Noisy Video Super-Resolution

Feng Liu
University of Wisconsin-Madison
fliu@cs.wisc.edu

Jinjun Wang, Shenghuo Zhu
NEC Laboratories America, Inc.
jjwang|zsh@sv.nec-labs.com

Michael Gleicher
University of Wisconsin-Madison
gleicher@cs.wisc.edu

Yihong Gong
NEC Laboratories America, Inc.
ygong@sv.nec-labs.com

ABSTRACT

Low-quality videos often not only have limited resolution, but also suffer from noise. Directly up-sampling a video without considering noise could deteriorate its visual quality due to magnifying noise. This paper addresses this problem with a unified framework that achieves simultaneous de-noising and super-resolution. This framework formulates noisy video super-resolution as an optimization problem, aiming to maximize the visual quality of the result. We consider a good quality result to be fidelity-preserving, detail-preserving and smooth. Accordingly, we propose measures for these qualities in the scenario of de-noising and super-resolution. The experiments on a variety of noisy videos demonstrate the effectiveness of the presented algorithm.

Categories and Subject Descriptors

I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Algorithms

1. INTRODUCTION

Super-resolution is the problem of creating high-resolution (HR) images from low-resolution (LR) inputs. Many methods have been presented on this topic as surveyed in [4], [17] and [7]. This paper focuses on generating a HR video from a single LR input. For this special purpose, most existing methods can be categorized into two classes: reconstruction-based methods and learning-based methods. Reconstruction based methods assume a simplified continuous imaging process, under which LR images are created, and usually formulate the imaging process as a linear system. Many methods, such as maximum likelihood [22], a maximum a posteriori [19] and projection onto convex sets [20], have been presented to solve the linear system for the HR result. Reconstruction-based methods require the LR input to form a good sampling of HR one to obtain a good result. Assisted with special hardware, these methods can achieve compelling results [1, 12]. However, most videos are captured by regular devices. Learning-based super-resolution methods have recently provided attractive results(c.f. [8, 3, 6, 15]). These methods usually learn a co-occurrence prior between HR and LR image patches or coefficients, or image feature statistics, and process the LR input along with appropriate smoothness constraint to generate HR image. Learning-based methods work well when applied to a specific domain. A good example is face hallucination(c.f. [6, 15]). However, for general images/videos, learning-based methods require a large amount of training data.

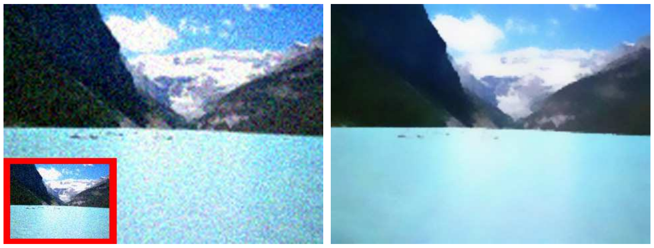

(a) Bicubic ×3 (b) Ours ×3

Figure 1: Motivating example. Super resolution can amplify noise. For example, (a) shows a noisy result created by using bi-cubic interpolation to up-sample the image indicated by the red rectangle in (a). Our algorithm achieves de-noising and super-resolution simultaneously.

Existing approaches did not explicitly consider the issue of noise in the original video. This is significant because low-resolution video is often plagued with noise, and super-resolution techniques can magnify this noise, leading to results with lower visual quality than the source as illustrated in Fig. 1. It has been our observation that the requirement of de-noising and super-resolution is similar. Both aim to achieve a result with the following three properties. First, the result should be visually similar to the input as much as possible. Second, the result should preserve or enhance the image detail. Third, the result should avoid introducing undesirable high-frequency content.

Based on these observations, in this paper we present a unified framework which achieves simultaneous de-noising

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-303-7/08/10 ...$5.00.

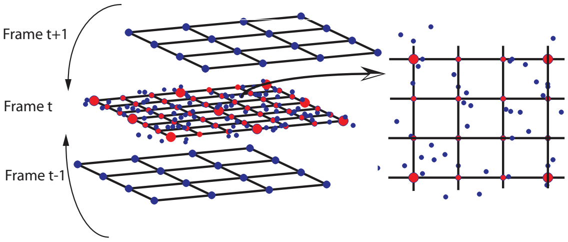


Figure 2: Neighboring frames are aligned to frame t. The big red points are low-resolution pixels in frame t, and the small red points are high-resolution pixels there. The blue points are low-resolution pixels in the neighboring frames.

and super-resolution. We extend our previous work of image super-resolution [14] and formulate noisy video super-resolution as an optimization problem aiming to maximize the overall visual quality. While video quality assessment is generally a difficult task, we approach it in the scenario of de-noising and super-resolution. Specifically, we measure the video quality in terms of the three criteria above. The key to the success of the proposed approach is to design effective measures for these criteria as detailed in § 2.

2. OUR APPROACH

The goal of this work is to create a high-resolution video from a noisy low-resolution input. Given only a single low-resolution noisy video, we necessarily miss important information to "reconstruct" a physically correct high-quality video. So instead of trying to create a "correct" high-resolution result, we aim to create a visually high-quality one.

Measuring visual quality is difficult. Most existing methods use original images to measure the visual quality of their transformed counterparts(c.f. [24]). These reference-based methods are unsuitable for our purpose. Current non-reference methods focus on measuring distortion caused by compression, thus unsuitable to our purpose either(c.f. [23]). We approach this problem in the scenario of super-resolution and de-noising, and assess the visual quality with respect to the following three aspects.

- Fidelity preserving. The super-resolution result should be visually similar to the low-resolution input.
- Detail preserving. The super-resolution result should keep and enhance details as much as possible.
- Spatial-temporal smoothness. Normally the human visual system prefers piecewise smooth results, and is very sensitive to undesirable high-frequency content, especially to the temporal jittering. The super-resolution result should avoid introducing and remove these undesirable high-frequency artifacts.

We encode violations against each of the above quality aspects as cost, and formulate super-resolution and de-noising as an optimization problem that aims to minimizing the total cost (equivalently maximizing the visual quality). The key is to define effective quality measures. We will describe perception-based measures in the following subsections.

2.1 Fidelity preserving

A straightforward metrics to measure the similarity between the low-resolution input and its super resolution result is the sum of the difference between each low-resolution pixel and its high-resolution counterpart as follows:

$$E_{fd} = \sum_{(x,t)\in I^l} \|I^h(x*scale,t) - I^l(x,t)\|_2^2$$

where E_{fd} is the difference between super-resolution result I^h and its low-resolution input I^l, (x,t) is the index of pixel x at frame t, and $scale$ is the magnification rate. This simple metrics is problematic and wastes the spatial-temporal information in the low-resolution video. First, the input suffers from noise, so this simple method will create a noisy result, possibly even magnifying the noise. Second, neighboring frames can provide information both for de-noising and super-resolution.

We have developed a robust motion estimation algorithm that will be described in the following subsection 2.1.1. Using the motion estimation result, we can align neighboring frames to every frame $I^l(t)$, as illustrated in Figure 2. Thus each pixel of the corresponding high-resolution frame $I^h(t)$ has a certain amount of space-time neighboring pixels. These neighboring pixels provide valuable information both for de-noising and super-resolution. We can estimate an approximation of super-resolution result from them.

Denote the approximation as $\tilde{I^h}$. We estimate it using a space-time median-bilateral filer to estimate the noise-reduced high-resolution video as follows.

$$\tilde{I^h}(x,t) = \frac{\sum_{(x',t')\in V_{x,t}} w(x',t',x,t)I(x',t')}{\sum_{x'\in V_{x,t}} w(x',t',x,t)} \qquad (1)$$

where $\tilde{I^h}(x,t)$ is value of pixel x at frame t of noise-reduced high-resolution video and $V_{x,t}$ is the space-time neighborhood of pixel (x,t) aligned to frame t. $w(x',t',x,t)$ is the adapted space-time bilateral filtering weight, where the impact of each pixel (x',t') is related to its distance to (x,t), its value, and the confidence of motion estimation as follows:

$$w(x',t',x,t) = w_{mv}(x',t',t)g_s(x'-x,\sigma_s)g_t(t'-t,\sigma_t) \qquad (2)$$
$$g_v(I(x',t') - med(x,t),\sigma_v)$$

where g_s, g_t and g_v are gaussian distributions, w_{mv} is the motion estimation quality measure defined in Equation 5, and $med(x,t)$ is the median pixel value in $V_{x,t}$.

To keep the fidelity, we encourage the super-resolution result I^h similar to the above noise-reduced high-resolution result $\tilde{I^h}$.

$$E_{fdv} = \sum_{(x,t)\in I^l} \|I^h(x*scale,t) - \tilde{I^h}(x*scale,t)\|_2^2 \qquad (3)$$

Research in visual perception and neuro-science suggests that the human visual system is more sensitive to contrast rather than pixel values [9, 16]. For images, local contrast can be approximated by image gradient. Accordingly, we encourage the gradient field of the high-resolution image to be close to that of $\tilde{I^h}$ as follows:

$$E_{fdg} = \sum_{(x,p)} \frac{\|G^h(x,t) - \tilde{G^h}(x,t)\|_2^2}{\|\tilde{G^h}(x,t)\|_2^2 + \epsilon} \qquad (4)$$

where G^h and $\tilde{G^h}$ are the gradient fields of I^h and $\tilde{I^h}$ respectively, and ϵ is a constant. The denominator of this equation is used to account for the “masking effect” [11], which states that changing in high gradient is less obvious than the same amount of changing in low gradient.

2.1.1 Motion estimation

Accurate and consistent pixel-wise motion estimation is important both to estimate a sharp noise-reduced high-resolution image and create temporally jittering-free high-resolution result. Traditional optical flow methods suffer from noise. We adopt the idea of the recent piecewise image registration method to estimate accurate optical flow [2]. The basic idea is to estimate some motion models in the scene and assign each pixel to one of the estimated motion model with the corresponding disparity using a multi-label graph-cut optimization. Since it is a time-consuming process, we simplify it as follows: We first estimate the global background motion, and assign it to the pixels that can be well predicated by the global motion. Then we estimate other pixels' motion using block matching method guided by the global motion model. Besides accuracy, one important advantage of using this method is that the estimated motion is more spatial-temporarily consistent than standard optical flow methods. According to the research in visual perception(c.f. [18]), the human visual system is very sensitive to local motion contrast than the motion itself. The above method provides consistent motion estimation, avoiding jittering. We measure the quality of the motion estimation based on the block matching error residuals as follows:

$$w_{mv}(x,t,t') = \exp(-\sum_{y \in B_x} \|I(y,t) - I(y + M_{x,t,t'})\|) \quad (5)$$

where $M_{x,t,t'}$ is the motion vector of pixel x between frame t and t', y is a pixel in B_x, the block centered at x, and $I(y,t)$ is the pixel value at y in frame t. Using these strategies, we can obtain a robust motion estimation along with its quality measurement.

2.2 Detail preserving

Preserving and enhancing image details are one of the major focuses of image super-resolution and de-nosing methods. Many existing methods enhance image details by sharping image edges [10, 13, 21, 5]. Edge directed methods [10, 21] estimate high-resolution edges from the low-resolution input, and use the edge information to guide super-resolution operations, such as interpolation and image reconstruction. The performance of these methods are subject to the quality of high-resolution edge estimation. A fundamental problem with edge-guided methods is that edges are often not good representations of image details. For example, details in image regions with rich fine textures are hard for edges.

According to the research in visual perception, details manifest themselves through local contrast to the low-level human visual system. Hence, we preserve/enhance image details by enhancing local contrast instead of edges. We calculate the local contrast in each image patch as the sum of difference between every two pixels. We define the following measure to encourage enhancing the detail:

$$E_{dt} = -\sum_{patch_k \in I^h} w_k \sum_{x_i, x_j \in patch_k} \|I^h(x_i) - I^h(x_j)\|_2^2 \quad (6)$$

where $patch_k$ denotes an image patch in the high-resolution image I^h. w_k is a weight. There are several options to set the weights. w_k can be a binary variable, set as 1 when there is an edge passing through the patch. The edge is estimated from the noise-reduced high resolution image $\tilde{I}^h$. To set w_k, no accurate information about the edge location inside the patch is required, which improves the tolerance of our method against error in edge estimation and is robust to noise.

2.3 Smooth

Smooth results are often favored by the human visual system. We encourage minimizing the modified space-time Laplacian of the high-resolution result to achieve smoothness:

$$E_{sm} = \sum_{(p,t)} \|\frac{\partial^2 I^h}{\partial p_x^2} + \frac{\partial^2 I^h}{\partial p_y^2} + w_{mv}(p,t,t+1)\frac{\partial^2 I^h}{\partial p_t^2}\|_2^2 \quad (7)$$

where $w_{mv}(p,t,t+1)$ is the motion estimation quality measure to accounting for the accuracy of motion estimation as defined in Equation 5, and $\frac{\partial^2 I^h}{\partial p_t^2}$ is calculated in the video where adjacent frames are aligned to t to account for the dynamics of the scene.

2.4 Optimization

Based on the quality measures defined in the above subsections, we formulate video de-noising ad super-resolution as an optimization problem by linearly combining all the measures as follows:

$$E = \lambda_{fidv}E_{fidv} + \lambda_{fidg}E_{fidg} + \lambda_{dt}E_{dt} + \lambda_{sm}E_{sm} \quad (8)$$

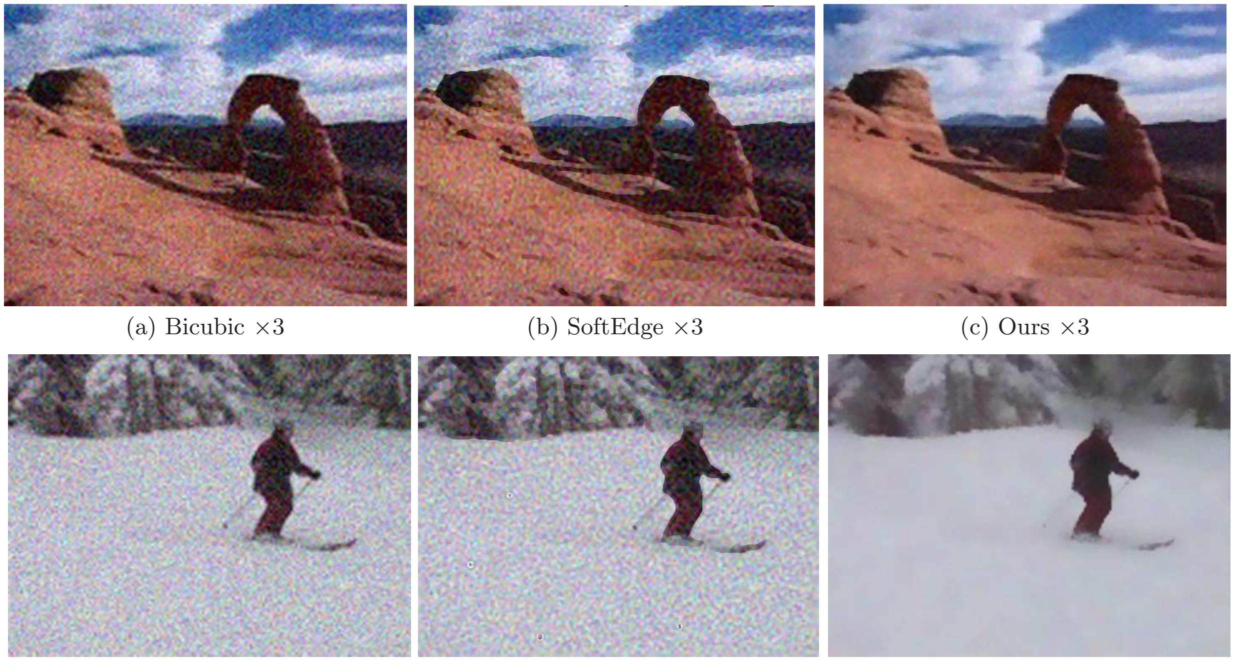
where $\lambda_?$ is weight for each term. We calculate image gradient and Laplacian using finite difference methods. Since all the measures are at most quadratic, the above problem is a quadratic minimization problem, and can be solved efficiently using a standard sparse linear system solver.

3. RESULTS

We experimented with our algorithm on some videos recorded by amateur users. To evaluate the performance of our algorithm against noise, we manually added white noise into these videos. Typical examples are shown in Figure 3. We did an informal user study to further evaluate our algorithm by comparing our results to those of the bicubic upsampling and the recent soft-edge algorithm [5]. Totally 6 users participated the study. Each user was shown 8 pairs, and was asked to select the one they consider better. The feedbacks from the participants consistently show that our algorithm creates more visually pleasant high-resolution videos than the traditional bicubic method and the recent soft-edge method [5]. Particularly, our results are much temporally smoother and less noisy than the others.

4. CONCLUSION

In this paper, we presented a unified framework that achieves simultaneous de-noising and super-resolution. This framework formulates noisy video super-resolution as an optimization problem, aiming to maximize the visual quality of the result. We consider a good quality result to be fidelity-preserving, detail-preserving and smooth. Accordingly, we presented measures for these qualities in the scenario of de-noising and super resolving. Our experiment shows the initial success of our method.



(a) Bicubic ×3
(b) SoftEdge ×3
(c) Ours ×3
(d) Bicubic ×3
(e) SoftEdge ×3
(f) Ours ×3

Figure 3: Examples. The top row shows a frame from a video where the camera pans around, and the bottom row shows a frame where the camera follows the skier.

Acknowledgements. The copyright of original videos in Fig.1 and the top row of Fig. 3 belongs to Youtube users: Michal2cze and yacro respectively.

5. REFERENCES

[1] M. Ben-Ezra, A. Zomet, and S. Nayar. Jitter camera: High resolution video from a low resolution detector. In *IEEE CVPR*, pages 135–142, 2004.

[2] P. Bhat, K. C. Zheng, N. Snavely, A. Agarwala, M. Agrawala, M. F. Cohen, and B. Curless. Piecewise image registration in the presence of multiple large motions. In *IEEE CVPR*, 2006.

[3] C. M. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. In *AISTATS*, 2003.

[4] S. Borman and R. L. Stevenson. Super-resolution from image sequences - a review. In *MWSCAS*, pages 374–378, 1998.

[5] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong. Soft edge smoothness prior for alpha channel super resolution. In *IEEE CVPR*, 2007.

[6] G. Dedeoglu, T. Kanade, and J. August. High-zoom video hallucination by exploiting spatio-temporal regularities. In *IEEE CVPR*, pages 151–158, 2004.

[7] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004.

[8] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *IJCV*, 40:25–47, 2000.

[9] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[10] K. Jensen and D. Anastassiou. Subpixel edge localization and the interpolation of still images. *IEEE Trans. on Image Processing*, 4:285–295, Mar. 1995.

[11] S. Karunasekera and N. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE Trans. on Image Proc.*, 1995.

[12] J. Kopf, M. Uyttendaele, O. Deussen, and M. Cohen. Capturing and viewing gigapixel images. *ACM Transactions on Graphics*, 26, 2007.

[13] X. Li and M. Orchard. New edge-directed interpolation. *IEEE Trans. on Image Processing*, 10(10):1521–1527, 2001.

[14] F. Liu, J. Wang, S. Zhu, M. Gleicher, and Y. Gong. Visual-quality optimizing super resolution. *Computer Graphis Forum*, to appear.

[15] W. Liu, D. Lin, and X. Tang. Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *IEEE CVPR*, pages 478– 484, 2005.

[16] H. Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40, 2000.

[17] 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.

[18] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39(19):3157–3163, 1999.

[19] R. Schultz and R. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996 – 1011, June 1996.

[20] H. Stark and P. Oskoui. High-resolution image recovery from image-plane arrays, using convex projections. *Journal of the Optical Society of America A*, 6:1715–1726, 1989.

[21] Y.-W. Tai, W.-S. Tong, and C.-K. Tang. Perceptually-inspired and edge-directed color image super-resolution. In *IEEE CVPR*, pages 1948–1955, 2006.

[22] B. Tom and A. Katsaggelos. Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. In *IEEE ICIP*, pages 539–542, 1995.

[23] H. Tong, M. Li, H.-J. Zhang, and C. Zhang. No-reference quality assessment for jpeg2000 compressed images. In *IEEE ICIP*, pages 24–27, 2004.

[24] Z. Wang, G. Wu, H. Sheikh, E. Simoncelli, E.-H. Yang, and A. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680 – 1689, 2006.