Towards a new video compression scheme using super-resolution

R. Molina[a], A.K. Katsaggelos[b], L.D. Alvarez[a] and J. Mateos[a]

[a]Departamento de Ciencias de la Computación e I.A.
Universidad de Granada. 18071 Granada, España.;
[b]Department of Electrical Engineering and Computer Science
Northwestern University, Evanston, Illinois 60208-3118, USA

ABSTRACT

The term *super-resolution* is typically used in the literature to describe the process of obtaining a high resolution (HR) image or a sequence of HR images from a set of low resolution (LR) observations. This term has been applied primarily to spatial and temporal resolution enhancement. However, intentional pre-processing and downsampling can be applied during encoding and super-resolution techniques to upsample the image can be applied during decoding when video compression is the main objective.

In this paper we consider the following three video compression models. The first one simply compresses the sequence using any of the available standard compression methods, the second one pre-processes (without downsampling) the image sequence before compression, so that post-processing (without upsampling) is applied to the compressed sequence. The third model includes downsampling in the pre-processing stage and the application of a super resolution technique during decoding. In this paper we describe these three models but concentrate on the application of super-resolution techniques as a way to post-process and upsample a compressed video sequences. Experimental results are provided on a wide range of bitrates for two very important applications: format conversion between different platforms and scalable video coding.

Keywords: super-resolution, compression, MPEG-4, pre-processing of video sequences, post-processing of compressed video sequences, downsampling, scalable video coding.

1. INTRODUCTION

Video compression has been the enabling technology behind the multimedia revolution we are experiencing. There have been a number of successful video compression standards (i.e., MPEG2, MPEG4, H.264), while there are on-going standardization efforts (i.e., SVC and the recently announced H.265). All existing standards define the decoder, i.e., the syntax of the bitstream which will make it decodable by the decoder. Pre- and post-processing techniques are therefore not considered part of the standard and represent the add-on value provided by the implementer of the standard. More importantly they are considered as separate steps, for example, the pre-processor typically does not interact with the rate controller (an exception is presented in Karunaratne *et al.*[4]) . Furthermore, as is justifiable in most cases, an open-loop system is considered, i.e., the pre-processor does not have knowledge of the port-processor, and vice versa. When a codec providing spatial scalability is considered (i.e., MPEG4, H..264, SVC), then the upsampling process represents the normative part.

In this paper we provide some initial investigation of a pre- and post-processing scheme (without down- and up-sampling) and concentrate on the application of super-resolution techniques as an integral part of a video codec. For ease of presentation, we consider three models. The first one considers the compression of the sequence using any of the standard compression methods, the second one pre-processes (without downsampling) the image sequence before compression, so that post-processing (without upsampling) is applied to the compressed sequence. The third one considers downsampling during pre-processing and upsampling during post-processing. Note that, although all three systems can be considered as special cases of a compression scheme with pre- post-processing forming an integral part of it (in an open or closed system consideration), we chose the consideration of three systems for easy of presentation. Thus the second model covers, for instance, the case of removing blur and noise from an observed video sequence, while the third could aim at removing blur and noise as well as viewing image sequences or regions of interest over different platforms. Although we describe in this paper all three models above, we will concentrate mainly on the third one and in particular on the use of super-resolution techniques as a post-processing step in video compression. We believe that by addressing a number of questions related, for example, to rate control, and the optimal design of the down-sampling and up-sampling factors and filters, a new paradigm can emerge in developing new video compression schemes.

The term *super-resolution* is typically used in the literature to describe the process of obtaining an HR image or a sequence of HR images from a set of LR observations. This term has been applied primarily to spatial and temporal resolution enhancement (a comprehensive classification of spatio-temporal super-resolution problems is provided by Borman[2]) and more recently to spectral together with spatial resolution enhancement (see Molina *et al.*[6] for a short review). In these two contexts super-resolution is used as a means to ameliorate the undesirable reduction in resolution introduced by the imaging system. However, the intentional downsampling during pre-processing and the application of super-resolution techniques during post-processing can be utilized as a mechanism in controlling the bitrate so that the overall quality of the reconstructed video is optimized.

Two quite apart (in terms of bits) scenarios are well suited for super-resolution for compression. As Segall *et al.*[10] describe, although the bitrates of current high definition systems ensure fidelity in representing an original video sequence, they preclude widespread availability of high definition programming. For example, satellite and Internet based distribution systems are poorly suited to deliver a number of high rate channels, and video on demand applications must absorb a significant increase in storage costs. Furthermore, pre-recorded DVD-9 stores less than an hour of high definition video. Additionally, with the current proliferation of HDTV sets, there is a growing number of DVD players and DVD recorders that can now upscale standard DVD playback output to match the pixel count of HDTV in order to display the signal on a HDTV set with higher definition. In this context, super-resolution can be applied at the output of the DVD player to increase not only the resolution of the video sequence but, also, to increase its quality.

Within the same high bitrate scenario, DVD recorders cannot record in HDTV standards. This is due to the spatial resolution used in DVD format and the limited space of current DVDs for the storage needs of HDTV signals. An alternative for storing HDTV programs with DVD recorders is to pre-process and downsample the HDTV signal to a resolution supported by the standard DVD recorder prior to the compression and storing process. Then, when playing back the compressed video, a super-resolution process can be applied to recover the signal at the original HDTV resolution. Note that the pre-processing and dowsampling operations can be different depending on the characteristics of the frame being processed. In this case it will be highly desirable that both the coder and the super-resolution algorithm share the information about how the low resolution DVD video sequence has been obtained from the HDTV signal.

Another scenario for super-resolution for compression is as follows. As described by Bruckstein *et al.*,[3] the use of DCT at low bitrates introduces disturbing blocking artifacts. It appears that at low bitrates an appropriately down-sampled image compressed using DCT and later interpolated (or filtered using super-resolution techniques), can be visually better than the high resolution image compressed directly with a DCT scheme at the same number of bits.

Together with these two examples, spatial scalability in video represents an area where downsampling, compression and upsampling play a key role in designing systems that split a single video source (or video object plane) into a base layer (lower spatial resolution) and enhancement layers (higher spatial resolution).

In this paper we analyze the effect of pre-processing and downsampling a video sequence prior to using the MPEG-4 video compression standard and the posterior use of super-resolution techniques to spatially upsample the compressed sequence. We will compare the use of different filtering methods, and also compare the quality of the compressed sequence using a “standard” approach and using pre-processing, downsampling, compression and super-resolution techniques during decoding, at the same bitrate.

The paper is organized as follows. In section 2 we mathematically formulate the three different described models for video compression. The first one uses hybrid motion compensation and transform based coding, while the second one pre- and post-process the compressed video sequence in order to obtain a better estimate of the original sequence. The goal is to design the best post-processor assuming that the pre-processor is known and also to design the best pre-processor once the post-processor is fixed. The third model extends the second one by including downsampling and upsampling at the pre- and post-filtering step, respectively. In section 3 we study several pre- and post- filtering approaches proposed in the literature and propose a post-processing and upsampling method based on the use of super-resolution techniques when downsampling is involved. Section 4 demonstrates the benefit of using the filtering techniques described in the paper. Finally, Section 5 concludes the paper.

2. PROBLEM FORMULATION

Let us denote by $f(x, y, t)$ the continuous in time and space dynamic scene which is being imaged. If sampling according to the Nyquist criterion in time and space, such scene is represented by the sequence $f_l(m, n)$, where $l = 1, \ldots, L$, $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$ represent respectively the discrete temporal and two spatial coordinates. We use $\mathbf{f}_l$ to denote the vector corresponding to the $l - th$ image frame which is lexicographically ordered by rows. In order to represent images as vectors, all images will be lexicographically ordered by rows.

To compress the original sequence $\mathbf{f} = \{\mathbf{f}_1, \ldots, \mathbf{f}_L\}$ we can use hybrid motion compensated and transform based coding at a given rate R^I, to obtain $\mathbf{f}^I = \{\mathbf{f}_1^I, \ldots, \mathbf{f}_L^I\}$, an estimate of the original sequence $\mathbf{f}$, as

$$\mathbf{f}^I = C_{R^I}[\mathbf{f}], \tag{1}$$

where C_{R^I} denotes a compression method at bitrate R^I. This model is henceforth referred to as Model I.

To reduce the bitrate we can also first spatially pre-process (no downsampling included) each frame in the original sequence to obtain

$$\mathbf{b}_l = H_l^{II}\mathbf{f}_l, \qquad l = 1, \ldots, L, \tag{2}$$

where H_l^{II} is a $(M \times N) \times (M \times N)$ pre-processing matrix, and then compress the sequence $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_L\}$ at a bitrate R^{II} to obtain

$$\mathbf{b}^I = C_{R^{II}}[\mathbf{b}]. \tag{3}$$

This sequence can now be post-processed O_l^{II} to obtain a second estimate of the original $l - th$ frame, that is,

$$\mathbf{f}_l^{II} = O_l^{II}[\mathbf{b}^I], \qquad l = 1, \ldots, L. \tag{4}$$

Note that the pre- and post-processing filters may be image dependent, the whole sequence $\mathbf{b}^I$ or a subset of it (not just $\mathbf{b}_l^I$) may be used to obtain $\mathbf{f}_l^{II}$ and O_l^{II} does not have to be a linear filter on $\mathbf{b}^I$. This model is henceforth referred to as Model II.

We can control (reduce) the bitrate needed to compress the sequence $\mathbf{f}$ by first downsampling in addition to pre-processing, then using a video compression method, and finally post-processing with upsampling the reconstructed downsampled sequence in order to obtain the reconstructed sequence at the original spatial resolution. This model, which extends Model II, will be referred to as Model III. More specifically, an original frame $\mathbf{f}_l$ in the sequence $\mathbf{f}$ is pre-processed and downsampled by a factor of P using the $(M/P \times N/P) \times (M \times N)$ matrix $H_{D\,l}^{III}$, producing the low resolution frame

$$\mathbf{c}_l = H_{D\,l}^{III}(\mathbf{f}_l), \qquad l = 1, \ldots, L. \tag{5}$$

The sequence $\mathbf{c} = \{\mathbf{c}_1, \ldots, \mathbf{c}_L\}$ is compressed at a bitrate R^{III} to obtain

$$\mathbf{c}^I = C_{R^{III}}[\mathbf{c}]. \tag{6}$$

Finally, the sequence $\mathbf{c}^I$ is post-processed and upsampled using $K_{U\,l}^{III}$ to obtain a third estimate of the original $l - th$ frame in the sequence

$$\mathbf{f}_l^{III} = O_{U\,l}^{III}[\mathbf{c}^I], \qquad l = 1, \ldots, L. \tag{7}$$

Note that, as it was mentioned in reference to Model II, the pre- and post-processing may be image dependent, the whole sequence $\mathbf{c}^I$ or a subset of it (not just $\mathbf{c}_l^I$) may be used to obtain $\mathbf{f}_l^{III}$, and $O_{U\,l}^{III}$ does not have to be linear on $\mathbf{c}^I$.

In the next section we first describe, at a high level, the model used to compress the sequence, that is, we describe $C_R[\ ]$. Then we describe some of the models used in the literature to define H_l^{II} and O_l^{II}. For Model III we will describe the normative upsampling procedure defined by the working draft (WD) of MPEG-4 and will propose an upsampling method based on super-resolution techniques.

3. VIDEO COMPRESSION METHODS

3.1. Model I. Hybrid motion compensated and transform based coding

Let us provide a brief and high level description of a hybrid motion-compensated video compression system. The sequence $\mathbf{f}$ is compressed with a video compression system resulting in $\mathbf{f}^I = \{\mathbf{f}_1^I, \ldots, \mathbf{f}_L^I\}$, as per Equation (1). Each compressed image $\mathbf{f}_l^I$ has components $f_l^I(i,j)$, where i and j are integer numbers that indicate spatial location and l is the time index. The size of the compressed images is also $M \times N$. Using matrix-vector notation, each compressed image is denoted by the $(M \times N) \times 1$ vector $\mathbf{f}_l^I$. The compression system also provides the motion vectors $v^I(i,j,l,m)$ that predict pixel $\mathbf{f}_l(i,j)$ from some previously coded frame $\mathbf{f}_m^I$ (more than one past frames can be used for prediction in certain standards, such as, H.264 and a combination of past and future frames are used to predict a bi-directional frame). These motion vectors that predicts $\mathbf{f}_l$ from $\mathbf{f}_m^I$ are represented by the $(2 \times M \times N) \times 1$ vector $\mathbf{v}_{l,m}^I$ that is formed by stacking the transmitted horizontal and vertical offsets.

During compression, frames are divided into blocks that are encoded with one of two available methods, intracoding or intercoding. For the first one, a linear transform such as the DCT (Discrete Cosine Transform) is applied to each block (usually of size 8×8 pixels). The operator decorrelates the intensity data and the resulting transform coefficients are independently quantized and transmitted to the decoder. For the second method, predictions for the blocks are first generated by motion compensating previously transmitted image frames. The compensation is controlled by motion vectors that define the spatial and temporal offset between the current block and its prediction. Computing the prediction error, transforming it with a linear transform, quantizing the transform coefficients, and transmitting the quantized information refine the prediction.

Using all this information, the relationship between compressed and uncompressed frames becomes

$$\mathbf{f}_l^I = T^{-1}Q\left[T\left(\mathbf{f}_l - MC_l(\mathbf{f}_l^{I\,P}, \mathbf{v}_l^I)\right)\right] + MC_l(\mathbf{f}_l^{I\,P}, \mathbf{v}_l^I), \qquad l = 1, \ldots, L, \tag{8}$$

where $Q[.]$ represents the quantization procedure, T and T^{-1} are the forward and inverse-transform operations, respectively, $MC_l(\mathbf{f}_l^{I\,P}, \mathbf{v}_l^I)$ is the motion compensated prediction of $\mathbf{f}_l$ formed by motion compensating previously decoded frame(s) as defined by the encoding method, and $\mathbf{f}_l^{I\,P}$ and $\mathbf{v}_l^I$ denote the set of decoded frames and motion vectors that predict $\mathbf{f}_l$, respectively. We want to make clear here that MC_l depends only on a subset of $\mathbf{v}_l^I$ and $\mathbf{f}^I$. For example, when the bitstream contains a sequence of P-frames then $\mathbf{f}_l^{I\,P} = \mathbf{f}_{l-1}^I$ and $\mathbf{v}_l^I = \mathbf{v}_{l,l-1}^I$. However, as there is a trend towards increased complexity and non-causal predictions within the motion compensation procedure, we keep the above notation for generality.

3.2. Model II. Preprocessing and Postprocessing of the video sequence

Let us assume that the $l-th$ frame in the original video sequence is processed before compression using the filter H_l^{II} in Equation (2). We want to obtain the best post-processed image $\mathbf{f}_k^{II}$ using the whole compressed sequence $\mathbf{b}^I$ in Equation (3).

Let us assume that the acquired images $\mathbf{f}_1, \ldots, \mathbf{f}_L$ satisfy

$$f_l(a,b) = f_k(a + d_{l,k}^x(a,b), b + d_{l,k}^y(a,b)) \;, \tag{9}$$

where $d_{l,k}^x(a,b)$ and $d_{l,k}^y(a,b)$ denote respectively the horizontal and vertical components of the displacement, that is, $d_{l,k}(a,b) = (d_{l,k}^x(a,b), d_{l,k}^y(a,b))$.

Equation (9) can be rewritten using matrix-vector notation as

$$\mathbf{f}_l = \mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k \;, \tag{10}$$

where $\mathbf{C}(\mathbf{d}_{l,k})$ is the $(M \times N) \times (M \times N)$ matrix that maps frame $\mathbf{f}_l$ to frame $\mathbf{f}_k$, and $\mathbf{d}_{l,k}$ is the $2(M \times N)$ column vector defined by lexicographically ordering the values of the displacements between the two frames. We will use $\mathbf{d}_k$ to denote the whole set of motion vectors, mapping frames $\mathbf{f}_l, \ldots, \mathbf{f}_L$ onto frame $\mathbf{f}_k$, that is,

$$\mathbf{d}_k = \{\mathbf{d}_{1,k}, \ldots, \mathbf{d}_{L,k}\}. \tag{11}$$

Using now Equation (8) we obtain $\mathbf{b}_l^I$, the $l-th$ compressed frame, in Equation (3) and substituting $H_l^{II}\mathbf{f}_l$ by $H_l^{II}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k$ we have

$$\mathbf{b}_l^I = T^{-1}Q\left[T\left(H_l^{II}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - MC_l(\mathbf{b}_l^{I\,P}, \mathbf{v}_l^{II})\right)\right] + MC_l(\mathbf{b}_l^{I\ P}, \mathbf{v}_l^{II}), \qquad l = 1, \ldots, L, \tag{12}$$

where $MC_l(\mathbf{b}_l^{I\,P}, \mathbf{v}_l^{II})$ is the motion compensated prediction of $\mathbf{b}_l$ formed by motion compensating previously decoded frame(s) as defined by the encoding method, and $\mathbf{b}_l^{I\ P}$ and $\mathbf{v}_l^{II}$ denote the set of decoded frames and motion vectors that predict $\mathbf{b}_l$, respectively.

Then $\mathbf{f}^{II}$, the best post-processed sequence, could, for instance, be calculated as

$$\mathbf{f}_k^{II} = \arg\min_{\mathbf{f}_k} \sum_{l=1}^{L} \parallel T^{-1}Q\left[T\left(H_l^{II}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - MC_l(\mathbf{b}_l^{I\,P}, \mathbf{v}_l^{II})\right)\right] + MC_l(\mathbf{b}_l^{I\ P}, \mathbf{v}_l^{II}) - \mathbf{b}_l^I \parallel^2 \tag{13}$$

where the motion vectors $\mathbf{d}_{l,k}, l = 1, \ldots, L$ are assumed known. Simple modeling of $T^{-1}QT$ as Gaussian with constant variance leads to the minimization problem

$$\mathbf{f}_k^{II} = \arg\min_{\mathbf{f}_k} \sum_{l=1}^{L} \parallel H_l^{II}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - \mathbf{b}_l^I \parallel^2 . \tag{14}$$

This represents an ill-posed problem. Segall[8] proposes the simultaneous estimation of the $k-th$ post-processed image and the corresponding motion vectors using the methodology developed by Segall *et al*[11] for super-resolution image reconstruction.

Let us now assume that O_l^{II} in Equation (4) is a known post-processing $(M \times N)^2$ matrix. We want to design the best pre-processing matrix H_l^{II} for each frame l such that

$$H_l^{II} = \arg\min_{H} \parallel \mathbf{f}_l^{II} - \mathbf{f}_l \parallel^2 . \tag{15}$$

Clearly $\mathbf{f}_l^{II}$ depends on H_l^{II}. By simply modeling the compression process as an independent Gaussian noise process in each frame we have

$$H_l^{II} = \arg\min_{H} \parallel O_l^{II}H^{II}\mathbf{f}_l - \mathbf{f}_l \parallel^2, \tag{16}$$

which is again an ill-posed problem. A more realistic modeling of the compression process that also takes into account the previously estimated H_n^{II}, $n < l$ is studied in Segall[8] and Segall *et al.*[10].

3.3. Super resolution for video compression

As already mentioned earlier, the basic idea in this work it to intentionally reduce the resolution of the original video data before encoding, encode the low resolution data, and then bring the decoded data back to the original resolution by utilizing a super-resolution approach. In doing so, a number of parameters will need to be optimally estimated in the rate-distortion sense, such as, the horizontal and vertical downsampling factor (which should vary per frame or even per region within the frame), and the parameters of the downsmapling and upsampling filters. A number of optimization problem can be formulated depending on what is considered to be known or unknown. As a first step towards this direction we will assume that the decoder that will be employing a super-resolution approach has knowledge of the filter used for downsampling at the encoder and the downsampling factor (this information can be provided to the decoder as part of the bitstream). Since this represents the same situation encountered in a scalable video coder, we will draw from the current standardization effort in developing a scalable video coder (SVC). In it, the upsampling operation is normative and defined to be an interpolation process (see Reichel *et al.*[7]). As described by Segall[12] the procedure defined by the working draft (WD) has as follows.

1. Map pixel values in the low-resolution grid to the high-resolution grid according to

   $$\mathbf{f}_l^{III}(2x, 2y) = \mathbf{c}^I(x, y), \tag{17}$$

   where $\mathbf{f}_l^{III}$ and $\mathbf{c}^I$ have been defined in Equations (7) and (5), respectively.

2. Convolve the high-resolution grid in the horizontal and vertical directions with the kernel

$$[0\ 0\ 1\ 0\ -5\ 0\ 20\ 32\ 20\ 0\ -5\ 0\ 1\ 0\ 0]/32. \tag{18}$$

Inspection of the interpolation kernel shows that the pixel locations defined by step 1 are unmodified while the remaining pixels are interpolated using a six-tap FIR filter.

The downsampling operation is a non-normative procedure in the current SVC WD. However, only a small number of downsampling filters are currently utilized for SVC testing and development. These filters are found in the MPEG-4 verification model (VM) (see Li *et al.*[5]). They possibly originate from Bjontegaard and Lillev[1] . The specific selection of these filters is described by Sullivan and Sun[13] .

We study the case of dyadic scalability. In this case, the majority of test sequences utilize the following procedure to generate a low-resolution frame: first, filter the high-resolution frame with the 13-tap separable kernel

$$[0\ 2\ 0\ -4\ -3\ 5\ 19\ 26\ 19\ 5\ -3\ -4\ 0\ 2\ 0]/64. \tag{19}$$

Then, discard every-other pixel in the horizontal direction and every-other line in the vertical direction. This will produce the matrix $H^{III}_{D\,l}$ corresponding to spatial scalability coding.

Let then $H^{III}_{D\,l}$ be in general the matrix defining the pre-processing and downsampling process. Then using the motion compensated image defined in Equation (9) in Equation (5) we have

$$\mathbf{c}_l = H^{III}_{D\,l}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k \tag{20}$$

Then using Equation (8) to obtain $\mathbf{c}^I_l$ defined in Equation (6) and substituting $H^{III}_{D\,l}\mathbf{f}_l$ by $H^{III}_{D\,l}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k$ we have

$$\mathbf{c}^I_l = T^{-1}Q\left[T\left(H^{III}_{D\,l}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - MC_l(\mathbf{c}^{I\,P}_l, \mathbf{v}^{III}_l)\right)\right] + MC_l(\mathbf{c}^{I\ P}_l, \mathbf{v}^{III}_l), \qquad l = 1, \ldots, L, \tag{21}$$

where $MC_l(\mathbf{c}^{I\,P}_l, \mathbf{v}^{III}_l)$ is the motion compensated prediction of $\mathbf{c}_l$ formed by motion compensating previously decoded frame(s) as defined by the encoding method, and $\mathbf{c}^{I\ P}_l$ and $\mathbf{v}^{III}_l$ denote the set of decoded frames and motion vectors that predict $\mathbf{c}^I_l$, respectively.

Any of the super-resolution techniques described in Segall *et al.*[9] can be used to recover the original high resolution sequence from $\mathbf{c}^I$ defined in Equation (6) as the compressed version of $\mathbf{c}$. In this paper we will use as super-resolution model the one proposed by Segall *et al.*[11] . Adapting this method to our problem, our goal becomes finding

$$\begin{aligned}
\hat{\mathbf{f}}_k, \hat{\mathbf{d}} &= \arg\max_{\mathbf{f}_k, \mathbf{d}} \\
&\quad (H^{III}_{D\,l}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - MC_l(\mathbf{c}^{I\ P}_l, \mathbf{v}^{III}_l))^T \mathbf{K}^{-1}_{Q,MV}(H^{III}_{D\,l}\mathbf{C}(\mathbf{d}_{l,k})\mathbf{f}_k - MC_l(\mathbf{c}^{I\,P}_l, \mathbf{v}^{III}_l)) \\
&+ \lambda_1 \parallel \mathbf{Q}_1\mathbf{f}_k \parallel^2 + \lambda_2 \parallel \mathbf{Q}_2 H^{III}_{D\,l}\mathbf{f}_k \parallel^2 + \lambda_3 \sum_l \parallel \mathbf{Q}_3\mathbf{d}_l \parallel^2
\end{aligned} \tag{22}$$

where $\mathbf{K}_{Q,MV}$ is the covariance matrix for the prediction errors, $\mathbf{Q}_1$ represents a linear high-pass operation that penalizes super-resolution estimates that are not smooth, $\mathbf{Q}_2$ represents a linear high-pass operator that penalizes estimates with block boundaries, $\mathbf{Q}_3$ is a linear high-pass operator, and λ_1, λ_2 and λ_3 control the influence of the associated norms.

4. EXAMPLES

In order to asses the validity of the proposed approach a set of experiments were performed on high resolution sequences. Results are reported on the "Rush-hour" sequence, taken at rush-hour in Munich, available at the Technische Universität München (`http://www.ldv.ei.tum.de/liquid.php?page=70`). The scene is being observed with a fixed camera, and has a high depth of field. The sequence has 500 frames of size 1920 × 1080 pixels and a frame rate of 25 fps stored in progressive format. A 352 × 288 part of frame 17 is displayed in Fig. 2a. The raw sequence needs a bandwidth of more than 1.15 Gbps to be transmitted or stored. The ATSC

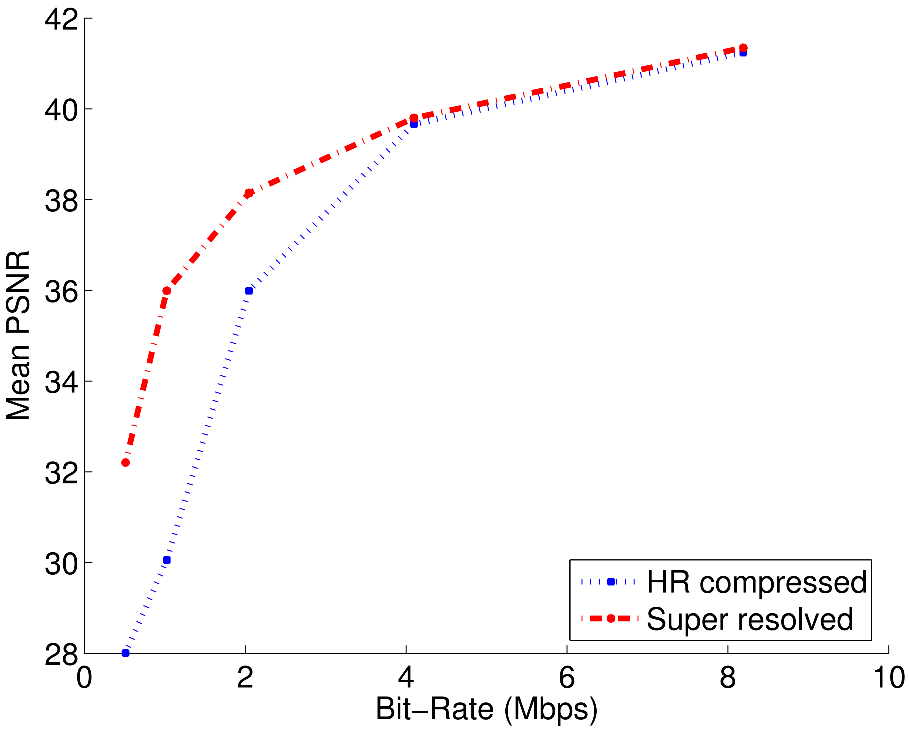


Figure 1. Rate distortion curves for the "Rush-hour" sequence. The super-resolution for video compression approach leads to significant quality improvement at low bitrates.

standard for digital television (DTV) applications defines video broadcasting at 19.4 Mbits/second, that is, a compression rate of 60:1 is required.

For compressing the video sequences, we utilize the MPEG-4 bitstream syntax, which describes a hybrid motion-compensated and block DCT compression system. Apart from the first image in the sequence, which is intra-coded, each frame is compressed as a P-frame. This restricts the reference frame for the motion vectors to be the temporally previous frame. The rate control VM5+ mechanism is utilized for bit allocation.

To asses the quality of a single reconstructed frame l, $\hat{\mathbf{f}}_l$, together with the visual quality inspection, the PSNR defined as

$$PSNR_l = 10\log_{10}\frac{255^2 \times M \times N}{\parallel \mathbf{f}_l - \hat{\mathbf{f}}_l \parallel^2},$$

is used. When evaluating the quality of the whole sequence, we use the mean PSNR, defined as

$$PSNR = \frac{1}{L}\sum_{l=1}^{L} PSNR_l.$$

Our first experiment addresses the problem of compressing high-resolution video sequences at low to medium bitrates. In this experiment the original sequence was compressed at different bitrates ranging from 0.5 Mbps to 8 Mbps. This range includes typical DVD (MPEG-1 or MPEG-2) and MPEG-4 compressed video bitrates.

The quality of the resulting compressed sequence, depicted in Fig. 1, ranges from a mean PSNR 28.0 dB at 0.5Mbps to 41.24 dB at 8 Mbps. Fig. 2b depicts a 352×288 part of the resulting compressed frame 17 when the sequence has been compressed to a bitrate of 2 Mbps. The PSNR of this frame is 35.95 dB.

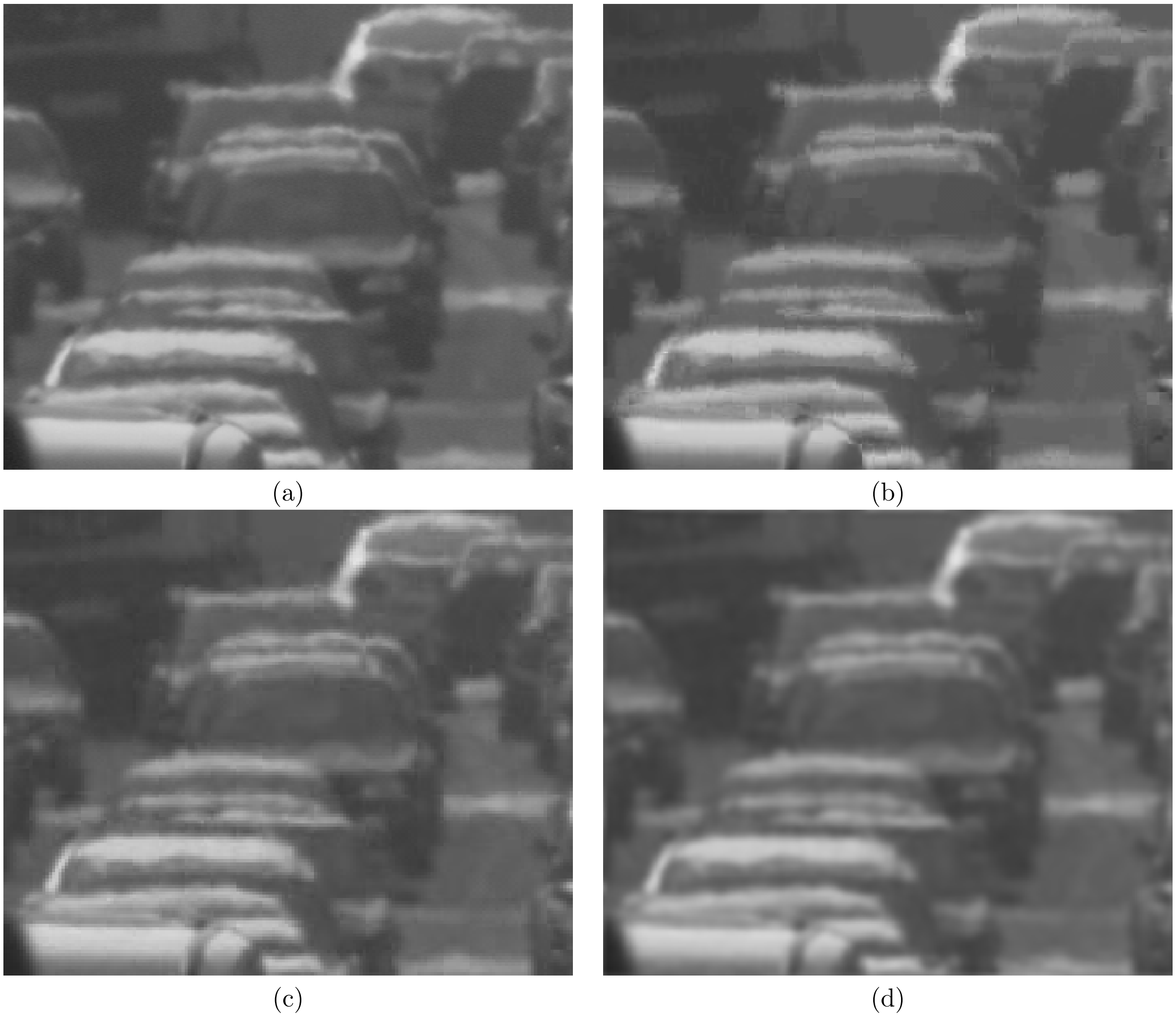


Figure 2. Cropped frame 17 of the "Rush-hour" sequence. (a) Original frame, (b) High resolution frame of the sequence coded at 2 Mbps (PSNR = 35.95 dB). (c) Low resolution frame of sequence coded at 2 Mbps. The frame has been enlarged for visualization purposes. (d) Corresponding super-resolved frame (PSNR = 37.84 dB)

In this first experiment, each frame of the original sequence was downsampled to a size of 960 × 540 pixels by processing it with the filter $0.25\mathbf{I}_{2\times 2}$ and then discarding every-other pixel in the horizontal direction and every-other line in the vertical direction. The resulting low resolution sequence was then coded using the same set of previously used bitrates. An 176 × 144 part of the resulting compressed frame 17 for the 2 Mbps compressed sequence is displayed in Fig. 2c. The image has been enlarged by pixel replication to 352 × 288 pixels for visualization purposes.

The super-resolution method corresponding to Eq. (22) was then applied to each low resolution compressed frame, resulting in sequences with mean PSNRs ranging from 32.20 dB at 0.5 Mbps to 41.34 dB at 8 Mbps, as shown in Fig. 1. A 352 × 288 part of the reconstructed frame corresponding to the frame in Fig. 2c is displayed in Fig. 2d. The visual quality of the reconstructed images is greatly improved. The PSNR of this frame is 37.84 dB thus obtaining an increase of the PSNR with respect to the high resolution compressed image of 1.95 dB. Figure 1 shows that super-resolution for compression techniques significantly increase the quality of the resulting video sequences at low bitrates, obtaining an increase of the mean PSNR of as much as 4.2 dB. However, at higher bitrates, where there is enough bandwidth to obtain good quality images when compressing the original

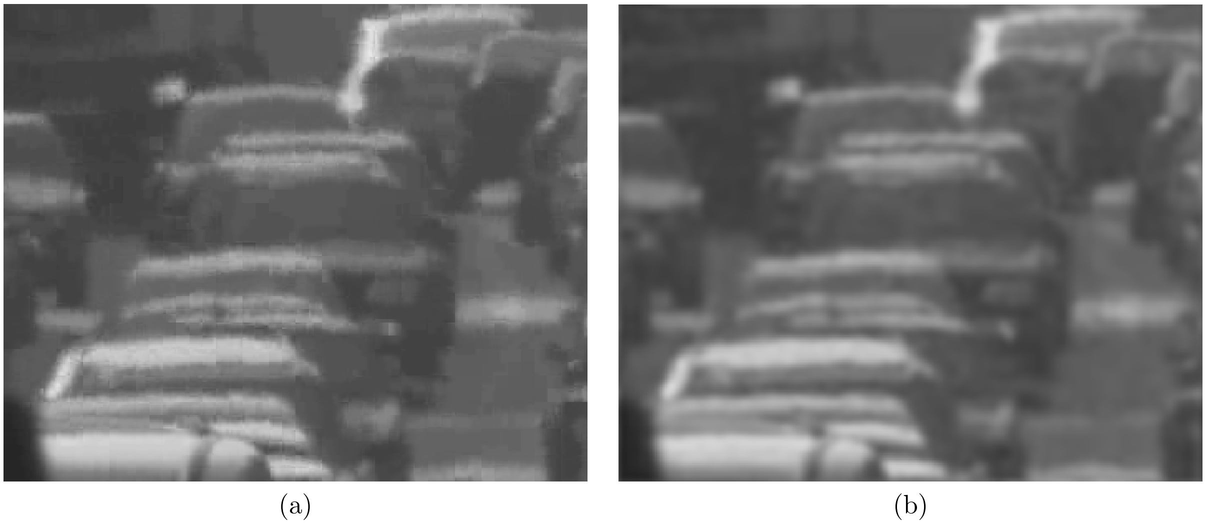


Figure 3. Cropped 13 frame of the “Rush-hour” sequence. (a) high resolution frame of the sequence coded at 2 Mbps (PSNR = 36.06 dB). (b) Super resolved frame from the sequence coded at 1 Mbps (PSNR = 36.56 dB). Similar visual quality is achieved with a 50% reduction of the bitrate.

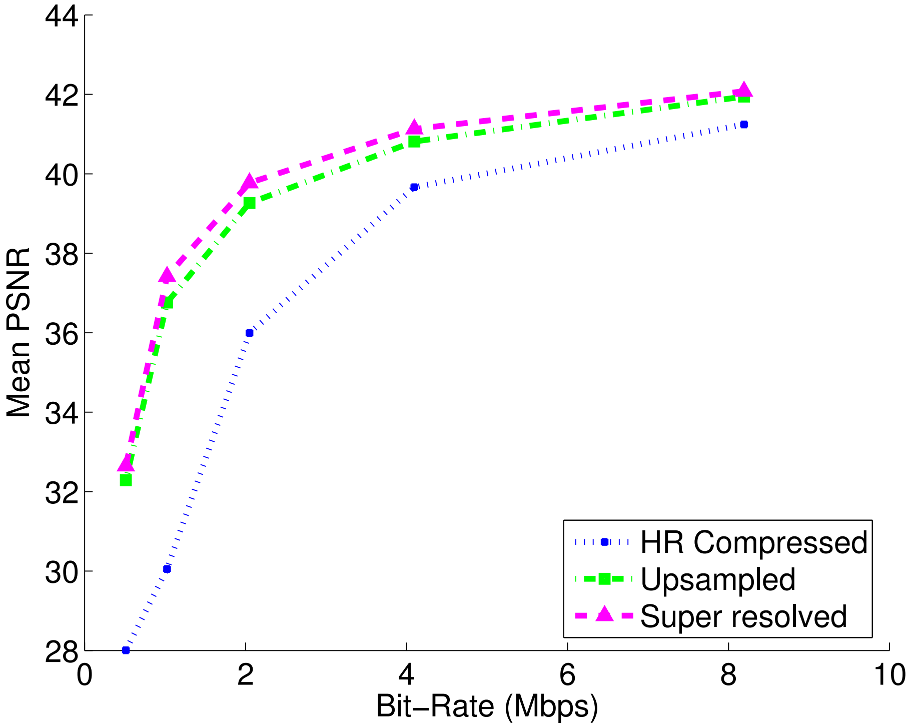
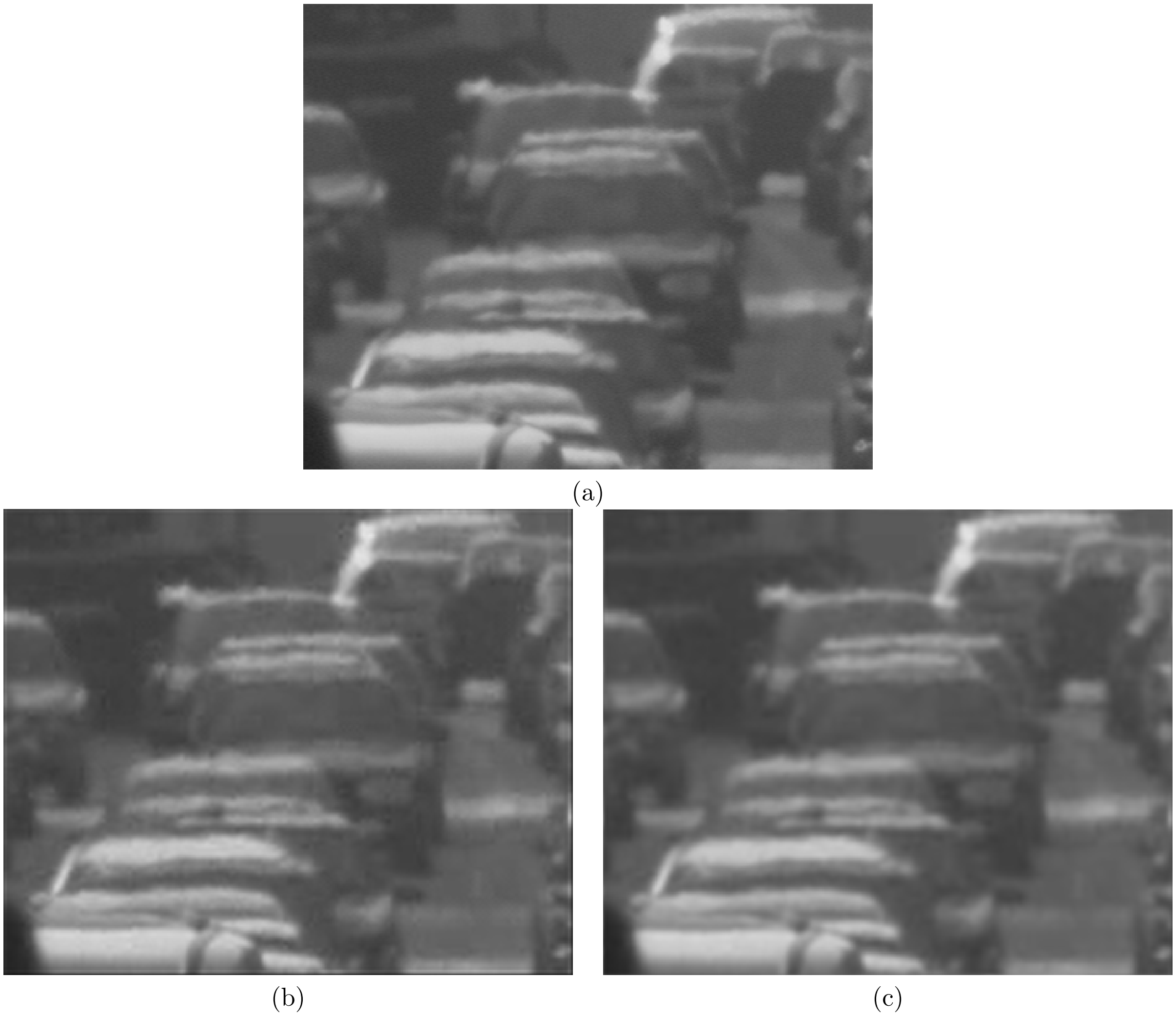


Figure 4. Rate distortion curves for the “Rush-hour” sequence. The super-resolution for video compression approach leads to significant quality improvement over the normative upsampling method at all the considered bitrates.



(a)

(b)

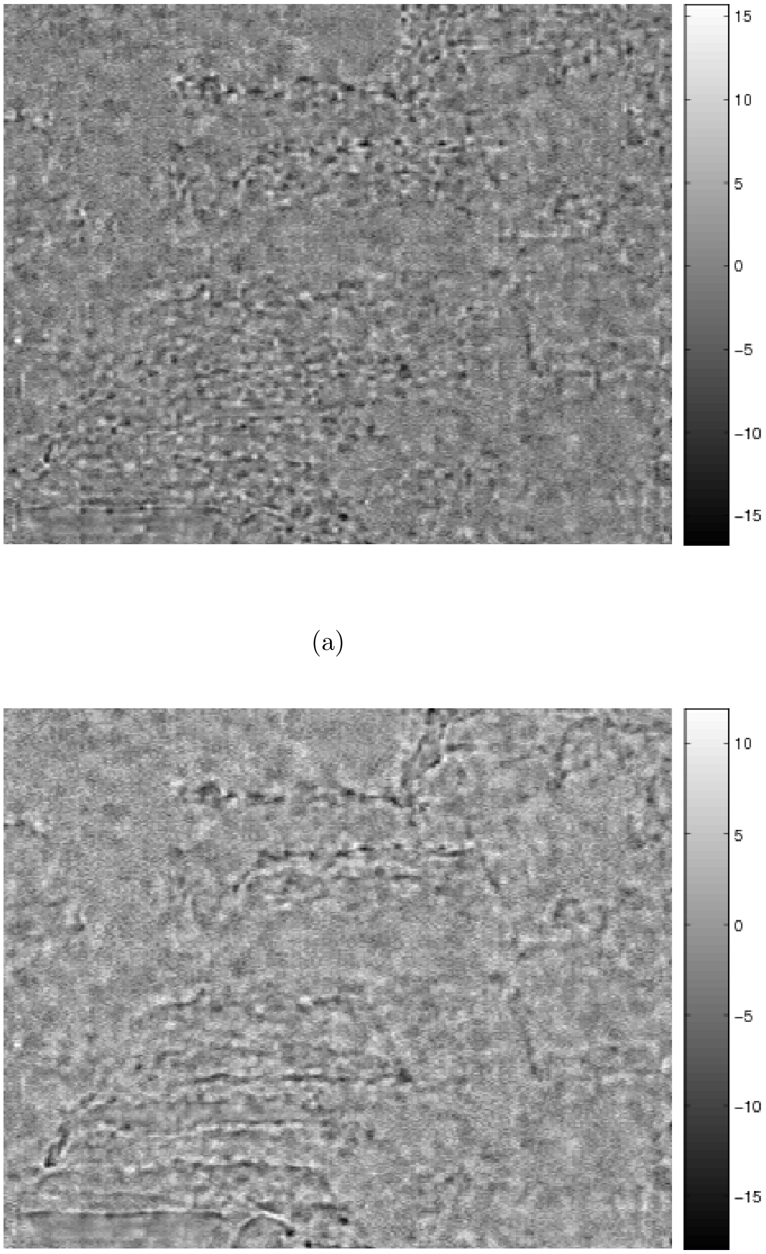
(c)

Figure 5. Frame 12 of the “Rush-hour” sequence. (a) Original frame, (b) Reconstructed frame using the normative upsampling. (c) Reconstructed frame using super-resolution.

high resolution sequence, the proposed method does not improve much the PSNR the obtained by compressing the high resolution sequence directly. This is due to the pre-filtering employed that, although it is widely used to model the integration produced by a low resolution CCD, it produces high textured differences between the image and the predicted frame, making such differences difficult to be compressed. However, when compressing the original sequence the residuals are, in general, smaller and not as textured and, as a consequence, they can be more efficiently compressed.

To conclude this experiment we note that it is possible for a range of bitrates to achieve with the proposed method the same PSNR quality as with the direct compression of the original image, but at about half the bitrate. To illustrate this statement, consider Fig. 3, where a part of frame 13 for the sequence compressed at 2 Mbps and its corresponding reconstructed frame from the down-sampled sequence compressed at 1 Mbps are depicted. Both frames have a similar PSNR while only 50% of the bits are needed to code the frame depicted in Fig. 3(b).

A second experiment is reported on the same video sequence to illustrate the use of the proposed technique in a super-resolution for scalable video compression environment. In this experiment, the high resolution sequence



(a)

(b)

Figure 6. Differences of the original frame 15 of the "Rush-hour" sequence and (a) the reconstructed frame using the normative upsampling depicted in figure 5b, (b) the reconstructed frame using super-resolution depicted in figure 5c.

was downsampled by convolving each frame with the 13-tap separable filter kernel described in Eq. (19) and then, every-other pixel in the horizontal direction and every-other line in the vertical direction were discarded. This low resolution sequence, taken as the base layer for SVC, is then compressed using MPEG-4 at different bitrates.

Reconstruction of the high resolution frame is then performed by two different methods. First, we use the normative upsampling operation defined by the WD described in section 3.3. We also applied the proposed super-resolution method on the low resolution base layer sequence. Results are summarized in Fig. 4 where it is shown that the proposed technique slightly outperforms the normative upsampling method, specially at lower bitrates. Better PSNRs imply smaller differences between the reconstructed and the original frames thus yielding lower bit needs when coding them and, hence, overall bitrate saving when coding the whole sequence. Figure 5 depicts a 352 × 288 part of the original frame 15 and the reconstructions of the base layer sequence compressed at 2 Mbps using the normative upsampling (figure 5b) and the super-resolution proposed method (figure 5c). Although both reconstructed frames are quite similar visually and in terms of PSNR, a study of the differences with the original image (see Figs. 6a and 6b, for the upsampled and super-resolved frames, respectively) shows that the super resolved image has lower difference values and it is less textured, thus making possible the more efficient compression of the residuals.

5. CONCLUSIONS

In this paper we have studied three different cases towards video compression. The first one simply compresses the sequence using any of the standard compression methods, the second one pre-processes (without downsampling) the image sequence before compression, so that post-processing (without upsampling) is performed at the decoder, and the third one includes downsampling in the pre-processing stage and downsampling in the post-processing stage. These three models have led to a new paradigm that analyses the performance of a video compression system. In this paper, we have concentrated on the application of super-resolution techniques as a way to post-process and upsample a compressed video sequences. Examples have been provided on a wide range of bitrates for two very important applications: format conversion between different platforms and scalable video coding.

The obtained results are promising and show the potential enhancements of video compression with the use of super-resolution techniques. The optimal estimation of a number of parameters in a rate distortion sense is currently under investigation.

ACKNOWLEDGMENTS

This work has been supported by the “Comisión Nacional de Ciencia y Tecnología” under contract TIC2003-00880.

REFERENCES

1. G. Bjontegaard and K. O. Lillevo, “H.263 Anchors Technical Description”, MPEG95/0322, Dallas, November 1995
2. S. Borman,“Topics in Multiframe Superresolution Restoration”, Ph.D Thesis, University of Notre Dame, 2004, Notre Dame, IN.
3. A. Bruckstein, M. Elad and R. Kimmel, “Down Scaling for Better Transform Compression”, IEEE Trans. on Image Processing, Vol. 12, No. 9, pp. 1132-44, Sept. 2003.
4. P. V. Karunaratne, C.A. Segall and A.K. Katsaggelos, “Rate Distortion Optimal Video Pre-Processing Algorithm,” IEEE Int. Conf. on Image Processing, Thessaloniki, Greece, October 7-11, 2001.
5. W. Li, J-R Ohm, M. van der Schaar, H. Jiang, S. Li, “Verification Model 18.0 of MPEG-4 Visual”, N3908, Pisa, January 2001.
6. R. Molina, J. Mateos, and A. K. Katsaggelos, “Super Resolution Reconstruction of Multispectral Images” in Virtual observatories: Plate Content Digitization, Archive Mining and Image Sequence processing, edited by Henron Press, Sofia (Bulgary), April 2005.
7. J. Reichel, H. Schwarz and M. Wien, “Scalable Video Coding - Working Draft 3”, JVT-P201, Poznan, PL, 24-29 July, 2005.

8. C. A. Segall,“Framework for the Post-Processing, Super-Resolution and Deblurring of Compressed Video”, Ph.D Thesis, Northwestern University, 2002, Evanston, IL.
9. C. A. Segall, R. Molina, and A. K. Katsaggelos. “High-Resolution Images from Low-Resolution Compressed Video”. IEEE Signal Processing Magazine, vol. 20, 37-48, 2003.
10. C. A. Segall, M. Elad, P. Milanfar, R. Webb and C. Fogg, “Improved High-Definition Video by Encoding at an Intermediate Resolution,” Proceedings of the SPIE Conference on Visual Communications and Image Processing, Jan. 18-22, 2004, San Jose, CA.
11. C. A. Segall, A.K. Katsaggelos, R. Molina, and J. Mateos, “Bayesian Resolution Enhancement of Compressed Video”, IEEE Transactions on Image Processing, vol. 13, no. 7, 898-911, 2004.
12. C. A. Segall, “Study of Upsampling/Down-Sampling for Spatial Scalability”, JVT-Q083, Nice, FR, PL, 14-21 October, 2005.
13. G. Sullivan and S. Sun, “Ad Hoc Report on Spatial Scalability Filters”, JVT-P007, Poznañ, PL, 24-29 July, 2005.