Rolling Shutter Super-Resolution

Abhijith Punnappurath, Vijay Rengarajan, and Rajagopalan A.N.
Department of Electrical Engineering
Indian Institute of Technology Madras, Chennai, India
jithuthatswho@gmail.com, {vijay.ap,raju}@ee.iitm.ac.in

Abstract

Classical multi-image super-resolution (SR) algorithms, designed for CCD cameras, assume that the motion among the images is global. But CMOS sensors that have increasingly started to replace their more expensive CCD counterparts in many applications do not respect this assumption if there is a motion of the camera relative to the scene during the exposure duration of an image because of the row-wise acquisition mechanism. In this paper, we study the hitherto unexplored topic of multi-image SR in CMOS cameras. We initially develop an SR observation model that accounts for the row-wise distortions called the "rolling shutter" (RS) effect observed in images captured using non-stationary CMOS cameras. We then propose a unified RS-SR framework to obtain an RS-free high-resolution image (and the row-wise motion) from distorted low-resolution images. We demonstrate the efficacy of the proposed scheme using synthetic data as well as real images captured using a hand-held CMOS camera. Quantitative and qualitative assessments reveal that our method significantly advances the state-of-the-art.

1. Introduction

With the evergrowing capabilities of high-resolution displays, super-resolution (SR) continues to be an active area of research as it offers a signal processing solution to the inherent problem of resolution limitation in low-cost imaging sensors (e.g., cell phone cameras). The goal of multi-image SR methods is to recover a high-resolution (HR) image from a set of low-resolution (LR) input images. The basic principle being that changes in the LR images caused by the motion of the camera and/or the scene provides additional information that can be utilized to reconstruct the HR image. Based on prior knowledge about the observation model that maps the HR image to the LR ones, multi-image SR algorithms attempt to recover the HR image under the constraint that the recovered image should result in the observed LR images upon applying the same model. Classical SR algorithms formulate this observation model so as to explain the image formation process in CCD cameras where the data is acquired by the sensor array all at once during the exposure time. Hence, CCD cameras are also called global shutter (GS) cameras because all elements in the sensor array are exposed at the same time. However, cameras employing CMOS sensors have a significantly different capture mechanism as compared to GS cameras. Each row in the CMOS sensor array has its own unique exposure duration, and the acquired data is read out at different times using a common circuit. As a result, the amount of circuitry, and thereby the cost, is reduced. In fact, this lower cost is the reason for the increasing popularity of CMOS sensors in many imaging applications.

Traditional GS-SR algorithms [4, 5] assume that the camera is stationary *during* the exposure time itself, and that the motion is only between one LR image capture to the next i.e., there is no blur in any of the captured images. Therefore, SR methods usually include two parts: (i) registration, where the motion *between* LR images is estimated, and (ii) image reconstruction, where the HR image is recovered from the LR images. However, camera shake is a common occurrence in hand-held imaging devices such as cell phones which have now become ubiquitous. Motivated by this fact, recent works [16, 20] address the more general situation of camera motion during the exposure time of a single image itself. This manifests as blur in the LR captured images for a GS camera, and the problem becomes one of blind deconvolution and super-resolution. However, in the case of CMOS cameras, the task of SR becomes significantly more challenging because motion during exposure leads to what is called the 'rolling shutter' (RS) effect[1] (see Fig. 1). This distortion of the captured image results from each row of sensors observing a different warp of the scene, and is a direct consequence of the row-wise acquisition principle in CMOS cameras. Clearly, the classical GS motion estimation step that assumes a global motion be-

[1]For large motion, the observed images can also incur blurring in addition to being RS affected. However, analysing such a compounded scenario is beyond the scope of this paper.

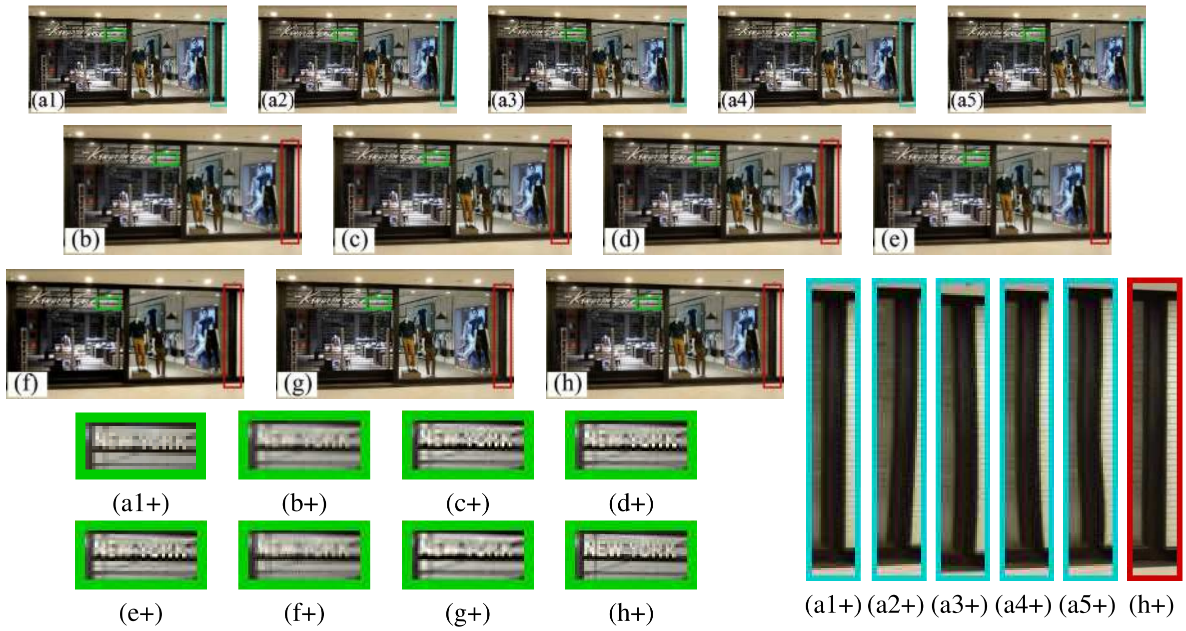

Figure 1. (a1) An image captured using a stationary CMOS camera, while (a2-a5) have RS effect due to motion during exposure. (b-h) are super-resolved (by a factor of two) outputs of various methods for the input LR image(s) (a1-a5). (b) Bicubic interpolation, (c) Glasner et al. [7], (d) Yang et al. [19], (e) Zhu et al. [21], (f) Ringaby and Forssén [15] + TV-SAR [18], (g) Ringaby and Forssén [15] + CDSR [17], and (h) our result. The notation '+' refers to zoomed-in patches. For display, the LR patches have been scaled up by a factor of two to match the size of the HR patches. The distortion of the vertical pillar is evident from the blue LR patches (a2+) to (a5+), while our red output patch (h+) is RS-free. It can be observed from the green patches that the text "New York" is clearly readable in our SR result (h+).

tween two LR images is not valid for RS cameras because the motion can vary from one row to the next. In this work, we develop an RS-SR observation model that incorporates the row-wise acquisition of data and a joint SR framework to recover the HR image from RS affected LR images.

We present below a brief overview of relevant works in super-resolution and rolling shutter.

Super-resolution: A large number of papers have addressed the classical multi-image SR problem for GS cameras when the LR images are not blurred. A good survey can be found in [13]. In general, spatial-domain SR approaches are more popular than their frequency-domain counterparts for their ability to incorporate complex image priors [13]. A robust SR framework based on the use of the l_1-norm both in the regularization and the measurement terms of the penalty function was proposed in [5]. A variational Bayesian formulation for joint image registration and SR has been mooted in [1]. This work was later improved in [18] using a combination of total variation and simultaneous auto regressive priors. A coordinate-descent approach for simultaneously obtaining registration and super-resolution from a sequence of LR images has been proposed in [17].

Example-based SR (also termed "image hallucination") techniques [19, 21] that seek an HR image from a *single* LR image have also been proposed. These methods learn correspondences between LR and HR image patches from a database of LR and HR image pairs, and then apply them to a new LR image to recover its most likely HR version. However, these techniques are known to *hallucinate* HR details that may not be present in the true HR image. Based on the observation that patches in a natural image tend to recur within the same image, both at the same as well as at different scales, [7] sought to combine the strengths of both classical as well as example-based SR. We would like to point out that all the above methods assume that the LR images are captured using a stationary camera, and there is no motion during exposure. [16, 20] relax this constraint and allow for camera motion during the exposure duration of individual images. While [16] limit the motion to pure in-plane translational motion and use the convolution model for blur, the method in [20] is based on the projective motion blur model for general camera motion. It is important to note that while both these methods allow for motion during capture, they limit their discussion to fronto-parallel scenes. LR images being related to each other through local motion was studied in [3, 9] in the context of a 3D scene.

Rolling Shutter: The study of the RS effect is on the rise during recent years. This effect is considered as an artifact and the main goal of existing works is to remove this undesired effect. To this end, removing RS effect from a video is a commonly dealt application. In [10], a global motion was modeled between pairs of frames in a video using motion vectors based on photometric energy and the motion for every row was refined through interpolation on a Bézier curve. In [2], frame-to-frame camera motion was considered as a low frequency component and row-wise motion as a high frequency component, and the motion was estimated by l_1 regularization. Key rows were defined within a frame in [6, 15] using which the motion of other rows are linearly interpolated. The camera motion was modeled as 3D rotations for general scenes with the assumption of known camera intrinsics. A homography mixture model was proposed in [8] in which the homography of each row was modeled as a weighted combination of homographies of fixed blocks of rows within an image. Both RS effect and motion blur were handled in [12] assuming a uniform camera velocity, and a minimization problem was posed based on photometric intensities. RS and motion blur artifacts for the application of change detection has been addressed in [14].

1.1. Contributions

1. To the best of our knowledge, this is the first attempt of its kind for the task of SR in CMOS cameras. Our work neatly generalizes the largely prevailing single global warp notion for SR to multiple local warps; the local warps being confined to rows due to the very nature of the acquisition process.
2. We propose an RS-SR observation model that explains the image formation process in CMOS cameras. The model also reveals how blocks of rows in the HR image experience different motions as a result of the row-wise capture mechanism.
3. Given multiple LR images that are RS affected, we develop a unified framework that alternates between solving for the HR image and the row-wise motion to obtain an undistorted and super-resolved image.

2. RS-SR image formation model

We first briefly review the super-resolution model for a GS camera before proceeding to an in-depth analysis of the RS case. Let $\mathbf{F}$ represent the discrete HR image of size $\epsilon M \times \epsilon N$ pixels where ϵ (> 1) is the super-resolution factor. Then its lexicographically ordered version $\mathbf{f}$ will be a column vector of size $\epsilon^2 MN \times 1$. Let $\mathbf{W}$ denote a warping matrix of size $\epsilon^2 MN \times \epsilon^2 MN$ that multiplies $\mathbf{f}$ to produce a geometrically warped instance of the HR image. Each row of $\mathbf{W}$ is associated with a pixel in $\mathbf{f}$ and contains a single non-zero entry if we consider only in-plane translation of the camera and integer pixel shifts. In the case of general camera motion with sub-pixel shifts, each row contains at most four non-zero entries which are the bilinear interpolation coefficients obtained by applying the motion that relates $\mathbf{f}$ and its warped instance on that particular pixel location. Thus, $\mathbf{W}$ is a highly sparse matrix with only $4\epsilon^2 MN$ entries. We define $\mathbf{D}_\epsilon$ as the decimation matrix (of size $MN \times \epsilon^2 MN$) which averages ϵ^2 neighboring pixels in the

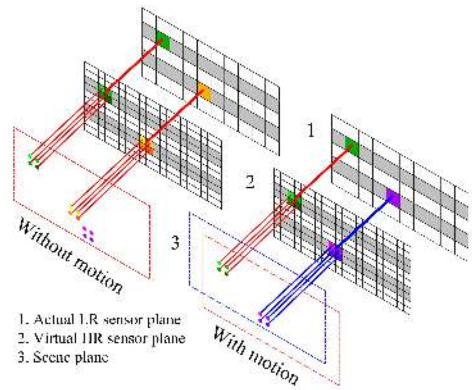

Figure 2. Image formation model for an RS camera - static versus moving. Best viewed in PDF.

HR image to produce the intensity value at a particular location in the LR image. As in the case of $\mathbf{W}$, the matrix $\mathbf{D}_\epsilon$ too is sparse since each row contains only ϵ^2 non-zero values. Then the classical super-resolution equation for a GS camera can be expressed as

$$\mathbf{g} = \mathbf{D}_\epsilon \mathbf{W} \mathbf{f} \tag{1}$$

where $\mathbf{g}$ is a column vector of size $MN \times 1$ which is the lexicographically ordered LR image $\mathbf{G}$ of size $M \times N$.

Consider now the case of a CMOS camera that is not static during the exposure duration. As discussed in Section 1, each row of pixels in the image experiences a different motion due to sequential acquisition. Fig. 2 illustrates the differences in the image formation process in the case of a static RS camera and a moving RS camera for a super-resolution factor of 2. The acquisition mechanism in the case of a static RS camera is shown on the left. On the right, the effect of motion during exposure for an RS camera is depicted. The scene plane (marked 3) is shown as dotted lines, while the actual LR sensor plane of the camera (marked 1) is shown at the top. For the sake of explanation, we have introduced a *virtual* HR sensor plane (marked 2) between the scene plane and the actual LR sensor plane. This is the HR representation of the scene that an HR camera with a higher resolution would have captured, and it is this HR image that we aim to recover. Let us first examine the static case. The actual scene intensities have been represented by dots on the scene plane, and the path traced by the rays by thick lines. A group of four rays (thick red lines) emerging from the scene plane (dotted red lines) are incident at four individual pixels on the virtual HR sensor plane. However, only their average intensity is recorded by the LR sensor at a single pixel location. Now consider the image on the right of an RS camera which is in motion during capture time. For simplicity of representation, in the figure, we have moved the scene plane and not the camera since the two motions are equivalent i.e., it is the relative motion between the scene and the camera that is relevant. When the second row is being recorded by the LR sensor, the scene plane (red dotted lines) is in the same position relative to the camera as the image shown on the left for the static case. Thus, the green pixel is recorded as before by the LR sensor plane at the same location. However, when the fourth row is being acquired, the camera has undergone an in-plane translation (one LR pixel down and to the right) with respect to the scene plane (marked by blue dotted lines). As a result, the purple pixel is recorded in place of the yellow pixel.

An important observation to be made from Fig. 2 is that for an SR factor of 2, although *each* row of the LR sensor has its own motion due to sequential capture, a *pair* of rows in the HR plane experience the same motion. For an SR factor of ϵ, this corresponds to a block of ϵ rows in the virtual HR sensor plane having the same motion associated with them. While it is true that every row of an *actual* HR RS camera might experience a different warp, our goal is only to estimate the image as seen by a static HR global shutter camera. Hence, it suffices to estimate the warp experienced by the *virtual* HR plane. Therefore, unlike in a GS camera where all rows of $\mathbf{W}$ are associated with a *single* camera motion, in RS cameras, the motion varies depending on which particular block of rows in the HR image the pixel belongs to. However, despite this difference in construction, the matrix retains its sparse nature since each row continues to have no more than four non-zero entries. (1) can be rewritten for a CMOS camera as

$$\mathbf{g} = \mathbf{D}_{\boldsymbol{\epsilon}} \boldsymbol{\mathcal{W}} \mathbf{f} \tag{2}$$

where $\boldsymbol{\mathcal{W}}$ is the warping matrix that multiplies $\mathbf{f}$ to produce an RS image. Note that while both $\mathbf{W}$ and $\boldsymbol{\mathcal{W}}$ operate on the HR image, there is only a single camera pose associated with $\mathbf{W}$ as against M camera poses for $\boldsymbol{\mathcal{W}}$.

3. Optimization problem

Suppose we have K different LR images $\{\mathbf{g_k}\}$ each of size $M \times N$ pixels. Our goal is to estimate $\mathbf{f}$, the HR representation of the original scene, using these K LR images. We assume that the first LR image $\mathbf{g_1}$ is free from RS effect and has only undergone a downsampling operation with respect to $\mathbf{f}$. The remaining images may have incurred an RS effect due to motion of the camera during exposure. The assumption that the first image is not RS affected is necessitated by the fact that one cannot determine the presence or absence of this effect from the LR images themselves. What this means is that we cannot say that a curve in the image is due to RS effect unless we have a reference image that has a straight line corresponding to it.

Equation (2) can be rewritten with the added subscript k for the image number as $\mathbf{g_k} = \mathbf{D}_{\boldsymbol{\epsilon}} \boldsymbol{\mathcal{W}}_\mathbf{k} \mathbf{f}$, for $k = 1$ to K. Since the warping matrices are also unknown, given $\{\mathbf{g_k}\}$, we have to estimate the HR image $\mathbf{f}$ as well as $\boldsymbol{\mathcal{W}}_\mathbf{k}$. We propose an alternating minimization (AM) strategy to solve for the two variables wherein we fix one unknown and compute the other in an iterative manner. The minimization sequence $(\mathbf{f}_p, \boldsymbol{\mathcal{W}}_{\mathbf{k}_p})$, where p indicates the iteration number, can be built by alternating between two minimization subproblems. Starting with an initial estimate $\mathbf{f}_0$, the two alternating steps are: step 1) estimate $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}$ using the previous iterate $\mathbf{f}_{p-1}$, step 2) use the current estimate $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}$ to compute $\mathbf{f}_p$. We elaborate on these two steps below.

3.1. Warp estimation

For every row i, where $1 \leq i \leq M$, in the LR images $\{\mathbf{g_k}\}_{k=2}^{K}$, our goal is to estimate a single camera motion during the exposure time of that particular row. We use $\mathbf{g_1}$ which we treat as the reference image (and free from RS effect) to initialize our algorithm. We upsample $\mathbf{g_1}$ by the desired SR factor ϵ and use this as our initial estimate $\mathbf{f}_0$ of the HR image. Since each LR image is an independent observation, we can consider a pair of images at a time ($\mathbf{f}_p$ and $\mathbf{g_k}$ for $k = 2$ to K) to estimate the camera motion and build $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}$. Feature-based approaches cannot be employed to estimate the camera motion (unlike in the GS case) because the camera motion can change from one row to the next, and a sufficient number of features may not even exist for a given row. Following works such as [16, 20] that relate images through the camera motion, we assume a fronto-parallel scene and known camera intrinsic matrix. This allows us to define a 6D discrete camera pose space $\mathbf{S}$ around the origin in which the motion of each row is sought. The six degrees of freedom arise from 3D rotation and translation vectors, $\mathbf{R}$ and $\mathbf{T}$. The origin corresponds to zero translations and rotations. However, the search space cannot be very finely binned due to memory constraints. For example, we can choose the bin intervals such that the difference in the displacements of a point light source due to two different motions from the discrete set $\mathbf{S}$ is at least one pixel. But in real scenarios, the true camera pose may not be present in the discrete space $\mathbf{S}$, so we formulate our cost function such that a few camera poses around the actual pose are selected from the search space for each row, and the centroid of these poses yields the true motion for that row.

Formally, our optimization problem for the ith row of the kth LR image at the pth iteration takes the following form

$$\widehat{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)} = \underset{\boldsymbol{w}_\mathbf{k}^{(i)}}{\operatorname{argmin}} \{ ||\mathbf{g_k}^{(i)} - \mathbf{D}_{\boldsymbol{\epsilon}}{}^{(i)} \boldsymbol{\mathcal{F}}_{p-1}^{(\epsilon i)} \boldsymbol{w}_\mathbf{k}^{(i)}||_2^2 + \lambda ||\boldsymbol{w}_\mathbf{k}^{(i)}||_1 \} \tag{3}$$
$$\text{subject to } \; \boldsymbol{w}_\mathbf{k}^{(i)} \geq \mathbf{0}.$$

Here $\mathbf{g_k}^{(i)}$ denotes the ith row of the LR image $\mathbf{g_k}$, while $\boldsymbol{w}_\mathbf{k}^{(i)}$ is its corresponding weight vector of size $|\mathbf{S}| \times 1$ (where $|.|$ denotes cardinality) which chooses the required set of poses from the search space $\mathbf{S}$. To construct the matrix $\boldsymbol{\mathcal{F}}_{p-1}^{(\epsilon i)}$, we apply the poses from the space $\mathbf{S}$ on $\mathbf{f}_{p-1}$, take the ϵi block of rows corresponding to the ith row of $\mathbf{g_k}^{(i)}$ from these transformed images, lexicographically order the subimages of size $\epsilon \times N$ pixels as vectors, and stack these vectors as columns of $\boldsymbol{\mathcal{F}}_{p-1}^{(\epsilon i)}$. As compared to (2), $\mathbf{D}_{\boldsymbol{\epsilon}}{}^{(i)}$ is a reduced decimation matrix of size $N \times \epsilon N$ since it operates on $\boldsymbol{\mathcal{F}}_{p-1}^{(\epsilon i)}$ of size $\epsilon N \times |\mathbf{S}|$. Since $\boldsymbol{w}_{\mathbf{k}}^{(i)}$ is sparse, we impose the l_1-norm with non-negativity so as to choose a sparse set of camera poses with corresponding weights to calculate the centroid. The regularization factor λ controls the number of chosen poses. The optimization problem in (3) is solved using the *nnLeastR* function of the Lasso algorithm [11]. The estimated weight vector $\widehat{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)}$ is used to determine the required single camera pose for the ith row of $\mathbf{g_k}$. We calculate the centroid pose using this weight vector which is first normalized to have a unit sum $\widetilde{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)} = \widehat{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)} / \|\widehat{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)}\|_1$, and the weighted average of the rotations and translations in the search space is found to give the centroid motion; $\mathbf{R}_c = \widetilde{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)} \circ \{\mathbf{R}_j\}_{j=1}^{|S|}$ and $\mathbf{T}_c = \widetilde{\boldsymbol{w}}_{\mathbf{k}_p}^{(i)} \circ \{\mathbf{T}_j\}_{j=1}^{|S|}$, where $\circ$ represents element-wise multiplication.

For textureless rows that could occur occasionally, the camera motion is interpolated from neighboring rows. Also, note that distinguishing an out-of-plane rotation about horizontal axis and a translation along vertical axis is not possible based on row correspondences alone. Hence, we consider only the 5D space of camera motion by excluding the horizontal out-of-plane rotation from the search space.

To estimate $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}$ for the first iteration of the AM approach, we start with the motion estimation of the middle row, and proceed to top and bottom rows. This procedure helps in choosing the camera pose space of a particular row as a neighborhood around the estimated centroid of its neighboring row. This adaptation of camera pose space additionally imposes smoothness of camera motion. After the first iteration of AM, a coarse camera motion would be estimated. For subsequent iterations, the search space for each row is chosen as the neighborhood of its previous estimate. Thus, from the second iteration onward, motion estimation for each row is independent and can be parallelized.

3.2. HR image estimation

In order to solve for the HR image $\mathbf{f}$, we minimize the following energy function

$$E(\mathbf{f}) = \sum_{k=1}^{K} ||\mathbf{D}_{\boldsymbol{\epsilon}} \boldsymbol{\mathcal{W}}_{\mathbf{k}_p} \mathbf{f} - \mathbf{g_k}||_2^2 + \alpha \mathbf{f}^T \mathbf{L} \mathbf{f} \qquad (4)$$

where the first term measures the fidelity to the data and emanates from our acquisition model (2). The second is a regularization term with a positive weighting constant α that attracts the minimum of E to an admissible set of solutions. Here $\mathbf{L}$ is the discrete form of the variational prior and is a positive semidefinite block tridiagonal matrix [16] constructed of values depending on the gradient of $\mathbf{f}$. The rationale behind the choice of this smoothness prior is to constrain the local spatial behavior of the images. While it exhibits isotropic behavior similar to the Laplacian in smooth areas, it also preserves edges. To solve (4),

$$\mathbf{f}_p = \underset{\mathbf{f}}{\operatorname{argmin}} E(\mathbf{f}) \Rightarrow \frac{\partial E}{\partial \mathbf{f}} = 0$$

$$\Leftrightarrow \left(\sum_{k=1}^{K} \boldsymbol{\mathcal{W}}_{\mathbf{k}_p}^T \mathbf{D}_{\boldsymbol{\epsilon}}^T \mathbf{D}_{\boldsymbol{\epsilon}} \boldsymbol{\mathcal{W}}_{\mathbf{k}_p} + \alpha \mathbf{L} \right) \mathbf{f} = \sum_{k=1}^{K} \boldsymbol{\mathcal{W}}_{\mathbf{k}_p}^T \mathbf{D}_{\boldsymbol{\epsilon}}^T \mathbf{g_k}.$$

The matrix $\mathbf{D}_{\boldsymbol{\epsilon}}^T$ spreads equally the intensity in LR to ϵ^2 pixels in HR. While $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}$ warps each row to introduce the RS effect, $\boldsymbol{\mathcal{W}}_{\mathbf{k}_p}^T$ corrects distortion to make the image RS-free. We use the method of conjugate gradients (function *pcg* in Matlab) to solve the above equation. We used a fixed value for α in all our experiments based on a visual assessment.

To perform RS correction as a spatial operation, we use a scattered interpolation approach to undistort the RS images using the estimated camera motion. We apply the inverse motion (negative rotation and negative translation) from the estimated camera motion on every pixel in the RS image. A forward mapping of a regular coordinate grid from the RS image to an empty image would result in an irregular grid due to the inverse camera motion applied on each row. We define a triangular mesh over this irregular grid and resample the mesh over the regular grid using bicubic interpolation. This procedure results in RS-free images.

Our AM scheme does not guarantee global convergence because of the way the variables are coupled. However, the two sub-problems (of computing the camera trajectory by fixing the HR image, and of estimating the HR image by fixing the camera motion) are individually convex even though the overall problem is not. By alternately descending in the HR image subspace and the camera pose subspace, the algorithm is guaranteed to hit at least a local minimum. We have tested our method extensively on many real examples and found that it exhibits good convergence properties.

4. Experiments

This section consists of two parts. In the first part, we perform a set of experiments on quasi-synthetic data to evaluate the performance of our algorithm and compare the reconstructed output, both qualitatively and quantitatively, with other competing methods. The second part demonstrates the applicability of the proposed method to real images captured using a hand-held CMOS camera.

We compare our results with various state-of-the-art single and multi-image SR algorithms. We provide the first LR image which is free from RS effect as input to the single-image techniques. Specifically, we use the example-based

Figure 3. Quasi-synthetic example: *Lena*.

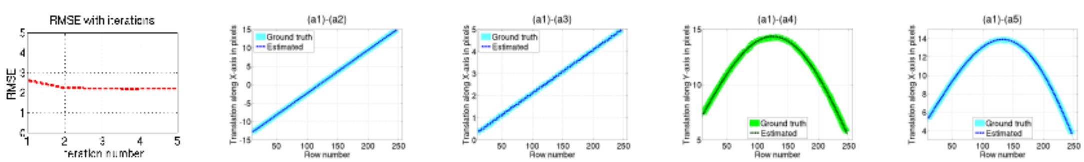

Figure 4. The first plot shows the RMSE with iterations for the *Lena* example. The last four plots show ground truth and estimated camera trajectories. The RMSE between ground truth and computed motion is 0.0217, 0.0178, 0.0574, 0.0249 for images two to five, respectively.

methods in [7, 19, 21] for comparison. The codes for the same were downloaded from the webpage of the authors. A bicubic spline interpolated version of the first image serves as baseline in these comparisons.

Since the remaining LR images may be distorted by RS effect that classical multi-image SR approaches are incapable of handling, we first rectify them with the help of the reference image using our own implementation of the method in [15]. (Please refer to the supplementary material for details on validation of the correctness of our implementation.) The reference image and the rectified images are then given as input to standard GS-based multi-image SR methods in [17, 18], the codes for which were downloaded from the webpage of the authors. This two-step strategy is necessary because the motion compensation step in contemporary SR algorithms cannot account for the row-wise distortions. Our method, on the other hand, does not require any pre-rectification step because the RS effect is modeled *within* the motion estimation step itself.

We begin with a quasi-synthetic example in Fig. 3 where the RS effect is due to pure in-plane translations of the camera. Fig. 3(a1-a5) shows the LR images obtained by downsampling the HR image $\mathbf{f}$ in Fig. 3(A) by a factor of two. The first LR image in Fig. 3(a1) is simply a downsampled version of Fig. 3(A) and is undistorted. To simulate the RS effect, we generate smooth camera trajectories by varying, within an interval, the values of in-plane translations about the X and Y axes. The remaining four LR images in Fig. 3(a2-a5) were generated by warping the rows of $\mathbf{f}$ using these camera paths, and then downsampling. We used Mat-

Figure 5. Quasi-synthetic example: *Butterfly*.

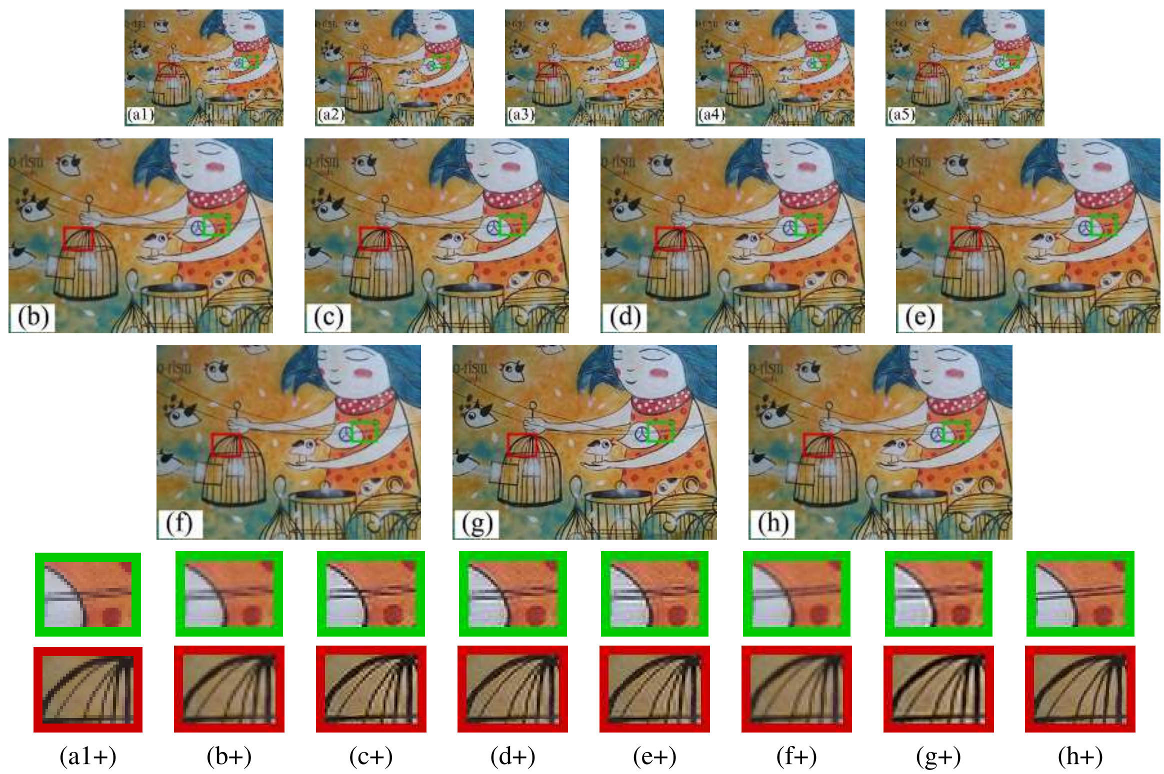

Figure 6. Real example of a wall painting.

Table 1. PSNR (in dB) values of our method, along with comparisons.

Image	Bicubic	[7]	[19]	[21]	[18]	[15] + [18]	[17]	[15] + [17]	Ours #5	Ours #4	Ours #3	Ours #2
Lena	31.81	32.83	33.18	32.95	28.69	30.75	27.55	29.90	**35.45**	35.39	35.11	34.15
Butterfly	30.48	31.71	32.33	32.01	26.35	28.06	24.34	29.38	**34.89**	34.82	34.53	33.66

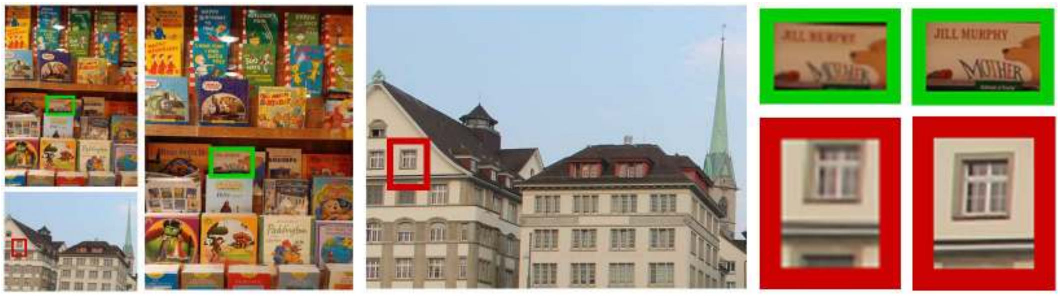

Figure 7. Real examples of a bookshelf and a building. Column one: RS affected LR frames, columns two and three: our SR results, column four: LR patches from column one, and column five: HR patches from columns two and three.

lab's *imresize* command (with the resize option set to *bicubic*) for downsampling. Hence the name *quasi-synthetic* because the data does not strictly mimic the the decimation operator in (2). Note that we have chosen an SR factor of two for this experiment. Hence, the same motion is applied on two adjacent rows of the HR image. The comparison results in Figs. 3(b-e) are obtained from the first LR image which is free from RS effect. Fig. 3(b) is bicubic spline interpolated, while Figs. 3(c), (d) and (e) are the outputs of state-of-the-art single image SR methods [7], [19], and [21], respectively. The four RS affected LR images are rectified with the first LR image as reference using the technique in [15]. The rectified results along with the first image are then provided as input to the classical GS-SR algorithms in [18] and [17] to obtain the outputs shown in Figs. 3(f) and (g), respectively. The output of the proposed method is shown in Fig. 3(h). Zoomed-in patches have also been provided in Figs. 3(a1+), (A+), (b+) to (h+) to better access the quality of the reconstructed output. (The LR patch in Fig. 3(a1+) has been scaled to twice the size for display. Patches from the remaining LR images have not been shown due to space constraints.) It can be observed that our output is sharp and free from distortions. The RMS error in the estimation of the HR image with iterations is shown in the first plot of Fig. 4. Note that our algorithm converges within a few iterations. The synthetically generated camera paths and the final estimated trajectories are also shown in the plots of Fig. 4. The RMS error between the ground truth and the estimated camera motion is a good indicator of our algorithm's ability to accurately estimate the row-wise distortions.

Another quasi-synthetic example with in-plane translations and rotations for an SR factor of 2 is shown in Fig. 5. The PSNR values are provided in Table 1 for quantitative assessment. The notation #X refers to the number of input LR images. It can be seen that our joint RS-SR framework outperforms contemporary methods even with as few as two LR images. Although the PSNR values of [17, 18] which were provided with five LR images as input improve when preceded by a rectification step [15], residual RS effect degrades their performance.

Fig. 1 shows a real example captured using a hand-held mobile phone camera. While the first LR image shown in Fig. 1(a1) was captured without any motion, the other LR images are RS affected. The text on the nameboard is clearly readable only in our super-resolved patch. The second real experiment in Fig. 6 for an SR factor of 2 consists of a wall painting. As can be seen from the zoomed-in patches, the cage and the wires are sharp and free from artifacts in our output. Fig. 7 shows two more real examples for an SR factor of 2 which used as input frames extracted from videos captured using a mobile phone camera. For each example, we have shown only one RS affected LR frame and our output in Fig. 7. More real results have been provided in the supplementary material.

5. Conclusions

In this paper, we dealt with the challenging task of SR from RS affected images. Through our observation model, we mapped the row-wise camera motion of LR images to the HR image, and proposed an AM scheme to recover the camera motion and the HR image. We compared our output with contemporary single as well as multi-image SR techniques. Experiments reveal that our proposed scheme yields a higher PSNR than competing methods. That our method advances the state-of-the-art is evident from the striking visual quality of our RS compensated and reconstructed HR image. Relaxing the constraints of no-blur and availability of an undistorted reference image would be an interesting direction for future work.

References

[1] S. Babacan, R. Molina, and A. Katsaggelos. Variational bayesian super resolution. *Image Processing, IEEE Transactions on*, 20(4):984–999, April 2011. 2

[2] S. Baker, E. Bennett, S. B. Kang, and R. Szeliski. Removing rolling shutter wobble. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2392–2399. IEEE, 2010. 3

[3] A. Bhavsar and A. Rajagopalan. Resolution enhancement in multi-image stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1721–1728, Sept 2010. 2

[4] D. P. Capel. Image mosaicing and superresolution, 2004. 1

[5] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, 2004. 1, 2

[6] P.-E. Forssén and E. Ringaby. Rectifying rolling shutter video from hand-held devices. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 507–514. IEEE, 2010. 3

[7] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356, Sept 2009. 2, 6, 8

[8] M. Grundmann, V. Kwatra, D. Castro, and I. Essa. Calibration-free rolling shutter removal. In *Computational Photography (ICCP), 2012 IEEE International Conference on*, pages 1–8. IEEE, 2012. 3

[9] H. S. Lee and K. M. Lee. Simultaneous super-resolution of depth and images using a single camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 281–288, June 2013. 2

[10] C.-K. Liang, L.-W. Chang, and H. H. Chen. Analysis and compensation of rolling shutter effect. *IEEE Transactions on Image Processing*, 17(8):1323–1330, 2008. 2

[11] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009. 5

[12] M. Meilland, T. Drummond, and A. I. Comport. A unified rolling shutter and motion blur model for 3D visual registration. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2016–2023. IEEE, 2013. 3

[13] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine, IEEE*, 20(3):21–36, May 2003. 2

[14] V. Pichaikuppan, R. Narayanan, and A. Rangarajan. Change detection in the presence of motion blur and rolling shutter effect. In *Computer Vision - ECCV 2014*, volume 8695 of *LNCS*, pages 123–137. Springer, 2014. 3

[15] E. Ringaby and P.-E. Forssn. Efficient video rectification and stabilisation for cell-phones. *International Journal of Computer Vision*, 96(3):335–352, 2012. 2, 3, 6, 8

[16] F. Sroubek, G. Cristobal, and J. Flusser. A unified approach to superresolution and multichannel blind deconvolution. *Image Processing, IEEE Transactions on*, 16(9):2322–2332, Sept 2007. 1, 2, 4, 5

[17] A. Snchez-Beato. Coordinate-descent super-resolution and registration for parametric global motion models. *J. Visual Communication and Image Representation*, 23(7):1060–1067, 2012. 2, 6, 8

[18] S. Villena, M. Vega, D. Babacan, R. Molina, and A. Katsaggelos. Bayesian combination of sparse and non sparse priors in image super resolution. *Digital Signal Processing*, 23(2):530–541, 2013. 2, 6, 8

[19] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *Image Processing, IEEE Transactions on*, 19(11):2861–2873, Nov 2010. 2, 6, 8

[20] H. Zhang and L. Carin. Multi-shot imaging: Joint alignment, deblurring, and resolution-enhancement. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2925–2932, June 2014. 1, 2, 4

[21] Y. Zhu, Y. Zhang, and A. Yuille. Single image super-resolution using deformable patches. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2917–2924, June 2014. 2, 6, 8