ADAPTIVE INCREMENTAL VIDEO SUPER-RESOLUTION WITH TEMPORAL CONSISTENCY

Heng Su[1], *Ying Wu*[2], *Jie Zhou*[3]

[1,3]Dept. of Automation, Tsinghua Univ., China; [2]Dept. of EECS, Northwestern Univ., US

ABSTRACT

Video super-resolution can be generally divided into two categories: incremental video super-resolution and simultaneous video super-resolution. Incremental video super-resolution algorithms are usually faster, but their results cannot be guaranteed to be visually consistent to the human vision system. An adaptive incremental video super-resolution framework with the temporal consistency constraint is proposed in this paper. The temporal consistency among the video frames is enforced by imposing the similarity between the adjacent reconstructed HR frames. The variances of the potential functions, which affect the weights of the different terms in the utility function, are adaptively determined so that the algorithm is robust to various motion and image content situations. Some considerations, such as the incremental motion estimation, are also incorporated to improve the efficiency of the algorithm, which makes the proposed algorithm near-realtime. The experimental results show that the proposed algorithm can generate HR video with high quality while saving the computational time as well.

Index Terms— Video super-resolution, temporal consistency, human vision system, adaptive framework

1. INTRODUCTION

Given a set of successive low-resolution(LR) frames, the purpose of reconstruction-based super-resolution(SR) technique is to recover the high-resolution(HR) version of a specific frame or the entire HR video. The super-resolution technique is first introduced in [1], which is essentially a highly ill-posed signal reconstruction problem. When the output is restricted to HR videos, the problem can be addressed as video super-resolution[2][3][4][5][6], while in this paper we call it image super-resolution when only one HR frame is required (e.g. [7][8]). The application of video SR includes video enhancement, video compression, visual surveillance, etc.

Intuitively, video SR can be performed by super-resolving the video frames one-by-one using image SR methods, which is the idea of most *incremental* video SR methods (e.g. [4][5]) (see Fig.1(a)). In this way video SR is merely a batch process of image SR algorithms. The question is, what are the differences between image SR and video SR? What can we benefit from extending from image SR to video SR? The answer could be in the following two aspects:

1. Video SR could be performed faster than applying image SR to each input frame, by saving the redundant computation;
2. Video SR could produce HR videos with better quality, by adopting the information in the video frames more effectively.

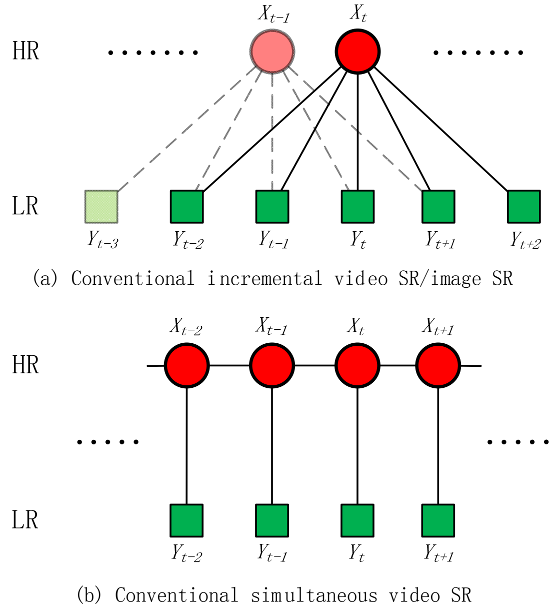


Fig. 1. The paradigms of conventional video SR algorithms. Y_t and X_t represent the input LR image and the unknown HR image at time t, respectively.

One crucial knowledge about video SR is that, even if we assume image SR generates the optimum HR image for each frame, the video composed by these images is not guaranteed to be the best HR video for the human vision system (HVS), since the video quality assessment criterion is different from image quality assessment. This makes the second aspect above reasonable. For example, people tend to appreciate videos with temporal consistency between frames, even when some frames in the video are not clear enough. *Simultaneous* video SR methods (e.g. [2][3]) (see Fig.1(b)) generate the entire HR video simultaneously within a comprehensive framework, which imposes the temporal consistency between frames. However, simultaneous video SR algorithms are usually very computationally demanding. Moreover, all LR input frames are required to perform simultaneous video SR algorithms, which limits their application in online/realtime processing.

On the other hand, various types of image contents and motion styles are likely to occur in the same video, so robustness and adaptiveness[8] are significant to a practical video SR algorithm. For example, it would handle the situation of motion registration errors[5], or complicated video contents to maintain a persistent high

The authors would like to thank the support from the Computational Vision Group in Northwestern University. This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, and US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504.

978-1-4577-1303-3/11/$26.00 ©2011 IEEE

quality performance.

Based on the considerations above, we propose an adaptive and practical incremental video SR framework, which both reduces the computational time and improves the visual quality of the output video. The generation of a certain HR frame not only depends on the input LR frames within the corresponding temporal window, but is also related to the reconstructed previous HR frame to enforce the temporal consistency. The variances of the joint Gaussian distributions (which affect the weights of the different terms in the utility function) are adaptively determined. The proposed algorithm provides a practical realtime or near-realtime solution for video super-resolution, which can be utilized in video display devices or online video play applications.

The main contribution of this paper is in three folds: First, the temporal accordance is ensured by adopting the temporal consistency term in the incremental video SR framework. Second, we introduce several adaptive processes to increase the robustness of the proposed algorithm, such as adaptively adjusting the potential functions and removing the motion registration errors without introducing much computational burden. Third, incremental motion estimation is proposed to reduce the computational cost.

The rest of this paper is organized as follows. The model adopted in the proposed algorithm is introduced in Section 2, and the algorithm details are described in Section 3. In Section 4 we show the experimental results and finally conclusions are made in Section 5.

2. THE MODEL

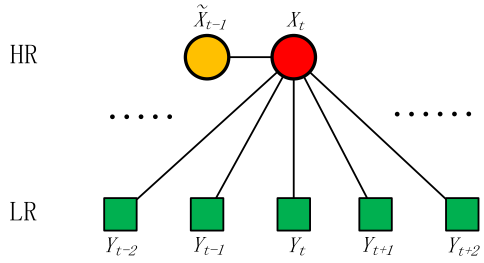


Fig. 2. The proposed framework. Note the differences from the paradigms in Fig.1. The posterior distribution of the HR image X_t depends on both the input LR images and the previous reconstructed HR image $\tilde{X}_{t-1}$.

The graph model used in the proposed framework is depicted in Fig.2. Note the differences between the paradigms in Fig.2 and Fig.1. In the figures, X_t and Y_t respectively represent the required HR image and the input LR image at time t, and $\tilde{X}_{t-1}$ is the estimated HR image of the previous frame at time $t-1$. The flow chart in Fig.2 shows that the posterior distribution of the HR image X_t in the proposed algorithm depends on both the input LR images and the previous reconstructed HR image $\tilde{X}_{t-1}$. This enforces the resulting HR video to be temporally consistent. The posterior probability can be written as:

$$p(X_t|\cdot) \propto \prod_{i=t-r}^{t+b} \Psi(X_t, Y_i) \cdot \Phi(X_t, \tilde{X}_{t-1}) \mathcal{L}(X_t), \qquad (1)$$

where r and b specify the number of the corresponding LR images within the temporal window around t, and $\Psi(\cdot,\cdot)$ and $\Phi(\cdot,\cdot)$ are the *potential functions*. Because of the ill-posed-ness of the problem, a *regularization term* $\mathcal{L}(\cdot)$ is incorporated to impose the spatial smoothness of the result HR image. Accordingly, the three functions $\Psi(\cdot,\cdot)$, $\Phi(\cdot,\cdot)$ and $\mathcal{L}(\cdot)$ are referred to as the *fidelity* term, the *consistency* term and the *regularization* term, respectively. Thus the maximum *a posteriori* (MAP) estimate of the HR frame X_t can be obtained via:

$$\tilde{X}_t = \arg\max p(X_t|\cdot). \qquad (2)$$

3. THE ADAPTIVE VIDEO SR ALGORITHM

3.1. The fidelity term Ψ

The fidelity term $\Psi(X_t, Y_i)$ measures the similarity between the HR frame X_t and the input LR image Y_i, which is defined as:

$$\Psi(X_t, Y_i) \triangleq \exp(-\frac{1}{2\sigma_i^2}\|DBM_t^i X_t - Y_i\|^2), \qquad (3)$$

where M_t^i represents the motion compensation matrix from the HR frame t to the HR frame i, B and D respectively denote the blurring matrix and the downsample matrix, and σ_i^2 is the variance of the joint distribution, which controls the confidence level of the corresponding component.

In order to remove the influence of the motion registration outliers, we adopt the intuitive idea of applying the confidence map W_t^i to the matrix M_t^i, which is defined as a diagonal binary matrix whose diagonal entries w_t^i are:

$$w_t^i(\vec{x}) = \begin{cases} 1, & |B(\tilde{M}_t^i \bar{X}_t - \bar{X}_i)|\Big|_{\vec{x}} < \eta_{\text{map}}, \\ 0, & \text{otherwise}, \end{cases} \qquad (4)$$

where $\tilde{M}_t^i$ is the original motion compensation matrix, and η_{map} is a pre-defined threshold. Note that we don't know the real HR image X_t, so an estimate $\bar{X}_t$ is used instead. In the experiments, $\bar{X}_t$ is directly interpolated from the input Y_t. Thus the new motion matrix M_k filtered by W_k is formulated as:

$$M_t^i = W_t^i \tilde{M}_t^i. \qquad (5)$$

The variance σ_i^2 can be measured by the variance of the corresponding motion registration error within the confidence map: Larger motion registration error indicates less reliable σ_i. Thus we define:

$$\sigma_i^2 = \sigma\tilde{\sigma}_i^2 = \sigma \cdot \text{var}\Big(B(\tilde{M}_t^i \bar{X}_t - \bar{X}_i)\Big| w_t^i = 1\Big), \qquad (6)$$

where σ is a constant.

According to (3), we need to estimate motion fields $\tilde{M}_t^i$ from frame t to all the frames i within the temporal window in order to calculate the potential functions for X_t, which means the motion estimation process has to run $r+b-1$ times to compute X_t. This could be simplified by adopting incremental motion estimation, i.e., the motion matrix $\tilde{M}_t^i$ is obtained by $\tilde{M}_t^i = \tilde{M}_{i-1}^i \tilde{M}_{i-2}^{i-1} \ldots \tilde{M}_{t+1}^{t+2} \tilde{M}_t^{t+1}$ when $i > t$, or $\tilde{M}_t^i = \tilde{M}_{i+1}^i \tilde{M}_{i+2}^{i+1} \ldots \tilde{M}_{t-1}^{t-2} \tilde{M}_t^{t-1}$ when $i < t$. Therefore for a new frame t, we only estimate the motion between frame $t+b-1$ and frame $t+b$: $\tilde{M}_{t+b}^{t+b-1}$ and $\tilde{M}_{t+b-1}^{t+b}$. Note this not only saves the computational burden, but also increases the motion estimation accuracy, as the composite motion field is usually more reliable with the help of the intermediate frames.

3.2. The consistency term Φ

The consistency term $\Phi(X_t, \tilde{X}_{t-1})$ controls the temporal accordance of the output HR video, which is defined as:

$$\Phi(X_t, \tilde{X}_{t-1}) \triangleq \exp(-\frac{1}{2\rho_0^2}\|M_t^{t-1}X_t - \tilde{X}_{t-1}\|^2), \qquad (7)$$

where:

$$\rho_0^2 = \rho \cdot \mathrm{var}\Big(B(\tilde{M}_t^{t-1}\bar{X}_t - \tilde{X}_{t-1})\Big| w_t^{t-1} = 1\Big), \qquad (8)$$

and ρ is a scalar. The HR frame $\tilde{X}_{t-1}$ is the previous estimated HR frame at time $t-1$.

Note that the consistency term is also adaptive. The weight ρ_0 depends on the motion registration error and the quality of the previous reconstructed HR image $\tilde{X}_{t-1}$. Smaller ρ_0 indicates that the consistency term is less reliable.

3.3. The regularization term $\mathcal{L}$

The regularization term $\mathcal{L}(X_t)$ is defined as the exponential of the bilateral total variation (TV) prior[7]:

$$\begin{aligned}\mathcal{L}(X_t) &\triangleq \exp(-\gamma(X_t)) \qquad (9)\\ &= \exp(-\underbrace{\sum_{l=-P}^{P}\sum_{j=0}^{P}}_{l+j\geq 0}\alpha^{|l|+|j|}\|X_t - S_x^l S_y^j X_t\|^2)\end{aligned}$$

where S_x^l and S_y^j are operators that shift the image X_t by l and j pixels in the horizontal and the vertical direction, respectively, and P is the local window size. The bilateral TV prior tends to preserve edges in the resulting images when $P > 1$. In the experiments we use $P = 2$.

3.4. The optimization process

To summarize, the maximization (2) is equivalent to minimizing the minus log-likelihood in (10):

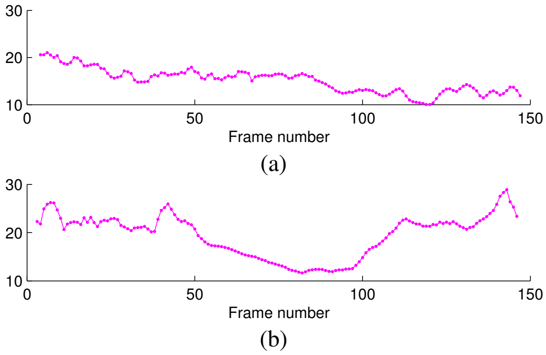
$$\begin{aligned}\tilde{X}_t &= \arg\min\Big(\gamma(X_t) + \sum_{i=t-r}^{t+b}\frac{1}{2\sigma_i^2}\|DBM_t^i X_t - Y_i\|^2 + \\ &\quad \frac{1}{2\rho_0^2}\|M_t^{t-1}X_t - \tilde{X}_{t-1}\|^2\Big). \qquad (10)\end{aligned}$$

The Equation (10) is a standard least squares problem with respect to X_t. The conjugate gradient (CG) method is applied to solve (10) for fast convergence.

4. EXPERIMENTAL RESULTS

We prepare two videos in the experiments. Each video contains 150 frames, with the frame rate 25 fps. The LR video resolution is 320×240. As we perform a 2×2 video super-resolution, the destination spatial resolution is 640×480. The original HR videos are captured by a handheld Fujifilm F601 Zoom digital camera. The LR videos are generated by blurring (with a 3×3 Gaussian blur kernel with the deviation 1-pixel wide) and downsampling the original videos, and additive Gaussian white noise is added to simulate the imaging noise. The videos contain both large-scale local motion and relatively small background global motion. The average motion registration errors within the temporal window of each frame are shown in Fig.3.

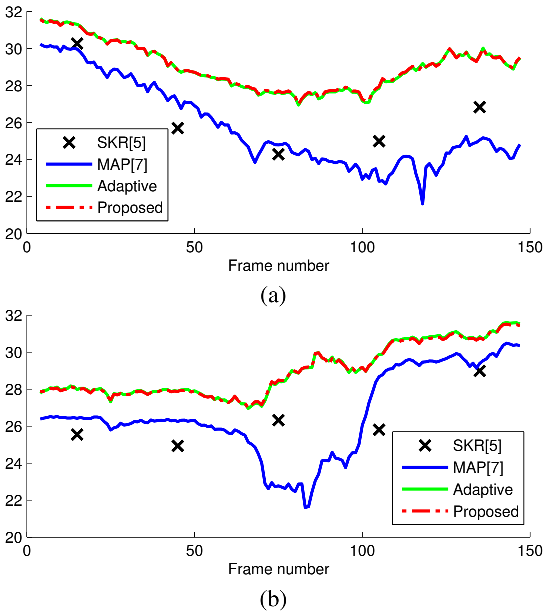
There are mainly 3 parameters σ, ρ and η_{map} in the proposed algorithm. As a rule of thumb, the two scalars σ and ρ are set as $0.01/\tilde{\sigma}_t^2$ and $0.05/\tilde{\sigma}_t^2$, respectively, which makes $\sigma_t^2 = 0.01$. The threshold $\eta_{\text{map}} = 10$. The other settings of the algorithm are as follows: $r = b = 3$ which means 7 adjacent frames are used for the reconstruction of each frame, and B is set as a spatially invariant 3×3 Gaussian blur kernel with the deviation of 1-pixel wide. The number of CG iterations is set as 3.



(a)

(b)

Fig. 3. The average motion registration errors (in PSNR, dB) within the temporal window of each frame in the two experimental videos



(a)

(b)

Fig. 4. The PSNR results (in dB) of the two experiments. The "Adaptive" algorithm in the figures is the same as the proposed except that "Adaptive" is without the consistency term (Φ). Note the curves of the "Adaptive" algorithm and the proposed are very similar to each other under the PSNR criteria, but the proposed generates HR videos with more visual consistency (see Table 1).

The PSNR results of the two experimental videos are shown in Fig.4. Note that the first and last 3 frames are directly interpolated from the corresponding input LR frames, so only the results of the middle frames are illustrated. The algorithm "MAP"[7] is similar as the proposed algorithm, but without the consistency term Φ and the adaptive weight (6)(8). We use $\sigma_i^2 = 0.01$ and $\rho_0^2 = 0.05$, and adopt

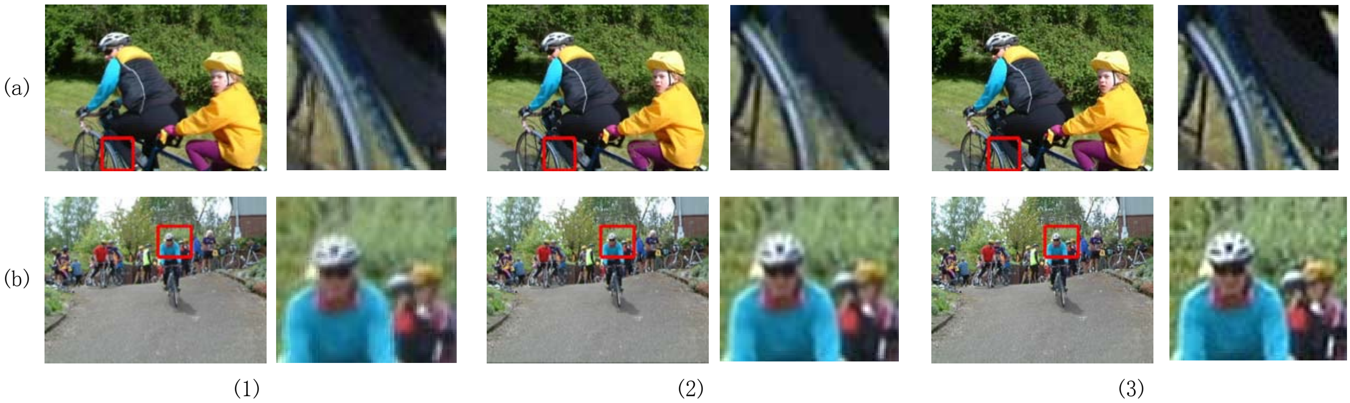


Fig. 5. Sample reconstructed image results. Row (a) shows the 75th frame in the first experimental video, while Row (b) is the 45th frame in the second video. Columns: (1) The results of SKR[5]; (2) MAP[7]; (3) The proposed.

Table 1. The MOVIE[9] scores[1] of the SR results

Exp.#	1	2
MAP[7]	2.42×10^{-3}	1.60×10^{-3}
Adaptive[2]	5.38×10^{-4}	5.17×10^{-4}
Proposed	5.10×10^{-4}	4.94×10^{-4}

[1] Smaller score indicates better video quality.

[2] The "Adaptive" algorithm is the same as the proposed except that "Adaptive" is without the consistency term (Φ).

*Note we don't show the MOVIE result of "SKR"[5] because the entire video is needed to calculate MOVIE.

the l^2-norm in "MAP" for a fair comparison. The algorithm "Adaptive" is the same as the proposed except that "Adaptive" is without the consistency term Φ. We can see the "Adaptive" method and the propose algorithm produce similar results under the single image PSNR criteria. However, adopting the consistency term improves the visual consistency in the video. This can be seen in Table 1 by comparing their MOVIE[9] scores. The MOVIE[9] is a video quality assessment criteria, which examines both the spatial and temporal consistency. Smaller score indicates better video quality.

With our computer with a 2.93GHz dual kernel CPU and a 3GB RAM, it takes around 2 second to process one frame in the proposed algorithm, while it takes around 10 hours using "SKR"[9] (The software is provided by the authors). Thus in Fig.4, we only show the results of 5 frames by the method "SKR", and we don't calculate the MOVIE score for "SKR", as the entire video is needed.

The sample reconstructed image results are shown in Fig.5, from which we can see the proposed algorithm sharpens the image edges and removes artifacts brought by motion registration outlier effectively. This is mainly because of the introduction of the adaptive weights σ_i and ρ_0.

5. CONCLUSION

In this paper, an adaptive incremental video SR algorithm is proposed. We enforce both the temporal and the spatial consistency in the framework, which generates better quality HR video results. The adaptive weights of the terms in the utility functions enable the robustness of the proposed algorithm. Some considerations to save the computational cost are also incorporated, which makes the framework near-realtime, which could be easily adopted in video display devices and video play softwares in the future.

6. REFERENCES

[1] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Advances in Computer Vision and Image Processing*, vol. 1, pp. 317–339, 1984.

[2] M. V. W. Zibetti and J. Mayer, "Simultaneous super-resolution for video sequences," in *Proc. IEEE Int. Conf. Image Processing ICIP 2005*, 2005, vol. 1.

[3] Dan Kong, Mei Han, Wei Xu, Hai Tao, and Yihong Gong, "A conditional random field model for video super-resolution," in *Proc. 18th Int. Conf. Pattern Recognition ICPR 2006*, 2006, vol. 3, pp. 619–622.

[4] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, 2007.

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

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

[7] S. Farsiu, M.D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super-resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[8] Heng Su, Liang Tang, D. Tretter, and Jie Zhou, "A practical and adaptive framework for super-resolution," in *Proc. 15th IEEE International Conference on Image Processing ICIP 2008*, 2008, pp. 1236–1239.

[9] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.