BAYESIAN HIGH-RESOLUTION RECONSTRUCTION OF LOW-RESOLUTION COMPRESSED VIDEO

C. Andrew Segall[†], *Rafael Molina*[*], *Aggelos K. Katsaggelos*[†] *and Javier Mateos*[*]

[†] Department of Electrical and Computer Engineering
Northwestern University, Evanston, IL 60208, USA
{asegall,aggk}@ece.nwu.edu

[*] Departamento de Ciencias de la Computación e I. A.
University of Granada, 18071 Granada, Spain
{rms, jmd}@decsai.ugr.es

ABSTRACT

A method for simultaneously estimating the high-resolution frames and the corresponding motion field from a compressed low-resolution video sequence is presented. The algorithm incorporates knowledge of the spatio-temporal correlation between low and high-resolution images to estimate the original high-resolution sequence from the degraded low-resolution observation. Information from the encoder is also exploited, including the transmitted motion vectors, quantization tables, coding modes and quantizer scale factors. Simulations illustrate an improvement in the peak signal-to-noise ratio when compared with traditional interpolation techniques and are corroborated with visual results.

1. INTRODUCTION

A number of applications require high-resolution images for success. Examples include scientific imaging and a variety of consumer and entertainment products. In the first case, medical, astronomical and weather images are processed, and the increased spatial accuracy provides the ability to resolve small anomalies in the data. In the second case, the conversion of low-resolution images into higher-resolution content is required. For example, desktop video, security monitoring and other low-rate applications are restricted from transmitting full frame images. Instead, the data is enlarged at the receiver.

In this paper, we are concerned with the estimation of high-resolution images from a series of low-resolution observations. When observations are corrupted by additive noise, then techniques such as [3,9,11] address the resolution enhancement problem successfully. However, in many modern systems, images are compressed to reduce the cost of transmission. This renders an additive noise model inaccurate. Instead, the degradation model must reflect knowledge of the compression system, which includes the design of the compression algorithm as well as the parameters provided in the compressed bit-stream. With the ITU or MPEG family of coding methods, these parameters include the coding modes, motion vectors, quantization tables and quantizer scale factors.

Methods that exploit either all or part of these parameters for resolution enhancement have appeared in the literature [1,2,5,6,7]. While the techniques reflect different approaches to the problem, each relies on the pre-calculation of an accurate, sub-pixel motion field. This field establishes a correspondence between pixels at different time instances. Here, we will consider an alternative approach. Instead of pre-computing the motion field prior to resolution enhancement, we simultaneously estimate the motion while enhancing the images. Thus, we are able to jointly estimate the spatial and temporal correlation between each frame of the sequence. This is done utilizing complete knowledge of the compressed bit-stream.

The rest of the paper is organized as follows. In section 2, we provide the necessary background for the approach and formulate the problem within the Bayesian framework. In section 3, we introduce an iterative algorithm for motion estimation and resolution enhancement. The algorithm relies on several models for the image sequence and incorporates information derived from the compressed bit-stream. Finally, experimental results are presented in section 4. A video sequence is decimated and compressed with the MPEG-4 standard. Simulations illustrate the performance of the algorithm.

2. BACKGROUND

Consider two frames of a high-resolution sequence at time instances *l* and *k*. When the frames are closely spaced in time, then it is reasonable to assume that frame *l* can be accurately predicted from frame *k* through the use of a motion vector. Thus, we can say that

$$f_l(x,y) = f_k\left(x + d_{l,k}^x(x,y),\, y + d_{l,k}^y(x,y)\right), \qquad (1)$$

where $f_l(x,y)$ and $f_k(x,y)$ are spatial locations in the high-resolution images at times *l* and *k*, respectively, and $(d_{l,k}^x(x,y),\, d_{l,k}^y(x,y))^T$ is the motion vector that relates the pixel at time *k* to the pixel at time *l*.

The expression in (1) can be rewritten in matrix-vector notion. In this form, the relationship between two images becomes

$$\mathbf{f}_l = \mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k\,, \qquad (2)$$

This work was partially supported by Spanish CICYT project TIC2000-1275.

where $\mathbf{f}_l$ and $\mathbf{f}_k$ are *PMPN*x1 column vectors that are formed by lexicographically ordering each *PM*x*PN* image into a one-dimensional vector, $\mathbf{d}_{l,k}$ is the 2*PMPN*x1 column vector that is formed by lexicographically ordering the motion vectors, and $\mathbf{C}(\mathbf{d}_{l,k})$ is a two-dimensional matrix that describes the motion compensation procedure for the entire frame. If the motion compensation is represented with pixel accuracy, then $\mathbf{C}(\mathbf{d}_{l,k})$ is a permutation matrix, with each entry equal to either zero or one. When sub-pixel motion is considered, then the entries of $\mathbf{C}(\mathbf{d}_{l,k})$ are real numbers that define an interpolation scheme.

Before transmission, the high-resolution images are filtered, down-sampled and compressed. The conversion of a high-resolution frame to its lower resolution observation is therefore expressed as

$$\mathbf{g}_k = \mathbf{T}_{DCT}^{-1} Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHf}_k - \mathbf{g}_k^{Pred}\right)\right] + \mathbf{g}_k^{Pred}, \qquad (3)$$

where $\mathbf{g}_k$ is a vector containing the low-resolution images with dimension *MN*x1, $\mathbf{f}_k$ is the high-resolution data, $\mathbf{g}_k^{Pred}$ is the low-resolution image predicted from a reference compressed frame by utilizing the transmitted motion vectors (for an intra-coded region the prediction is zero), $\mathbf{A}$ is an *MN*x*PMPN* matrix that sub-samples the high-resolution image, $\mathbf{H}$ is an *PMPN*x*PMPN* matrix that filters the high-resolution image, $\mathbf{T}_{DCT}$ and $\mathbf{T}_{DCT}^{-1}$ are the forward and inverse-DCT calculations, and $Q[\cdot]$ represents the quantization procedure. Combining (2) and (3), we state the relationship between any low and high-resolution frame as

$$\mathbf{g}_l = \mathbf{T}_{DCT}^{-1} Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHC}(\mathbf{d}_{l,k})\mathbf{f}_k - \mathbf{g}_l^{Pred}\right)\right] + \mathbf{g}_l^{Pred}. \qquad (4)$$

Looking at the relationship in (4), we see that every low-resolution observation contains information about multiple high-resolution images. We exploit this relationship by considering the reverse logic – that information about every high-resolution frame is stored in multiple observations. One obstacle complicates the approach, as we must identify and define the motion field that relates the high-resolution frame to different observations. In previous approaches for the resolution enhancement of compressed video, the task of finding the motion field is treated as an independent problem. Enhancement is applied assuming that the motion field is completely known. In this work, we treat the motion field as an unknown value and estimate it as we perform the resolution enhancement.

The maximum *a posterior* (MAP) estimate allows us to realize the goal of simultaneously estimating the high-resolution image and its corresponding motion field. Within this framework, the estimate of the high-resolution frame at time *k*, $\hat{\mathbf{f}}_k$, and the matrix of the motion vectors, $\hat{\mathbf{D}}$, are obtained as

$$\begin{aligned}\hat{\mathbf{f}}_k, \hat{\mathbf{D}} &= \arg\max_{\mathbf{f}_k,\mathbf{D}}\left\{p\left(\mathbf{f}_k, \mathbf{D} \mid \mathbf{G}, \mathbf{D}^{Encoder}\right)\right\} \\ &= \arg\max_{\mathbf{f}_k,\mathbf{D}}\left\{\frac{p\left(\mathbf{G}, \mathbf{D}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}\right)p\left(\mathbf{f}_k, \mathbf{D}\right)}{p\left(\mathbf{G}, \mathbf{D}^{Encoder}\right)}\right\},\end{aligned} \qquad (5)$$

where $\mathbf{D}$ is the matrix defined as $(\mathbf{d}_{k-TB,k}, \ldots, \mathbf{d}_{k+TF,k})^T$, $\mathbf{G}$ is the matrix defined as $(\mathbf{g}_{k-TB}, \ldots, \mathbf{g}_{k+TF})^T$, and $\mathbf{D}^{Encoder}$ is the matrix that contains the transmitted motion vectors. In the definitions, *TF* and *TB* describe the number of frames utilized along the forward and backward directions of the temporal axis. Taking the logarithm of (5) and recognizing that the optimization is independent of $p(\mathbf{G}, \mathbf{D}^{Encoder})$, the MAP estimates become

$$\begin{aligned}\hat{\mathbf{f}}_k, \hat{\mathbf{D}} = \arg\max_{\mathbf{f}_k,\mathbf{D}} &\left\{\log p\left(\mathbf{G} \mid \mathbf{f}_k, \mathbf{D}\right) + \log p\left(\mathbf{f}_k\right)\right. \\ &\left. + \log p\left(\mathbf{D}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}\right) + \log p\left(\mathbf{D}\right)\right\}\end{aligned} \qquad (6)$$

where $\mathbf{f}_k$ and $\mathbf{D}$ as well as $\mathbf{G}$ and $\mathbf{D}^{Encoder}$ are assumed to be independent.

3. PROPOSED ALGORITHM

To realize the estimate described in (6), we must first define the probability density functions that appear in the expression. After that, we can construct an appropriate optimization technique for the maximization. In this section, we accomplish both tasks. The first sub-section discusses the probability density models required for the estimate. Specifically, we define $p(\mathbf{G}|\ \mathbf{f}_k,\mathbf{D})$, $p(\mathbf{D}^{Encoder}|\ \mathbf{f}_k,\mathbf{D})$, $p(\mathbf{f}_k)$ and $p(\mathbf{D})$. The second sub-section describes an algorithm that simultaneously estimates the motion field and high-resolution data.

3.1 Probability Density Models

The compression system motivates all of the probability definitions of the sub-section. The first probability density in (6) is $p(\mathbf{G}|\mathbf{f}_k,\mathbf{D})$. This density function describes the noise process introduced during compression, errors introduced during the conversion of the high-resolution image to the lower resolution observation, and any fluctuation in intensity values along the motion field. The conditional density is modeled as Gaussian distributed and expressed as

$$p\left(\mathbf{G} \mid \mathbf{f}_k, \mathbf{D}\right) \propto \exp\left\{-\sum_{l=k-TB}^{k+TF}\beta_l \left\|\mathbf{AHC}\left(\mathbf{d}_{l,k}\right)\mathbf{f}_k - \mathbf{g}_l\right\|^2\right\}, \qquad (7)$$

where β_l is proportional to the inverse of the noise variance at time instant *l*, and the summation includes all of the observations contributing to the estimate of the current high-resolution frame.

When developing a model for the high-resolution images, expressed as $p(\mathbf{f}_k)$, we realize that standard compression techniques introduce two major types of artifacts. Blocking artifacts arise from the independent processing of the image blocks, while ringing artifacts are introduced by the coarse quantization of high frequency information. Since these visual artifacts are rarely part of the original image sequence, they should not be present in a high-resolution estimate. These artifacts are penalized with the density

$$p\left(\mathbf{f}_k\right) \propto \exp\left\{-\left(\lambda_1\left\|\mathbf{Q}_1\mathbf{f}_k\right\|^2 + \lambda_2\left\|\mathbf{Q}_2\mathbf{f}_k\right\|^2\right)\right\}, \qquad (8)$$

where $\mathbf{Q}_1\mathbf{f}_k$ represents the high frequency content within each block, $\mathbf{Q}_2\mathbf{f}_k$ represents the differences across the horizontal and vertical block boundaries, and λ_1 and λ_2 control their relative importance [10]. Frames with significant high frequency energy or large variations across the block boundaries are assigned a low probability with the model in (8).

The distribution of the motion vectors also relies on the compression mechanism. During encoding, video compression algorithms identify motion vectors by comparing the down-sampled high-resolution images to previously encoded frames.

The down-sampled images are not available to the resolution enhancement procedure, as they are not transmitted to the decoder. Therefore, the motion vectors in the bit-stream provide an additional observation of the high-resolution frame and should be incorporated into the enhancement method. This is accomplished with the density function

$$p\left(\mathbf{D}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}\right) \propto \exp\left\{ -\sum_{l=k-TB}^{k+TB} \gamma_l \left\| \mathbf{d}_{l,k} - \mathbf{d}_{l,k}^{Encoder} \right\|^2 \right\}, \quad (9)$$

where $\mathbf{d}_{l,k}^{Encoder}$ is a vector that contains the transmitted motion vector, up-sampled to the higher resolution, and γ is a positive value that expresses a confidence in the transmitted vectors. As the value for γ increases, the estimates become closer to the transmitted motion vectors.

The last distribution appearing in (6) provides an *a priori* model for the motion field. In this paper, we utilize the non-informative prior

$$p(\mathbf{D}) \propto K, \quad (10)$$

where *K* is a constant. This assigns equal probability to every estimate for the motion field. More sophisticated models could also be utilized, such as those explored in [8].

The encoder provides one final piece of information that should be incorporated into the resolution enhancement algorithm. When compressing the low-resolution images, an encoder calculates the DCT for each image block (or error residual) and quantizes the transform coefficients. These quantized values are transmitted to the encoder as a quantization step size and quantization index. When transmitting the information in this format, the decoder is completely aware of the maximum difference between the actual coefficient and the quantized value. Incorporating this constraint into the algorithm requires that

$$\begin{aligned}\hat{\mathbf{f}}_k \in \Big\{\mathbf{f}_k : Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHf}_k - \mathbf{g}_k^{Pred}\right)\right] \\ = Q\left[\mathbf{T}_{DCT}\left(\mathbf{g}_k - \mathbf{g}_k^{Pred}\right)\right]\Big\},\end{aligned} \quad (11)$$

be satisfied. Simply stated, this constraint requires that all estimates for the high-resolution image, after filtering and down-sampling, quantize to the coefficients appearing in the bit-stream.

3.2 Optimization Procedure

By substituting the models presented in (7)-(11) into the estimate described in (6), we obtain

$$\begin{aligned}\hat{\mathbf{f}}_k, \hat{\mathbf{D}} = \arg\min_{\mathbf{f}_k, \mathbf{D}} \Bigg\{ & \sum_{l=k-TB}^{k+TF} \beta_l \left\| \mathbf{AHC}\left(\mathbf{d}_{l,k}\right)\mathbf{f}_k - \mathbf{g}_l \right\|^2 + \lambda_1 \left\| \mathbf{Q}_1 \mathbf{f}_k \right\|^2 \\ & + \sum_{l=k-TB}^{k+TF} \gamma_l \left\| \mathbf{d}_{l,k} - \mathbf{d}_{l,k}^{Encoder} \right\|^2 + \lambda_2 \left\| \mathbf{Q}_2 \mathbf{f}_k \right\|^2 \Bigg\} \\ s.t\ & \mathbf{T}_{DCT}\left(\mathbf{AHf}_k - \mathbf{g}_k^{Pred}\right) \in DCT_{Allowable},\end{aligned} \quad (12)$$

where $DCT_{Allowable}$ denotes the allowable set of DCT coefficients as signaled by the encoder.

The minimization of (12) is accomplished with a cyclic coordinate-decent optimization procedure [4]. In this approach, an estimate for the motion field is found while the high-resolution image is assumed known. Then, the high-resolution image is estimated with the recently found motion field. The motion field is then re-estimated using the current solution for the high-resolution frame, and the process iterates by alternatively finding the motion field and high-resolution images. Treating the high-resolution image as a known parameter, the estimate for the motion field in (12) becomes

$$\begin{aligned}\hat{\mathbf{D}} = \arg\min_{\mathbf{D}} \Bigg\{ & \sum_{l=k-TB}^{k+TF} \beta_l \left\| \mathbf{AHC}\left(\mathbf{d}_{l,k}\right)\mathbf{f}_k - \mathbf{g}_l \right\|^2 \\ & + \sum_{l=k-TB}^{k+TF} \gamma_l \left\| \mathbf{d}_{l,k} - \mathbf{d}_{l,k}^{Encoder} \right\|^2 \Bigg\},\end{aligned} \quad (13)$$

which is minimized with a motion estimation algorithm. Any algorithm is allowable within the framework, and an example is the well-known block matching technique.

Once the estimate for the motion field is found, then the high-resolution image is computed. For a fixed **D**, the minimization of (12) is accomplished by the method of successive approximations, expressed as

$$\begin{aligned}\hat{\mathbf{f}}_k^{n+1} = \mathrm{P}_{\mathrm{DCT}} \Bigg[\hat{\mathbf{f}}_k^{n} + \alpha \Bigg\{ & \lambda_1 \mathbf{Q}_1^T \mathbf{Q}_1 \hat{\mathbf{f}}_k^{n} + \lambda_2 \mathbf{Q}_2^T \mathbf{Q}_2 \hat{\mathbf{f}}_k^{n} \\ & + \sum_{l=k-TB}^{k+TF} \beta_l \left(\mathbf{C}^T\left(\mathbf{d}_{l,k}\right) \mathbf{H}^T \mathbf{A}^T \left(\mathbf{AHC}\left(\mathbf{d}_{l,k}\right) \hat{\mathbf{f}}_k^{n} - \mathbf{g}_l \right)\right) \Bigg\} \Bigg],\end{aligned} \quad (14)$$

where $\hat{\mathbf{f}}_k^{n+1}$ and $\hat{\mathbf{f}}_k^{n}$ are the enhanced frames at the *n*th and (*n*+1)th iteration, $\mathrm{P}_{\mathrm{DCT}}$ is a projection operator that constrains the solution to the valid set of DCT coefficients, α is a relaxation parameter that determines the convergence and rate of convergence of the algorithm, $\mathbf{C}^T(\mathbf{d}_{k,l})$ compensates an image backwards along the motion vectors and $\mathbf{A}^T$ defines the up-sampling operation.

4. EXPERIMENTAL RESULTS

To explore the performance of the proposed super-resolution algorithm, forty frames of the *Mobile* sequence are decimated and compressed. The original images are 704x576 pixel arrays, which are decimated by sub-sampling by a factor of two. Only the central portion of the images is considered, resulting in a low-resolution image that is 176x144 pixels in extent. The sequence is then compressed with an MPEG-4 encoder operating at 256kbps and utilizing the VM5+ rate control mechanism. The resolution enhancement algorithm then processes the compressed frames. In the algorithm, *TB*=3, *TF*=5, $\mathbf{Q}_1$ is a 3x3 discrete Laplacian, $\mathbf{Q}_2$ is a difference operation across the block boundaries, λ_1=λ_2=.1, β=.45, and γ=1. The algorithm is stopped when $\| \hat{\mathbf{f}}_{k+1} - \hat{\mathbf{f}}_k \|^2 < 50$, and a new estimate for the motion field is computed whenever $\| \hat{\mathbf{f}}_{k+1} - \hat{\mathbf{f}}_k \|^2 < 100$.

Visual results from the experiments are shown in Figure 1. In the figure, (a) displays a portion of an uncompressed high-resolution image, (b) shows the result of bi-linearly interpolating the decoded low-resolution image and (c) illustrates the result of the proposed approach. When comparing the images, notice the performance of the algorithm within the numbers of the calendar. For example, the numbers 29 and 30 are not discernible in the bi-linear result. However, they are readable in the super-resolution estimate. As a second example, notice the improvement of the number 18. In the bi-linear estimate, the number is severely distorted. The proposed method provides significant improvement and recovers much of the spatial detail.

Improving the legibility of the numbers is just one example of the performance of the algorithm. Improvements over the entire sequence are quantified with the peak signal-to-noise (PSNR) metric. In the figure, the PSNR of the bi-linear and super-resolution results are 28.79dB and 29.58dB, respectively. This is an improvement of .79dB, which is representative of the entire sequence. The average PSNR of bi-linearly interpolated result is 28.72dB, while the average PSNR of the proposed algorithm is 29.53dB.

5. REFERENCES

[1] Y. Altunbasak and A.J. Patti, "A Maximum A Posteriori Estimator for Higher Resolution Video Reconstruction from MPEG Video," *Proc. of the IEEE ICIP*, Vancouver, BC, Canada, Sept. 10-13, 2000.

[2] D. Chen and R.R. Schultz, "Extraction of High-Resolution Still from MPEG Sequences," *Proc. of the IEEE ICIP*, pp.465-69, Chicago, IL, Oct. 4-7, 1998.

[3] R.C. Hardie, K.J. Barnard and E.E. Armstrong, “Joint MAP Registration and High-Resolution Image Estimation Using a Sequence of Undersampled Images,” *IEEE Trans. IP*, vol.6, no.12, pp.1621-1633, 1997.

[4] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, 1984.

[5] A.J. Patti and Y. Altunbasak, “Super-Resolution Image Estimation for Transform Coded Video with Application to MPEG,” *Proc. of IEEE ICIP*, Kobe, Japan, Oct. 25-28, 1999.

[6] B. Martins and S. Forchhammer, "A Unified Approach to Restoration, De-interlacing and Super-resolution of MPEG-2 Decoded Video," *Proc. of the IEEE ICIP*, Vancouver, CA, Sept. 10-13, 2000.

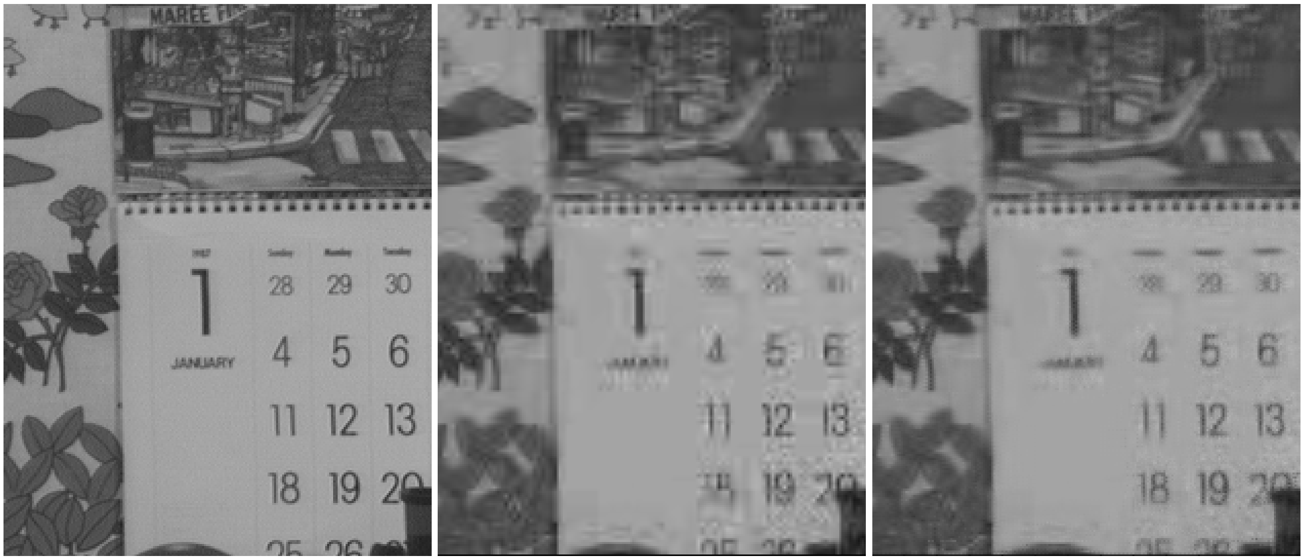
[7] J. Mateos, A.K. Katsaggelos and R. Molina, "Resolution Enhancement of Compressed Low Resolution Video," *Proc. IEEE ICASSP*, Istanbul, Turkey, June 5-9, 2000.

[8] T. Ozcelik, J.C. Brailean, and A.K. Katsaggelos, “Image and Video Compression Algorithms Based on Recovery Techniques using Mean Field Annealing,” *Proc. of the IEEE*, vol.83, no.2, pp.304-316, 1995.

[9] R.R. Schultz and R.L. Stevenson, "Extraction of High Resolution Frames from Video Sequences," *IEEE Trans. IP*, vol.5, no.6, pp.996-1011, 1996.

[10] C.A. Segall and A.K. Katsaggelos, “Enhancement of Compressed Video using Visual Quality Metrics,” *Proc. of the IEEE ICIP*, Vancouver, BC, Canada, Sept. 10-13, 2000.

[11] B. Tom and A.K. Katsaggelos, "Resolution Enhancement of Monochrome and Color Video Using Motion Compensation," *IEEE Trans. IP*, vol.10, no.2, pp.278-287, Feb. 2001.



(a) (b) (c)

Figure 1. Experimental results illustrate the potential of the proposed procedure: (a) original image, (b) bi-linearly interpolating the low-resolution compressed image and (c) result of the proposed approach. The proposed algorithm introduces improvement throughout the image sequence. (Notice the improvement in the number 18.) The PSNR values for the bi-linear and proposed technique are 28.79dB and 29.53dB, respectively.