SPACE-TIME SUPER-RESOLUTION FROM MULTIPLE-VIDEOS

Esmaeil Faramarzi, Dinesh Rajan, and Marc P. Christensen

Department of Electrical Engineering, Southern Methodist University, 6251 Airline Road, Dallas, Texas 75275-0338
efaramarzi, rajand, mpc @ lyle.smu.edu

ABSTRACT

In this paper, a new method for reconstructing a video of higher spatial and temporal resolutions from multiple low-resolution video sequences is introduced. The proposed cost function includes a new space-time regularization term based upon Huber-Markov random field (HMRF) model which is convex but non-quadratic. However, this prior can be efficiently and accurately approximated by a quadratic form through an iterative process. This regularization form is a type of variational integral that exploits the piecewise smoothness nature of high-resolution images. We also address in detail the main reason for appearance of the so called "ghosting effect" in the temporal super-resolution reconstruction and explain how it can be resolved.

1. INTRODUCTION

In many applications such as medical imaging, astronomy, surveillance, and remote sensing, capturing high-quality images and videos is very desirable. Traditional high-resolution (HR) imaging systems require high-cost and bulky optical elements, the physical sizes of which dictate the light-gathering capability and the resolving power of the imaging system. This is a serious constraint that has persisted since their invention [1, 2]. In contrast, computational imaging systems combine the power of digital processing with data gathered from optical elements to extract HR information.

Image super-resolution (ISR) refers to signal processing algorithms that produce a HR image of higher spatial resolution by fusing information from one or a series of low-resolution (LR) images degraded by various artifacts such as aliasing, blurring and noise. The same concept extends to the area of video super-resolution (VSR), which is the process of reconstructing a HR video of higher spatial and/or temporal resolution from one or multiple video sequences.

The spatial resolution of an imaging system depends on the spatial density of the detector (sensor) array and the point spread function (PSF) of the lens (optics). The temporal resolution, on the other hand, is influenced by the frame rate and exposure time of the camera [3]. Spatial aliasing appears in video frames when the cut-off frequency of the detector (sensor array) is lower than that of the lens. By contrast, temporal aliasing results when the frame-rate of the camera is not high enough to capture the high frequencies caused by fast camera/scene movement.

Most of the proposed works in the literature on VSR operate on a single input video at a time. In one class of these techniques, a video of higher spatial resolution is constructed from a single LR video by defining a sliding window around each frame and applying an ISR algorithm to the frames inside the window to build a HR frame [4, 5]. For this system to work, usually a local registration algorithm (such as optical flow, block-based, pel-recursive, or Bayesian [6]) is needed to estimate the motion vectors of all pixels or small patches. However, local registration may not be reliable in some cases especially when there are complex dynamic changes (e.g. complex 3D motions), non-rigid deformations (e.g. flowing water, flickering fire), or changes in illumination [7].

Another class of single-video super-resolution (SVSR) techniques that has received attention in recent years, is the one known as learning-based, patch-based or example-based VSR [8, 9]. Here the basic idea is that small space-time patches within a video are repeated many times inside that video or other videos, at multiple spatio-temporal scales. By replacing LR patches with their corresponding HR patches, the resolution can be enhanced. The major advantage of patch-based methods is that motion estimation and/or segmentation are not required. However, techniques of this type often have high computational complexity and most of them need offline database training. Also in most cases, the results have been only tested on video sequences with low amount of aliasing.

Multiple-video super-resolution (MVSR) is proposed to increase both spatial and temporal resolutions through fusing a number of LR video sequences having subpixel displacements in space and time. A big challenge here is to find the alignment parameters between the corresponding video frames, which is a three-dimensional (3D) registration problem in general. However as in [7], we restrict the problem to the case that 1) the internal parameters of cameras (such as exposure time, frame-rate, and focus) are fixed but unknown, 2) the inter-

978-1-4673-0382-8/12/$31.00 ©2012 IEEE

camera external parameters (such as zooming) are relative, 3) the cameras are either stationary or moving together, and 4) the inter-camera distances are negligible relative to the camera-scene distances, or the scene is nearly planar. With these restrictions, spatial displacement between two sequences is modeled as a parametric 2D projective transformation (homography), and temporal misalignment is modeled as a 1D affine transformation. Hence, only one set of space-time registration parameters needs to be computed between the two sequences.

SVSR has some unique features compared to other SR problems which have been thoroughly explained in [3], such as 1) no need for complex "inter-frame" alignments, 2) the potential of combining different space-time inputs, 3) the feasibility of producing different space-time outputs, and 4) the possibility of handling severe motion aliasing and motion blur without the need of doing motion segmentation.

To the best of our knowledge, MVSR has been considered in only two works so far. In [3] a MVSR system is introduced for the first time by applying super-resolution simultaneously in space and in time. The defined cost function consists of one quadratic fidelity term and three quadratic directional space-time regularization terms. Directional smoothing is performed which does not smooth across space-time edges. The weights in the regularization terms are determined by the location, orientation, and magnitude of space-time edges and are approximated using space-time derivatives in the low resolution sequences. In the other work [10], a HR video is modeled as a Markov random field (MRF) and maximum *a posteriori* (MAP) estimate is applied to obtain the final solution of higher spatio-temporal resolution using graph-cut optimization technique.

In this paper, a new MVSR method is proposed. The cost function in our work includes a spatio-temporal prior for the HR image based on a Huber-Markov random field (HMRF) model. The HMRF prior suppresses noise and artifacts while preserving edges and fine structures effectively without causing any noticeable ringing. The HMRF prior is convex but non-quadratic; however under the majorization-minimization (MM) framework, we drive a quadratic upper-bound for this prior which makes the whole cost function quadratic and hence allows for employing a fast iterative optimization method such as conjugate gradient (CG) to solve the optimization problem. Another novelty of our work is the successful suppressing of so-called "ghosting effect", an annoying artifact that is intrinsic to MVSR systems and appears when there are fast relative movements between the camera and the scene (or an object in the scene).

This paper is organized as follows: Section 2 explains the SR observation (forward) model used in our work. The proposed space-time SR method is introduced in Section 3 and the optimization procedure is explained in Section 4. In section 5 we discuss how the ghosting effect can be suppressed from the reconstructed output. Finally experimental results are presented in Section 6.

2. OBSERVATION MODEL

The linear forward imaging model which illustrates the process of generating the kth (of total N) observed LR video sequence g_k from a HR video sequence f, is defined in the space-time domain as:

$$g_k(x_\downarrow, y_\downarrow, c, t_\downarrow) = [f_w(x, y, c, t) * h_k(x, y, 1, t)]_{\downarrow_{L_k}} + n_k(x_\downarrow, y_\downarrow, c, t_\downarrow) \quad (1)$$

In (1) g_k is of size $N_x^g \times N_y^g \times C \times N_t^g$ where N_x^g and N_y^g are the spatial vertical and horizontal dimensions of LR frames, respectively, c is the color-channel number from totally C color channels (one for gray-scale sequences and three for color sequences), and N_t^g is the number of LR frames. Also f_w is the kth warped version of f of size $N_x^f \times N_y^f \times C \times N_t^f$, h_k is the kth PSF caused by the overall effects of spatial blurring (caused by factors such as lens and atmosphere blurring, detector's light integration and scene depth) and temporal blurring (resulted by exposure-time of the camera), and n_k is the noise which is commonly modeled as AWGN. The symbol $*$ is the convolution operator and the symbol $\downarrow_{L_k}$ indicates downsampling in both space and time. The positions of space-time points in the HR and LR domains are shown, respectively, by (x, y, c, t) and $(x_\downarrow, y_\downarrow, c, t_\downarrow)$ which are related by a spatial inter-camera homography and a temporal affine transformation [7]. The sequences g and f are defined in YIQ color space.

In matrix notation, (1) is rewritten as:

$$\boldsymbol{g}_k = \boldsymbol{D}_k\boldsymbol{H}_k\boldsymbol{M}_k\boldsymbol{f} + \boldsymbol{n}_k = \boldsymbol{W}_k\boldsymbol{f}_k + \boldsymbol{n}_i \quad (2)$$

where $\boldsymbol{g}_k$ and $\boldsymbol{f}$ are LR and HR sequences in lexicographical notation, matrix $\boldsymbol{M}_k$ is the motion (warping) operator, matrix $\boldsymbol{H}_k$ is the blur convolution operator, and $\boldsymbol{D}_k$ is the downsampling matrix.

The model in (2) can be expressed in a compact form as:

$$\boldsymbol{g} = \boldsymbol{W}\boldsymbol{f} + \boldsymbol{n} \quad (3)$$

where $\boldsymbol{g} = [\boldsymbol{g}_1^T, \dots, \boldsymbol{g}_N^T]^T$, $\boldsymbol{n} = [\boldsymbol{n}_1^T, \dots, \boldsymbol{n}_N^T]^T$ and $\boldsymbol{W} = diag\{\boldsymbol{W}_1, \dots, \boldsymbol{W}_N\}$.

3. SPACE-TIME SUPER-RESOLUTION

Our proposed cost function for estimating the HR video sequence is:

$$J(\boldsymbol{f}) = \sum_{k=1}^{N} \|\boldsymbol{O}_k(\boldsymbol{g}_k - \boldsymbol{W}_k\boldsymbol{f})\|_2^2 + \lambda \left\| \boldsymbol{\rho}\left(\sqrt{\sum_{i=1}^{5} (\boldsymbol{B}_i\boldsymbol{f})^2} \right) \right\|_1 \quad (4)$$

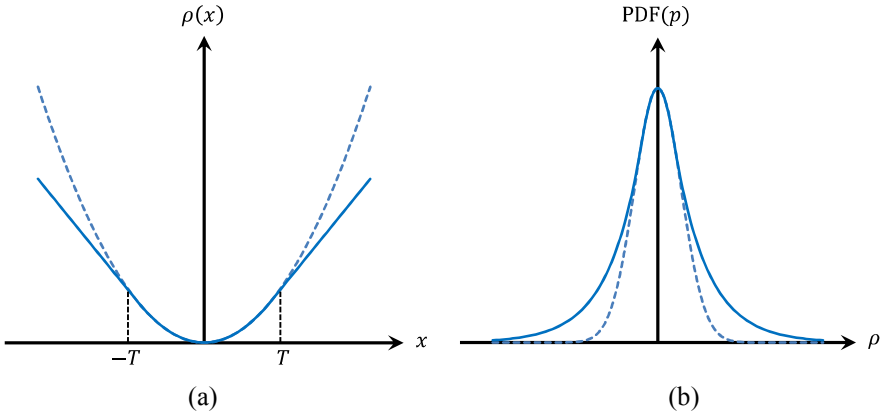

Fig. 1. (a) Huber (solid line) and quadratic functions (dashed line). (b) Their corresponding Gibbs PDFs.

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote $L1$ and $L2$ norms, respectively. The first term in (4) represents the norm of the residual error between the observed and simulated LR sequences and is known as the data fidelity or the data fusion term. The logical matrices $\boldsymbol{O}_k$ ($\boldsymbol{O}_1$ = Identity matrix)**,** which we call them motion validity maps, will be explained in detail in Section 5. The second term in (4) is a regularization term whose role is to remove or at least alleviate the illposedness of SR reconstruction and λ is the regularization coefficient. It should be noted that the root and square operations in the regularization term are element-wise; i.e. they are applied to each element of their underlying vectors separately (see [11] for better understanding of such a notation). Each element of vector function $\boldsymbol{\rho}(\cdot)$ is the scalar Huber function defined as:

$$\rho(x) = \begin{cases} x^2 & |x| \le T \\ 2T|x| - T^2 & |x| > T \end{cases} \tag{5}$$

and is shown in Fig. 1(a). This Huber function $\rho(x)$ is a convex function that has a quadratic form for values less than or equal to a threshold T and a less-than-quadratic (linear) growth for values larger than T. Consequently, edge contents in the video sequences are less penalized with this prior than with a Gaussian (quadratic) prior. Fig. 1(b) shows the Gibbs PDF of the Huber function which is heavier in the tails than a Gaussian.

The matrices $\boldsymbol{B}_1, \dots, \boldsymbol{B}_4$ in (4) are the convolution operators of first-order derivative (FOD) masks in the spatial directions of 0, 45, 90 and 135 degrees. Moreover, the matrix $\boldsymbol{B}_5$ is the convolution kernel of zero-degree FOD mask in the temporal direction.

4. OPTIMIZATION ALGORITHM

In order to minimize the cost function in (4), we use the majorization minimization (MM) technique to iteratively replace the regularization term, which is non-quadratic, with its corresponding majorization function and hence make the cost functions' derivatives linear. By following the work in [11], we obtain the majorization functions for the regularization term in (4) at the nth iteration as:

$$\boldsymbol{Q}(\boldsymbol{f}^n;\, \boldsymbol{f}^{n-1}) = \sum_{i=1}^{5} \|\boldsymbol{B}_i \boldsymbol{f}^n\|^2_{\boldsymbol{\Lambda}^n} \tag{6}$$

where $\|\boldsymbol{A}\|^2_{\boldsymbol{B}} = \boldsymbol{B}^T\boldsymbol{A}\boldsymbol{B}$ for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and:

$$\boldsymbol{\Lambda}^n = \operatorname{diag}\left(\begin{cases} 1 & \boldsymbol{\Delta}^n \le T \\ \dfrac{T}{\boldsymbol{\Delta}^{n-1}} & \boldsymbol{\Delta}^n > T \end{cases} \right) \tag{7}$$

where $\boldsymbol{\Delta}^n = \sqrt{\sum_{j=1}^{5} \left(\boldsymbol{B}_j \boldsymbol{f}^n\right)^2}$. Now by substituting (6) into (4), the minimization with respect to $\boldsymbol{f}$ leads to the following linear system:

$$\left(\sum_{k=1}^{N} \boldsymbol{W}_k^{n\,T} \boldsymbol{O}_k \boldsymbol{W}_k^n + \lambda \sum_{i=1}^{5} \boldsymbol{B}_i{}^T \boldsymbol{\Lambda}^n \boldsymbol{B}_i \right) \boldsymbol{f}^n = \sum_{k=1}^{N} \boldsymbol{W}_k^{n\,T} \boldsymbol{O}_k \boldsymbol{g}_k \tag{8}$$

5. SUPPRESSING THE GHOSTING EFFECT

The PSF h_k in (1) includes the effect of integrating light during the exposure time of the video camera. This temporal blurring effect has a rectangular shape and for this reason it is more likely to produce a temporal artifact called ghosting effect. This effect appears in temporal super-resolved video sequences as replicas of a fast moving object, as shown in [3].

It is explained in [3] that this problem emanates from the inability of the linear system of equations in recovering the frequencies that have been set to zero by the temporal rectangular blur. So if such frequencies are in some manner born in the reconstruction process due to noise, they will remain in the estimated solution. Applying the regularization term reduces this effect but cannot fully suppress it. However, it is not stated in [3] that how the frequencies set to zero by the temporal PSF waken in the reconstruction process. We now explain the reason of appearing this effect below.

To be able to obtain temporal enhancement, there should be sub-frame displacements between the LR video sequences. For a successful reconstruction, we need to precisely know or estimate spatial and temporal inter-camera registration parameters. When the motion in the scene is global, the inter-camera registration is accurate to align the video sequences at every pixel location. However, when an object in the scene is moving fast specifically with a non-linear speed, global inter-camera registration is not adequate to accurately align that object along the corresponding frames of different cameras.

Fig. 2. The "wagon wheel effect" appeared as a result of motion aliasing. (a)-(d) Three successive frames from one of four PAL video recordings of a fan rotating clockwise. The fan seems to be moving counter-clockwise. (e) Six successive (left to right, up to down) frames of the reconstructed video in which the spatial and temporal resolutions were increased by factors of 2 and 3, respectively. Now the fan is rotating in the true direction. (f) Comparison between a reconstructed frame with the proposed method (left) and with the result reported in [3] (right; converted to gray-scale).

These small misalignments are amplified in the reconstruction process because the linear system in (8) cannot apply any constraint on the frequencies that are set to zero by the temporal PSF.

In order to suppress the ghosting effect in the output sequence, we replace the rectangular temporal PSF with a soft filter such as Gaussian or hamming that drops off smoothly over time and has approximately the same width as the original rectangular PSF. Also we add the motion validity map matrices $\boldsymbol{O}_k$ to the fidelity term of the cost function to exclude from the reconstruction process those pixels of $\boldsymbol{g}_k$ that have inaccurate motion vectors. To do this we propose the following method:

I. Compute the kth displaced frame difference (DFD) matrix $\boldsymbol{A}_k$ between the kth LR video sequence $\boldsymbol{g}_k$ and a reference video sequence $\boldsymbol{g}_r$ as:

$$\boldsymbol{A}_k = |\boldsymbol{f}_k - \boldsymbol{M}_k \boldsymbol{f}_r| \tag{9}$$

II. Compute $\boldsymbol{A}'_k$ by truncating $\boldsymbol{A}_k$ in space and time to remove the boundary pixels since their corresponding pixels in other frames may drop out of the frame of the video due to motion.

III. Estimate Threshold T_k as:

$$T_k = \mu_k + 2\sigma_k \tag{10}$$

where

$\mu_k = \left\| \frac{1}{N_g} \boldsymbol{A}'_k \right\|_1$ and $\sigma_{ik}^{DFD} = \left\| \frac{1}{N_g - 1} (\boldsymbol{A}'_k - \mu_k) \right\|_2$ are the mean and standard deviation of all elements within $\boldsymbol{A}'_k$.

IV. Finally, matrix $\boldsymbol{O}_k$ results by applying the following condition:

$$\boldsymbol{O}_k = \begin{cases} 1 & \text{if } \boldsymbol{A}_k < T_k \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

The enhancement obtained by this proposed technique is shown in the next section.

6. EXPERIMENTAL RESULTS

As the first experiment, Reduction in motion aliasing by increasing temporal resolution of a real video is demonstrated in Fig. 2. Four PAL (25 frame/sec) video sequences (available at the demo web page of [3]) of a fan rotating clockwise very fast are used as the input sequences. Three successive frames of each of these videos are shown in Figs. (a)-(d). Due to inadequate temporal sampling, motion aliasing is appeared as the familiar wagon wheel effect (rotation in the opposite direction). We used our SR method to increase the spatial resolution by a factor of 2 in each direction and temporal resolution by a factor of 3. The resulting HR sequence shown in Fig. 2(e) displays the true forward (clockwise) motion of the fan as if recorded by a high-speed camera (in this case, 75 frames/sec). Fig. 2(f) illustrates a comparison between one reconstructed frame obtained by our proposed method (left image) with one achieved by the method in [3].

For the second experiment, we synthetically generate 18 LR video sequences from the standard "Mobile" sequence by applying a special 2D Gaussian filter with standard deviation of 1 pixel, temporal rectangular blur of length 3 pixels, spatial downsampling of 2, temporal downsampling of 6, and noise with $PSNR$ of 30 dB. One LR frame is shown in Fig. 3(a). We increase the special resolution by a factor of 2 and temporal resolution by a factor of 6. Two frames of the ground truth HR sequence are shown in Figs. 3(a) and 3(b) whereas the ball is nearly fixed in Fig. 3(a) and rolling in Fig. 3(b). The corresponding LR frames are illustrated in Figs. (c) and (d). Figs. 3(e) and 3(f) show the reconstructed HR frames without applying the regularization term. As seen here, the only artifact appearing in Fig. 3(c) is noise, but Fig. 3(d) also suffers from severe ghosting effect. In Figs. 3(g) and 3(h), the effect of regularization term on diminishing noise and ghosting effect are shown, though the ghosting effect is not fully suppressed yet. Finally Figs. 3(i) and 3(j) demonstrated the results of applying the method introduced in Section 5. The improvement in Fig 3(j) is clear compared to Fig. 3(h).

7. CONCLUSION

We propose a SR method to reconstruct one video sequence of higher spatial and temporal resolutions from multiple LR video sequences. Multiple-video SR (MVSR) has two important advantages over single-video SR (SVSR): 1) Temporal SR is feasible in addition to spatial SR; (2) The challenge of local pixel-to-pixel alignment between the successive frames in SVSR is reduced to global sequence-to-sequence alignment which is much more reliable specifically for the case that the displacements between the input sequences can be modeled as 2D spatial homography and 1D affine transformation. The main drawback is the appearance of ghosting effect in the reconstructed result. We address this issue in detail and propose a technique to suppress it.

ACKNOWLEDGEMENT

The authors are very grateful to the members of the PANOPTES group at SMU for valuable discussions. This research was funded in part through a collaborative technology agreement with the U.S. Army Research Laboratory under award W911NF-06-2-0035.

REFERENCES

[1] P. Milojkovic, J. Gill, D. Frattin, K. Coyle, K. Haack, S. Myhr, D. Rajan, S. Douglas, P. Papamichalis, M. Somayaji, M. P. Christensen, and K. Krapels, "Multichannel, agile, computationally enhanced camera based on PANOPTES architecture," in *Computational Optical Sensing and Imaging*. Optical Society of America, 2009, p. CTuB4.

[2] M. P. Christensen, V. Bhakta, D. Rajan, T. Mirani, S. C. Douglas, S. L. Wood, and M. W. Haney, "Adaptive flat multiresolution multiplexed computational imaging architecture utilizing micromirror arrays to steer subimager fields of view," *Appl. Opt.*, vol. 45, no. 13, pp. 2884–2892, May 2006.

[3] E. Shechtman, Y. Caspi, and M. Irani, "Space-time super-resolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 4, pp. 531 –545, Apr 2005.

[4] E. Faramarzi, D. Rajan, and M. P. Christensen, "Blind video deconvolution and super-resolution using spatio-temporal Huber-Markov priors," *Submitted to 2012 International Conference on Image Processing (ICIP 2012)*.

[5] R. R. Schultz, L. Meng, and R. L. Stevenson, "Subpixel motion estimation for super-resolution image sequence enhancement," *Journal of Visual Communication and Image Representation*, pp. 38–50, Mar. 1998.

[6] A. Tekalp, *Digital video processing*, ser. Prentice-Hall signal processing series. Prentice Hall PTR, 1995.

[7] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1409–1424, 2002.

[8] V. Cheung, B. J. Frey, and N. Jojic, "Video epitomes." *International Journal of Computer Vision*, pp. 141–152, 2008.

[9] O. Shahar, A. Faktor, and M. Irani, "Space-time super-resolution from a single video," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pp. 3353–3360, 2011.

[10] U. Mudenagudi, S. Banerjee, and P. K. Kalra, "Space-time super-resolution using graph-cut optimization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 995–1008, 2011.

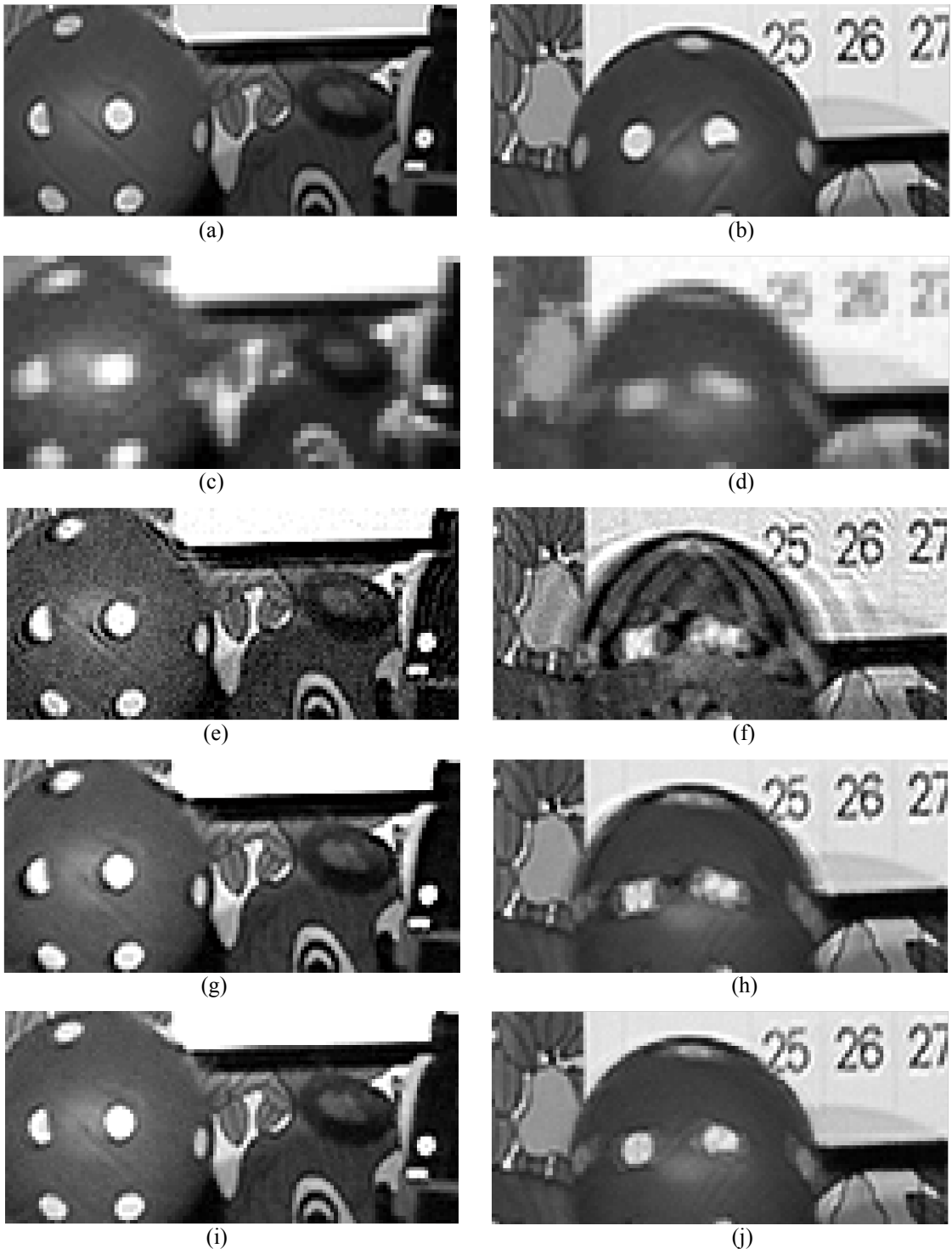

Fig. 3. Suppressing the ghosting effect, a common problem in multiple-video SR. (a) Two frames of the ground truth mobile video sequence. In frame (a) the ball is nearly fixed while in frame (b) it is rolling. (c), (d) Corresponding frames of a LR video synthetically generated by applying spatial Gaussian blur with variance 1, temporal rectangular blur of support 3, spatial downsampling of 2, temporal downsampling of 6, and $PSNR$ of $30dB$. (e), (f) reconstruction results without applying the regularization term. The ghosting effect is severely appeared in (b). (g), (h) Reconstruction result after applying the regularization term. The ghosting effect is diminished but not completely suppressed. (i), (j) Reconstruction results after applying the technique introduced in Section 5.

[11] J. P. Oliveira, J. M. Bioucas-dias, and M. A. T. Figueiredo, "Adaptive total variation image deblurring: A majorization-minimization approach," *Signal Processing*, vol. 89, pp. 1683–1693, 2009.