Anantrasirichai, N., & Canagarajah, C. N. (2010). Spatiotemporal super-resolution for low bitrate H.264 video. In IEEE International Conference on Image Processing. (pp. 2809 - 2812). 10.1109/ICIP.2010.5651088

SPATIOTEMPORAL SUPER-RESOLUTION FOR LOW BITRATE H.264 VIDEO

N. Anantrasirichai, C.N. Canagarajah

University of Bristol, UK

ABSTRACT

Super-resolution and frame interpolation enhance low resolution low-framerate videos. Such techniques are especially important for limited bandwidth communications. This paper proposes a novel technique to up-scale videos compressed with H.264 at low bit-rate both in spatial and temporal dimensions. A quantisation noise model is used in the super-resolution estimator, designed for low bitrate video, and a weighting map for decreasing inaccuracy of motion estimation are proposed. Results show improvement both in rate-distortion and perceived image quality.

Index Terms**—** Video enhancement, Interpolation,

1. INTRODUCTION

Telecommunications in rural areas suffer from problems of poor communications. Therefore the solution which is increasingly being turned to is the use of wireless communication systems such as satellite and GSM based technologies. The advantage is clearly that these are cost-effective to set up, but at the price of limiting the bandwidth available.

One of the potential uses for such telecommunication systems is for transmission of video, for applications such as remote medicine, agricultural advice and education. These applications require the video to be of high enough resolution and quality – e.g. attempting to diagnose a medical problem from a harshly compressed QCIF video is not possible. Due to the massively limited bandwidth, existing video solutions (such as H.264) of acceptable quality are not suitable in this scenario. Moreover, many such rural areas of interest are in developing countries, which therefore mean that high resolution cameras are not available, removing the possibility of using scalable video coding and other associated techniques.

Two possible solutions to this problem are super-resolution and frame rate up conversion which both fit around an existing compression system. Super-resolution is a spatial upsampling technique employed at the decoder which attempts to better estimate the high-resolution (from the camera) source video using information from multiple low resolution frames. Frame rate conversion is an equivalent technique which works in the temporal domain: extra frames are interpolated between the low frame rate decoded video in an attempt to smooth the motion.

Super-resolution is an area which is of considerable interest, however the majority of work uses a downsampled version of the high-resolution video source, possibly with added noise – very little work uses source video which has been compressed. Therefore although the work has lots of valid concepts, some of the modelling and assumptions (e.g. that any noise present is Gaussian) are ill-formed.

This paper proposes a novel technique for upsampling low bitrate video compressed using H.264, which uses information which is present in an H.264 bitstream to perform spatiotemporal super-resolution. We also introduce a quantisation noise model used in the super-resolution estimator, particularly for low bitrate video, and a weighting map for decreasing inaccuracy of motion estimation.

The remainder of this paper is organised as follows: section 2 discusses related work in both super-resolution estimation and frame rate conversion, before the proposed methodology is presented in section 3. Results are reported in section 4 and the paper is concluded and further work suggested in section 5.

2. RELATED WORK

As mentioned in the introduction, little work has been carried out on the application of super-resolution to compressed video. However a lot of the background to super-resolution is of interest. Research on super-resolution can be divided into two main streams – spatial and temporal. Section 2.1 introduces related work on the former whilst section 2.2 discusses the latter.

2.1. Spatial Super-Resolution Estimation

Super-resolution (SR) image reconstruction increases the resolution of an image by observing multiple low-resolution (LR) images. In video sequences, successive frames are employed to construct a high-resolution (HR) frame. Although the basic concept of the SR algorithm is simple, there are many problems related to perceptual quality and restriction of available data. The LR images are typically aliased and have sub-pixel shifts between one another, and the different data are possibly located at the same point in the HR image and some points in the HR image do not correspond to any information from the LR images.

The first proposed SR technique was formulated in the frequency domain due to simplicity and cost [1], but exhibits high sensitivity to model errors. Later the spatial domain has been considered but the early approaches are generally complicated. Consequently maximum likelihood estimation has been introduced with proper prior information and the assumption of additive Gaussian noise [2]. For the reconstruction process, Bayesian interpretation and the Tikhonov regularisation is employed as statistical estimation problem. The fundamental SR techniques have been reviewed in [3]. The newest technique developed in [7] introduces a spatiotemporal video super-resolution using a multidimensional kernel regression.

Most existing SR techniques have been proposed for uncompressed images. Unfortunately, super-resolution algorithms designed for uncompressed data do not perform well when directly applied to decompressed image sequences, especially for high compression rates. The reason is that the quantization error introduced during the compression process is often the dominant source of error when the compression rate is high and this error is not modelled. In [4], a method for SR MPEG video is introduced using quantisation information embedded in the compressed video stream. A Bayesian SR reconstruction technique is used to model compression and exploit the quantization step size information for reconstruction [5]. This technique shows promising results for high bitrate; however, the image quality is insignificantly improved at low bitrate (256kbps). The method applied to compressed video has been further developed by improving the registration precision and the prior model [6].

2.2. Frame Rate Up-Conversion

Decreasing frame rate generally saves bits transmitted/stored but produces artefacts of temporal visual degradation. Frame rate up-conversion (FRUC) is then required at the receiver.

Motion compensated frame interpolation (MCFI) is the most common technique for interpolating frame as it provided simplicity with desirable performance. The concept uses the motion vectors estimating motion between the existing frames, normally forward and backward frames, to construct the interval frame. The interpolated frame located between these two frames consequently uses motion directions with fractional values depending on temporal distance between the interpolated frame and two reference frames. This simple approach cannot deal with occlusions where holes are present in the interpolated frame. Authors in [8] proposed the bi-directional motion estimation by dividing the interpolated frame into non-overlapping blocks. Then the motion vector for each block is obtained by positioning the forward and backward reference frames in opposite directions. Later, they proposed a variable block size to achieve better performance [9].

3. PROPOSED SCHEME

The proposed scheme enhances a decoded LR low-framerate video by applying SR method to construct a HR low-framerate video, whilst applying a FRUC method to construct a LR high-framerate video in parallel. Afterward the FRUC method enhances the HR low-framerate video using the prepared LR high-framerate video to generate a HR high-framerate video.

We apply spatial super-resolution before frame interpolation (SR → FI) since it processes faster than the alternative order (FI → SR) as the more complex SR is working with fewer frames. Generally applying FI → SR yields better quality compared to SR → FI because the iterations of the SR process (will be explained in 3.1) is able to correct some of the errors produced from the FI process. However, in our scheme the FI for the HR video exploits the interpolated frame of the LR video as a guide thereby achieving comparable quality (as shown in the result section).

3.1. Spatial super-resolution

The decoded low-resolution frame $\mathbf{y}_k$ (formed by lexicographical ordering) is reconstructed using motion vectors $\mathbf{v}_{k,i}$ and reference frames $\mathbf{y}_i$ as shown in equation 1 [5].

$$\mathbf{y}_k = \mathbf{T}^{-1}Q\left[\mathbf{T}\left(\mathbf{g}_k - \sum_{\forall i} C(\mathbf{v}_{k,i})\mathbf{y}_i\right)\right] + \sum_{\forall i} C(\mathbf{v}_{k,i})\mathbf{y}_i \qquad (1)$$

where $C(\cdot)$ and $Q[\cdot]$ represent the prediction and quantisation processes, respectively, and $\mathbf{T}$ and $\mathbf{T}^{-1}$ represent the forward and inverse transform matrix, respectively. The LR image $\mathbf{g}_k$ relates to its HR version $\mathbf{f}_k$ as $\mathbf{g}_k = \mathbf{AHf}_k + \mathbf{n}_k$, where $\mathbf{H}$ is a blur filter, $\mathbf{A}$ is a decimation operation and $\mathbf{n}_k$ is acquisition noise. The relationship between $\mathbf{y}_k$ and $\mathbf{f}_k$ is irreversible. The estimator is therefore employed to estimate $\mathbf{f}_k$. Here we assume that the quantisation noise is dominant and the compression parameters can be extracted from bitstream. A Bayesian approach is consequently used as a posteriori probability density function (PDF) of $\mathbf{f}_k$ can be established. The estimated HR frame $\hat{\mathbf{f}}_k$ is then found from a maximum a-posteriori (MAP) estimator.

$$\hat{\mathbf{f}}_k = \arg\max_{\mathbf{f}_k}\left\{\log\prod_l p(\mathbf{y}_l|\mathbf{f}_k) + \log p(\mathbf{f}_k)\right\} \qquad (2)$$

$$\begin{aligned} p(\mathbf{y}_l|\mathbf{f}_k) \propto \exp\Big\{ -\frac{1}{2}(\mathbf{y}_l - \mathbf{AH}C(\mathbf{d}_{l,k})\mathbf{f}_k)^T \\ \times \mathbf{K}_l^{-1}(\mathbf{y}_l - \mathbf{AH}C(\mathbf{d}_{l,k})\mathbf{f}_k)\Big\} \end{aligned} \qquad (3)$$

$$p(\mathbf{f}_k) \propto \exp\{-\lambda_1 \|\nabla\mathbf{f}_k\|\} \qquad (4)$$

where $\mathbf{K}_l^{-1}$ is an inverse covariance matrix and $\mathbf{d}_{l,k}$ is a displacement in the HR frames corresponding to the motion vectors $\mathbf{v}_{l,k}$ in equation 1. Ringing artefacts introduced by the coarse quantisation is used for modelling the first density function $p(\mathbf{f}_k)$. Note that H.264 has deblocking filter, so the blocking artefact is not included in $p(\mathbf{f}_k)$. The $\nabla\mathbf{f}_k$ can be seen as the regularization term defined as $\nabla\mathbf{f}_k = (f_{xx}f_y^2 - 2f_xf_yf_{xy} + f_{yy}f_y^2)/(f_x^2 + f_y^2)$ [6], where λ_1 is a weighting constant (Here $\lambda_1 = 0.1$). Then equation 2 can be rewritten as

$$\begin{aligned} \hat{\mathbf{f}}_k = \arg\min_{\mathbf{f}_k}\Big\{ \sum_{l=k-TB}^{k+TF}(\mathbf{y}_l - \mathbf{AH}C(\mathbf{d}_{l,k})\mathbf{f}_k)^T \\ \times \mathbf{K}_l^{-1}(\mathbf{y}_l - \mathbf{AH}C(\mathbf{d}_{l,k})\mathbf{f}_k) + \lambda_1\|\nabla\mathbf{f}_k\|\Big\} \end{aligned} \qquad (5)$$

In the following sections the proposed quantisation noise model for low bitrate video is described before the technique for using the motion estimation error to decide how much the displacement can be trusted is outlined.

3.1.1. *Quantisation Noise Model*

The previous work on spatial super-resolution for H.264 compressed video has been primarily concerned with high bit-rate scenarios and as such has modelled the quantisation noise as uniform for each frequency. This assumption does not hold for low bit-rate video, where the quantisation noise is distributed in a Laplacian fashion. The variance of these distributions varies across different frequency bands, with higher frequencies yielding lower variance.

The covariance in frequency domain K_T for each 4×4 block transform is found here experimentally by looking at five standard test sequences (Foreman, Mobile, Akiyo, Bus & Carphone), and is shown in equation 6.

$$K_T = \begin{bmatrix} 4 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \frac{q^2}{100} + \begin{bmatrix} 21 & 16 & 21 & 7 \\ 16 & 21 & 14 & 5 \\ 21 & 14 & 5 & 2 \\ 7 & 5 & 2 & 0.4 \end{bmatrix} \frac{q}{10} + \begin{bmatrix} -16.1 & -10.4 & -16.1 & 1.1 \\ -10.4 & -16.1 & -7.2 & 3.2 \\ -16.1 & -7.2 & 3.2 & 6.1 \\ 1.1 & 3.2 & 6.1 & 3.5 \end{bmatrix} \tag{6}$$

where q is a quantisation step size embeded in the decoded bitstream. Then the covariance matrix in equation 3 will be $\mathbf{K}_l = \mathbf{T}^{-1}\mathbf{K}_T\mathbf{T}^{-1^T}$, where $\mathbf{K}_T$ is a diagonal covariance matrix of repeat repositioned K_T associating to $\mathbf{y}_l$.

3.1.2. *Motion Estimation Error*

The prediction process for the displacement $C(\mathbf{d}_{l,k})$, shown in equation 5, employs block-based matching with initial motion vectors extracted from H.264 video. That is, $\mathbf{d}_{l,k} = m \cdot \mathbf{v}_{l,k}$, where m is the proportion between the size of the HR frame and the LR frame. Subsequently refining displacement is operated in a small search window.

To reduce the error incurred for instances of displacement of homogeneous blocks, the grid of blocks is offset on even iterations by half a block.

The macroblock mode and decoded residual from the H.264 bitstream are used to imply the areas of $\mathbf{y}_l$ which are difficult to match with areas in the reference frames. That is, the areas with high energy ϵ_l^r decoded residual, or encoded with intra mode are likely to be occlusions. Similarly the areas with high energy of the compensated error $\epsilon_l^c = \mathbf{f}_l - C(\mathbf{d}_{l,k})\mathbf{f}_k$ could be occlusions.

ϵ_l^r and ϵ_l^c are employed to create the weighting maps to be used in equation 5 as follows:

$$W_l^x = 1 - \frac{\lambda_2 \cdot (\epsilon_l^x)^2}{1 + \lambda_2 \cdot (\epsilon_l^x)^2}, \; x \in \{r, c\} \tag{7}$$

where λ_2 is a decreasing rate. Here λ_2 is chosen as $\frac{1}{100}$ so W_l^x is defined as 0.5 when the intensity error is 10. Using the steepest descent algorithm, the minimisation of equation 5 can be found as

$$\mathbf{f}_k^{n+1} = \mathbf{f}_k^n - \alpha \left[\sum_{l=k-TB}^{k+TF} W_l^c C(\mathbf{d}_{k,l}) \mathbf{H}^T \mathbf{A}^T \times \mathbf{K}_l^{-1} W_l^r \left(\mathbf{y}_l - \mathbf{AH}C(\mathbf{d}_{l,k}) \mathbf{f}_k^n \right) \right] + \lambda_1 \left\| \nabla \mathbf{f}_k \right\| \tag{8}$$

where $\mathbf{f}_k^{n+1}$ and $\mathbf{f}_k^n$ are the estimated HR frame at the $(n+1)$th and nth interations, α is a relaxation parameter controlling rate of convergence (Here $\alpha = 1$).

The same maps are used to generate the initial value, $\mathbf{f}_k^0$, i.e. using the weighted average of the adjacent frames as shown in equation 9 and the W_k^c is 1. Note that if $\hat{\mathbf{f}}_l$ does not yet exist, bilinear interpolation is employed.

$$\mathbf{f}_k^0 = \frac{\sum_{l=k-TB}^{k+TF} W_l^c C(\mathbf{d}_{l,k}) \hat{\mathbf{f}}_l}{\sum_{i=k-TB}^{k+TF} W_i^c} \tag{9}$$

3.2. Frame Rate Up-Conversion

H.264 is a motion compensated codec, which means there is temporal information available within the bitstream which can be used in the motion estimation process. In the proposed method the motion vectors and macroblock modes present in the H.264 bitstream are used as the initial parameters for bilateral motion estimation.

A frame to be interpolated is first divided into blocks each of which in turn is estimated as the average of blocks of pixels in the neighbouring frames y_F, y_B which produce the minimum weighted sum of absolute differences, calculated as shown in equation 10 and 11:

$$w_d = \frac{\lambda_3 \left| \mathbf{d} \right|^2}{1 + \lambda_3 \left| \mathbf{d} \right|^2} \tag{10}$$

$$SAD_d = w_d \sum_{(i,j) \in Bl} \Big| y_F(i + n_F v_i + d_i, j + n_F v_j + d_j) - y_B(i - n_B v_i - \frac{n_B}{n_F} d_i, j + n_B v_j - \frac{n_B}{n_F} d_j) \Big| \tag{11}$$

where the weight w_d is used to promote motion vectors which are close to the initial one. This prevents a distant block with homogeneous detail from being selected as a match between frames. $\mathbf{d} = (d_i, d_j)$ is a distance from the initial motion vectors v_i, v_j extracted from H.264 bitstream, n_F and n_B are the temporal distances to y_F and y_B, respectively. λ_3 is a scaling parameter. Here λ_3 is chosen as $\frac{1}{25}$ so w is defined as 0.5 when the distance is 5 pixels from the initial motion vector. Afterward an adaptively weighted vector-median filter is employed to suppress outliers [9].

The frame rate of the LR video is enhanced using proposed method explained above. The interpolated frames are then upsampled to high resolution using bilinear interpolation. Subsequently as the current frame the motion compensation is applied using the neighbouring HR frames pre-generated via SR method as references. Finally the motion is smoothed in the same way applied to LR frames.

4. RESULTS

The proposed scheme was tested with three standard test sequences: *Foreman*, *Mobile*, *Carphone*. The LR video sequences (QCIF format) were coded with H.264 at 15 frames per second (fps) and were enhanced to CIF format at 30 fps. The results were compared with the full resolution videos coded with H.264 and the existing algorithms introduced in [5] and [7].

Figures 1 and 2 show the results of the proposed technique for three different sequences compared to the other algorithms discussed. It can be seen that at low bitrate (50kbps) the proposed scheme exhibits a 30kbps (37.5%) saving over the full resolution version compressed with H.264 whilst maintaining the same PSNR. The proposed scheme also outperforms the scheme proposed in [5], which employs a uniform distribution for modelling quantisation noise, by upto approximately 0.5 dB consistently across all bitrates. The algorithm proposed in [7] is included in the experiment in order to demonstrate the fact that algorithms which obtain promising results for uncompressed videos don't necessarily perform well with heavily compressed data.

For *Mobile* sequence, the proposed scheme achieves less impressive improvement in terms of PSNR over other methods compared the the other sequences (figures 1 & 2). This is due to the high level of detail it contains, which is much more senstive to the the downsampling stage present at the encoder and hence makes recovery far more difficult.

Note that using PSNR to evaluate the performace of FI is possibly not suitable because the reason interval frames are inserted is to produce temporal smoothness, to increase the subjective quality. The algorithm isn't attempting to estimate the original frames so much as it is attempting to make the video less irritating for the observer.

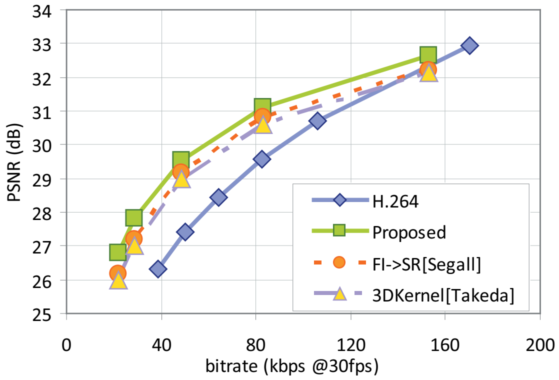

Fig. 1. Rate-distortion performance of *Foreman*

5. CONCLUSIONS & FUTURE WORK

This paper presents a novel technique for upsampling low bitrate video compressed using H.264. The coding parameters embeded in the bitstreams are used to up-scaling both in spatial and temporal dimensions. A quantisation noise model used in the SR estimator for low bitrate video is proposed. Also the accuracy of the motion estimation is weighted to avoid error accumulation produced in the iterative process. The proposed scheme achieves a bitrate reduction compared to transmitting HR video and gains an improvement of image quality compared to the existing schemes. The more accurate quantisation noise model will be considered using the details of each video in the future.

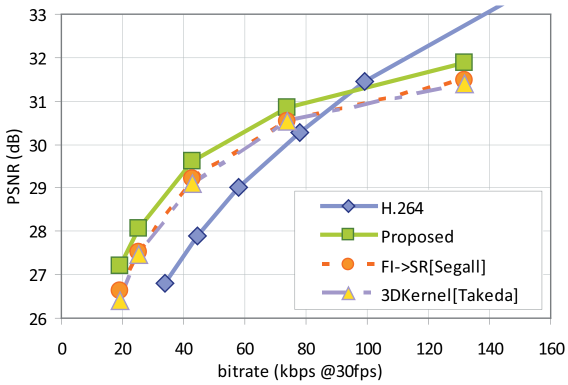

Fig. 2. Rate-distortion performance of *Carphone*

6. REFERENCES

[1] T Huang and R Tsai, "Multi-frame image restoration and registration," *Adv. Comput. Vis. Image Process*, vol. 1, pp. 317339, 1984.

[2] M Elad and Y Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur," *IEEE Trans. Image Processing*, vol. 10, no. 8, pp. 1187–1193, Aug 2001.

[3] S. C Park, M. K Park, and M. G Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[4] Y Altunbasak, A Patti, and R Mersereau, "Super-resolution still and video reconstruction from mpeg-coded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 4, pp. 217–226, Apr 2002.

[5] C Segall, A Katsaggelos, R Molina, and J Mateos, "Bayesian resolution enhancement of compressed video," *IEEE Trans. Image Processing*, vol. 13, no. 7, pp. 898–911, July 2004.

[6] C Wang, P Xue, and W Lin, "Improved super-resolution reconstruction from video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1411–1422, Nov. 2006.

[7] H Takeda, P Milanfar, M Protter, and M Elad, "Superresolution without explicit subpixel motion estimation," *IEEE Trans. Image Processing*, vol. 18, no. 9, Sep 2009.

[8] B.-T Choi, S.-H Lee, and S.-J Ko, "New frame rate up-conversion using bi-directional motion estimation," *IEEE Trans. Consumer Electron.*, vol. 46, no. 3, pp. 603–609, Aug 2000.

[9] B.-D Choi, J.-W Han, C.-S Kim, and S.-J Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 407–416, April 2007.