Fast super-resolution reconstructions of mobile video using warped transforms and adaptive thresholding

Sandeep Kanumuri, Onur G. Guleryuz and M. Reha Civanlar

DoCoMo Communications Laboratories USA, Inc.
3240 Hillview Ave, Palo Alto CA 94304, USA
{skanumuri,guleryuz,rcivanlar}@docomolabs-usa.com

ABSTRACT

Multimedia services for mobile phones are becoming increasingly popular thanks to capabilities brought about by location awareness, customized programming, interactivity, and portability. With mounting attraction to these services there is desire to seamlessly expand the mobile multimedia experience to stationary environments where high-resolution displays can offer significantly better viewing conditions. In this paper, we propose a fast, high quality super-resolution algorithm that enables high resolution display of low-resolution video. The proposed algorithm, SWAT, accomplishes sparse reconstructions using directionally warped transforms and spatially adaptive thresholding. Comparisons are made with some existing techniques in terms of PSNR and visual quality. Simulation examples show that SWAT significantly outperforms these techniques while staying within a limited computational complexity envelope.

Keywords: Warped transforms, sparse signal models, sparse reconstructions, adaptive thresholding, iterated denoising, shrinkage

1. INTRODUCTION

Mobile phones and other wireless devices are rapidly becoming an important destination for multimedia services such as wireless video broadcasts, video streaming, and video conferencing just to name a few. Some of these services have already been firmly established in countries such as Japan and South Korea, and they are on their way to getting deployed in the US and the rest of the world. Improvements in the radio access network and emergence of highly capable wireless terminals have led to a much improved user-experience with better quality, lower startup delay, and an increased variety of programming.[3, 16, 17] Capabilities such as location awareness, customized programming, interactivity, and portability are further popularizing these services and there is desire to seamlessly expand the mobile multimedia experience to stationary environments where high-resolution displays can offer significantly better viewing conditions.

In this paper, we propose a fast, high quality super-resolution algorithm that enables high resolution display of low-resolution video received by a mobile device. While our formulation can easily be extended to use multiple frames, for complexity and memory bandwidth reasons we restrict ourselves to operate on single frames. Hence in some sense we solve an interpolation problem*. Our work is a continuation of early work[3] but significantly improves on it by providing better quality at a much reduced computational complexity envelope. We accomplish these improvements by using fast directionally warped transforms, which enable even better performance around image singularities, and spatially adaptive nonlinearities, which enable high quality results without the significant number of iterations required in.[3] The resulting technique based on sparse warped transforms and adaptive thresholding (SWAT) is compact and fast, allowing one to obtain the improvements of recent sparse reconstruction techniques[8–10] with a low computational burden.

Figure 1 illustrates our system model. Original video frames are spatially low pass filtered using a low pass filter, $H_{LL}(\omega_1, \omega_2)$, and downsampled. The low-resolution frames are compressed and the decompressed video frames are processed by a super-resolution module to form a high-resolution estimate of the original video. For generalizations of this model we refer the reader to.[1, 4, 7, 15] For a discussion of expected quality improvements

*Single frame super-resolution is often times considered to be an interpolation problem. However, since our technique effectively predicts missing high frequency information, it is in effect also solving an extrapolation problem.

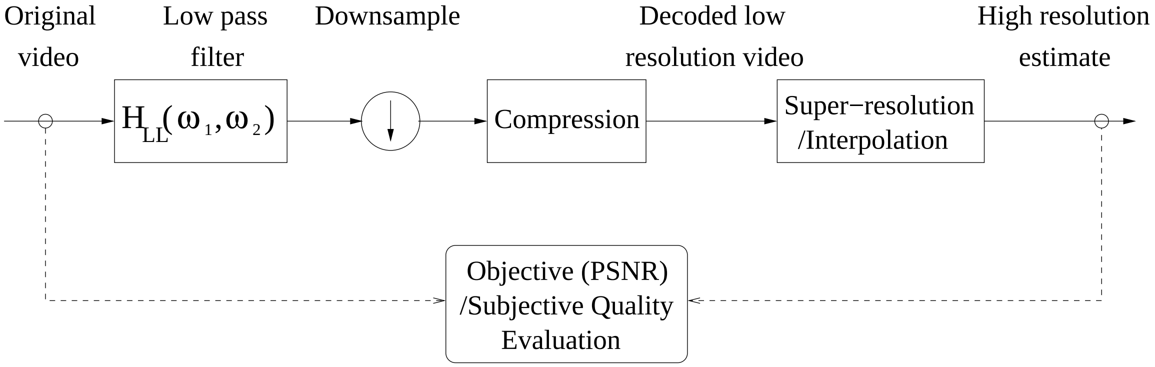

Original video
Low pass filter
Downsample
Decoded low resolution video
High resolution estimate
H_LL(ω_1,ω_2)
Compression
Super−resolution /Interpolation
Objective (PSNR) /Subjective Quality Evaluation

Figure 1. Super-resolution system model: Original video frames are spatially low pass filtered and downsampled. The low-resolution frames are compressed and the decompressed video frames are processed by a super-resolution module to form a high-resolution estimate of the original video.

and limitations of single-frame resolution algorithms under this model,[3] can be consulted. Analysis of limitations in general setups can be found in.[2,14] In this paper, we evaluate the quality of high-resolution estimates produced by different super-resolution modules via visual comparisons and in terms of the PSNR metric.

The paper is organized as follows: Section 2 provides a basic formulation of the core idea and the assumptions made. Section 3 describes SWAT in detail and section 4 compares the performance of SWAT to existing algorithms. Section 5 provides the conclusion.

2. BASIC FORMULATION

Similar to our early work[3] on iterated denoising (ID), the formulation of this paper uses a set of sparsity enforcing transforms that allow the recovery of the high resolution frame. Let u denote the unknown, original high resolution frame. Suppose u is lexicographically ordered into a $(N \times 1)$ vector and assume that we are given a linear orthonormal transform $\mathbf{G}$ $(N \times N)$. Let g_i^T, the i^{th} row of $\mathbf{G}$, denote the i^{th} transform basis function and let $c_i = g_i^T u$ be the corresponding transform coefficient. We have

$$u = \sum_{i=1}^{N} c_i g_i. \tag{1}$$

We assume that $\mathbf{G}$ generates a sparse decomposition of u so that most of the transform coefficients of u are zero or close to zero. Define the significant set $\mathcal{S}(u, K)$ as the indices of the K largest in magnitude coefficients of u with $K << N$. Then our assumption can be stated as

$$u = \sum_{i \in \mathcal{S}(u,K)} c_i g_i + \sum_{i \notin \mathcal{S}(u,K)} c_i g_i \approx \sum_{i \in \mathcal{S}(u,K)} c_i g_i, \tag{2}$$

that is nonlinear approximation of u using transform $\mathbf{G}$ with the largest K coefficients closely approximates u. Since the transform is orthonormal, (2) amounts to assuming that $|c_i| \approx 0,\ i \notin \mathcal{S}(u, K)$.

Our formulation starts with an approximation y to the original frame u. We take this approximation as the simple inverse shown in Figure 3. Based on this approximation we obtain a significant set $\mathcal{S}(y, K_0)$ which allows us to establish sparsity constraints of the form $|c_i| \approx 0,\ i \notin \mathcal{S}(y, K_0)$. Enforcing these constraints followed by enforcing the available information, i.e., that the high resolution frame when low pass filtered and downsampled should result in the low resolution frame, allows us to obtain the next approximation y' using two consecutive projections. We then repeat the process one more time by obtaining a new significant set, establishing new sparsity constraints, and enforcing these sparsity constraints and the available information in turn to arrive at our final estimate. The algorithm we propose here deviates from the ID algorithm used in[3,9,10] in terms of significant set construction, which we do with an adaptive thresholding strategy, and the utilized transforms,

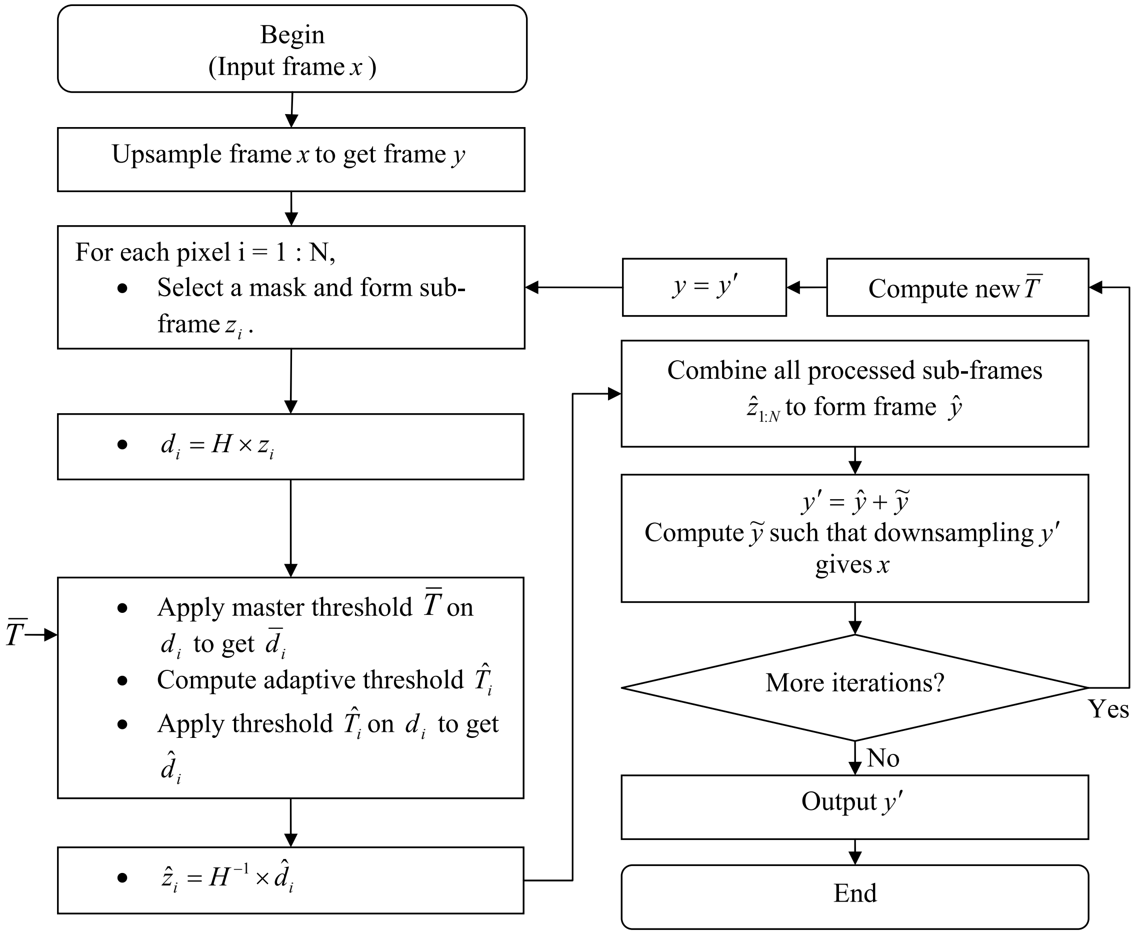

Begin
(Input frame x)
Upsample frame x to get frame y
For each pixel i = 1 : N,
• Select a mask and form sub-frame z_i.
• $d_i = H \times z_i$
$\overline{T}$
• Apply master threshold $\overline{T}$ on d_i to get $\overline{d}_i$
• Compute adaptive threshold $\hat{T}_i$
• Apply threshold $\hat{T}_i$ on d_i to get $\hat{d}_i$
• $\hat{z}_i = H^{-1} \times \hat{d}_i$
$y = y'$
Compute new $\overline{T}$
Combine all processed sub-frames $\hat{z}_{1:N}$ to form frame $\hat{y}$
$y' = \hat{y} + \tilde{y}$
Compute $\tilde{y}$ such that downsampling y' gives x
More iterations?
Yes
No
Output y'
End

Figure 2. Flowchart of SWAT. Only two iterations are carried out.

which we have designed to be an expansive set of directionally warped DCT transforms. While ID uses many iterations, as we illustrate, our SWAT technique allows us to obtain similar caliber performance in just two rounds. SWAT also takes advantage of directional regularity with *separable* DCT kernels via directional warping. The details of SWAT are provided in Section 3.

In related work that obtains sparse reconstructions,[13] uses multiscale geometric representations and[6] employs the EM algorithm. Since the problem can be posed as a general denoising/recovery problem, other sparse recovery techniques based on l_1 and l_p regularization can also be used.[5] In general, we expect SWAT to be significantly faster than these techniques as they require many iterations in order to achieve similar quality levels.

3. SWAT DESCRIPTION

In this section, we describe the details of the proposed super-resolution algorithm. SWAT takes a video frame at low-resolution (LR) as input and outputs a video frame at high-resolution (HR). A flowchart of SWAT is shown in Figure 2. SWAT operates separately on each frame of video to avoid the complexity of precise motion estimation in multi-frame super-resolution. For convenience purposes, we represent all video frames as vectors with the pixel values listed in raster-scan order.

Let x denote the input LR video frame. The input frame is upsampled and filtered using an inverse filter, $\tilde{H}_{LL}(\omega_1, \omega_2)$, to form frame y as shown in Figure 3. We refer to the frame y as 'Simple Inverse' and use it as an initial estimate for the rest of the algorithm which is iterative and repeated twice. The inverse filter, $\tilde{H}_{LL}(\omega_1, \omega_2)$,

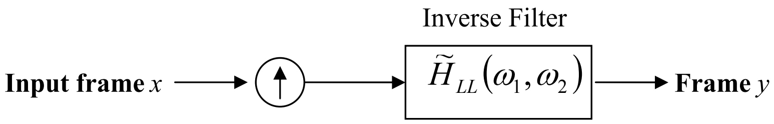


Figure 3. Formation of 'Simple Inverse' (y).

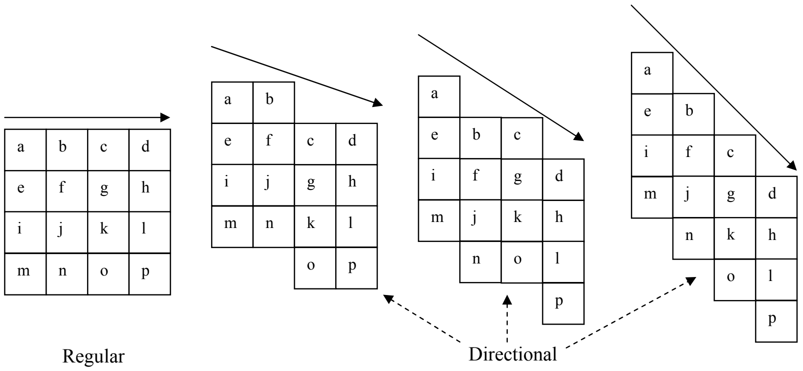


Figure 4. Example regular and directional masks on two dimensional image plane.

is chosen based on the low pass filter $H_{LL}(\omega_1, \omega_2)$. If the low pass filter is part of a wavelet filter bank, then the inverse filter is the corresponding filter from the dual filter bank. Otherwise, the inverse filter can be obtained as a Wiener filter that accomplishes the minimum mean squared error (MSE) for a given constraint on the number of filter taps.

In the first step of the two round iterative process, for each pixel i in frame y, an $M^2 \times 1$ vector (z_i) of pixel values called a sub-frame is formed using a mask that defines the pixel positions. The mask is selected from a library of masks where each mask corresponds to a set of M^2 pixels warped in a particular direction. The sub-frame z_i is transformed using a separable 2D-DCT matrix H to get transform coefficients d_i. A smooth region with a horizontal or vertical edge can be compactly represented using the separable 2D-DCT. However, this property is violated in regions with slant edges. The idea is to select a mask that can warp the slant edge into a horizontal or vertical edge. Figure 4 shows example masks covering four different directions (shown as solid arrows) when $M = 4$. The regular mask corresponds to the horizontal direction while the directional masks correspond to non-trivial directions that are neither horizontal or vertical. The pixel positions are labeled from 'a' to 'p' and the sub-frame z_i is formed from pixel values corresponding to these positions where position 'a' overlaps pixel i. Since the support of the transform is warped and adapts to the local direction, we call this an adaptive warped transform. One can also deal with slant edges by always choosing the regular mask and using direction-adaptive transforms such as non-separable DCTs, 2-D Gabor wavelets, curvelets and contourlets. However, these transforms are non-separable and incur significantly higher computational complexity when compared to 4×4 separable block DCTs used here.

Adaptive Mask Selection: For the purpose of mask selection, the frame y is divided into groups of pixels (4×4 blocks in our case) and a mask is chosen for the entire group based on majority choice by the pixels in the group. Each pixel votes for the mask that offers the least signal variance along the mask's direction. The idea is that signal variance along the direction of an edge will be small compared to other directions.

Adaptive Thresholding: The transform coefficients d_i are initially hard-thresholded using a master threshold $\bar{T}$ to get $\bar{d}_i$. The thresholding operation sets coefficients with magnitude less than the threshold to zero while the other coefficients are left unaltered. In smooth regions where the sparse model assumption is satisfied, the energy lost due to the thresholding operation is quite small. However, in regions with a lot of detail, the sparsity property is violated and the energy lost due to the thresholding operation is significant. So we compute a spatially adaptive threshold $\hat{T}_i$ given by $\hat{T}_i = \bar{T} \times f(\|d_i - \bar{d}_i\|_2^2)$, where $f(x)$ is a monotonic decreasing staircase function with

$f(0) = 1$ and $f(\infty) = 0$. The transform coefficients d_i are thresholded again using the adaptive threshold $\hat{T}_i$ to get $\hat{d}_i$. The idea behind adaptive thresholding is to preserve the detail in clutter and edge regions by reducing the threshold.

Subframe Combination: The thresholded coefficients $\hat{d}_i$ are inverse transformed to get processed sub-frames $\hat{z}_i$. The processed sub-frames $\hat{z}_{1:N}$ (corresponding to all pixels) are combined to form the frame $\hat{y}$. Since each pixel is involved in a multitude of sub-frames, each pixel has multiple estimates for its value. All these estimates are combined in a weighted fashion to get its value in frame $\hat{y}$, where the weight contribution from a sub-frame is proportional to a sparsity metric for the sub-frame and is obtained similarly to.[11]

Data Consistency: The frame $\hat{y}$ is processed by a data consistency step that outputs y', which is the super-resolution estimate for the current iteration. This step ensures that the downsampling of y' using the modeled low pass filter results in the input frame x. We refine our estimate y' with one more iteration by decreasing the master threshold $\bar{T}$ and copying the current estimate into frame y.

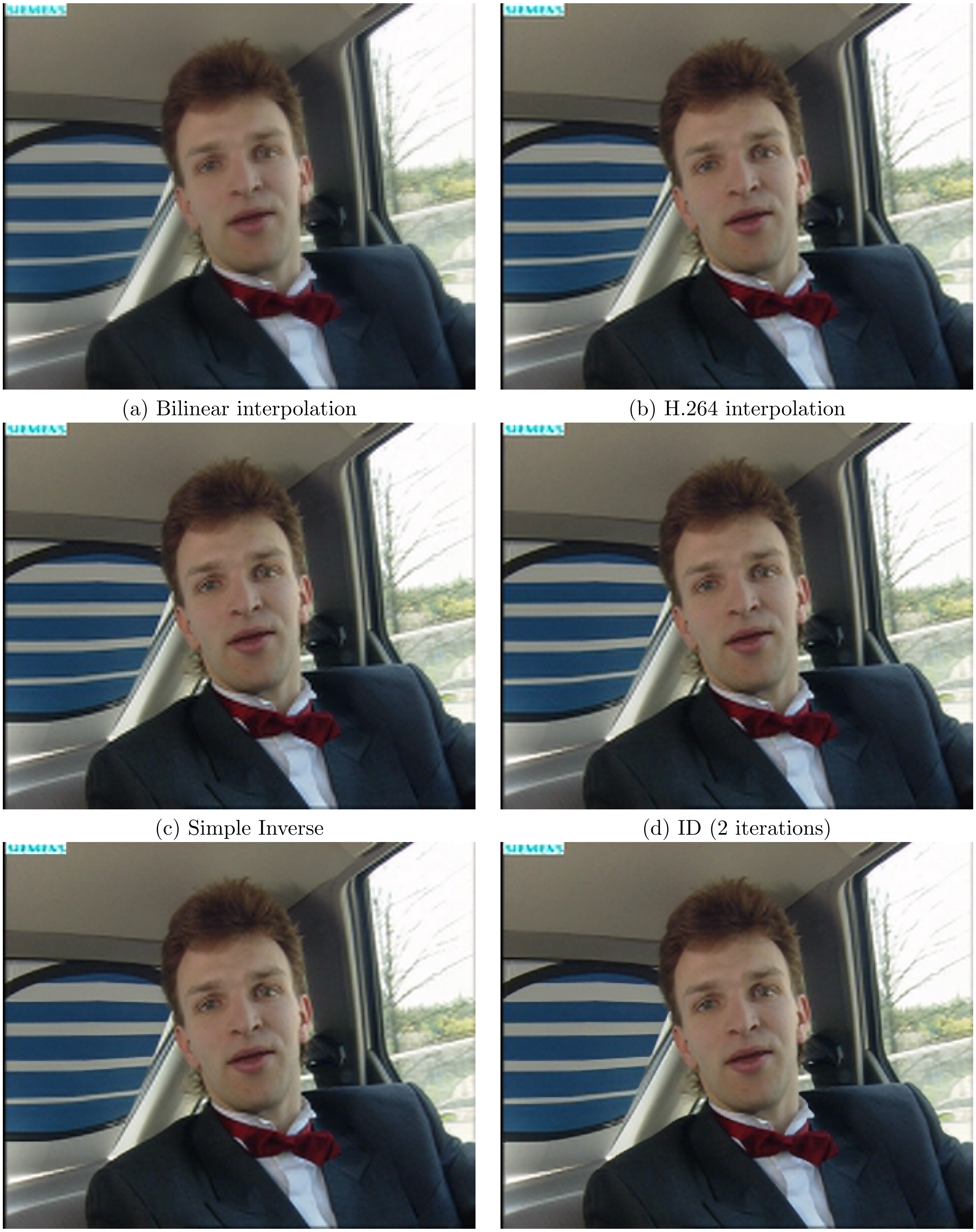
4. RESULTS

In this paper, we use the 9×9 low pass filter from the Daubechies 7/9 wavelet filter bank as the filter used in obtaining the low-resolution (QCIF) video frames. We compare SWAT to 'Simple Inverse' (SI) and to existing interpolation techniques such as Bilinear interpolation, H.264 interpolation (interpolation via the filter used in H.264/AVC codec[12]) and to ID (using 4×4 DCTs with 2 and 10 iterations). The high-resolution estimates generated by different techniques are compared in terms of visual quality and PSNR (with respect to the original video). Figures 5-7 show the high-resolution (CIF) estimates formed by different super-resolution algorithms on the first frame of the sequences carphone, akiyo and foreman respectively. When compared to the other algorithms, we can see that the estimate from SWAT is much sharper and has much reduced ringing. It is worth noting that SWAT does a better job than ID (10 iterations) on slant edges while the computational complexity of SWAT is similar to ID (2 iterations).

The PSNR values obtained by different techniques are plotted as a function of frame number in Figure 8 for the sequences carphone, akiyo and foreman. SWAT outperforms bilinear interpolation by up to 3dB, H.264 interpolation by up to 1.7dB and ID (2 iterations) by up to 0.7dB. As illustrated, SWAT has a similar PSNR performance compared to ID (10 iterations) while its computational complexity is similar to that of ID (2 iterations). A non-optimized implementation of SWAT can accomplish the QCIF to CIF conversion well under a second per frame.

We also compare the performance of SWAT to bilinear and H.264 interpolation methods on low-resolution videos that underwent compression. The H.264/AVC reference codec (JM 12.0) is used to compress low-resolution (QCIF) videos at a constant QP. The first frame is coded as an I picture while the other frames are coded as P pictures. The performance comparison is done at two QP values of 20 and 25. Figure 9 shows high-resolution estimates formed by different super-resolution algorithms from carphone sequence at the two different QPs. From the figure, we can see that SWAT provides sharper pictures with reduced ringing when compared to the other techniques. The PSNR plots shown in Figure 10 for sequences akiyo, carphone and foreman also show that SWAT significantly outperforms bilinear and H.264 interpolation for these examples.

Finally, to demonstrate the effect of warped transforms, we compare SWAT to a non-directional version of SWAT that is forced to always use the regular mask. Since none of the directional masks are used in the non-directional SWAT, the support of the transform is not warped. The improvement of SWAT over non-directional SWAT is mainly localized to regions with slant edges where warped transforms are most effective. To quantify this, we compute a local PSNR value for each pixel in the high-resolution estimate over a 3×3 block centered at that pixel. The pixels that show at least 0.5dB and 1dB local PSNR improvement with SWAT when compared to non-directional SWAT are marked as red in Figure 11. From the figure, we can see that there is significant improvement along slant edges such as the edge of the rear windshield in carphone, the shirt collar of the woman in akiyo and the grooves of building in foreman.



(a) Bilinear interpolation

(b) H.264 interpolation

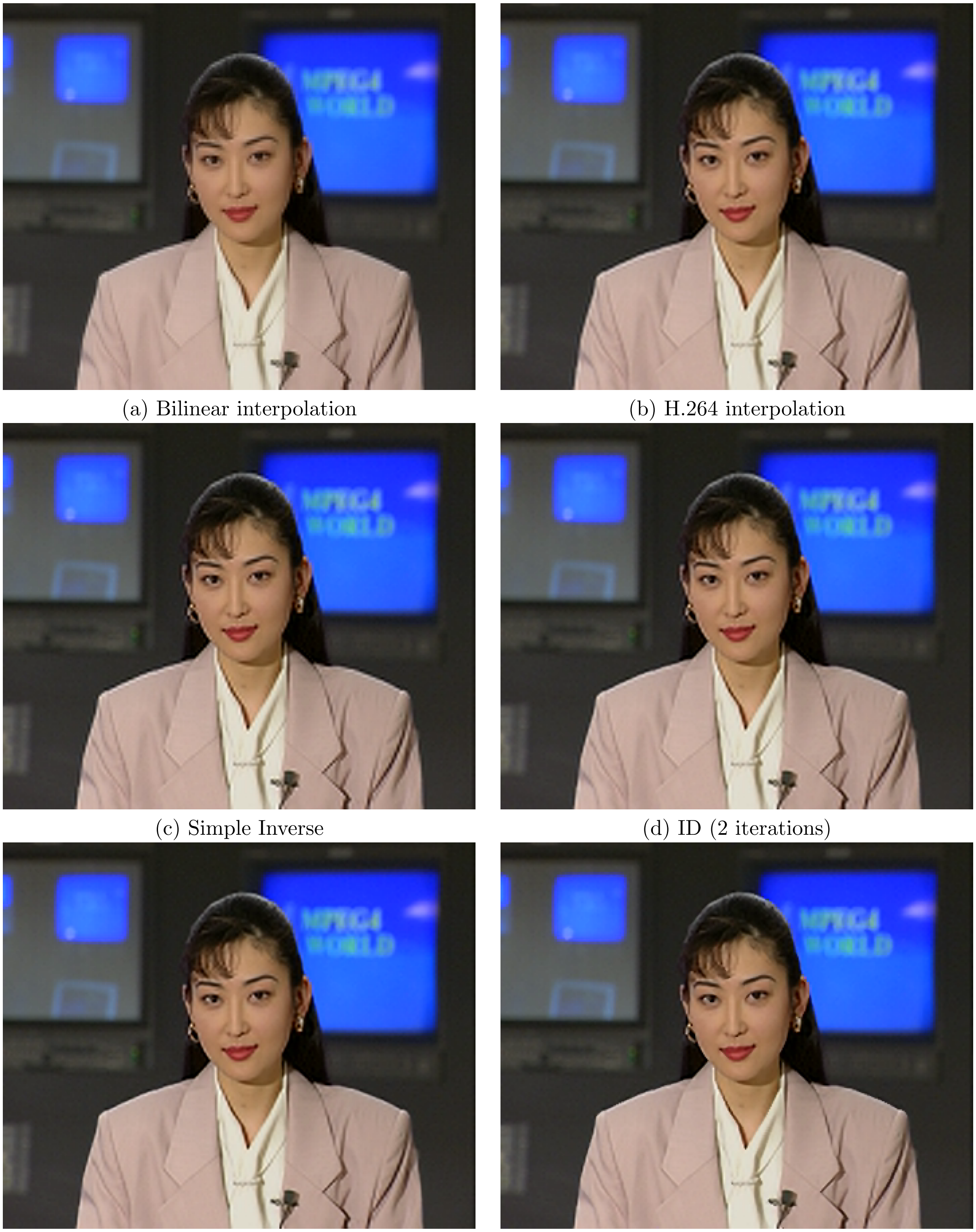
(c) Simple Inverse

(d) ID (2 iterations)

(e) ID (10 iterations)

(f) SWAT

Figure 5. Visual comparison. Carphone sequence, frame 1, uncompressed input. The SWAT algorithm reduces the ringing artifacts much better than all other algorithms and results in a sharper image.



(a) Bilinear interpolation

(b) H.264 interpolation

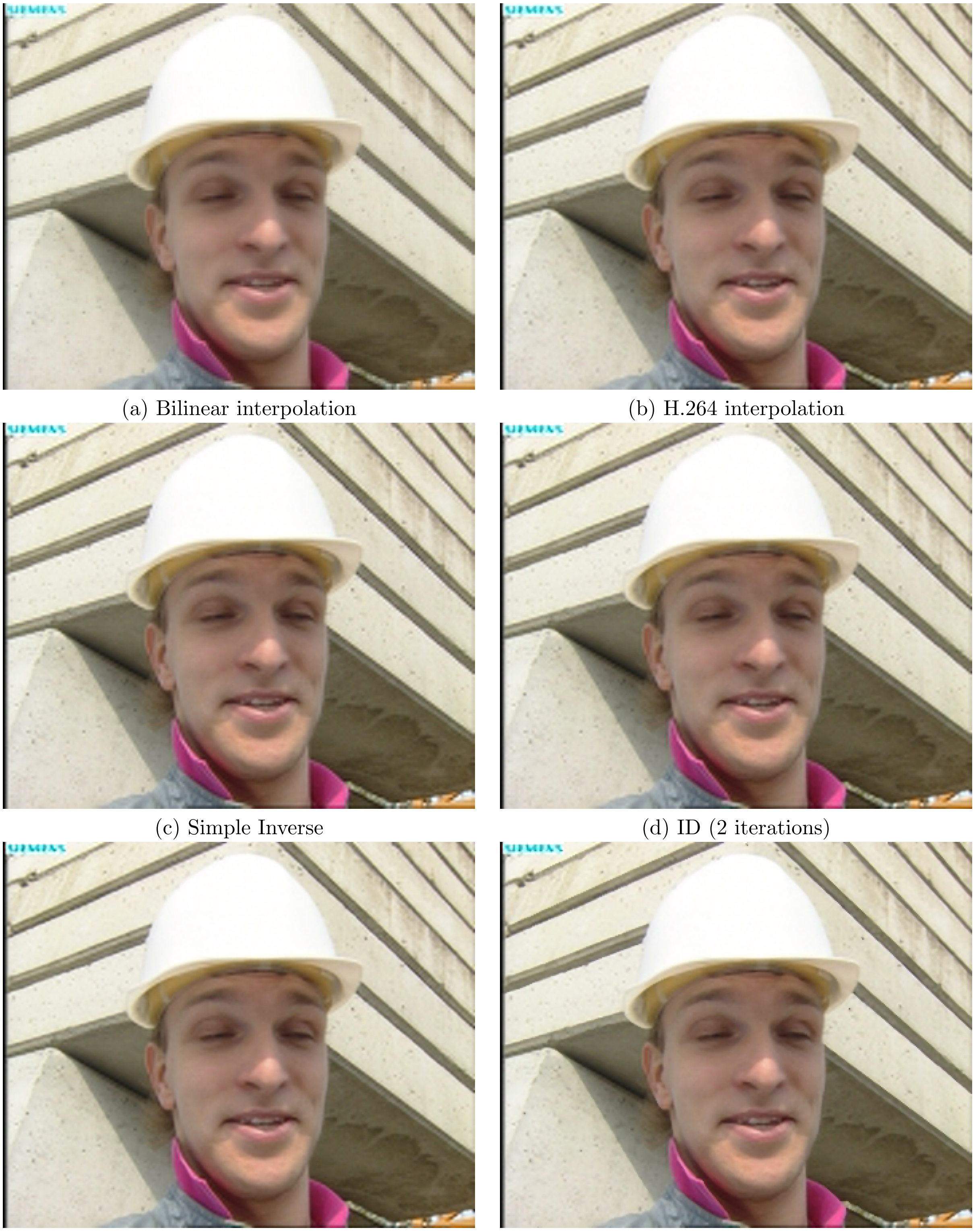
(c) Simple Inverse

(d) ID (2 iterations)

(e) ID (10 iterations)

(f) SWAT

Figure 6. Visual comparison. Akiyo sequence, frame 1, uncompressed input.



(a) Bilinear interpolation

(b) H.264 interpolation

(c) Simple Inverse

(d) ID (2 iterations)

(e) ID (10 iterations)

(f) SWAT

Figure 7. Visual comparison. Foreman sequence, frame 1, uncompressed input.

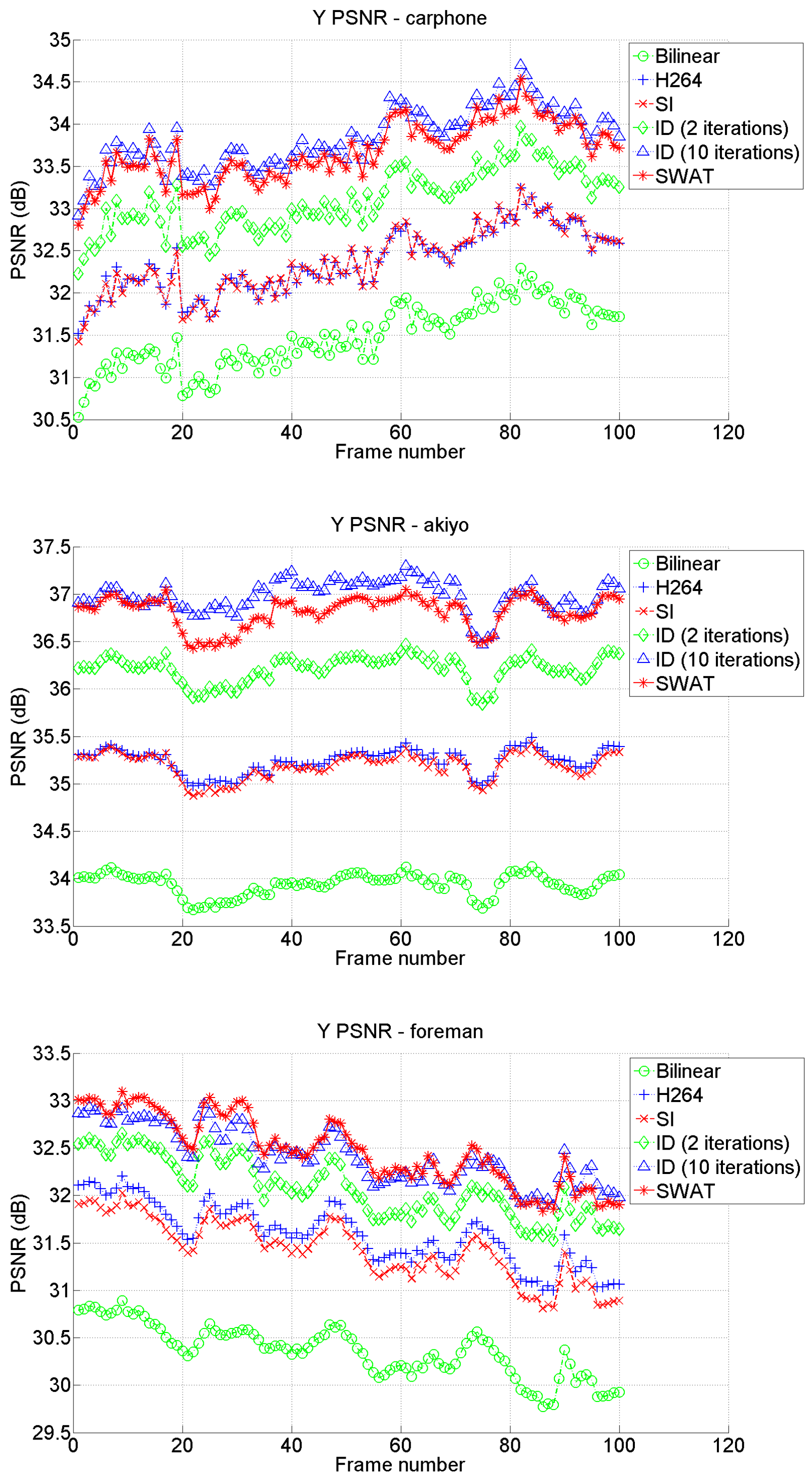
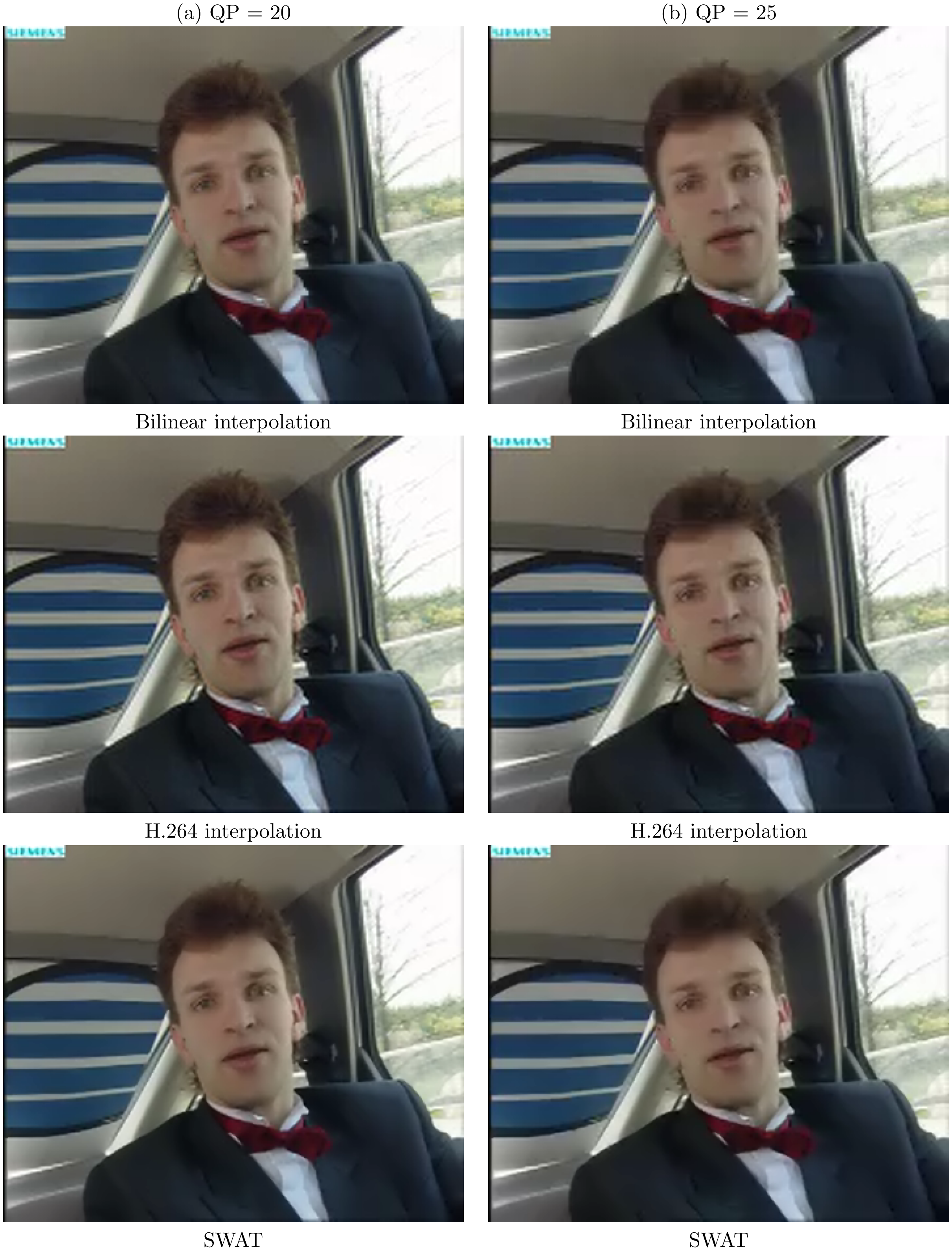


Figure 8. PSNR comparison of different super-resolution algorithms on the first 100 frames of uncompressed sequences Carphone, Akiyo and Foreman. SWAT outperforms bilinear interpolation by up to 3dB, H.264 interpolation by up to 1.7dB and ID (2 iterations) by up to 0.7dB. SWAT has a similar PSNR performance compared to ID (10 iterations) while its computational complexity is similar to that of ID (2 iterations). Simple Inverse (SI) performs very similar to H.264 interpolation.



(a) QP = 20

(b) QP = 25

Bilinear interpolation

Bilinear interpolation

H.264 interpolation

H.264 interpolation

SWAT

SWAT

Figure 9. Visual comparison. Carphone sequence, frame 1, compressed input. Compression is done using H.264/AVC reference codec (JM 12.0) with IPP.. encoding at (a) QP=20 and (b) QP=25. The SWAT algorithm not only reduces ringing significantly but also cleans some of the compression artifacts.

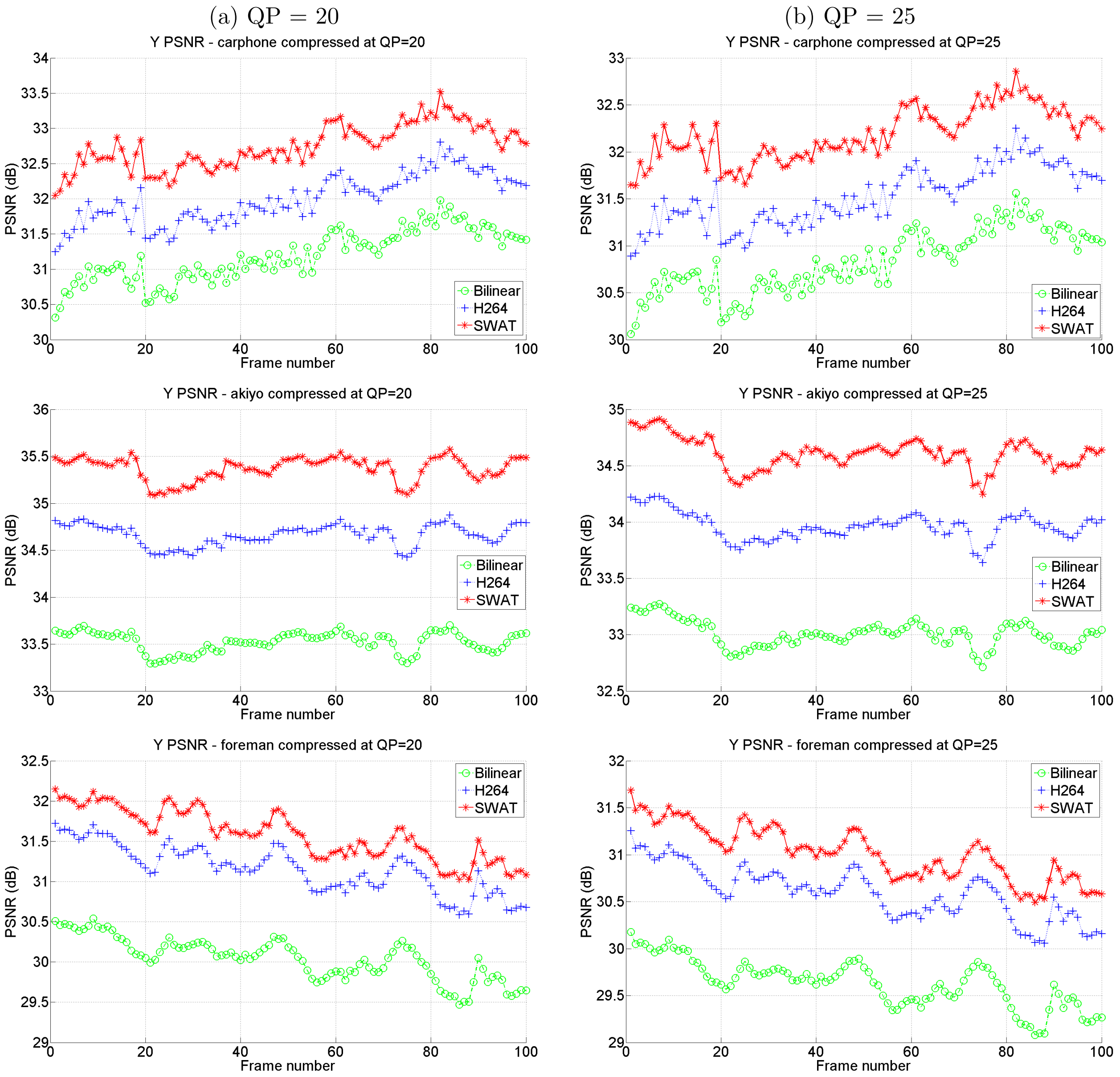


Figure 10. PSNR comparison of different super-resolution algorithms on the first 100 frames of compressed sequences Carphone, Akiyo and Foreman. Compression is done using H.264/AVC reference codec (JM 12.0) with IPP.. encoding at (a) QP=20 and (b) QP=25. Results and Improvements are similar to Figure 8, with some performance lost due to compression.

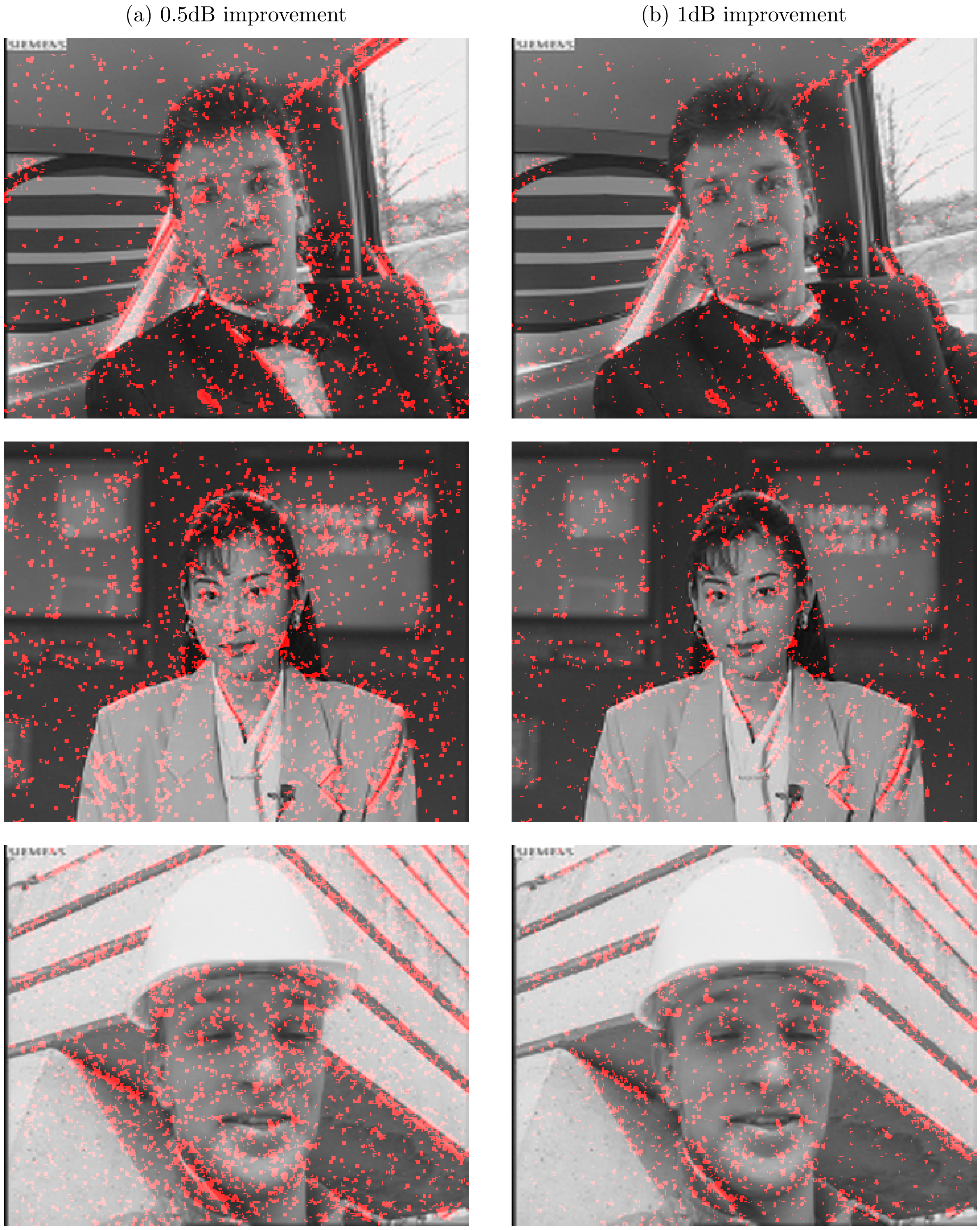


Figure 11. The pixels marked in red gain at least (a) 0.5dB and (b) 1dB in local PSNR when SWAT is used instead of non-directional SWAT.

5. CONCLUSION

Existing algorithms for super-resolution are either computationally expensive or do not provide good quality reconstructions. In this paper, we proposed a fast super-resolution algorithm called SWAT which uses sparse warped transforms and spatially adaptive thresholds. As our results show, the proposed algorithm renders high-resolution video with good subjective and objective quality starting from both uncompressed and compressed low-resolution video.

REFERENCES

1. Y. Altunbasak, A. Patti, and R. Mersereau. “Super-resolution still and video reconstruction from mpeg-coded video,” *IEEE Trans Circuits Systems for Video Tech.*, vol 12, 2002.
2. S. Baker and T. Kanade, “Limits on Super-Resolution and How to Break Them,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol: 24, No. 9, September 2002.
3. Choong S. Boon, Onur G. Guleryuz, Toshiro Kawahara, and Yoshinori Suzuki, “Sparse super-resolution reconstructions of video from mobile devices in digital TV broadcast applications,” Proc. SPIE Conf. on Applications of Digital Image Processing XXIX, in Algorithms, Architectures, and Devices, San Diego, Aug. 2006
4. M. Elad and A. Feuer, “Super-resolution reconstruction of image sequences,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol:21, pp:817-834, 1999.
5. M. Elad, J-L. Starck, P. Querre, and D.L. Donoho, “Simultaneous Cartoon and Texture Image Inpainting Using Morphological Component Analysis (MCA)”, Journal on Applied and Computational Harmonic Analysis, Vol. 19, pp. 340-358, November 2005.
6. M.J. Fadili, J.-L. Starck, and F. Murtagh, “ Inpainting and Zooming using Sparse Representations,” The Computer Journal, July 2007.
7. S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, “Fast and robust multi-frame super-resolution,” *IEEE Trans. Image Processing,* October, 2004.
8. O. G. Guleryuz, “Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions and Iterated Denoising: Part I - Theory,” *IEEE Trans. on Image Processing,* vol. 15, No. 3, pp. 539-554, March, 2006.
9. O. G. Guleryuz, “Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions and Iterated Denoising: Part II - Adaptive Algorithms,” *IEEE Trans. on Image Processing,* vol. 15, No. 3, pp. 555-571, March, 2006.
10. O. G. Guleryuz, “Predicting Wavelet Coefficients Over Edges Using Estimates Based on Nonlinear Approximants,” *Proc. Data Compression Conference*, IEEE DCC-04, April 2004.
11. O. G. Guleryuz, “Weighted Overcomplete Denoising,” *Proc. Asilomar Conf. on Signals and Systems,* Pacific Grove, CA, Nov. 2003.
12. Joint Video Team of ITU-T and ISO/IEC JTC 1, “Draft ITU T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC),” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, March 2003.
13. N. Mueller, Y. Lu, and M. N. Do, ‘ “Image interpolation using multiscale geometric representations,” Proc. of SPIE Symposium on Electronic Imaging, San Jose, 2007.
14. D. Robinson and P. Milanfar, Statistical Performance Analysis of Super-Resolution, *IEEE Trans. on Image Processing,* vol: 15, No. 6, June 2006.
15. C.A. Segall, R. Molina, A. Katsaggelos, and J. Mateos. “Bayesian highresolution reconstruction of low-resolution compressed video,” Proc. IEEE Int’l Conf. on Image Proc. (ICIP2001), Oct. 2001.
16. O. Yamamori, H. Atsumi, T. Mizoguchi, and K. Ishii, “Mobile Terminal Supporting Terrestrial Digital TV Broadcasting,” NTT DoCoMo Technical Journal, Vol. 8, No. 1, pp. 47-56.
17. http://www.nttdocomo.co.jp/product/concept_model/p901itv/