High Frequency Component Compensation based Super-resolution Algorithm for Face Video Enhancement

Junwen Wu, Mohan Trivedi, Bhaskar Rao
CVRR Lab, UC San Diego, La Jolla, CA 92093, USA

Abstract

This paper proposes a video-based high-frequency component compensation (HFCC) super-resolution algorithm. The lost high-frequency information is estimated by local MAP criteria, using the registered frames. By compensating the high frequency component iteratively, the high-resolution images are recovered. The algorithm has lower computational cost than the alternatives. Experimental evaluation verified the usefulness of the algorithm.

1 Introduction

Recently there has been considerable interests in high resolution video reconstruction. In general, existing video super-resolution algorithms can be classified into three categories: frequency domain algorithms [1]; spatial domain algorithms from image generative degrading model; interpolation methods. Frequency domain algorithms are limited by the underlying global translation motion assumption. Real world videos usually have multiple rigid motions as well as non-rigid motion. For such cases the performance will deteriorate. Spatial domain approaches are motivated from the generative degrading model of low resolution videos. Super-resolution reconstruction is modeled as an inverse problem of this generative model. The essential ill-condition has inspired many efforts in providing different priors as for solution [2, 3, 4]. Yet the performance is limited by the consistency between the prior and the data. For interpolation methods, registered low-resolution images are mapped onto a unique non-uniform high-resolution grid [5]. Interpolation is used to get the high-resolution image residing on the corresponding uniform grid. However, aliasing is a problem for such approaches.

The human face is different from other popular subjects in computer vision area due to its essential non-planarity and non-rigidity. In this paper, a novel algorithm is presented for human face video super-resolution reconstruction. High-resolution frames are reconstructed by high-frequency component compensation (HFCC). Experiments on substantive face videos verified its usefulness.

The paper is organized as follows. Section 2 presents the proposed super-resolution algorithm. In section 3, experimental results and comparisons are reported. Section 4 concludes the paper. In the following discussion, all the frames are registered.

2 Algorithm design

Fig. 1 shows the flowchart of the algorithm. The procedure is realized iteratively. Let the t-th original high-resolution frames be $\mathbf{I}_{h,t}$ and its k-th estimate be $\mathbf{I}^k_{h,t}$. $\mathbf{I}^k_{h,t}$ is a smoothed version of $\mathbf{I}_{h,t}$: $\mathbf{I}^k_{h,t} = h^k_t * \mathbf{I}_{h,t}$, where h^k_t is the blurring function for the current estimate. The estimation error is $\varepsilon_t = \mathbf{I}_{h,t} - \mathbf{I}^k_{h,t}$ Since true $\mathbf{I}_{h,t}$ is unknown, the estimate $\varepsilon^k_t = \mathbf{I}^k_{h,t} - h^k_t * \mathbf{I}^k_{h,t}$ is used instead. Compensate it back to the estimate of the high resolution image, we can refine it as:

$$\mathbf{I}^{k+1}_{h,t} = \mathbf{I}^k_{h,t} + \varepsilon^k_t. \quad (1)$$

This iteration equation is our basis for reconstruct $\mathbf{I}_{h,t}$. In the proposed algorithm, $h^k_t * \mathbf{I}^k_{h,t}$ is predicted without explicitly computing h^k_t.

Suppose every pixel on the original low-resolution grid will be projected onto a $q \times q$ grid of the high-resolution image domain. Different from [2, 3], we assume the point spread function (PSF) is unknown (which is more general for real data) and PSF is non-uniform. The degrading model of the low resolution images is:

$$\mathbf{I}_{l,t}(\mathbf{x}_i) = \sum_{m=0}^{q-1}\sum_{n=0}^{q-1} f_t(\mathbf{x}_{i,m,n}) \times \mathbf{I}_{h,t}(\mathbf{x}_{i,m,n}), \quad (2)$$

f_t shows the pixel weight in the $q \times q$ grid of the original high-resolution image. We refer to it as local point spread function (local PSF). The blurring function h^k_t is closely related with the estimation of f_t. Later in this section, we will see that $h^k_t * I^k_{h,t}$ is inferred heuristically from f_t.

This degrading model is solved directly in a simplified way. The current estimate $\mathbf{I}^k_{h,t}$ should also satisfy the degrading model from the MSE sense. Therefore:

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

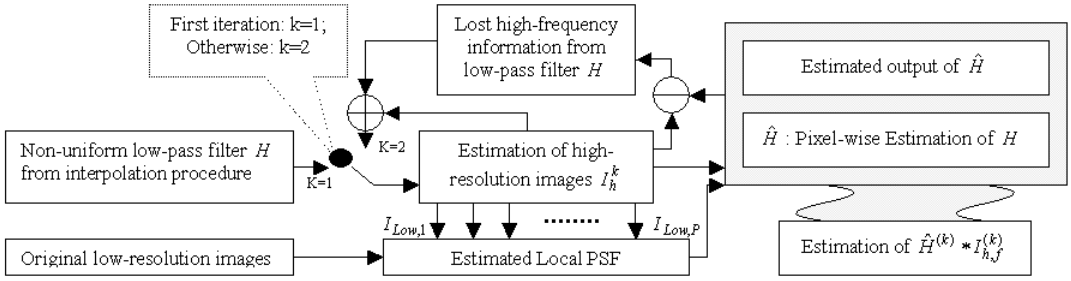


Figure 1. The flowchart of the algorithm. (All frames are registered.)

$$\mathbf{I}_{l,t}(\mathbf{x}_i) \models \sum_{m=0}^{q-1}\sum_{n=0}^{q-1} f_t(\mathbf{x}_{i;m,n}) \times \mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n}). \quad (3)$$

It is reasonable to assume that the local PSF keeps unchanged for successive $2p+1$ frames. Let $q \times q$ matrix $\mathbf{f}_t^k(\mathbf{x}_i) = [f_t^k(\mathbf{x}_{i;m,n})]_{q\times q}$ be the local PSF on the current $q \times q$ grid. k is the iteration index. The optimal function for estimating f_t on the current $q \times q$ grid is:

$$\begin{aligned} & \mathcal{J}(\mathbf{f}_t^k(\mathbf{x}_i)) \\ = & \sum_{t=-p}^{p} (\mathbf{I}_{l,t}(\mathbf{x}_i) - \sum_{m=0}^{q-1}\sum_{n=0}^{q-1} f_t^k(\mathbf{x}_{i;m,n})\mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n}))^2 \\ & +\lambda \nabla \mathbf{f}_t^k(\mathbf{x}_i), \end{aligned} \quad (4)$$

$$s.t. : \|\mathbf{f}_t^k(\mathbf{x}_i)\|_1 = 1;$$

where:

$$\nabla \mathbf{f}_t^k(\mathbf{x}_i) = \|\partial_x \mathbf{f}_t^k\|_2 + \|\partial_y \mathbf{f}_t^k\|_2 + \|\partial_{xy} \mathbf{f}_t^k\|_2 + \|\partial_{yx} \mathbf{f}_t^k\|_2.$$

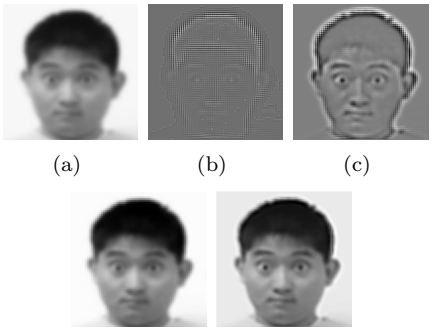
The first term of $\mathcal{J}(\mathbf{f}_t^k(\mathbf{x}_i))$ boosts $\mathbf{f}_t(\mathbf{x}_i)$ as an impulse function with non-zero at the pixel most similar to the known low-resolution pixel. The second term is a smoothing term which keeps the local PSF as uniform as possible. A simplified solution is provided for equation 4, which is a direct trade-off between these two terms. We choose a uniform function as the initial values for $\mathbf{f}_t^k(\mathbf{x}_i)$, and apply one-step steepest descendent update as follows:

$$\widetilde{f_t^k}(\mathbf{x}_{i;r,l}) = \frac{\sum_{t=-p}^{p}\{I_{l,t}(\mathbf{x}_i) - S_{t;r,l}\}}{\sum_{t=-p}^{p} I_{h,t}^k(\mathbf{x}_{i;r,l})}, \quad (5)$$

$$S_{t;r,l} = \frac{1}{q^2} \sum_{m=0;m\neq r}^{q-1} \sum_{n=0;n\neq l}^{q-1} I_{h,t}^k(\mathbf{x}_{i;m,n});$$

The optimal local PSF for current $q \times q$ grid at kth iteration is the obtained from the normalization:

$$(f_t^k(\mathbf{x}_{i;r,l}))^\star = \frac{\widetilde{f_t^k}(\mathbf{x}_{i;r,l})}{\|[\widetilde{f_t^k}(\mathbf{x}_{i;m,n})]_{q\times q}\|_1};$$



(a) (b) (c) (d) (e)

Figure 2. (a): initial input of the current frame. (b):estimated local point spread function after 1^{st} iteration. (c): estimated high-frequency components after 1^{st} iteration. (d): reconstructed high-resolution image after 1^{st} iteration. It is also the input for the 2^{nd} iteration. (e): final high-resolution reconstruction after 3 iterations.

$$0 \leq r \leq q-1, 0 \leq l \leq q-1.$$

Now we relate $(f_t^k)^\star$ with the estimation of $h_t^k * \mathbf{I}_{h,t}^k$. Equation 3 can also be written as:

$$\mathbf{I}_{l,t}(\mathbf{x}_i) \models \sum_{m=1}^{q}\sum_{n=1}^{q} f_t(\mathbf{x}_{i;m,n}) \times (h_t^k * \mathbf{I}_{h,t})(\mathbf{x}_{i;m,n}) \quad (6)$$

It's clear that h_t^k and f_t^k are reciprocally related. For simplification, assume h_t^k has a limited support of 3×3. Then at every pixel, the smoothed output of $h_t^k * \mathbf{I}_{h,t}^k$ is determined jointly by the intensity and the local PSF of its eight neighborhood. Model the distribution of $(h_t^k * \mathbf{I}_{h,t}^k)(\mathbf{x}_{i;m,n})$ by the following Gaussian mixture:

$$\begin{aligned} & Pr((h_t^k * \mathbf{I}_{h,t}^k)(\mathbf{x}_{i;m,n}) | \{\mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n,j})\}_{j=1,\cdots,8}) \\ \sim & \sum_{j=1}^{8} w_{i,m,n,j} \mathcal{N}(\mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n,j}); d_j^2); \end{aligned} \quad (7)$$

where $\mathbf{x}_{i;m,n,j} (j = 1, \cdots, 8)$ are pixel $\mathbf{x}_{i;m,n}$'s eight neighboring pixels; d_j (=1 or $\sqrt{2}$) is the Euclidean distance between pixel $\mathbf{x}_{i;m,n}$ and its neighbor $\mathbf{x}_{i;m,n,j}$. This model actually describes a low-pass procedure characterized by the mixing factor $w_{i;m,n,j}$. The filtered pixel

value $(h_t^k * \mathbf{I}_{h,t}^k)(\mathbf{x}_{i;m,n})$ will be as similar as its neighborhood with confidence $w_{i;m,n,j}$. This confidence is critical since it incorporates the prior knowledge about h_t^k: the reciprocal relationship between h_t^k and f_t^k. We use the following heuristic function to model it:

$$w_{i;m,n,j} = \exp\{-b_j^2((f_t(\mathbf{x}_{i;m,n}))^\star)^2\}, \quad (8)$$

where b_j is $\mathbf{x}_{i;m,n,j}$'s bias from its mean value over the successive p frames: $b_j = \mathbf{x}_{i;m,n,j} - \bar{\mathbf{x}}_{i;m,n,j}$. b_j assures that less weight be given to pixels with large deviations, which are most likely outliers.

Therefore, $(h_t^k * \mathbf{I}_{h,t}^k)(\mathbf{x}_{i;m,n})$ is the solution of:

$$\arg\max_{\mathbf{y}^\star} F(\mathbf{y}^\star),$$

where: $F(\mathbf{y}^\star) = \sum_{j=1}^{8} w_{i;m,n,j} \exp\{-\frac{(\mathbf{y}^\star - \mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n,j}))^2}{2d_j^2}\}$. This is actually a local MAP estimation. Steepest descendent algorithm is used to solve it. Denote $(h_t^k * \mathbf{I}_{h,t}^k)$ as Z. Steepest descendant algorithm gives:

$$Z^{(r+1)}(\mathbf{x}_{i;m,n}) = Z^{(r)}(\mathbf{x}_{i;m,n}) + \mu df \quad (9)$$

$$\begin{aligned} df &= -\frac{\partial}{\partial \mathbf{x}^\star} F(\mathbf{x}^\star)|_{Z^{(r)}(\mathbf{x}_{i;m,n})} \\ &= \sum_{i=1}^{8} \frac{w(\mathbf{x}_{i;m,n,j})(\mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n,j}) - Z^{(r)}(\mathbf{x}_{i;m,n}))}{d_j^2} \\ &\quad \exp\{-\frac{(Z^{(r)}(\mathbf{x}_{i;m,n}) - \mathbf{I}_{h,t}^k(\mathbf{x}_{i;m,n,j}))^2}{2d_j^2}\} \quad (10) \end{aligned}$$

The entire procedure is repeated. In this way, the high-resolution videos are recovered. Fig. 2 give an example of one iteration. Initial input of the algorithm are bilinear interpolation of the $(2p+1)$ successive frames (here $p = 2, q = 2$). In our experiment, the iteration times are all set to 3. When $k = 3$, the dynamic range of the obtained high-frequency component has been small enough.

3 Experimental evaluation

3.1 Videos under different settings

In this section we show experimental results from sequences with different content and sensors.

1. Videos with changing facial expressions. Subtle changes in facial expressions that could be lost in low resolution or blurred sequences are enhanced. Fig. 3(a) and 3(b) shows results of HFCC algorithm compared to those generated by bilinear interpolation. It is apparent that the bilinear interpolation results in Fig. 3(a) are severely blurred, whereas HFCC algorithm generates significantly clearer facial features, shown in Fig. 3(b).

2. Videos with large head motions. In Fig. 3(c) we compare our results with the super-resolution optical flow algorithm from [5] using the same video sequence from Dr. Baker. Our experiments show better performance in certain frames that are problematic for [5]. Specifically, super-resolution optical flow is vulnerable to artifacts, while HFCC algorithm successfully augments the original image with more perceptually appealing results.

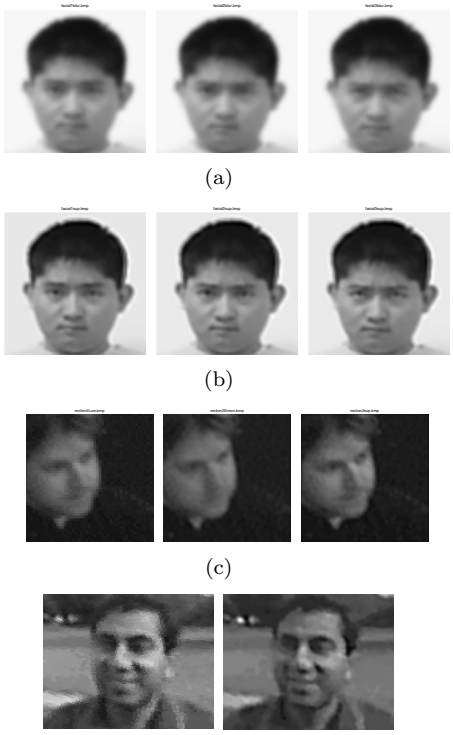


Figure 3. Examples of the experimental results. 3(a): bilinear interpolated high-resolution frames. 3(b): corresponding super-resolved results from the HFCC algorithm. In 3(c): example from sequence with large head motion. Leftmost: the original image. Middle: result from [5]. It is a problematic frame for [5]. Rightmost: corresponding result from the HFCC algorithm. The original video in 3(c) is the courtesy of Dr. Baker [5]. 3(d): example from omni-directional face video. Left: the input; right: the super-resolved result.

3. Videos from an omni-directional camera. Omni-directional video cameras are widely used for their 360 degree field of view [6, 7]. However, images from these cameras are typically low resolution and suffer from non-uniform distortion across the image. Our HFCC algorithm can be used to enhance the video quality, as shown in Fig. 3(d).

3.2 Quantitative comparison

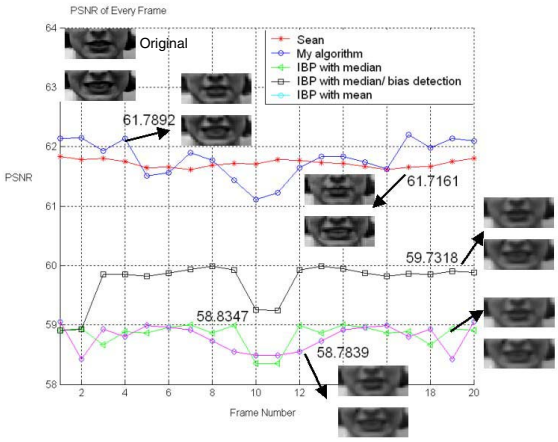


Figure 4. PSNR of different algorithms.

We compare the results of our HFCC algorithm with [2] and [3] using high resolution face videos from [8]. The videos contain substantive facial expressions from various subjects. We use sub-sampled video frames as the input. The original video sequences are used as the ground-truth for comparison. Examples of the perceptual results are shown in Fig. 4. Due to legal issues, only the lip patches are shown. Qualitatively, more details are resolved by our algorithm and Borman's algorithm than Zomet's IBP algorithms. However, Borman's algorithm produces blobby images that are perceptually unappealing. The examples in [4] also exhibit the same problems, possibly due to the Huber function prior used leading to excessive constraints on the high frequency components. The PSNR is computed for each algorithm on a frame-by-frame basis. Fig. 4 shows the PSNR curve for the first video sequence in the database. The mean PSNR over all frames is computed and displayed as well. It indicates that our HFCC algorithm exhibits the least distortion. Overall, these quantitative comparative results show the effectiveness of the proposed HFCC algorithm. Also, this algorithm has a lower computational cost than the others.

4 Conclusion

Due to the non-rigidity and non-planarity characteristics of face subject, a lot of existing high-resolution algorithms are not applicable for face video high-resolution reconstruction. This paper proposed a video based super-resolution algorithm based on high frequency component compensation. The lost high-frequency information is estimated by local MAP criteria, using the registered frames. By compensating the lost high-frequency information, the high-resolution frames are recovered. The computational cost for the algorithm is much lower than the alternatives. Also, although the start point for the algorithm is for face subjects, the algorithm is not limited to faces. One drawback of the high frequency compensation algorithm is that the overall brightness of the image may be altered. We are currently exploring ways to resolve this issue. Also, we are working on combining this promising algorithm with omni-directional face recognition [9] to get a better recognition rate.

ACKNOWLEDGEMENTS

We are thankful for the grant awarded by the Technical Support Working Group (TSWG) of the US Department of Defense which provided the primary sponsorship of the reported research. We also thank our colleagues from the UCSD Computer Vision and Research Laboratory for their contributions and support.

References

[1] S. Borman and R. Stevenson. Super-resolution from image sequences - a review. In *Proceedings of Midwest Symposium on Circuits and Systems.*, 1998.

[2] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super resolution. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, December. 2001.

[3] S. Borman and R. Stevenson. Simultaneous multi-frame map super-resolution video enhancement using spatio-temporal priors. In *Proceedings of IEEE International Conference on Image Processing.*, October. 1999.

[4] D. P. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2001.

[5] S. Baker and T. Kanade. Super-resolution optical flow . Technical report, Carnegie Mellon University., 1999.

[6] K. Huang and M.M.Trivedi. Networked omnivision arrays for intelligent environment. In *Proceedings of the Applications and Science of Soft Computing IV.*, August. 2001.

[7] K. Huang and M.M.Trivedi. Video arrays for real-time tracking of person, head, and face in an intelligent room. *Machine Vision and Applications.*, (2):103–111, June. 2003.

[8] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *The 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00).*, March 2000.

[9] K. Huang and M.M.Trivedi. Streaming face recognition using multicamera video arrays. In *Proceedings of the 16th International Conference on Pattern Recognition.*, pages 213–216, August. 2002.