon_i^x) \otimes T_m(\epsilon_i^y)$, where $T_n(\epsilon_i^x) and T_m(\epsilon_i^y)$ are the circulant matrices given by (3) with $(\epsilon_i^x L_x, \epsilon_i^y L_y)$ being the horizontal and vertical displacements of g_i, see Fig. 1a and [8]. In reality, the changes between different LR frames are much more complicated. It can involve illumination changes and other complex non-planar motions. We will discuss the formation of A_i in more detail in Sects. 3.1 and 3.3.

3 Nuclear-Norm Model

Given (4), a way to obtain $\mathbf{f}$ is to apply least-squares. However, because D is singular, the problem is ill-posed. Regularization is necessary to make use of some priori information to choose the correct solution. A typical regularizer for solving this problem is *total variation* (TV) [24]. The TV model is well known for edge preserving and can give a reasonable solution for MFSR problems. However, it assumes that the HR image is piecewise constant. This will produce some artifacts.

Instead we will develop a low-rank model for the problem. The main motivation is as follows: We consider each LR image $\mathbf{g}_i$ as a downsampled version of an HR image $\mathbf{f}_i$. If all these HR images $\mathbf{f}_i$ are properly aligned with the HR image $\mathbf{f}$, then they all should be the same exactly (as they are representing the same scene $\mathbf{f}$). W_i is the alignment matrix that aligns $\mathbf{f}_i$ with $\mathbf{f}$. For example, if $p = 2$, and

$$f_1 = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, f_2 = \begin{pmatrix} b \\ c \\ d \end{pmatrix},$$

then we can let

$$W_1 = \begin{pmatrix} 0\,1\,0 \\ 0\,0\,1 \end{pmatrix}, W_2 = \begin{pmatrix} 1\,0\,0 \\ 0\,1\,0 \end{pmatrix},$$

thence

$$W_1\mathbf{f}_1 = \begin{pmatrix} b \\ c \end{pmatrix} = W_2\mathbf{f}_2.$$

In general, $[W_1\mathbf{f}_1, W_2\mathbf{f}_2, \ldots, W_p\mathbf{f}_p]$ should be a low-rank matrix (ideally a rank 1 matrix). Thus, the rank of the matrix can be used as a prior.

In Sect. 3.1, we introduce our low-rank model in the case where the LR images are perturbed only by translations. Then in Sect. 3.2, we explain how to solve the model by the alternating direction method. In Sect. 3.3, we discuss how to modify the model when there are more complex motions or changes between the LR frames.

3.1 Decomposition of the Warping Matrices

In order to introduce our model without too cumbersome notations, we assume first here that the displacements of the LR images from the reference frame are translations only. Let $s_i^x L_x$ and $s_i^y L_y$ be the horizontal and vertical displacements of g_i from g_0. (How to obtain s_i^x and s_i^y will be discussed in Sect. 3.3.) Since the width and height of one HR pixel are L_x/r and L_y/r, respectively, the displacements are equivalent to rs_i^x HR pixel length and rs_i^y HR pixel width. We decompose rs_i^x and rs_i^y into the integral parts and fractional parts:

$$rs_i^x = l_i^x + \epsilon_i^x,\ rs_i^y = l_i^y + \epsilon_i^y, \tag{5}$$

where $l_i^x and l_i^y$ are the integers and $0 \le \epsilon_i^x, \epsilon_i^x < 1$. Then the warping matrix can be decomposed as

$$A_i = C_i B_i, \tag{6}$$

where $B_i = N_n(l_i^x) \otimes N_m(l_i^y)$ is given by (1) and $C_i = T_n(\epsilon_i^x) \otimes T_m(\epsilon_i^y)$ is given by (3) [6]. Thus, by letting $\mathbf{f}_i = B_i\mathbf{f}$, $1 \le i \le p$, (4) can be rewritten as

$$\mathbf{g}_i = DKC_i\mathbf{f}_i + \mathbf{n}_i,\ i = 1, 2, \cdots, p. \tag{7}$$

As mentioned in the motivation above, all these $\mathbf{f}_i$, which are equal to $B_i\mathbf{f}$, are integral shift from $\mathbf{f}$. Hence, if they are aligned correctly by an alignment matrix W_i, the overlapping entries should be the same. Figure 2 is the 1D illustration of this idea. W_i^x is the matrix that aligns $\mathbf{f}_i$ with $\mathbf{f}$ (in the x-direction) and the dark squares

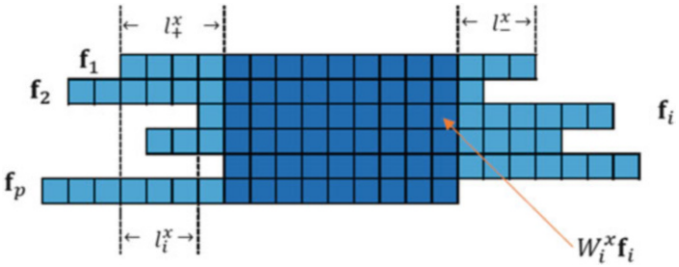


Fig. 2 1D signals with integer displacements

are the overlapping pixels and they should all be the same as the corresponding pixels in $\mathbf{f}$.

Mathematically, W_i is constructed as follows: Given the decomposition of rs_i^x and rs_i^y in (5), let $l_+^x = \max_i\{0, l_i^x\}$, $l_+^y = \max_i\{0, l_i^y\}$ and $l_-^x = \max_i\{0, -l_i^x\}$, $l_-^y = \max_i\{0, -l_i^y\}$. Then

$$W_i = W_i^x \otimes W_i^y, \tag{8}$$

where

$$W_i^x = \begin{bmatrix} 0_{l_+^x - l_i^x} & & \\ & I_{rn-l_+^x-l_-^x} & \\ & & 0_{l_-^x+l_i^x} \end{bmatrix},$$

$$W_i^y = \begin{bmatrix} 0_{l_+^y - l_i^y} & & \\ & I_{rm-l_+^y-l_-^y} & \\ & & 0_{l_-^y+l_i^y} \end{bmatrix}.$$

Note that W_i nullifies the entries outside the overlapping part (i.e., outside the dark squares in Fig. 2).

Ideally, the matrix $[W_1\mathbf{f}_1, W_2\mathbf{f}_2, \cdots, W_p\mathbf{f}_p]$ should be a rank-one matrix as every column should be a replicate of $\mathbf{f}$ in the overlapping region. In practice, it can be of low rank due to various reasons such as errors in measurements and noise in the given video. Since nuclear norm is the convexification of low-rank prior, see [5], this leads to our convex model

$$\min_{\mathbf{f}_1,\cdots,\mathbf{f}_p} \alpha\|W_1\mathbf{f}_1, W_2\mathbf{f}_2, \cdots, W_p\mathbf{f}_p\|_* + \frac{1}{2}\sum_{i=1}^{p}\|\mathbf{g}_i - DKC_i\mathbf{f}_i\|_2^2, \tag{9}$$

where $\|\cdot\|_*$ is the matrix nuclear norm and α is the regularization parameter. We call our model (9) the *nuclear-norm model*. We remark that here we use the 2-norm

data fidelity term because we assume the noise is Gaussian. It can be changed to another norm according to the noise type.

3.2 Algorithm for Solving the Nuclear-Norm Model

We use *alternating direction method of multipliers* (ADMM) [4] to solve the nuclear-norm model. We replace $\{W_i\mathbf{f}_i\}_{i=1}^p$ in the model by variables $\{\mathbf{h}_i\}_{i=1}^p$. Let $H = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p]$, $F = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_p]$, and $WF = [W_1\mathbf{f}_1, W_2\mathbf{f}_2, \cdots, W_p\mathbf{f}_p]$. The *augmented Lagrange* of model (9) is

$$\mathscr{L}_{\alpha\rho}(H, F, \Lambda) = \alpha\|H\|_* + \frac{1}{2}\sum_{i=1}^{p}\|\mathbf{g}_i - DKC_i\mathbf{f}_i\|_2^2 + \sum_{i=1}^{p}\langle\Lambda_i, \mathbf{h}_i - W_i\mathbf{f}_i\rangle + \frac{1}{2\rho}\|H - WF\|_{\mathscr{F}}^2,$$

where $\Lambda = [\Lambda_1, \Lambda_2, \cdots, \Lambda_p]$ is the matrix of Lagrange multipliers, $\|\cdot\|_{\mathscr{F}}$ is the Frobenius norm, and ρ is an algorithm parameter.

To solve the nuclear-norm model, it is equivalent to minimize $\mathscr{L}_{\alpha\rho}$, and we use ADMM [4] to minimize it. The idea of the scheme is to minimize H and F alternatively by fixing the other, i.e., given the initial value F^0, Λ^0, let $H^{k+1} = \arg\min_H \mathscr{L}_{\alpha\rho}(H, F^k, \Lambda^k)$ and $F^{k+1} = \arg\min_F \mathscr{L}_{\alpha\rho}(H^{k+1}, F, \Lambda^k)$, where k is the iteration number. These two problems are convex problems. The *singular value threshold* (SVT) gives the solution of the H-subproblem. The F-subproblem is reduced to solving p linear systems. For a matrix X, the SVT of X is defined to be

$$\mathrm{SVT}_\rho(X) = U\Sigma_\rho^+ V^T,$$

where $X = U\Sigma V^T$ is the singular value decomposition (SVD) of X and $\Sigma_\rho^+ = \max\{\Sigma - \rho, 0\}$. We summarize the algorithm in Algorithm 1. It is well-known that the algorithm is convergent if $\rho > 0$ [4].

In Algorithm 1, the SVT operator involves the SVD of a matrix $WF^k - \Lambda^k$. Its number of column is p, the number of LR frames, which is relatively small. Therefore, the SVT step is not time consuming. For the second subproblem, we need to solve p linear systems. The coefficient matrices contain some structures which help accelerating the calculation. The matrices $D^T D$ and $W_i^T W_i$ are diagonal matrices, while K and C_i can be diagonalized by either FFT or DCT depending on the boundary conditions we choose, see [23]. In our tests, we always use periodic boundary conditions.

In Algorithm 1, within each iteration, we should apply once singular value decomposition to an $r^2mn \times p$ matrix. The complexity of SVD is $O(r^2mnp^2)$. Then, we should solve p linear systems with r^2mn equations. By using FFT, the

Algorithm 1 $\mathbf{f} \leftarrow (\{g_i, W_i, C_i\}, K, \alpha, \rho, \Lambda^0, F^0)$

for $k = 1, 2, 3, \cdots$ **do**
  $H^{k+1} = \mathrm{SVT}_{\alpha\rho}(WF^k - \rho\Lambda^k)$;
  for $i = 1$ to p **do**
    $M_i = (DKC_i)^T DKC_i + \frac{1}{\rho} W_i^T W_i$;
    $\mathbf{f}_i^{k+1} = (M_i)^{-1}\left((DKC_i)^T \mathbf{g}_i + W_i^T \Lambda_i^k + \frac{1}{\rho} W_i^T \mathbf{h}_i^{k+1}\right)$;
  end for
  $\Lambda^{k+1} = \Lambda^k + \frac{1}{\rho}(H^{k+1} - WF^{k+1})$;
end for
Output: $\mathbf{f}$ as the average of the columns of F^k.

complexity for this step is $O(pr^2mn\log(r^2mn))$. Usually, $\log(r^2mn)$ is larger than p. Thence, the overall complexity for Algorithm 1 is $O(pr^2mn\log(r^2mn))$, where $m \times n$ is the size of LR images; r is the upsampling factor; and p is the number of frames.

3.3 Image Registration and Parameter Selection

In Algorithm 1, we assume that there are only translations between different LR frames. However, there can be other complex motions and/or illumination changes in practice. We handle these by using the *local all-pass* (LAP) optical flow algorithm proposed in [13]. Given a set of all-pass filters $\{\phi_j\}_{j=0}^N$ and $\phi := \phi_0 + \sum_{j=1}^{N-1} c_j\phi_j$, the optical flow $\mathscr{M}_i$ of g_i is obtained by solving the following problem:

$$\min_{\{c_1,\cdots,c_{N-1}\}} \sum_{l,k\in R} |(\phi * g_i)(k,l) - (\phi_- * g_0)(k,l)|^2,$$

where $*$ is the convolution operator, R is a window centered at (x, y), and $\phi_-(k,l) = \phi(-k,-l)$. In our experiments, we followed the settings in the paper [13], and let $N = 6$, $R = 16$ and

$$\begin{aligned}
\phi_0(k,l) &= e^{-\frac{k^2+l^2}{2\sigma^2}}, & \phi_1(k,l) &= k\phi_0(k,l),\\
\phi_2(k,l) &= l\phi_0(k,l), & \phi_3(k,l) &= (k^2+l^2-2\sigma^2)\phi_0(k,l),\\
\phi_4(k,l) &= kl\phi_0(k,l), & \phi_5(k,l) &= (k^2-l^2)\phi_0(k,l),
\end{aligned}$$

where $\sigma = \frac{R+2}{4}$ and ϕ is supported in $[-R, R] \times [-R, R]$. The coefficients c_n can be obtained by solving a linear system. The optical flow $\mathscr{M}_i$ at (x, y) is then given by

$$\mathscr{M}_i(x,y) = \left(\frac{2\sum_{k,l} k\phi(k,l)}{\sum_{k,l}\phi(k,l)}, \frac{2\sum_{k,l} l\phi(k,l)}{\sum_{k,l}\phi(k,l)}\right),$$

which can be used to transform g_i back to the grid of g_0. In order to increase the speed by avoiding interpolation, here we consider only the integer part of the flow. Hence, we get the *restored LR images*

$$\tilde{g}_i(x, y) = g_i([\mathscr{M}_i](x, y)),\ i = 1, 2, \cdots, p, \quad \forall (x, y) \in \Omega, \tag{10}$$

where $[\mathscr{M}_i]$ is the integer part of the flow $\mathscr{M}_i$ and Ω is the image domain.

The optical flow method can handle complex motions and illumination changes and will restore the positions of pixels in g_i w.r.t g_0. To enhance the accuracy of the image registration, we further estimate if there are any translations that are unaccounted for after the optical flow. In particular, we assume that $\tilde{g}_i$ may be displaced from g_0 by a simple translation

$$\mathscr{T}(x, y) = \begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} s_i^x \\ s_i^y \end{bmatrix}. \tag{11}$$

To estimate the displacement vector $[s_i^x, s_i^y]^T$, we use the Levenberg–Marquardt algorithm proposed in [15, 21], which is a well-known method for nonlinear least-squares problems. It aims to minimize the squared error

$$E(\tilde{g}_i, g_0) = \sum_{(x,y)\in\Omega} [\tilde{g}_i(\mathscr{T}(x, y)) - g_0(x, y)]^2. \tag{12}$$

The detailed implementation of this algorithm can be found in [8, Algorithm 3]. After obtaining $[s_i^x, s_i^y]$, then by (6) and (8), we can construct the matrices C_i and W_i for our nuclear-norm model (9).

Before giving out the whole algorithm, there remains the problem about parameters selection. There are two parameters to be determined: α, the regularization parameter, and ρ, the algorithm (ADMM) parameter. We need to tune these two parameters in practice such that the two subproblems can be solved effectively and accurately. Theoretically, ρ will not affect the minimizer of the model but only the convergence of the algorithm [4]. However, in order to get an effective algorithm, it should not be set very small. For α, we use the following empirical formula to approximate it in each iteration [16],

$$\alpha \approx \frac{1/2 \sum_{i=1}^{p} \|\widetilde{\mathbf{g}}_i - DKC_i\mathbf{f}_i^k\|^2}{\|W_1\mathbf{f}_1^k, W_2\mathbf{f}_2^k, \cdots, W_p\mathbf{f}_p^k\|_*}, \tag{13}$$

where $\mathbf{f}_i^k$ is the estimation of $\mathbf{f}_i$ in the kth iteration. The formula may not give the best α but can largely narrow its scope. We then use trial and error to get the best parameter. We give out the full algorithm for our model below.

4 Numerical Experiments

In this section, we illustrate the effectiveness of our algorithm by comparing it with 3 different variational methods on 7 synthetic videos and real videos. Chan et al. [6] applied wavelet analysis to MFSR problem and then developed an iterative approach by using tight-frame wavelet filters [8]. We refer their model as *tight-frame* (TF) model. Li et al. [16] proposed the *sparse directional regularization* (SDR) model where they used l_1 norm of the geometric tight-framelet coefficients as the regularizer and the adaptively mimicking l_1 and l_2 norms as the data fidelity term. Ma et al. [20] introduced an expectation-maximization (EM) framework to the Bayesian approach of Liu and Sun [18]. They also tackled motion blur in their paper. We refer it as the MAP model. We will compare our Algorithm 2 (the nuclear-norm model) with these three methods. The sizes of the videos we used are listed in Table 1. The CPU timing of all methods is also listed. It shows that our method is the fastest, with two exceptions (i.e., the "disk" video when $r = 2$ and the "text" video when $r = 2$). For other instances, our model is the best. We marked the fastest results with bold letters. These data show that, when dealing with small-size images, the SDR model is the fastest. When the size of the images gets larger, our nuclear-norm model is the fastest.

There is one parameter for the TF model—a thresholding parameter η which controls the registration quality of the restored LR images $\tilde{g}_i$ (see (10)). If the PSNR value between $\tilde{g}_i$ and the reference image g_0 is smaller than η, it will discard $\tilde{g}_i$ in the reconstruction. We apply *trial and error* method to choose the best η. For the SDR method, we use the default setting in the paper [16]. Hence, the parameters are selected automatically by the algorithm. The TF model, the SDR model, and the nuclear-norm model are applied to $\tilde{g}_i$, i.e., we use the same optical flow algorithm [13] for these three models. For the MAP model, it utilized an optical flow algorithm from Liu [17]. Following the paper, the optical flow parameter α is very small. We also apply *trial and error* method to tune it.

All the videos used in the tests and the results are available at http://www.math.cuhk.edu.hk/~rchan/paper/super-resolution/experiments.html.

Algorithm 2 $\mathbf{f} \leftarrow (\{g_i\}, i_0, K, \Lambda^0, F^0, \alpha, \rho)$

for $i = 0, 1, 2, \cdots p$ **do**
  Compute $\widetilde{g}_i(x, y)$ from (10);
  Compute s_i^x and s_i^y in (11) by using the Levenberg–Marquardt algorithm in [8, Algorithm 3]
  Compute the warping matrices C_i and W_i, according to (6) and (8);
end for
Apply Algorithm 1 to compute the HR images $\mathbf{f} \leftarrow (\{\widetilde{g}_i, W_i, C_i\}, K, \alpha, \rho, \Lambda^0, F^0)$;
Output **f**.

Table 1 Size of each data set and CPU time for all models

	Size of data			Factor	CPU time (in seconds)			
	Height	Width	Frame	r	TF	MAP	SDR	Nuclear
Boat	240	240	17	2	1251	198	336	**138**
Boat	120	120	17	4	7642	196	282	**94.4**
Bridge	240	240	17	2	3256	202	348	**142**
Bridge	120	120	17	4	9703	189	278	**92.3**
Disk	57	49	19	2	568	**6.4**	28	7.9
Disk	57	49	19	4	5913	21.4	53	**13.6**
Text	57	49	21	2	497	**6.2**	30	**8.5**
Text	57	49	21	4	4517	22.1	56	**14.5**
Alpaca	96	128	21	2	816	26.1	78	**24**
Alpaca	96	128	21	4	6178	172	250	**90.6**
Books	288	352	21	2	3943	1511	818	**689**

4.1 Synthetic Videos

We start from a given HR image $\mathbf{f}^*$, see, e.g., the boat image in Fig. 3f. We translate and rotate $\mathbf{f}^*$ with known parameters and also change their illuminations by different scales. Then we downsample these frames with the given factor $r = 2$ or $r = 4$ to get our LR frames $\{\mathbf{g}_i\}_{i=1}^{p}$. We take $p = 17$, and Gaussian noise of ratio 5% is added to each LR frame.

After we reconstruct the HR image $\mathbf{f}$ by a method, we compare it with the true solution $\mathbf{f}^*$ using two popular error measurements. The first one is *peak signal-to-noise ratio* (PSNR) and the second one is *structural similarity* (SSIM) [29]. For two signals $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$, they are defined by

$$\text{PSNR}(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \left(\frac{d^2}{\|\mathbf{x} - \mathbf{y}\|^2 / n} \right),$$

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

where d is the dynamic range of $\mathbf{x}, \mathbf{y}$; $\mu_x and \mu_y$ are the mean values of $\mathbf{x}$ and $\mathbf{y}$; $\sigma_x and \sigma_y$ are the variances of $\mathbf{x}$ and $\mathbf{y}$; σ_{xy} is the covariance of $\mathbf{x}$ and $\mathbf{y}$; and $c_i, i = 1, 2$, are the constants related to d, which are typically set to be $c_1 = (0.01d)^2 and c_2 = (0.03d)^2$. Because of the motions, we do not have enough information to reconstruct $\mathbf{f}$ near the boundary; hence, this part of $\mathbf{f}$ will not be accurate. Thus, we restrict the comparison within the overlapping area of all LR images.

Table 2 gives the PSNR values and SSIM values of the reconstructed HR images $\mathbf{f}$ from the boat and the bridge videos. The results show that our model gives much more accurate $\mathbf{f}$ for both upsampling factor $r = 2$ and 4, see the boldfaced values.

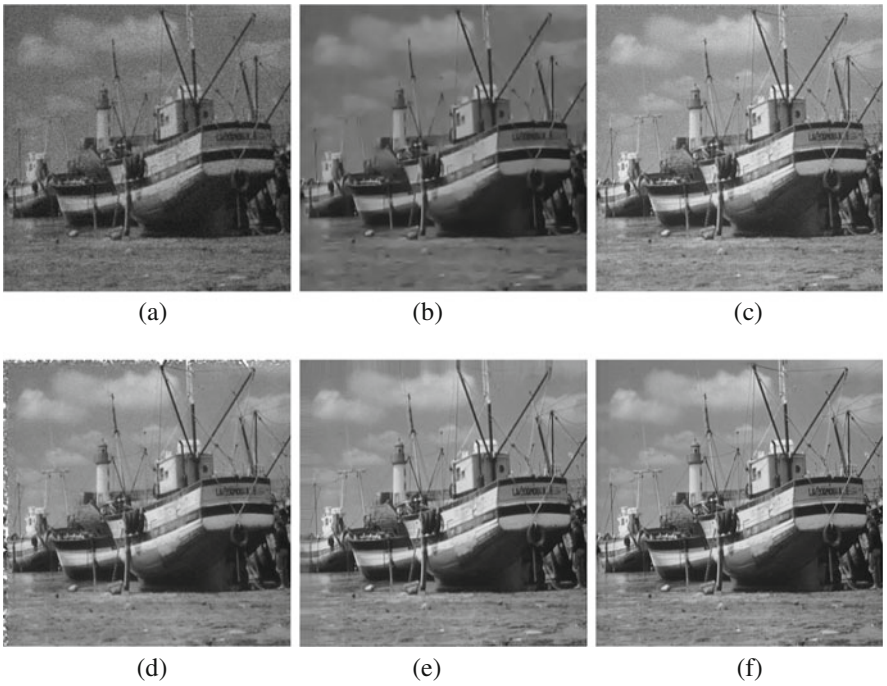


Fig. 3 Comparison of different algorithms on "boat" image with upsampling factor $r = 2$. (**a**) The reference LR image. (**b**) Result of the TF model [8]. (**c**) Result of the MAP model [20]. (**d**) Result of the SDR model [16]. (**e**) Result of our nuclear-norm model ($\alpha = 35.2924$ and $\rho = 3.379 \times 10^4$). (**f**) True HR image

Table 2 PSNR and SSIM values for the "boat" and "bridge" videos

		Upsampling factor $r = 2$				Upsampling factor $r = 4$			
		TF	MAP	SDR	Nuclear	TF	MAP	SDR	Nuclear
Boat	PSNR	18.7	25.3	28.2	**29.8**	20.7	23.6	27.0	**27.1**
	SSIM	0.69	0.70	0.80	**0.82**	0.69	0.67	0.72	**0.76**
Bridge	PSNR	20.7	23.6	**27.0**	26.9	20.1	22.4	24.6	**24.9**
	SSIM	0.69	0.67	0.72	**0.80**	0.53	0.57	0.65	**0.70**

The improvement is significant when comparing to the other three models, e.g., at least 1.6 dB in PSNR for the boat video when $r = 2$. All the PSNR values and SSIM values of our method for boat video are higher than that of other models. All the PSNR values and SSIM values of our method for bridge video are higher than that of other models except the PSNR value when $r = 2$, see the fifth column of the last row. It is comparable with the SDR method. However, the SSIM value is higher. This means the reconstructed structure is better for our method. The major cost of this algorithm is to solve the $\mathbf{f}_i$ subproblems in Algorithm 1. Since the resulting images

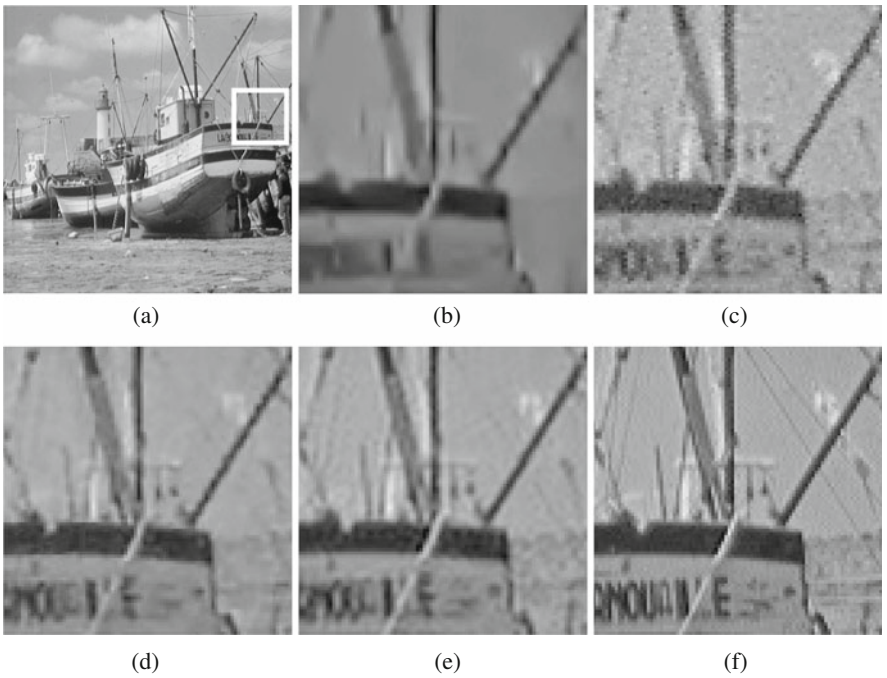


Fig. 4 Zoomed-in comparison of different algorithms on "boat" image for $r = 2$. (**a**) The zoom-in part in the HR image. (**b**) Result of the TF model [8]. (**c**) Result of the MAP model [20]. (**d**) Result of the SDR model [16]. (**e**) Result of our nuclear-norm model ($\alpha = 35.2924$ and $\rho = 3.379 \times 10^4$). (**f**) Zoomed-in original HR image

are with larger sizes, the sizes of coefficients of all subproblems in Algorithm 1 are larger. Thence, when $r = 4$, the cost is larger than that when $r = 2$.

To compare the images visually, we give the results and their zoom-ins for the boat video in Figs. 3, 4, 5. The results for the bridge video are similar and therefore omitted. Figure 3 shows the boat reconstructions for $r = 2$. We notice that the TF model loses many fine details, e.g., the ropes of the mast. The MAP model produces some distortion on the edges and is sensitive to the noise; and the SDR model contains some artifacts along the edges. One can see the difference more clearly from the zoom-in images in Fig. 4. We also give the zoom-in results for $r = 4$ in Fig. 5. We can see that the nuclear-norm model produces more details and less artifacts than the other three models.

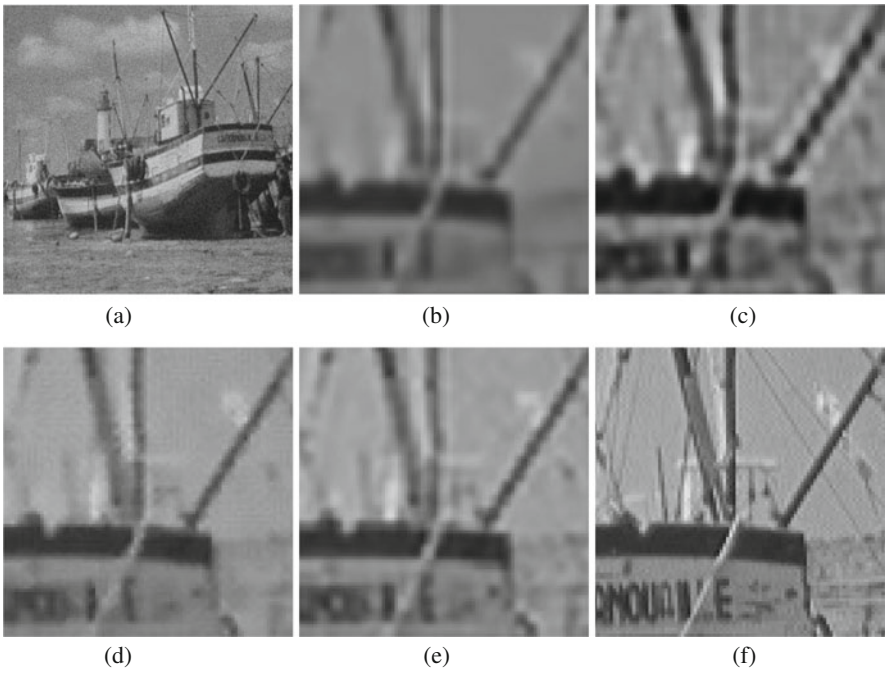


Fig. 5 Zoom-in comparison of different algorithms on "boat" image for $r = 4$. (**a**) The reference LR image. (**b**) Result of the TF model [8]. (**c**) Result of the MAP model [20]. (**d**) Result of the SDR model [16]. (**e**) Result of our nuclear-norm model ($\alpha = 32.0659$ and $\rho = 3.5841 \times 10^4$). (**f**) Zoomed-in original HR image

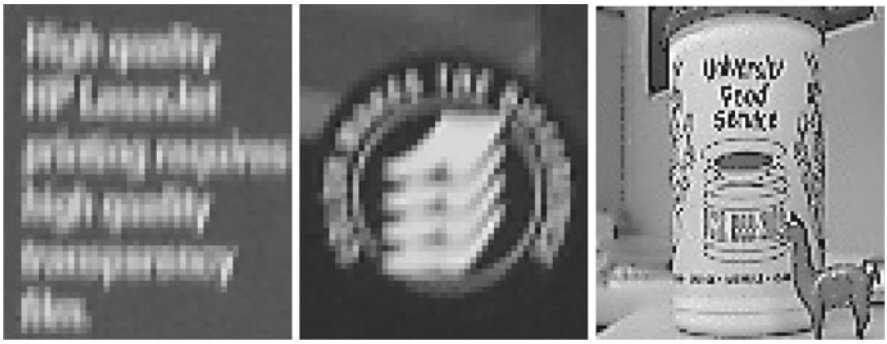
4.2 Real Videos

In the following, experiments on real videos are carried out. Three videos "text," "disk," and "alpaca" are downloaded from the website https://users.soe.ucsc.edu/~milanfar/software/sr-datasets.html.

The basic information of these videos are listed in Table 1. We see that they are very low-resolution videos. Figure 6 shows the reference LR images for these videos. It is difficult to discern most of the letters from the reference images.

The first test video is the "text video." The results are shown in Fig. 7. We see that the TF model produces blurry reconstructions. The images by the MAP model have obvious distortions. We also see that for the SDR model, some of the letters are coalesced, e.g., the word "film." The results of the nuclear-norm model are better. One can easily tell each word and there are no obvious artifacts for the letters.

The second video is the "disk video," which contains 26 gray-scale images with the last 7 ones being zoom-in images. Therefore, we only use the first 19 frames in our experiment. The results are shown in Fig. 8. The TF model again produces blurry reconstructions. The MAP results are better but still blurry. The SDR results



(a) (b) (c)

Fig. 6 The reference LR images of (**a**) "text," (**b**) "disk," and (**c**) "alpaca"

(a) (b) (c) (d)

Fig. 7 Comparison of different algorithms on "text video." Top row with upsampling factor $r = 2$ and second row with $r = 4$. (**a**) Result of the TF model [8]. (**b**) Result of the MAP model [20]. (**c**) Result of the SDR model [16]. (**d**) Result of our nuclear-norm model ($\alpha = 8.368$ and $\rho = 3.6236 \times 10^6$ for $r = 2$; $\alpha = 8.6391$ and $\rho = 4.5618 \times 10^5$ for $r = 4$)

have some artifacts, especially in the word "DIFFERENCE." Our results are the best ones with each letter being well reconstructed, especially when $r = 2$.

The third video is the "alpaca video," and the results are shown in Fig. 9. When $r = 2$, the word "service" is not clear from the TF model, the MAP model, and the SDR model. When $r = 4$, the resulting images from all models are improved and the phrase "university food service" is clear. However, we can see that our nuclear-norm model still gives the best reconstruction.

(a) (b) (c) (d)

Fig. 8 Comparison of different algorithms on "disk video." Top row with upsampling factor $r = 2$ and second row with $r = 4$. (**a**) Result of the TF model [8]. (**b**) Result of the MAP model [20]. (**c**) Result of the SDR model [16]. (**d**) Result of our nuclear-norm model ($\alpha = 6.6802$ and $\rho = 1.0701 \times 10^6$ for $r = 2$; $\alpha = 11.6185$ and $\rho = 8.6404 \times 10^5$ for $r = 4$)

(a) (b) (c) (d)

Fig. 9 Comparison of different algorithms on "alpaca video." Top row with upsampling factor $r = 2$ and second row with $r = 4$. (**a**) Result of the TF model [8]. (**b**) Result of the MAP model [20]. (**c**) Result of the SDR model [16]. (**d**) Result of our nuclear-norm model ($\alpha = 35.3704$ and $\rho = 2.7892 \times 10^4$ for $r = 2$; $\alpha = 45.6486$ and $\rho = 2.9798 \times 10^5$ for $r = 4$)

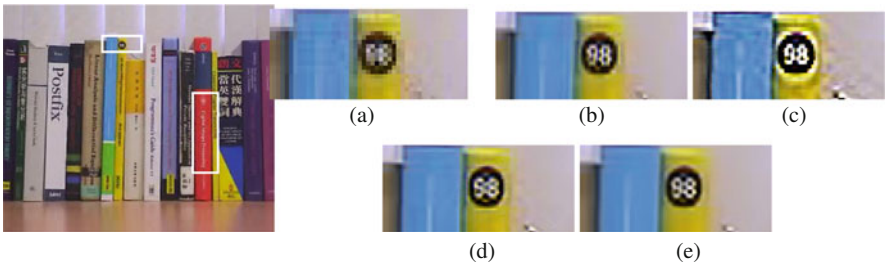


Fig. 10 Zoom-in comparison of different algorithms on "books video" with $r = 2$. Leftmost figure: the LR reference frame with zoom-in areas marked. (**a**) Zoomed-in LR image. (**b**) Result of the TF model [8]. (**c**) Result of the MAP model [20]. (**d**) Result of the SDR model [16]. (**e**) Result of our nuclear-norm model ($\alpha = 15.3958$ and $\rho = 5.6858 \times 10^5$ for $r = 2$)

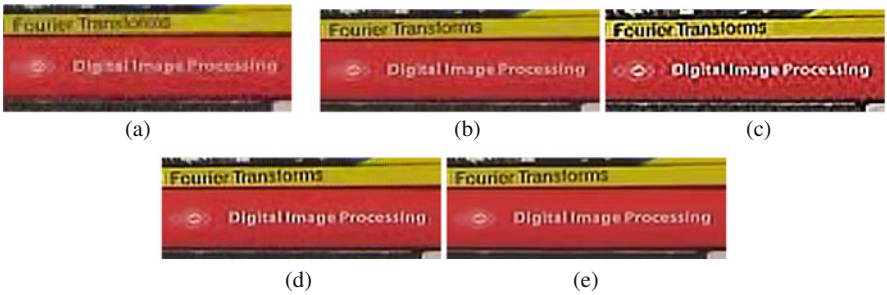


Fig. 11 Another zoom-in comparison on "books video" with $r = 2$. (**a**) Zoomed-in LR image. (**b**) Result of the TF model [8]. (**c**) Result of the MAP model [20]. (**d**) Result of the SDR model [16]. (**e**) Result of our nuclear-norm model ($\alpha = 15.3958$ and $\rho = 5.6858 \times 10^5$ for $r = 2$)

The last video is a color video which is used in the tests in [7, 8]. It contains 257 color frames. We take the 100th frame to be the reference frame, see the leftmost figure in Fig. 10. Frames 90–110 in the video are used as LR images to enhance the reference image. We transform the RGB images into the Ycbcr color space and then apply the algorithms to each color channel. Then we transform the resulting HR images back to the RGB color space. Figures 10 and 11 show the zoom-in patches of the resulting images by different models. In Fig. 10, the patch shows a number "98" on the spine of a book. We see that the TF model gives a reasonable result when compared with MAP and SDR. However, our nuclear-norm model gives the clearest "98" with very clean background. Figure 11 shows the spines of two other books: "Fourier Transforms" and "Digital Image Processing." Again, we see that our nuclear-norm model gives the best reconstruction of the words with much less noisy artifacts.

5 Conclusion

In this paper, we proposed an effective algorithm to reconstruct a high-resolution image using multiple low-resolution images from video clips. The LR images are first registered to the reference frame by using an optical flow. Then a low-rank model is used to reconstruct the high-resolution image by making use of the overlapping information between different LR images. Our model can handle complex motions and illumination changes. Tests on synthetic and real videos show that our model can reconstruct an HR image with much more details and less artifacts.

Acknowledgements This work was supported by HKRGC Grants Nos. CUHK14306316, HKRGC CRF Grant C1007-15G, and HKRGC AoE Grant AoE/M-05/12.

References

1. Altunbasak, Y., Patti, A., Mersereau, R.: Super-resolution still and video reconstruction from mpeg-coded video. IEEE Trans. Circuits Syst. Video Technol. **12**(4), 217–226 (2002)
2. Bishop, C.M., Blake, A., Marthi, B.: Super-resolution enhancement of video. In: Proc. Artificial Intelligence and Statistics, vol. 2. Key West, FL, USA (2003)
3. Bose, N., Boo, K.: High-resolution image reconstruction with multisensors. Int. J. Imaging Syst. Technol. **9**(4), 294–304 (1998)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
5. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(3), 11 (2011)
6. Chan, R.H., Chan, T.F., Shen, L., Shen, Z.: Wavelet algorithms for high-resolution image reconstruction. SIAM J. Sci. Comput. **24**(4), 1408–1432 (2003)
7. Chan, R.H., Riemenschneider, S.D., Shen, L., Shen, Z.: Tight frame: an efficient way for high-resolution image reconstruction. Appl. Comput. Harmon. Anal. **17**(1), 91–115 (2004)
8. Chan, R.H., Shen, Z., Xia, T.: A framelet algorithm for enhancing video stills. Appl. Comput. Harmon. Anal. **23**(2), 153–170 (2007)
9. Chen, X., Qi, C.: A single-image super-resolution method via low-rank matrix recovery and nonlinear mappings. In: 20th IEEE International Conference on Image Processing (ICIP), 2013, pp. 635–639 (2013). https://doi.org/10.1109/ICIP.2013.6738131
10. Duponchel, L., Milanfar, P., Ruckebusch, C., Huvenne, J.P.: Super-resolution and Raman chemical imaging: from multiple low resolution images to a high resolution image. Anal. Chim. Acta **607**(2), 168–175 (2008)
11. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Trans. Image Process. **13**(10), 1327–1344 (2004)
12. Farsiu, S., Elad, M., Milanfar, P.: Multiframe demosaicing and super-resolution of color images. IEEE Trans. Image Process. **15**(1), 141–159 (2006)
13. Gilliam, C., Blu, T.: Local all-pass filters for optical flow estimation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, IEEE (2015)
14. Jin, C., Nunez-Yanez, J., Achim, A.: Video super-resolution using low rank matrix completion. In: 20th IEEE International Conference on Image Processing (ICIP), 2013, Melbourne, Australia, pp. 1376–1380. (2013)

15. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Q. Appl. Math. **2**, 164–168 (1944)
16. Li, Y.R., Dai, D.Q., Shen, L.: Multiframe super-resolution reconstruction using sparse directional regularization. IEEE Trans. Circuits Syst. Video Technol. **20**(7), 945–956 (2010)
17. Liu, C.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Citeseer (2009)
18. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 346–360 (2014)
19. Lu, Y., Shen, L., Xu, Y.: Multi-parameter regularization methods for high-resolution image reconstruction with displacement errors. IEEE Trans. Circuits Syst. Regul. Pap. **54**(8), 1788–1799 (2007)
20. Ma, Z., Liao, R., Tao, X., Xu, L., Jia, J., Wu, E.: Handling motion blur in multi-frame super-resolution. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5224–5232 (2015). http://dx.doi.org/10.1109/CVPR.2015.7299159
21. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. **11(2)**, 431–441 (1962)
22. Narayanan, B., Hardie, R.C., Barner, K.E., Shao, M.: A computationally efficient super-resolution algorithm for video processing using partition filters. IEEE Trans. Circuits Syst. Video Technol. **17**(5), 621–634 (2007)
23. Ng, M.K., Chan, R.H., Tang, W.C.: A fast algorithm for deblurring models with Neumann boundary conditions. SIAM J. Sci. Comput. **21**(3), 851–866 (1999)
24. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1), 259–268 (1992)
25. Shankar, P.M., Neifeld, M.A.: Sparsity constrained regularization for multiframe image restoration. JOSA A **25**(5), 1199–1214 (2008)
26. Shen, L., Sun, Q.: Biorthogonal wavelet system for high-resolution image reconstruction. IEEE Trans. Signal Process. **52**(7), 1997–2011 (2004)
27. Takeda, H., Milanfar, P., Protter, M., Elad, M.: Super-resolution without explicit subpixel motion estimation. IEEE Trans. Image Process. **18**(9), 1958–1975 (2009)
28. Tsai, R., Huang, T.: Multiframe image restoration and registration. Adv. Comput. Vis. Image Process. **1**(2), 317–339 (1984)
29. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). http://dx.doi.org/10.1109/TIP.2003.819861
30. Wang, C., Xue, P., Lin, W.: Improved super-resolution reconstruction from video. IEEE Trans. Circuits Syst. Video Technol. **16**(11), 1411–1422 (2006)
31. Zibetti, M.V.W., Mayer, J.: A robust and computationally efficient simultaneous super-resolution scheme for image sequences. IEEE Trans. Circuits Syst. Video Technol. **17**(10), 1288–1300 (2007)