J. Vis. Commun. Image R. 30 (2015) 181–190

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Video super-resolution with registration-reliability regulation and adaptive total variation ☆

Xinfeng Zhang [a], Ruiqin Xiong [b,*], Siwei Ma [b], Ge Li [c], Wen Gao [b]

[a] Key Lab of Intelligent Information Processing, Institute of Computing technology, Chinese Academy of Sciences, Beijing 100190, China
[b] Institute of Digital Media, Peking University, Beijing 100871, China
[c] School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518005, China

ARTICLE INFO

Article history:
Received 28 July 2014
Accepted 5 April 2015
Available online 11 April 2015

Keywords:
Super-resolution
Registration reliability
Regularization
Total variation
Nonlocal similarity
Structure tensor
Interpolation
Optical flow

ABSTRACT

In super-resolution that constructs a high-resolution (HR) image from a set of low-resolution (LR) reference images, it is crucial to align the LR reference images in order to efficiently exploit the pixels therein. However, due to the existence of complex local motion, ideal registration is difficult to acquire. In this paper, we present a robust video super-resolution scheme with registration-reliability regulation and content adaptive total variation regularization, which make the scheme resilient to registration failures. In order to handle ill-registered pixels, we propose a registration-reliability regulated data-fidelity term, which assigns smaller weights to the pixels with larger locally-averaged registration residuals. In addition, a content adaptive total variation based on structure tensor, which is used to estimate image local structures, is proposed to regularize the super-resolved images. The structure tensor is derived not only from the gradients of local patches but also the nonlocal similar patches. Experimental results show that the proposed scheme can remarkably improve both the objective and subjective quality of the video super-resolution results.

1. Introduction

Super-resolution (SR) is a technique to generate high resolution (HR) images from one or multiple low resolution (LR) observation(s). Numerous algorithms have been proposed in the literatures [1–13]. They can be broadly classified into three categories, i.e., interpolation-based, learning-based and reconstruction-based methods. The interpolation-based methods, e.g., in [1–3], usually generate the HR image from only one LR image, utilizing the continuity or smoothness within a small neighborhood of any pixel. However, such assumption of local smoothness tends to blur the edges or high contrast textures in images. The learning-based methods, e.g., in [4–6], utilize a set of example images, organized in a form of low- and high- resolution pairs, to derive the missing high frequency information from the high resolution parts, whose corresponding low resolution parts are similar with input image. This kind of methods is efficient only if the input images are similar to the examples in terms of image structures. However, the performance of these methods usually deteriorates drastically when the input images are not similar to the examples. The reconstruction-based super-resolution algorithms, e.g., in [7–9,13], compute HR images by formulating and inversing the image formation process, using prior knowledge to regularize the solution. Most of reconstruction-based SR methods proposed in the literature consist of the three stages, i.e., image registration, interpolation and restoration (i.e., inverse procedure), where the last two steps are usually implemented jointly. In [11,12], Izadpanahi et al. jointly utilize the super-resolution and edge-directed interpolation to reconstruct high resolution images.

To recover high-frequency information reliably, it is crucial to exploit the relevant pixels in reference images to increase the effective sampling rate. For this purpose, accurate registration (e.g., with sub-pel accuracy) is required to determine the object displacement in a sequence of images, i.e., optical flow, to map the pixels from reference images to the current image. Although many image registration methods [14–18] have been proposed in the literature, video sequence registration is still a very challenging problem, because complex local motion usually exists in real scenes. Traditional SR schemes generally regard all the reference LR images as equally reliable and the different magnitude of registration errors are not taken into consideration. Therefore, undesirable visual artifacts may still appear at the region of complex motion in reconstructed images.

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: xfzhang@jdl.ac.cn (X. Zhang), rqxiong@pku.edu.cn (R. Xiong), swma@pku.edu.cn (S. Ma), gli@pkusz.edu.cn (G. Li), wgao@pku.edu.cn (W. Gao).

http://dx.doi.org/10.1016/j.jvcir.2015.04.002
1047-3203/

To cope with the inaccurate registration problem, the cost function measuring the conformance between the estimated image and the observations is extended to incorporate a set of weights which reflect the registration reliability of reference pixels in the SR process [8,19–22]. In [19], Kondi et al. employ a frame-wise weighting scheme in which LR frame with larger registration residuals is assigned smaller weight. Unfortunately, frame-wise weights cannot reflect the variation of the registration reliabilities within a frame. To address the issue, in [8], pixel-wise weights are proposed in super-resolution reconstruction. However, since the weights are calculated based on the registration residual of only one pixel, they are sensitive to noises or interpolation errors in registration. In [20], Kanaev and Miller apply a Gaussian filter to the single pixel residual based weights to depress the influence of random noise. In the work [21,22], they propose the region-based weight to improve its robustness, in which a weight is computed using all the registration errors in a region and is assigned to each pixel in the local region. The main problem of [21,22] is that the weights cannot reflect the motion difference in the same region, especially when parts of regions are occluded. In addition, it is also a difficult problem of image region segment. In [7], Farsiu et al. take L1 norm to constrain the registered LR images instead of L2 norm to improve the robustness of the SR reconstruction to outliers. In [23], Takeda et al. take a spatial–temporal steering kernel based on the local structures in image and motions between images to estimate high resolution image.

Due to the ill-condition of the super-resolution reconstruction, regularization methods are widely used in solving SR problem. Total variation (TV) [24] is a widely used regularization in image processing, which implicitly assumes the pixel differences follow identical and independent Laplacian distribution. However, the distribution of pixel difference changes for image with different structures, and the spatial invariant TV is unable to adapt different image structures. In [7], Farsiu et al. proposed a bilateral total variation (BTV) by assigning weights to pixel differences according to the pixel similarity. In [25], Yuan et al. designed the weight according to the image difference curvature. This method assigns smaller weights to the pixels around edges and larger weights to the pixels in smooth areas.

In this paper, we propose a robust video super-resolution scheme with a new reconstruction objective function consisting of registration-reliability regulated data-fidelity and content adaptive total variation (CATV) regularization. Firstly, to tackle pixels with different registration accuracy, we propose a registration-reliability regulated data-fidelity to differentiate reference pixels by assigning different weights to them according to their registration residuals. Instead of considering the "lack of fit" of a single pixel, we utilize the weighted registration residuals in a neighborhood to compute the registration reliability of a reference pixel. For pixels with large locally-averaged registration residuals, they may be ill-registered and are assigned to small weights, vice versa. Secondly, to regularize the super-resolved HR image, we propose a content adaptive total variation regularization which penalizes image pixel difference (or image variation) differentially based on the anisotropic local structure of reconstructed image. We take both the local and nonlocal similar patches to calculate structure tensor, which is used to reflect the image local structure. Based on the structure tensor, we assign different weights to pixel differences. Generally speaking, larger weights are assigned to image pixels along edges or in smooth areas, and smaller weights are assigned to those across edges or in texture areas, which is equal to penalizing pixel differences crossing edges weakly.

The remainder paper is organized as follows. In Section 2, we give the formulation of the video super-resolution problem. The proposed registration-reliability regulated data-fidelity is introduced in the Section 3. The proposed content adaptive total variation regularization is elaborated in Section 4. The proposed super-resolution reconstruction method is presented in Section 5. Experimental results on real video sequences are reported in Section 6 and some concluding remarks are made in Section 7.

2. Problem formulation

Super-resolution reconstruction is the inverse problem of low resolution observation formation. Many LR video observation models have been proposed in the literature e.g. [7,26,27]. In this paper, we utilize the observation model that assumes the neighboring HR frames in temporal domain describing the same scene and having complementary information to each other. The LR frames are acquired from the corresponding HR frames through blurring and down-sampling. In this process, the LR frames may be distorted by noise. Fig. 1 shows a LR video observation model, in which the HR frames of size $L_1N_1 \times L_2N_2$ written in lexicographical notation as the vector $\mathbf{X}_t = [x_{t,1}, x_{t,2}, x_{t,3}, \ldots, x_{t,N}]^T$ and the corresponding

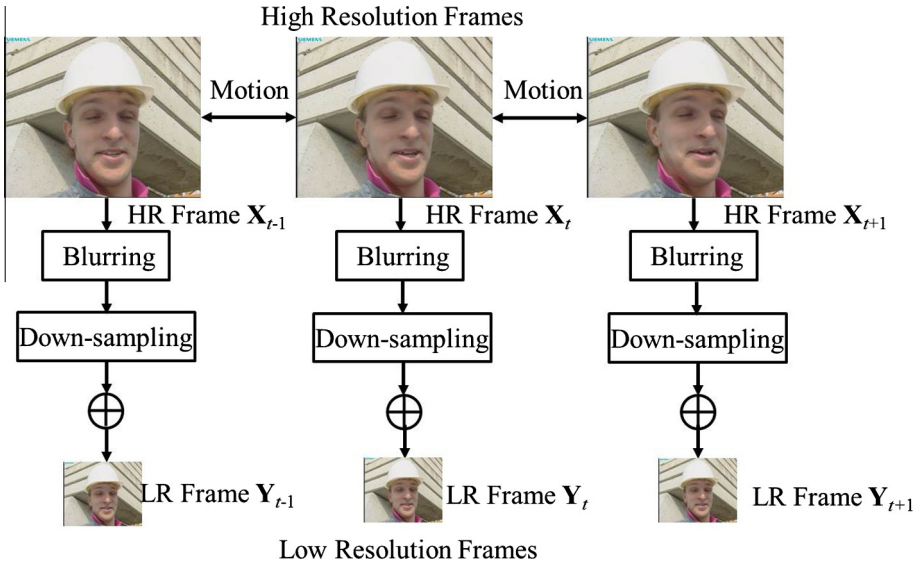

Fig. 1. Video sequence observation model.

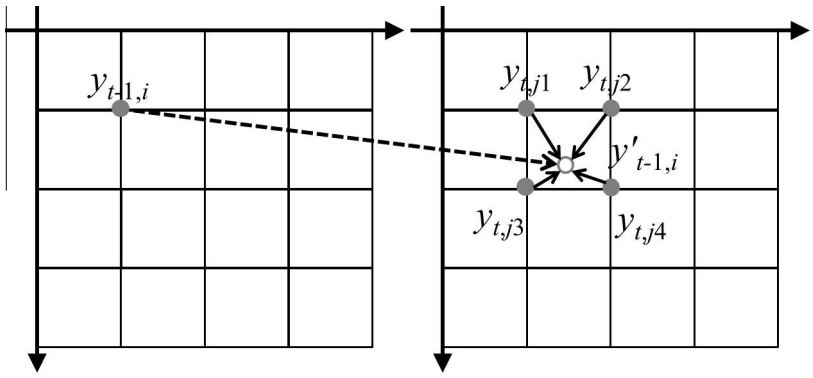

Fig. 2. The pixel $y_{t-1,i}$ is registered onto the subpixel position of the tth frame, $y'_{t-1,i}$ is its prediction value.

LR frames of size $N_1 \times N_2$ are denoted as $\mathbf{Y}_t = [y_{t,1}, y_{t,2}, y_{t,3}, \ldots, y_{t,M}]^T$, where t presents the temporal index, $N = L_1N_1 \times L_2N_2$ and $M = N_1 \times N_2$. The neighboring frames can be predicted from each other via image registration as follows,

$$\mathbf{X}_k = \mathbf{F}_{kt}\mathbf{X}_t. \tag{1}$$

Eq. (1) indicates that the frame $\mathbf{X}_k$ can be reproduced from $\mathbf{X}_t$ with the warping matrix $\mathbf{F}_{kt}$ which is figured out by registration. Therefore, all the LR frames can be related with one of the high resolution frames, e.g. $\mathbf{X}_t$, and the relationship can be formulated as,

$$\mathbf{Y}_k = \mathbf{DHF}_{kt}\mathbf{X}_t + \mathbf{n}_k. \tag{2}$$

Here $\mathbf{H}$ is the blurring matrix representing the point spread function (PSF) in the observation formation process, $\mathbf{D}$ is the down-sampling matrix and $\mathbf{n}_k$ is random noise.

In reconstruction-based SR problem, the HR frame $\mathbf{X}_t$ is estimated from a group of the LR frames, $\mathbf{Y}_k$, in a temporal neighborhood,

$$\{\mathbf{Y}_k | k = t - L, \ldots, t + L\}. \tag{3}$$

Herein, the frame $\mathbf{X}_t$ is referred to as *target frame*, and the neighboring LR frames, $\{\mathbf{Y}_k\}$, are referred to as *reference frames*. Traditional reconstruction-based SR methods seek a recovery of $\mathbf{X}_t$ by minimizing the deviation of the *target frame* from the observed frames $\{\mathbf{Y}_k\}$ according to the LR video observation model. In general, the SR reconstruction problem is an ill-posed inverse problem. Regularization techniques are frequently used to stabilize the solution. Therefore, the SR reconstruction can be formulated as,

$$\hat{\mathbf{X}}_t = \arg\min_{\mathbf{X}_t} \sum_{k=t-L}^{t+L} \|\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k\|_2^2 + \lambda\varphi(\mathbf{X}_t). \tag{4}$$

The first term is data-fidelity constraint, which makes the estimated high resolution frame conform to the low resolution video frames via the observation model. The second one is a regularization term, which constrains the estimated high resolution frame with certain image prior model. λ is the regularization parameter. Based on the formulation in Eq. (4), the registration plays an important role in efficiently utilizing temporal neighboring LR frames. The ill-registered pixels may deteriorate the reconstructed image in data fusion process. Therefore, it is necessary to differentiate the registered pixels reliability and depress the influence of ill-registered pixels on data-fidelity term. In addition, the regularization term also can help to decrease the negative influence of ill-registered pixels, which may not conform to image prior models well.

3. Registration-reliability regulated data-fidelity constraint

Considering that inaccurate registration may deteriorate the SR results dramatically, we propose a registration-reliability regulated data-fidelity to differentiate the reference pixels with different registration accuracy. Based on the optical flow assumption that the pixel intensity does not change along motion trajectories, i.e. brightness constancy, we take the registration residuals to reflect the pixel registration reliability. In general, a pixel in *reference frame* with large registration residuals may be ill-registered and unreliable. However, only a single pixel registration residual is sensitive to noise and is inefficient to reflect its registration reliability accurately. Therefore, the locally-averaged registration residuals are used to reflect the registration accuracy in our method. For a pixel, if its locally-averaged residual is large, a small weight is assigned to it, which may be ill-registered.

However, in super-resolution reconstruction problem, due to subpixel accuracy of image registration being needed, the registration residuals are actually the difference between the pixels in *reference frame* and the interpolated pixels in *target frame*. For example, the pixel $y_{t-1,i}$ in the $(t-1)$th *reference frame* is registered onto the subpixel position of *target frame*, just as illustrated in Fig. 2. The registration residual $r_{t-1,i}$ is the difference between $y_{t-1,i}$ and its prediction $y'_{t-1,i}$, which is generated by interpolation with neighboring pixels, e.g., $\{y_{t,j1}, y_{t,j2}, y_{t,j3}, y_{t,j4}\}$.

$$r_{t-1,i} = y_{t-1,i} - y'_{t-1,i} = y_{t-1,i} - f(y_{t,j1}, y_{t,j2}, y_{t,j3}, y_{t,j4}). \tag{5}$$

Here, f is usually a low pass filter. Therefore, the registration residuals may be generated from three sources, i.e., registration errors, interpolation errors and noise, as follows,

$$r = e_r + e_i + n_i. \tag{6}$$

Here, r represents registration residual, e_r, e_i and n_i are registration error, interpolation error and noise, respectively. Herein, e_r is assumed to be generated by mis-registration and reflects image

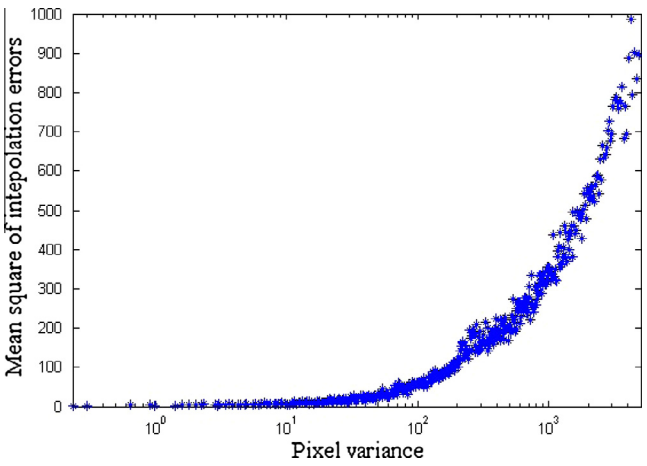

Fig. 3. The relationship of interpolation errors and image local variance.

registration reliability. In most cases, the registration residual, r, is proportional to the registration error, e_r, and the local-average operation can further depress the negative influence of the noise. However, the interpolation errors usually are large and correlated in edge or texture areas, where they may dominate the registration residuals. Fig. 3 illustrates the relationship between image interpolation errors and image local variance. The interpolation errors rapidly increase along with image local variance. Thus, the registration residuals may make the registration residuals irrelevant in measuring the registration reliability efficiently in these areas with high variance.

In order to depress the negative effect of interpolation errors, we propose to take the weighted average of local residuals to reflect registration reliability, where the weights decrease along with the increase of the image local variance. Therefore, we take the reciprocal of the variance of the pixels used in interpolation process as residual weights to cope with the influence of interpolation errors. The final registration reliability can be calculated via the following weight function Eqs. (7)–(10).

$$\mathbf{W}_k(i,j) = g(k,t)\cdot\exp\left(-\frac{R_{kt}(i,j)}{h}\right), \tag{7}$$

$$R_{kt}(i,j) = \frac{1}{Z}\sum_{(m,n)\in\mathcal{N}_k(i,j)}\frac{1}{\sigma_{kt}^2(m,n)+\varepsilon}r_{kt}(m,n), \tag{8}$$

$$Z = \sum_{(m,n)\in\mathcal{N}_k(i,j)}\frac{1}{\sigma_{kt}^2(m,n)+\varepsilon}, \tag{9}$$

$$g(k,t) = \exp(-\tau|k-t|). \tag{10}$$

Here, $r_{kt}(m,n)$ is the registration residual when registering pixel located at (m,n) in the kth frame to the tth frame and h is a smoothness factor. $\mathcal{N}_k(i,j)$ is the neighborhood centered at (i,j) in the kth frame. Z is a normalizing constant and ε is a small constant to avoid division by zero. $g(k,t)$ is a function to measure the reliability of reference frame based on temporal distance and τ is a constant. $\sigma_{kt}^2(m,n)$ is the variance of pixels used to predict the value in sub-pixel position corresponding to (m,n). Therefore, the proposed registration reliability regulated data-fidelity term in objective function of super-resolution is rewritten as

$$\hat{\mathbf{X}}_t = \arg\min_{\mathbf{X}_t}\left[\sum_{k=t-L}^{t+L}(\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k)^T\mathbf{W}_k(\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k) + \lambda\varphi(\mathbf{X}_t)\right]. \tag{11}$$

Here, $\mathbf{W}_k$ is the weighting matrix for the kth frame, which is used to reduce the negative influence of inaccurate registration. It is calculated based on the registration residuals in Eqs. (7)–(10).

4. Content adaptive total variation regularization

Regularization techniques are commonly used to deal with ill-posed problems (e.g., [7,24]). Total variation (TV) [24] and bilateral TV (BTV) [7], measuring the difference between pixels and their neighboring pixels with L1-norm, are often used as image regularization, which implicitly assumes the pixel difference following identical and independent Laplacian distribution, e.g., BTV in (12),

$$\varphi(\mathbf{X}_t) = \sum_{m=-P}^{P}\sum_{n=-P}^{P}\alpha^{|m|+|n|}\|\mathbf{X}_t - S_y^n S_x^m\mathbf{X}_t\|_1, \tag{12}$$

Here, S_x^m and S_y^n are the shifting operators in the horizontal and the vertical directions by m and n respectively and $0<\alpha<1$. P is the radius of neighborhood.

However, due to the diversity of image structures, the neighboring pixel differences are obviously anisotropic and have different distribution characteristic. Fig. 4 illustrates the histograms of pixel difference with neighboring pixels along horizontal and vertical directions, respectively. It shows that the pixel difference in smooth area follows more centralized distribution with smaller variance. However, it illustrates different characteristic in edge areas, a relatively loose distribution across the edge direction and centralized distribution along the edge direction. Therefore, traditional TV regularization penalizing pixel difference equivalently may blur image edges.

To regularize the ill problem while well preserving image structures, we propose a content adaptive total variation (CATV) regularization, which differentiates pixel difference according to local pixel correlation. In our proposed CATV, if the pixels have higher correlation, the differences between them are assigned larger weights, vice versa. The weights are estimated based on the image structure tensor [28,29], which is derived from image local gradients. In general, the structure tensor summarizes the predominant direction of the gradients in a specified neighborhood of a pixel, just as the formulation in Eq. (13).

$$\mathbf{C}_\mathbf{q} \approx \begin{bmatrix} \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial x} & \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial y} \\ \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial y} & \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial y}\frac{\partial\mathbf{X}_t(\mathbf{p})}{\partial y} \end{bmatrix}. \tag{13}$$

Here, $\mathbf{C}_\mathbf{q}$ is the structure tensor for pixel $\mathbf{q}$. and $\mathbf{p},\mathbf{q}$ are coordinate vectors. $\mathcal{N}(\mathbf{q})$ is a neighborhood centered at position $\mathbf{q}$. Due to the

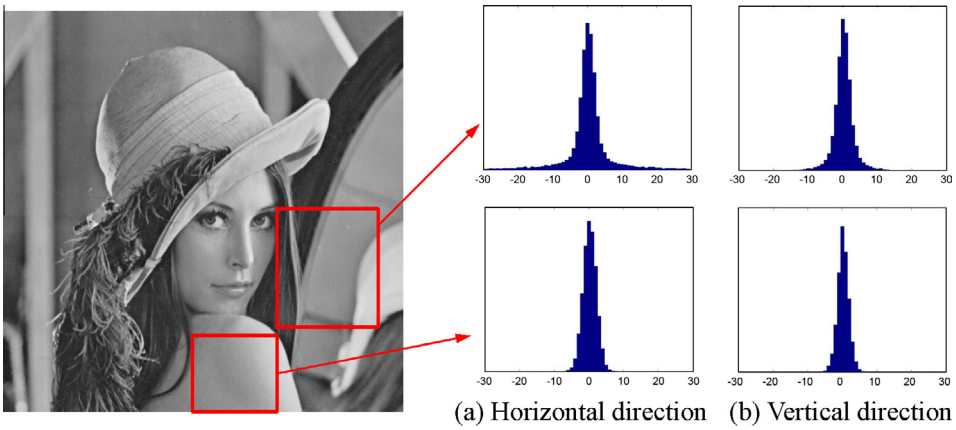

Fig. 4. The histograms of pixel difference in different areas.

structure tensor matrix $\mathbf{C_q}$ being real and symmetric, it can be further factorized into three matrices with singular value decomposition,

$$\mathbf{C_q} = \mathbf{U_q}\Lambda_\mathbf{q}\mathbf{U_q}^T, \tag{14}$$

$$\mathbf{U_q} = \begin{bmatrix} cos\theta_\mathbf{q} & sin\theta_\mathbf{q} \\ -sin\theta_\mathbf{q} & cos\theta_\mathbf{q} \end{bmatrix}, \quad \Lambda_\mathbf{q} = \begin{bmatrix} \lambda_{\mathbf{q},1} & 0 \\ 0 & \lambda_{\mathbf{q},2} \end{bmatrix}. \tag{15}$$

Here, $\lambda_{\mathbf{q},1}$ and $\lambda_{\mathbf{q},2}$ are the eigenvalues and $\lambda_{\mathbf{q},1} \geqslant \lambda_{\mathbf{q},2} \geqslant 0$. The corresponding eigenvectors $\mathbf{u}_{\mathbf{q},1} = [cos\theta_\mathbf{q}, -sin\theta_\mathbf{q}]^T$ and $\mathbf{u}_{\mathbf{q},2} = [sin\theta_\mathbf{q}, cos\theta_\mathbf{q}]^T$ summarize the distribution of gradients in the local area, $\mathcal{N}(\mathbf{q})$. The first eigenvector, $\mathbf{u}_{\mathbf{q},1}$, is maximally aligned with the local gradients and the corresponding eigenvalue, $\lambda_{\mathbf{q},1}$, defines the strength of local gradients bias to $\mathbf{u}_{\mathbf{q},1}$. The accuracy of the structure tensor is influenced by the neighborhood size. Decreasing the neighborhood size makes less sample gradients available for structure tensor calculation, which makes structure tensor estimation sensitive to noise. Although extending the neighborhood size can improve the robustness of structure tensor estimation to noise with more sample gradients available, it also may raise the risk of bring in pixels with different structures, which decreases the accuracy of local structure estimation.

Therefore, we propose to calculate the structure tensor not only with pixels in the local patch, but also with the pixels in nonlocal similar patches. In order to avoid the negative effect of the patches with different structures, we assigned each nonlocal patches a weight according to its similarity with the current processed patch. The weight is calculated with the following equation,

$$w_i = exp\left(-\frac{\|P_i - P_j\|_2}{h}\right), \tag{16}$$

where P_i and P_j are patches centered at the ith pixel and the jth pixel and h is a smoothness factor. The modified structure tensor calculation can be formulated as,

$$\mathbf{C_q} \approx \begin{bmatrix} \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})} w_\mathbf{p} \frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial x} & \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})} w_\mathbf{p} \frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial y} \\ \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})} w_\mathbf{p} \frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial x}\frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial y} & \sum_{\mathbf{p}\in\mathcal{N}(\mathbf{q})} w_\mathbf{p} \frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial y}\frac{\partial \mathbf{X}_t(\mathbf{p})}{\partial y} \end{bmatrix}. \tag{17}$$

The similar formulation of the structure tensor is also presented in [29], which determines the weights according to distance of gradient covariance of image patches. Based on the local structure estimation with modified structure tensor, we further formulate the content adaptive total variation in (18) and (19), respectively.

$$k(\mathbf{p}-\mathbf{q}) = \gamma exp\left(-\frac{(\mathbf{p}-\mathbf{q})^T\mathbf{C_q}(\mathbf{p}-\mathbf{q})}{\eta^2}\right), \gamma = \left(\frac{\lambda_{\mathbf{q},2}\lambda_{\mathbf{q},2}+\varepsilon}{T}\right)^{-\frac{1}{2}} \tag{18}$$

$$\varphi(\mathbf{X}_t) = \sum_{m=-P}^{P}\sum_{n=-P}^{P}\|\mathbf{K}_{m,n} * (\mathbf{X}_t - S_y^n S_x^m \mathbf{X}_t)\|_1 \tag{19}$$

Here, $*$ is element-wise multiplication and $\mathbf{K}_{m,n}$ is a weight matrix, which is comprised of $k(\mathbf{p}-\mathbf{q})$ when $(\mathbf{p}-\mathbf{q})$ is equal to (m,n). In Eq. (18), the exponential part adjusts the weights of pixel difference according to local structures, i.e., large weights assigned to pixels along the predominant direction where pixels have high correlation, vice versa. γ is an adaptive scaling factor, which is large for smooth area and small for texture area. T represents the number of samples used in structure tensor calculation, and η is a smoothness factor. Thus, the weights are large in smooth area and decrease slow along image edges where the pixels with high correlation, and they are small in texture regions and decrease fast across edges where the correlation of pixels is small. $\varphi(\mathbf{X}_t)$ in Eq. (19) is the proposed CATV regularization. By taking advantage of the content adaptive weight, the proposed CATV regularization penalizes pixel difference slightly for pixels in area with texture or across the predominant direction, which makes it preserve image details better than traditional TV.

5. Optimization solution of the proposed super-resolution reconstruction

Based on the above discussion, the proposed super-resolution scheme is formulated as the following minimization problem,

$$\hat{\mathbf{X}}_t = \arg\min_{\mathbf{X}_t} \sum_{k=t-L}^{t+L} \|\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k\|^2_{2,W_k} + \lambda\varphi(\mathbf{X}_t), \tag{20}$$

$$\sum_{k=t-L}^{t+L} \|\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k\|^2_{2,W_k} = \sum_{k=t-L}^{t+L} (\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k)^T\mathbf{W}_k(\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k), \tag{21}$$

$$\varphi(\mathbf{X}_t) = \sum_{m=-P}^{P}\sum_{n=-P}^{P}\|\mathbf{K}_{m,n} * (\mathbf{X}_t - S_y^n S_x^m \mathbf{X}_t)\|_1 = \sum_i \|\mathbf{G}_i\mathbf{X}_{t,i}\|_1. \tag{22}$$

Here, $\mathbf{G}_i\mathbf{X}_{t,i}$ denotes the weighted image pixel difference between the ith pixel and its neighboring pixels. To solve the problem in Eq. (20) which has L2-norm data-fidelity term and L1-norm regularization term, we utilize variable-splitting and penalty approach [30], which transform Eq. (20) to

$$\hat{\mathbf{X}}_t = \arg\min_{\mathbf{X}_t} \sum_{k=t-L}^{t+L} \|\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k\|^2_{2,W_k} + \frac{\beta}{2}\sum_i \|\mathbf{u}_i - \mathbf{G}_i\mathbf{X}_{t,i}\|_2^2 + \sum_i \|\mathbf{u}_i\|_1, \tag{23}$$

Table 1
PSNR comparison for different resolution enhancement methods on noise-free LR videos, enlarging 2 times. (Unit: dB).

Sequences	Bicubic	MASK	I-IBP	RMAP	BTV-L1	PRO-I	PRO-II
Akiyo	33.43	33.88	34.32	35.06	34.28	35.22	35.60
Carphone	30.89	31.19	31.57	31.99	32.05	32.39	32.48
City	28.58	28.81	29.15	29.75	30.27	30.85	30.84
Flower	22.08	22.07	22.55	22.95	22.40	23.38	23.40
Foreman	30.42	30.36	31.25	31.37	31.24	31.58	31.95
Mobile	21.87	22.41	22.45	23.31	23.69	24.58	24.61
Mthr_dotr	33.11	33.03	34.21	34.77	34.07	34.96	35.31
News	28.37	28.21	29.24	30.48	30.73	30.52	30.74
Salesman	29.83	29.46	30.44	31.00	29.87	31.07	31.20
Slient	30.15	30.00	30.41	30.99	30.39	30.91	31.67
Stefan	25.20	24.94	36.15	27.11	26.58	27.45	27.50
Students	29.62	29.49	30.19	30.73	30.93	30.78	30.91
Average	28.22	28.26	29.33	29.51	29.32	29.90	30.08

Table 2
PSNR comparison for different resolution enhancement methods on noisy LR videos (standard deviation of noise, $\sigma = 5$), enlarging 2 times. (Unit: dB).

Sequences	Bicubic	MASK	I-IBP	RMAP	BTV-L1	PRO-I	PRO-II
Akiyo	32.45	32.50	33.14	32.88	33.79	33.89	34.12
Carphone	30.14	30.58	30.59	30.37	31.53	31.06	31.24
City	28.16	28.53	28.57	28.44	28.44	29.56	29.63
Flower	21.79	21.88	22.13	22.02	22.14	22.70	22.88
Foreman	29.18	29.49	29.46	29.22	30.30	30.20	30.53
Mobile	21.73	22.27	22.34	22.09	23.23	23.57	23.91
Mthr_dotr	31.94	32.05	32.56	32.33	33.34	33.27	33.51
News	28.12	27.78	28.97	29.13	30.40	29.87	30.28
Salesman	29.56	29.16	30.09	30.17	29.75	30.68	30.75
Silent	29.36	29.15	29.59	29.55	29.94	29.94	30.57
Stefan	24.92	24.75	25.25	25.49	26.17	26.30	26.33
Students	29.21	29.00	29.74	29.65	30.61	30.25	30.35
Average	28.05	28.09	28.54	28.44	29.14	29.27	29.51

with the penalty parameter $\beta \to \infty$. The problem can be easily solved by iteratively minimizing the objective functions Eqs. (24) and (25) with respect to $\mathbf{u}$ and $\mathbf{X}_t$, alternatively.

$$\min_{\mathbf{u}} \frac{\beta}{2}\sum_i \|\mathbf{u}_i - \mathbf{G}_i\mathbf{X}_{t,i}\|_2^2 + \sum_i \|\mathbf{u}_i\|_1, \tag{24}$$

$$\min_{\mathbf{X}_t} \sum_{k=t-L}^{t+L} \|\mathbf{DHF}_{kt}\mathbf{X}_t - \mathbf{Y}_k\|_{2,W_k}^2 + \frac{\beta}{2}\sum_i \|\mathbf{u}_i - \mathbf{G}_i\mathbf{X}_{t,i}\|_2^2. \tag{25}$$

The first optimization problem in Eq. (24) can be solved by soft threshold operation [31] and the second optimization problem can be solved by least square method. We take the notation $\mathbf{C}$ as the structure tensor of the estimated high resolution image. Therefore, the complete iterative algorithm for our proposed SR method is outlined as follows.

Algorithm 1. The proposed video super-resolution.

Input: Low resolution sequences $\{\mathbf{Y}_k | k = t-L, \ldots t, \ldots, t+L\}$

Initialization:

Initialize $\hat{\mathbf{X}}_t^{(0)}$ from $\mathbf{Y}_t$ with Bicubic interpolation;

Initialize $\mathbf{C}^{(0)}$ with $\hat{\mathbf{X}}_t^{(0)}$ using Eq. (17), n = 0, Iter = 0;

Iteration:

While MaxIterNum > iter and MinDistortion < distortion **do**

(1) Solve optimization problem Eq. (24) with $\mathbf{u}$ for fixed $\mathbf{X}_t^{(n)}$

$$\mathbf{u}_i^{(n)} = \max\left\{\|\mathbf{G}_i\hat{\mathbf{X}}_{t,i}^{(n)}\| - \frac{1}{\beta}, 0\right\}\frac{\mathbf{G}_i\hat{\mathbf{X}}_{t,i}^{(n)}}{\|\mathbf{G}_i\hat{\mathbf{X}}_{t,i}^{(n)}\|}$$

(2) Solve optimization problem in Eq. (25) $\mathbf{X}_t^{(n)}$ for fixed $\mathbf{u}^{(n)}$

$$\hat{\mathbf{X}}_t^{(n+1)} = \left(\frac{1}{\beta}\sum_{t-L}^{t+L}(\mathbf{DHF}_{kt})^T\mathbf{W}_k(\mathbf{Y}_k - \mathbf{DHF}_{kt}\mathbf{X}_t^{(n)}) + \sum_{i=0}\mathbf{G}_i^T\mathbf{G}_i\right)^{-1} \times \left(\frac{1}{\beta}\sum_{t-L}^{t+L}(\mathbf{DHF}_{kt})^T\mathbf{W}_k\mathbf{Y}_k + \sum_{i=0}\mathbf{G}_i^T\mathbf{u}_i^{(n)}\right)$$

(3) Update Variables:

Calculate $\mathbf{C}^{(n+1)}$ with $\hat{\mathbf{X}}_t^{(n+1)}$

(4) iter++,distortion = $\|\hat{\mathbf{X}}_t^{(n+1)} - \hat{\mathbf{X}}_t^{(n)}\|_2$;

End

Output: High resolution image $\hat{\mathbf{X}}_t$

6. Experimental results

In this section, we evaluate the performance of the proposed SR method by comparing with Bicubic interpolation, SR method with registration reliability in [8] (denoted as RMAP), MASK [23], improved iterative back projection method (denoted as I-IBP) [13] and BTV-L1 ([7,9]), which is released since OpenCV-2.4.7 [32]. The MASK method takes the structure tensor to estimate the image local structures in high resolution pixel estimation process. The method BTV-L1 takes the L1-norm as the data-fidelity measurement to improve the robustness of super-resolution reconstruction to registration errors. To evaluate the proposed registration-reliability regulated data-fidelity and CATV regularization, we test the two cases, the proposed data-fidelity with BTV regularization and with the CATV regularization, denoted as PRO-I and PRO-II respectively.

The LR frames are generated firstly by blurring HR frames with 3×3 and 5×5 Gaussian filters and then down-sampling the blurred frames with decimation factor 2 and 3 in both horizontal and vertical directions, respectively. For different super-resolution methods, five LR frames are used as *reference frames* to reconstruct each high resolution frame. The pyramid Lucas–Kanade optical flow estimation method is utilized to register LR videos [15]. Considering the continuous of the optical flow between neighboring frames in videos, we take the coarse-to-fine strategy in temporal domain by first registering any two adjacent frames and then we connect the resulted optical flow fields to form an initial optical flow field estimation for nonadjacent frames, which is then further refined by traditional optical flow methods.

Table 1 lists the PSNR results of the reconstructed high resolution frames from different LR frames without noise and Table 2 lists the PSNR results from noisy LR frames, where the standard deviation of noise is 5. The resolution of the reconstructed frames is two times of the input LR videos both in horizontal and vertical directions. The values in the tables are the average PSNR results of 50 successive frames for each sequence. From these results, we can see that super-resolution results are usually better than Bicubic interpolation method since more temporal LR frames are utilized.

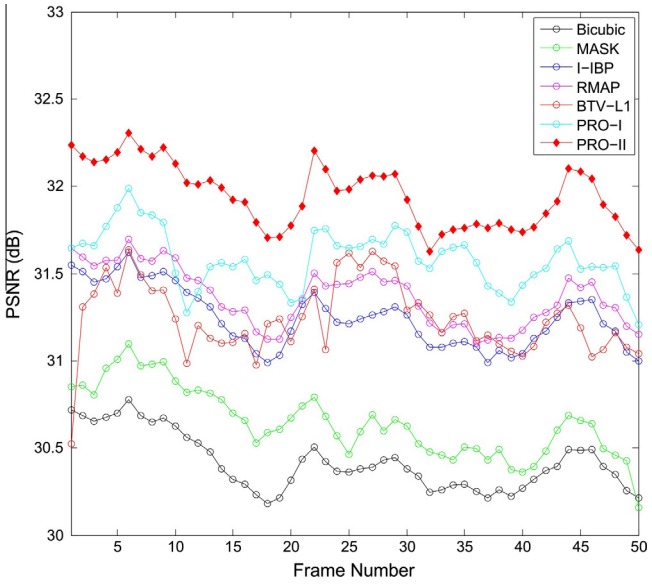

Fig. 5. The PSNR results with different resolution enhancement methods for each frame of Foreman, enlarging 2 times.

Table 3
PSNR comparison for different resolution enhancement methods on noise-free LR videos, enlarging 3 times (Unit: dB).

Sequences	Bicubic	MASK	I-IBP	RMAP	BTV-L1	PRO-I	PRO-II
Akiyo	30.93	31.28	30.96	31.50	31.04	31.61	31.86
Carphone	28.92	29.64	28.95	29.43	29.81	30.08	30.25
City	26.61	27.11	26.62	26.97	27.40	28.24	28.42
Flower	20.48	20.58	20.49	20.76	20.36	21.21	21.31
Foreman	28.31	28.59	28.31	28.84	29.10	29.04	29.53
Mobile	19.77	20.21	19.79	20.13	20.20	21.09	21.13
Mthr_dotr	30.73	30.87	30.77	31.45	31.26	31.43	31.55
News	25.64	25.57	25.68	26.40	25.86	26.41	26.53
Salesman	27.57	27.31	27.60	27.96	27.29	27.97	28.07
Slient	27.93	27.79	27.94	28.23	27.69	27.93	28.36
Stefan	22.42	22.52	22.44	22.88	22.64	23.25	23.53
Students	27.50	27.37	27.52	27.82	27.75	27.85	27.95
Average	26.40	26.57	26.42	26.86	26.70	27.17	27.37

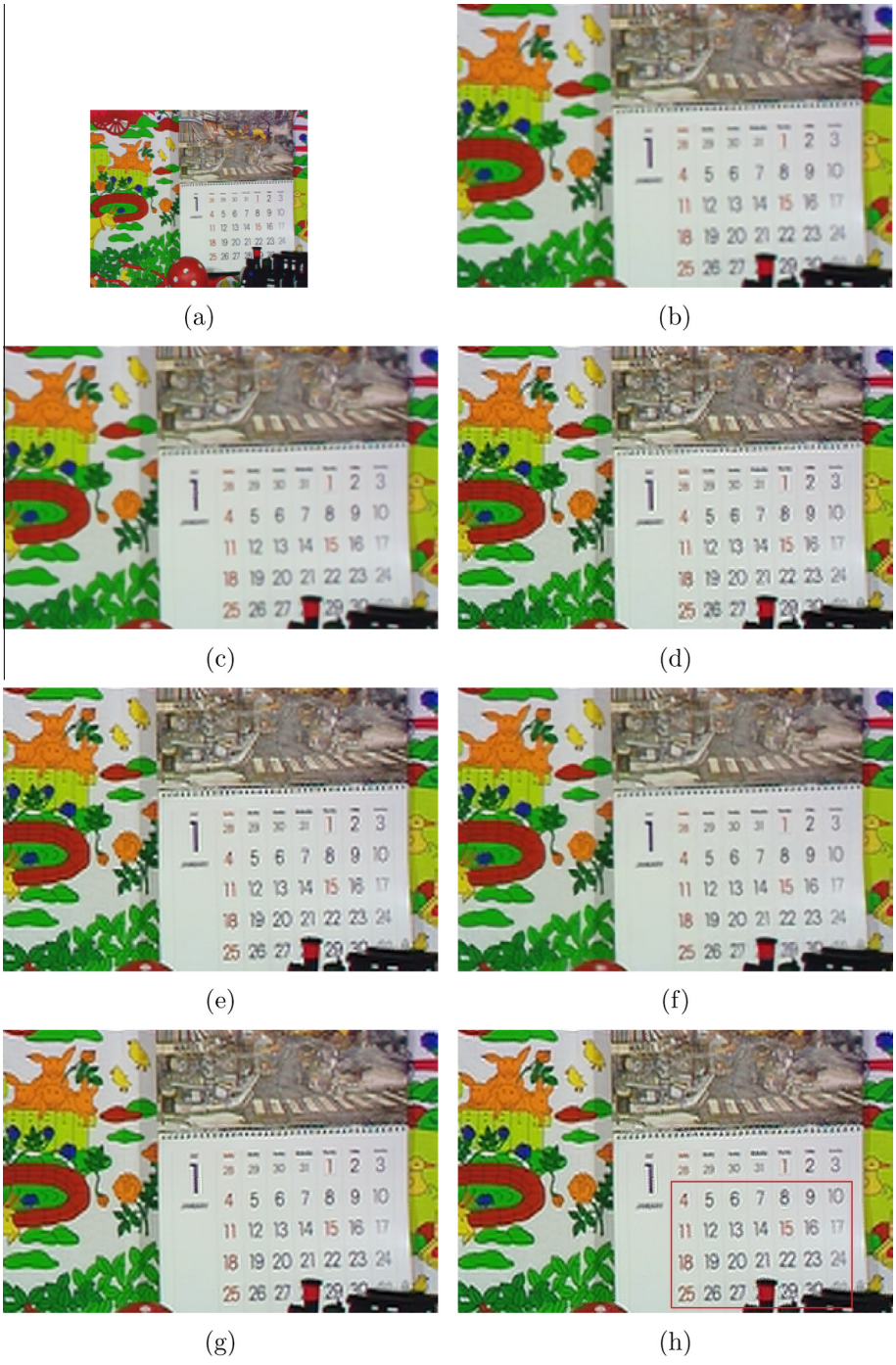

Fig. 6. The reconstruction results of the 31th frame in Mobile with different methods, (a) LR frame, (b) Bicubic, (c) MASK, (d) I-IBP, (e) RMAP, (f) BTV-L1, (g) PRO-I method and (h) PRO-II method.

Our proposed method, PRO-II, achieves PSNR gain up to 2.7 dB over Bicubic for Mobile. However, these schemes shows different performance for sequences with different characteristic. For some sequences with complex motions, e.g., Foreman and Stefan, the MASK scheme is inefficient, even inferior to Bicubic interpolation. The PRO-I method and RMAP achieve better results than MASK by differentiating the registration reliability of reference pixels with a weighted data-fidelity. The PRO-I method also outperforms RMAP, although they both take registration residual to differentiate the reliability of pixels in reference images. This verifies the proposed registration reliability regulation is more efficient. The I-IBP method introduces a bicubic filter to smooth out some outliers and achieves performance improvement. The BTV-L1 with L1 norm data-fidelity also can exclude most of the wrongly registered pixels and achieves better performance, but it is still inferior to our proposed method. Since the L1-norm data-fidelity can be regarded as a median filtering operation, it may filter out the content only appearing in *targe frame*, which is minority comparing with that in *reference frames*. The proposed data-fidelity term can efficiently distinguish the efficiency of reference pixels for super-resolution reconstruction by taking locally-averaged registration residuals. The proposed CATV regularization further improves the quality of reconstructed image by less penalizing pixel difference across edges than traditional TV regularization. It outperforms

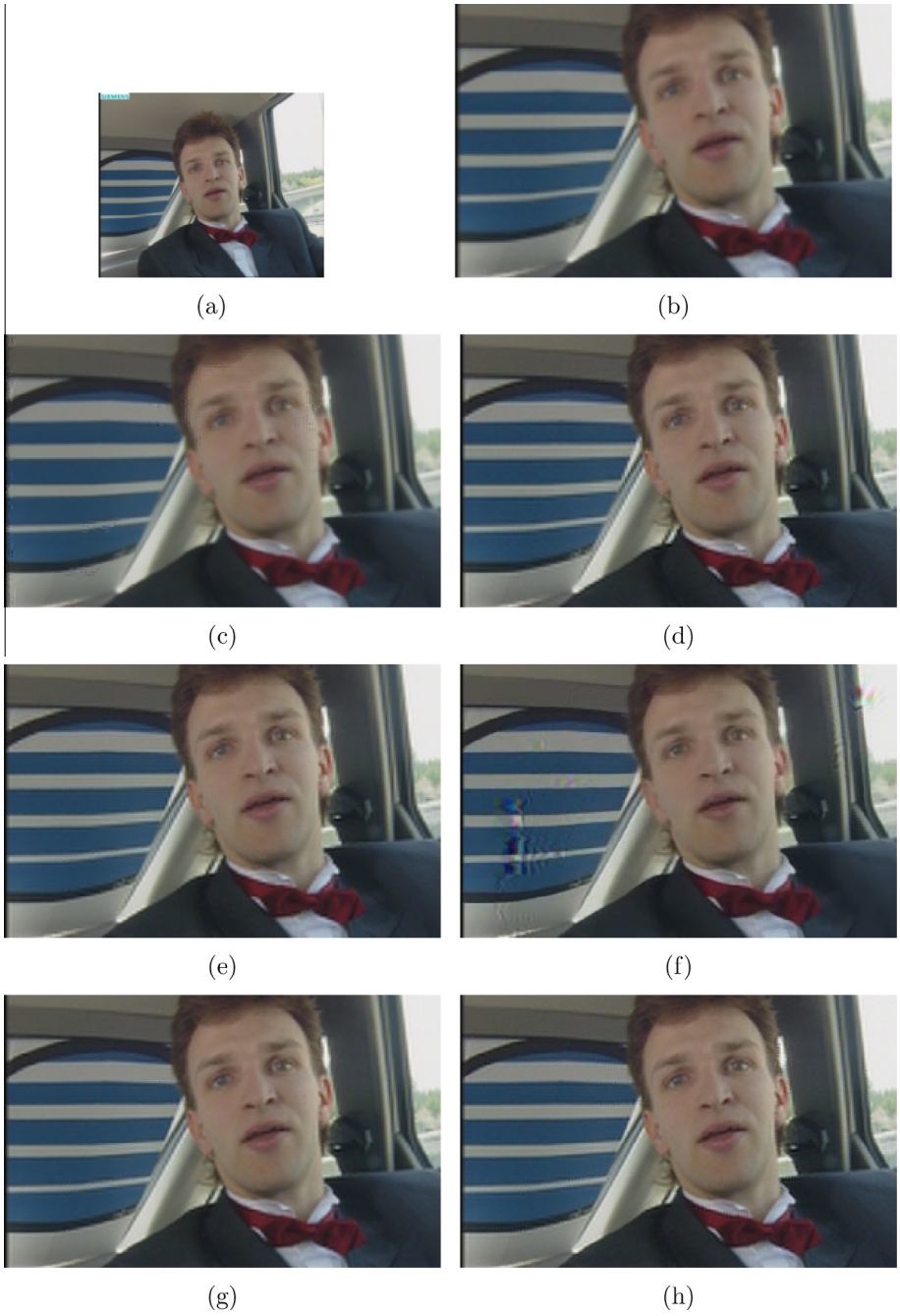

Fig. 7. The reconstruction results of the 18th frame in Carphone with different methods, (a) LR frame, (b) Bicubic, (c) MASK, (d) I-IBP, (e) RMAP, (f) BTV-L1, (g) PRO-I method and (h) PRO-II method.

other method obviously. Fig. 5 illustrates the frame-by-frame PSNR results for Foreman. Obviously the PRO-II method outperforms others for every frame. It shows that our proposed methods are more robust and achieve better results for each frames, while the BTV-L1 method is not so robust and its performance fluctuates obviously. Table 3 illustrates the similar experimental results on LR frames generated by blurring the HR frames with a 5×5 low-pass filter and then down-sampling them with decimation factor 3 in both directions. The similar conclusion can be made.

In Figs. 6 and 7, we demonstrate the subjective results of the reconstructed high resolution frame from simulated LR videos, which is blurred with known Gaussian filter and downsampled from high resolution videos. From the results, we can see that the PRO-II method produces more visually pleasing results than other methods, especially around the edges of the numbers in the red box region. In Fig. 8, we apply the super-resolution methods on the real videos, which are captured with mobile device. Since it is difficult to estimate the point spread function, we also take Gaussian filter in the observation model. From the results, we can see that the Bicubic and MASK blur the high resolution images. Although the RMAP and BTV-L1 methods improve the reconstruction quality to some extent, they are still inferior to our proposed method. Especially in the plate number area, the super-resolved image by the proposed method is much clearer and easier for recognition. We implement our proposed method with C++ program language and compile it by Microsoft Visual Studio 2010, 64-bit. We evaluate the computation complexity of the proposed method by the average running time. The tests

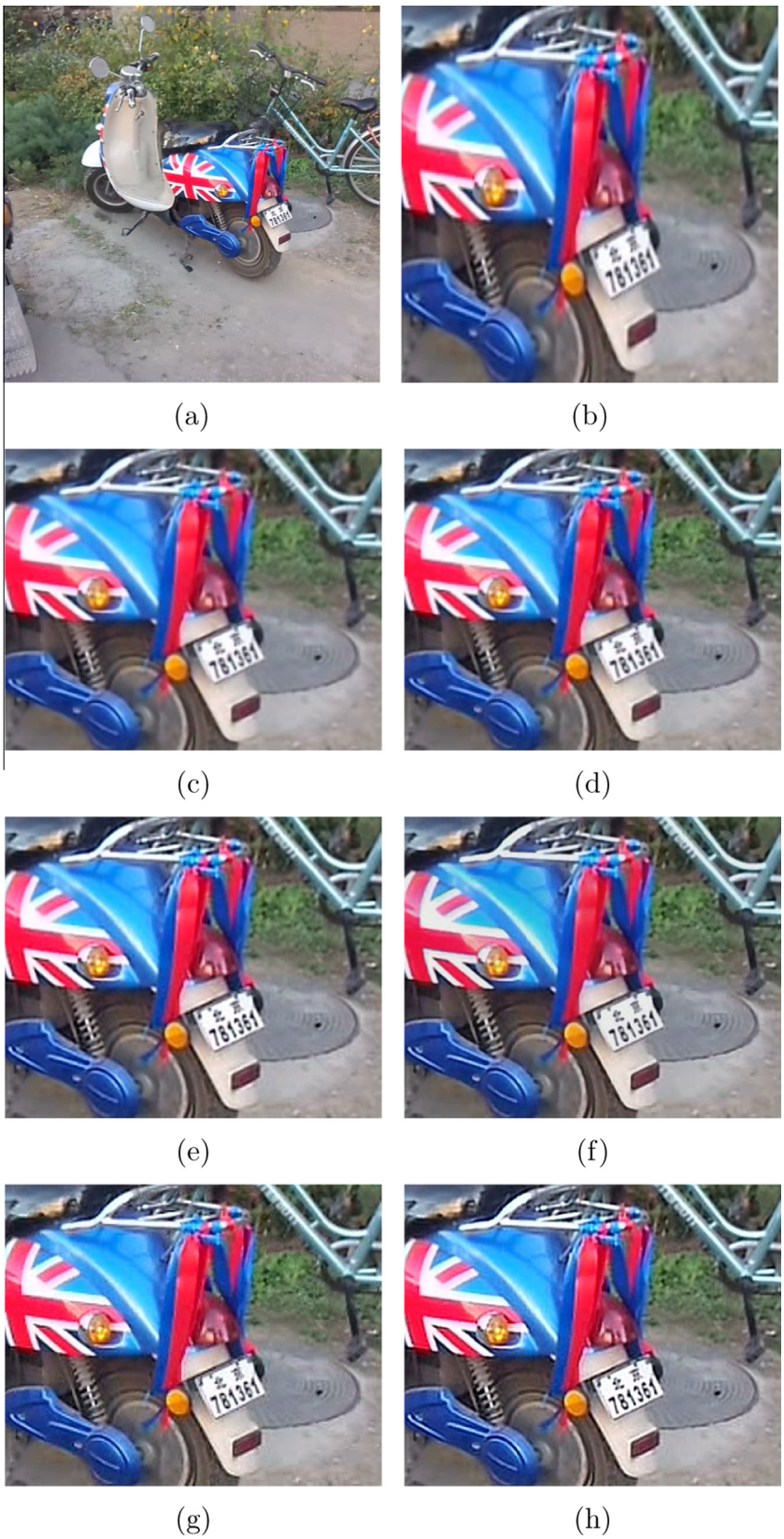

Fig. 8. The reconstruction results of the 4th frame in real video with different methods, (a) LR frame, (b) Bicubic, (c) MASK, (d) I-IBP, (e) RMAP, (f) BTV-L1, (g) PRO-I method and (h) PRO-II method.

are run on Windows 7 64-bit and the CPU in tests is Intel(R) Core(TM) i5-4570@3.20 GHz. For each QCIF frame, it needs about 5s to reconstruct a HR frame on average. The program is implemented without any optimization and the running time can be largely reduced by programming optimization and parallel techniques.

7. Conclusion

In this paper, we propose a robust video super-resolution method. The main contributions include registration-reliability regulated data-fidelity and content adaptive total variation regularization. Extensive experiments are conducted on simulated and real video sequences. Both objective and subjective results demonstrate that our proposed method outperforms the existing methods remarkably, especially for videos with complex motion.

Acknowledgments

This work was supported in part by the National Science Foundation of China (61322106, 61370114) and National Basic Research Program of China (973 Program, 2015CB351800).

References

[1] H. Takeda, S. Farsiu, P. Milanfar, Kernel regression for image processing and reconstruction, IEEE Trans. Image Process. 16 (2) (2007) 349–366.
[2] X. Zhang, X. Wu, Image interpolation by 2-d autoregressive modeling and soft-decision estimation, IEEE Trans. Image Process. 17 (6) (2008) 887–896.
[3] X. Zhang, S. Ma, Y. Zhang, L. Zhang, W. Gao, Nonlocal edge-directed interpolation, in: Proceedings of IEEE Pacific-Rim Conference International Conference on Multimedia.
[4] W. Freeman, T. Jones, E. Pasztor, Example-based super-resolution, IEEE Comput. Graph. Appl. 22 (2) (2002) 56–65.
[5] H. Chang, D.Y. Yeung, Y. Xiong, Super-resolution through neighbor embedding, in: IEEE Proceedings on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 1, 2004, pp. 275–282.
[6] J. Yang, J. Wright, T. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[7] S. Farsiu, M. Robinson, M. Elad, P. Milanfar, Fast and robust multiframe super-resolution, IEEE Trans. Image Process. 13 (10) (2004) 1327–1344.
[8] O. Omer, T. Tanaka, Multiframe image and video super-resolution algorithm with inaccurate motion registration errors rejection, in: Proceedings of the SPIE Conference on Visual Communication and Image Processing, vol. 1, 2008, pp. 275–282.
[9] M. Dennis, P. Thomas, S. Thomas, C. Danie, Video super resolution using duality based tv-l1 optical flow, in: Proceedings of the 31st DAGM Symposium on Pattern Recognition, vol. 1, 2009, pp. 432–441.
[10] X. Zhang, R. Xiong, S. Ma, W. Gao, A robust video super-resolution algorithm, in: Picture Coding Symposium (PCS), 2010, pp. 574–577.
[11] S. Izadpanahi, H. Demirel, Multi-frame super resolution using edge directed interpolation and complex wavelet transform, 2012, pp. 1–5.
[12] S. Izadpanahi, H. Demirel, Motion based video super resolution using edge directed interpolation and complex wavelet transform, Signal Process. 93 (7) (2013) 2076–2086.
[13] P. Rasti, H. Demirel, G. Anbarjafari, Improved iterative back projection for video super-resolution, in: Signal Processing and Communications Applications Conference (SIU), 2014 22nd, 2014, pp. 552–555.
[14] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, Int. Joint Conf. Artif. Intell. 3 (1981) 674–679.
[15] J.-Y. Bouguet, Pyramidal Implementation of the Affine Lucas Kanade Feature Tracker Description of the algorithm, vol. 5, Intel Corporation, 2001, pp. 1–10.
[16] S. Volz, A. Bruhn, L. Valgaerts, H. Zimmer, Modeling temporal coherence for optical flow, IEEE Int. Conf. Comput. Vis. (ICCV) (2011) 1116–1123.
[17] B.K. Horn, B.G. Schunck, Determining optical flow, Artif. Intell. 17 (1–3) (1981) 185–203.
[18] T. Brox, A.Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: Proceedings of the European Conference on Computer Vision, 2004, pp. 25–36.
[19] H. He, L. Kondi, An image super-resolution algorithm for different error levels per frame, IEEE Trans. Image Process. 15 (3) (2006) 592–603.
[20] A. Kanaev, C.W. Miller, Multi-frame super-resolution algorithm for complex motion patterns, Opt. Express 21 (17) (2013) 19850–19866.
[21] O. Omer, T. Tanaka, Region-based super resolution for video sequences considering registration error, in: Proceedings of the 3rd Pacific-Rim Symposium on Image and Video Technology, vol. 5414, 2009, pp. 944–954.
[22] O. Omer, T. Tanaka, Region-based weighted-norm approach to video super-resolution with adaptive regularization, in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 833–836.
[23] H. Takeda, P. van Beek, P. Milanfar, Spatio-temporal video interpolation and denoising using motion-assisted steering kernel (mask) regression, in: Proceedings of IEEE International Conference on Image Processing, 2008, pp. 637–640.
[24] L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (14) (1992) 259–268.
[25] Q. Yuan, L. Zhang, H. Shen, Regional spatially adaptive total variation super-resolution with spatial information filter and clustering, IEEE Trans. Image Process. 22 (6) (2013) 2327–2342.
[26] A.K. Katsaggelos, R. Molina, J. Mateos, Super resolution of images and video, Synth. Lect. Image Video Multimedia Process. 3 (1) (2007) 1–134.
[27] X. Zhang, R. Xiong, S. Ma, L. Zhang, W. Gao, Robust video super-resolution with registration efficiency adaptation, in: the SPIE Conference on Visual Communication and Image Processing, 2010, pp. 774432–774432.
[28] J. van de Weijer, L.J. van Vliet, P.W. Verbeek, M. van Ginkeli, Curvature estimation in oriented patterns using curvilinear models applied to gradient vector fields, IEEE Trans. Pattern Anal. Mach. Intell. 23 (9) (2001) 1035–1042.
[29] V. Dor, R. Farrahi Moghaddam, M. Cheriet, Non-local adaptive structure tensors: application to anisotropic diffusion and shock filtering, Image Vis. Comput. 29 (11) (2011) 730–743.
[30] Y. Wang, J. Yang, W. Yin, Y. Zhang, A new alternating minimization algorithm for total variation image reconstruction, J. Sci. Comput. 45 (1–3) (2010) 272–293.
[31] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (1) (2009) 183–202.
[32] Opencv 2.4.7, <http://sourceforge.net/projects/opencvlibrary>.