14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPATIO-TEMPORAL SELECTIVE EXTRAPOLATION FOR 3-D SIGNALS APPLIED TO CONCEALMENT IN VIDEO COMMUNICATIONS

Katrin Meisinger, Sandra Martin, and André Kaup

Chair of Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg, Cauerstr. 7, 91058 Erlangen, Germany
{meisinger,kaup}@LNT.de

ABSTRACT

In this paper we derive a frequency selective extrapolation method for three-dimensional signals. Extending a signal beyond a limited number of known samples is commonly referred to as signal extrapolation. We provide an extrapolation technique which enables to estimate image areas by exploiting simultaneously spatial and temporal correlations of the video signal. Lost areas caused by transmission errors are concealed by extrapolation from the surrounding. The missing areas in the video sequence are estimated conventionally from either the spatial or temporal surrounding. Our approach approximates the known signal by a weighted linear combination of 3-D basis functions from spatial as well as temporal direction and extrapolates it into the missing area. The algorithm is able to extrapolate smooth and structured areas and to inherently compensate motion and changes in luminance from frame to frame.

1. INTRODUCTION

Estimating image areas from the surrounding image or video signal is an important topic of various applications in image and video communications. The surrounding consists of spatial information, i.e. within the image, or temporal data, i.e. previous or following frames.

Extending a signal beyond a limited number of known samples is commonly referred to as signal extrapolation. For example, the problem of concealing corrupted video data caused by transmission errors in mobile video communications can be seen as an extrapolation of the surrounding available video signal into the missing area. In hybrid video coding, prediction of the video signal is applied in order to increase coding efficiency. This step can also be interpreted as extrapolation of the known signal in order to predict the following pixels.

Commonly, the unknown signal areas are either predicted spatially or temporally. In case of spatial prediction the block is predicted from surrounding data within the image. Motion compensated prediction exploits the similarity of subsequent frames. For block based techniques, the displacement for each block of the image from one frame to the next one is described by a motion vector. In [1] an overview of different spatial as well as temporal error concealment methods is given.

The Boundary Matching Algorithm (BMA) [2] takes advantage of temporal information in order to conceal lost blocks. The error across block boundaries for the block compensated with different motion vectors and its neighbouring correctly received blocks is measured. The zero motion vector, the vector of the block in the previous frame, the vectors of the neighbouring blocks, the median and the average of the neighbouring vectors are tested. The vector which results in a minimum boundary error is selected. The Extended Boundary Matching Algorithm (EBMA) is applied, if also the prediction error signal is lost. Additionally to the motion vectors the prediction error signals of the neighbouring blocks and an assumed zero prediction error are used. The combination of prediction error and motion vector is selected which is minimising the boundary error.

The temporal concealment method of the reference software of the most recent video coding standard H.264/AVC [3] is based on the BMA algorithm. As test vectors all motion vectors of the correctly received neighbouring blocks are used. A MB can be split into several blocks down to a size of 4×4 which is rate-distortion optimised. Thus, multiple vectors per MB can be transmitted. Further, five reference frames can be used for prediction. However, the prediction error signal is lost and the surrounding prediction error signals are not used for the motion vector selection.

In [4], missing image areas are estimated from the surrounding with help of 2-D frequency selective extrapolation applied to concealment problems in image communications. In [5] results are shown for uncoded data comparing the performance of different algorithms known from literature (amongst others [6], [7]). Our algorithm showed superior results due to its ability to extrapolate smooth areas, as well as edges and noise-like areas. This ability is further used to remove TV logos [8].

However, conventional techniques use *either* spatial *or* temporal information. Encouraged by the excellent extrapolation properties for two dimensions, we developed the 3-D algorithm. We introduce a mathematical description of the video signal in *spatial and temporal* direction *at the same time*. Therefore, we provide an extrapolation technique which enables to estimate image areas by exploiting *simultaneously* spatial *and* temporal correlations of the video signal. We apply this principle to concealment in video communications.

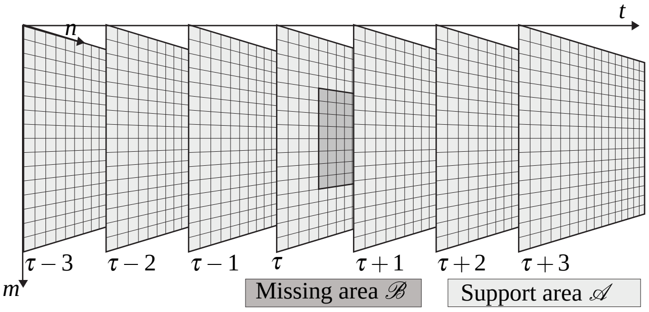

Figure 1: Image areas used for 3-D extrapolation consisting of the area to be estimated and its known surrounding.

2. SPATIO-TEMPORAL EXTRAPOLATION

The method of spatio-temporal extrapolation is applied to three-dimensional signals in order to gain additional information from previous and following frames about the image content to be estimated. In contrary to the two-dimensional approach [4], we do not only exploit the surrounding image content within the image of the signal to be predicted but also the image content of preceding and/or proceeding frames.

Fig. 1 shows a possible sequence of seven frames where the spatial dimensions are depicted by m,n and the temporal dimension by t. The block $\mathscr{B}$ shaded gray in image τ is to be extrapolated from the area $\mathscr{A}$. The support area $\mathscr{A}$ ranges from the three previous to the three subsequent frames including the surrounding of the area

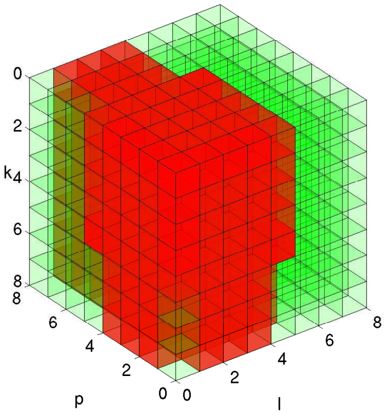

Figure 2: Red: Search area for 3-D-DFT basis functions. Green: conjugate complex values.

to be estimated in the actual frame. Only the respective areas which are used for the reconstruction of the missing area are shown. The entire region   - consisting of the region to be estimated $\mathscr{B}$ and the support area $\mathscr{A}$ - is described by a volume.

The known pixels $f[m,n,t]$ are approximated by the parametric model $g[m,n,t]$. The spatio-temporal description by $g[m,n,t]$ approximates the support area by a linear combination of basis functions $\varphi_{k,l,p}[m,n,t]$ weighted by expansion coefficients $c_{k,l,p}$

$$g[m,n,t] \quad = \sum_{(k,l,p)\in\mathscr{K}} c_{k,l,p}\cdot\varphi_{k,l,p}[m,n,t]. \tag{1}$$

The set   describes the basis functions used. The basis functions are defined in the entire area   and its number of basis functions $M\times N\times T$ equals the number of pixels in . Here, the principle is described for real valued basis functions and expansion coefficients.

In order to determine the expansion coefficients, the weighted error energy between the original signal and its approximation by the parametric model is evaluated with respect to the support area

$$E_{\mathscr{A}} = \sum_{(m,n,t)\in\mathscr{A}} w[m,n,t]\cdot\left(f[m,n,t]-g[m,n,t]\right)^2, \tag{2}$$

where the weighting function $w[m,n,t]$ has only positive amplitudes in the support area and is zero elsewhere

$$w[m,n,t] = \begin{cases} \rho[m,n,t], & (m,n,t)\in\mathscr{A} \\ 0, & (m,n,t)\in\mathscr{B} \end{cases} \tag{3}$$

Generally, $w[m,n,t]$ allows to emphasise pixels which are more important for the extrapolation and is specified for the application of concealment in Sec. 4.

The weighted error criterion is minimised by taking the derivative with respect to the unknown coefficients and setting it to zero. We approximate the known area successively because the considered problem is underdetermined. The coefficients are obtained by an iterative algorithm. The approximation in iteration v is given by

$$g^{(v)}[m,n,t] \quad = \sum_{(k,l,p)\in\mathscr{K}_v} c^{(v)}_{k,l,p}\cdot\varphi_{k,l,p}[m,n,t]. \tag{4}$$

The set $_v$ consists of all basis functions used for the approximation so far. With help of the window function $b[m,n,t]$

$$b[m,n,t] = \begin{cases} 1, & (m,n,t)\in\mathscr{A} \\ 0, & (m,n,t)\in\mathscr{B} \end{cases} \tag{5}$$

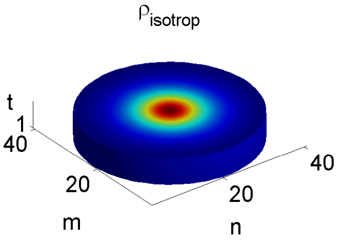

Figure 3: Volume of a 3-D isotropic function for 7 images ($\hat{\rho}$=0.8).

the residual error signal in the support area is calculated in this step

$$r^{(v)}[m,n,t] = b[m,n,t]\cdot\left(f[m,n,t]-g^{(v)}[m,n,t]\right). \tag{6}$$

Assuming an appropriate basis function $\varphi_{u,v,w}[m,n,t]$ is already selected, the residual error signal is further approximated by

$$r^{(v+1)}[m,n,t] \quad = \quad r^{(v)}[m,n,t] - b[m,n,t]\left(\Delta c\cdot\varphi_{u,v,w}[m,n,t]\right) \tag{7}$$

Per iteration, we choose that basis function $\varphi_{u,v,w}[m,n,t]$ which leads to a maximum reduction of the residual error criterion

$$\Delta E^{(v)}_{\mathscr{A}} = \sum_{(m,n,t)\in\mathscr{A}} w[m,n,t]\left(\Delta c\cdot\varphi_{u,v,w}[m,n,t]\right)^2 \tag{8}$$

$$= \quad (u,v,w) = \arg\max_{(k,l)} \Delta E^{(v)}_{\mathscr{A}}. \tag{9}$$

Then, the respective coefficient is computed by minimising (2)

$$\Delta c = \frac{\sum\limits_{(m,n,t)\in\mathscr{L}} w[m,n,t]r^{(v)}[m,n,t]\ \varphi_{u,v,w}[m,n,t]}{\sum\limits_{(m,n,t)\in\mathscr{L}} w[m,n,t]\ \varphi_{u,v,w}[m,n,t]\varphi_{u,v,w}[m,n,t]} \tag{10}$$

and subsequently updated

$$c^{(v+1)}_{u,v,w} \quad = \quad c^{(v+1)}_{u,v,w} + \Delta c. \tag{11}$$

The index of the selected basis function is included in the set of used basis functions

$$_{v+1} = {}_{v}\cup\ u,v,w \qquad if \quad u,v,w \notin {}_{v}. \tag{12}$$

The algorithm terminates if the reduction of the residual error energy drops below a pre-specified threshold.

Summarising the main points of the algorithm, the image content in the spatio-temporal volume is described *simultaneously* in *spatio-temporal* direction by dominant features in terms of weighted basis functions.

The basis functions are defined in the entire volume, therefore each approximation provides at the same time an estimation of the missing samples. Finally, the extrapolated area is cut out of the parametric model.

3. 3-D FREQUENCY SELECTIVE EXTRAPOLATION

We use 3-D DFT basis functions for the approximation since they are especially suited to extrapolate monotonous areas, edges and noise-like areas

$$\varphi_{k,l,p}[m,n,t] = e^{j\frac{2\pi}{M}mk}\cdot e^{j\frac{2\pi}{N}nl}\cdot e^{j\frac{2\pi}{T}tp}. \tag{13}$$

For real valued video signals the expansion coefficients or DFT coefficients, respectively, fulfil the following conjugate complex symmetry

$$c^{(v)}_{M-k,N-l,T-t} \quad = \quad c^{(v)*}_{k,l,t} \text{ as well as} \tag{14}$$

$$\varphi_{M-k,N-l,T-t}[m,n,t] \quad = \quad \varphi^{*}_{k,l,t}[m,n,t]. \tag{15}$$

Sequence Param. set	Flower A	Flower B	Crew A	Crew B
$\Delta E_{\text{min}} = 0.1$	27.94 dB 177.80 It.	26.03 dB 177.24 It.	32.40 dB 147.40 It.	30.82 dB 148.72 It.
$\Delta E_{\text{min}} = 1.0$	27.35 dB 116.68 It.	25.48 dB 103.96 It.	31.50 dB 30.0 It.	30.60 dB 30.76 It.
$\Delta E_{\text{min}} = 2.0$	26.57 dB 73.8 It.	24.84 dB 66.48 It.	30.98 dB 17.44 It.	30.19 dB 17.40 It.

Table 1: PSNR results and average number of iterations per block for block losses. Parameter set A: 2 previous and 2 subsequent frames. Parameter set B: 2 previous frames.

Taking the symmetry properties into account, we can rewrite the parametric model to

$$g^{(v)}[m,n,t] = \frac{1}{2MNT} \sum_{(k,l,p)\in\mathscr{K}_v} \left(c^{(v)}_{k,l,p} \cdot \varphi_{k,l,p}[m,n,t] + c^{(v)}_{M-k,N-l,T-p} \cdot \varphi_{M-k,N-l,T-p}[m,n,t]\right) \quad (16)$$

Using 3-D DFT basis functions allows us to express the computationally expensive equations as (10 in the frequency domain. This enables an efficient implementation of the extrapolation algorithm. The multiplication of the weighting function with the complex exponential $\varphi_{u,v,w}[m,n,t]$ is equivalent to a shift of its DFT by u,v,w

$$\sum_{(m,n,t)\in\mathscr{A}} w[m,n,t]\varphi_{k,l,p}[m,n,t]\varphi_{u,v,w}[m,n,t] = W^*[k+u,l+v,p+w]$$

Hence, the update equation for Δc (10) can be expressed with help of $r^{(v)}_w[m,n,t] = w[m,n,t]\cdot r^{(v)}[m,n,t]$ by

$$\Delta c = \begin{cases} MNT \cdot \frac{R^{(v)}_W[u,v,w]}{W[0,0,0]}, & \text{if } (u,v,w) \in \\ 2MNT \cdot \frac{R^{(v)}_W[u,v,w]\cdot W[0,0,0] - R^{(v)*}_W[u,v,w]\cdot W[2u,2v,2w]}{W[0,0,0]^2 - |W[2u,2v,2w]|^2}, & \text{else} \end{cases} \quad (17)$$

with a real-valued spectrum for the set consisting of the discrete frequencies $(0,0,0)$, $(\frac{M}{2},0,0)$, $(0,\frac{N}{2},0)$, $(0,0,\frac{T}{2})$, $(\frac{M}{2}\frac{N}{2},0)$, $(\frac{M}{2},0,\frac{T}{2})$, $(0,\frac{N}{2},\frac{T}{2})$ and $(\frac{M}{2},\frac{N}{2},\frac{T}{2})$ resulting from symmetry requirements (14),(15) and the definition of $g^{(v)}[m,n,t]$. For Δc^* we obtain a conjugate complex equation.

We select that basis function which is maximising

$$\Delta E^{(v)}_{\mathscr{A}} = \frac{1}{2M^2N^2T^2} \cdot \left(|\Delta c|^2 \cdot W[0,0,0] + \Re\left\{\Delta c^2 \cdot W^*[2u,2v,2w]\right\}\right) \quad (18)$$

Due to the symmetry properties of the DFT the search area is limited to approximately half of the volume as illustrated in red in Fig. 2. This holds also for the number of expansion coefficients which have to be updated.

The residual error signal can then be expressed by

$$R^{(v+1)}_W[k,l,p] = R^{(v)}_W[k,l,p] - \quad (19)$$

$$\frac{1}{2MNT} \cdot \left(\Delta c \cdot W[k-u,l-v,p-w] + \Delta c^* \cdot W[k+u,l+v,p+w]\right)$$

Finally, the parametric model is obtained by an inverse DFT

$$g[m,n,t] = IDFT_{M,N,T} \quad G[k,l,p] \quad (20)$$

and the missing block cut out. Since all equations are expressed in the frequency domain, there is only one DFT transform required in the beginning and an inverse DFT in the end.

	Flower	Foreman	Table Tennis	Crew
TR	17.79 dB	28.77 dB	19.75 dB	19.55 dB
BMA	28.99 dB	36.40 dB	25.19 dB	26.46 dB
EBMA	22.95 dB	32.13 dB	22.25 dB	24.44 dB
2-D	17.04 dB	27.43 dB	19.94 dB	27.62 dB
3-D	27.94 dB	37.06 dB	30.07 dB	32.40 dB

	Flower	Foreman	Table Tennis	Crew
TR	17.99 dB	28.64 dB	20.40 dB	19.25 dB
BMA	26.16 dB	33.22 dB	25.16 dB	25.89 dB
EBMA	25.22 dB	33.46 dB	24.06 dB	24.02 dB
2-D	16.12 dB	22.47 dB	19.25 dB	26.15 dB
3-D	28.60 dB	36.11 dB	30.56 dB	32.09 dB

Table 2: Comparison of different concealment techniques. Top: Isolated block losses. Bottom: Consecutive block losses.

4. WEIGHTING OF THE ERROR CRITERION

As already mentioned, the weighting function (3) allows to emphasise important regions for the extrapolation by weighting the error criterion in (2). We incorporate a model taking pixels closer to the missing area more into account as more distant pixels. Hence, we choose a radial symmetric decaying isotropic 3-D model

$$\rho_{isotrop}[m,n,t] = \rho^{\sqrt{\left(m-\frac{M-1}{2}\right)^2+\left(n-\frac{N-1}{2}\right)^2+\left(t-\frac{T-1}{2}\right)^2}}, \quad \hat{\rho} < 1 \quad (21)$$

where the loci of constant correlation are globes as depicted for a cut-out of 7 frames in Fig. 3.

5. RESULTS FOR CONCEALMENT

The application of 3-D extrapolation to concealment of MB block losses is investigated for uncoded YUV-sequences. The PSNR is evaluated only for the lost areas. First of all, the impact of different parameters is analysed for images of the sequences Flowergarden, Foreman, Table Tennis and Crew in CIF-format (352×288 pixels). We simulated isolated and consecutive MB losses. In the following, we only show the analysis of a few parameters, the uncritical parameters are kept fixed.

The correlation coefficient of the 3-D windowing function in (21) showed for all sequences a similar behaviour and was therefore set to $\hat{\rho} = 0.8$. The support area consists of 13 pixels in spatial direction surrounding the block to be estimated which coincides with the results reported in [5]. In spatial direction the next larger FFT size of 64×64 was applied. The FFT size in temporal direction is set to 32 but could be reduced to 16 at the sacrifice of a marginal loss in PSNR. Hence, the FFT size was chosen to $64 \times 64 \times 32$. If a maximum number of 200 iterations is reached the algorithm terminates in order to limit the computational cost.

The iterative algorithm terminates if either the reduction of the residual error energy drops below a prespecified threshold ΔE_{min} per pixel or the maximum number of 200 iterations is exceeded. In the following, we evaluate the effect of the parameter ΔE_{min} on the performance and the average number of iterations per block. Parameter set A takes two previous $N_v = 2$ and two subsequent frames $N_N = 2$ into account and parameter set B only two previous frames $N_v = 2$ in case no following frames are available in order to estimate the parametric model. Table 1 shows the PSNR result and the average number of iterations for $\Delta E_{\text{min}} = 0.1, 1.0, 2.0$. The PSNR is the better the lower the threshold is because more iterations lead to a better approximation of the parametric model. In the first iteration, the DC component is chosen. Then, subsequently higher frequencies are selected with respect to the image content. The average number of iterations per block depends on the image content. The detailed flowering meadow needs more iterations than smoother areas. For our further investigations, we choose $\Delta E_{\text{min}} = 0.1$. However, reducing the number of iterations and thus the load provides

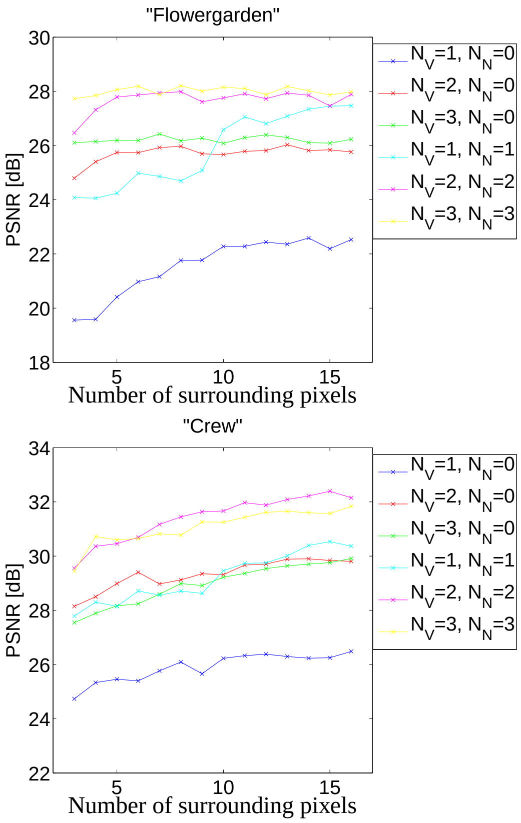

Figure 4: PSNR with respect to involved number of frames.

still satisfying results. Obviously, PSNR increases if not only the causal previous but also subsequent frames are involved.

Hence, we take a closer look at the impact of the number of previous and/or subsequent frames involved for extrapolation. Fig. 4 depicts on the left hand side the results for the sequence Flowergarden evaluated for block losses and on the right hand side for the sequence Crew evaluated for consecutive losses. The best results can be obtained for two or three previous and subsequent frames depending on the motion of the sequence. In case only previous frames are available, one previous frame yields the worst result. Two or three previous frames lead to a similar result depending on the motion of the sequence obtaining still satisfying results. In the following, we choose $N_V = N_N = 2$.

In the sequel, we evaluate the concealment performance with respect to other concealment techniques. Therefore, we implemented the *Temporal Blockreplacement (TR)* algorithm copying simply the block from the previous frame. Further, the *Boundary Matching Algorithm (BMA)* and the *Extended Boundary Matching Algorithm (EBMA)* described in [2] are used for comparison. Motion compensated prediction with a search area of 8 pixels, full search and pixel accuracy is performed in order to obtain the motion vectors and the prediction error. In case of BMA the motion vectors of the blocks to be concealed and in case of EBMA additionally the prediction error signal are discarded. Further, the results of the spatial 2-D extrapolation [5] are displayed.

In Table 3 the comparison for block losses is shown at the top and for consecutive losses at the bottom. The 3-D extrapolation yields the best results for all sequences and loss cases except for isolated block losses in Flowergarden. There, the BMA algorithm is 1 dB better than the 3-D extrapolation. The vector of the homogenous motion can be reconstructed very well from the vectors of *all* surrounding MBs. To receive all surrounding vectors is unlikely in a real-world scenario because several blocks in a row are commonly coded in a packet. Additionally, usually the prediction error signal is lost and not reconstructed. Otherwise, the concealment performance of 3-D extrapolation outperforms the other algorithms up to several dBs in PSNR. The performance of 3-D extrapolation does not decrease for consecutive losses even if less data is available for estimation.

Next, we want to confirm the gained insights subjectively with Fig. 6 and Fig. 7, respectively. Fig. 6 depicts on top the concealed image Flowergarden by EBMA and on bottom by 3-D extrapolation for isolated block losses shown in Fig. 5. Obviously, the EBMA algorithm is not able to compensate the motion which is apparent at the right edge of the tree. In contrary to the 3-D extrapolation which inherently compensates the motion. Also details like the branches of the tree or the flowers in the meadow can be extrapolated.

The top image of Fig. 7 shows the concealment of consecutive losses of Fig. 5, right hand side, by BMA. The algorithm is not able to compensate the change in luminance to the previous frame caused by a flash of a camera. Therefore, wrong motion vectors are found. As opposed to the spatio-temporal extrapolation which can compensate the variations in luminance and additionally reconstruct detailed areas like badges on the uniforms proven by the bottom image of Fig. 7.

6. CONCLUSION

We derived a 3-D extrapolation algorithm which allows to exploit *spatial and temporal* correlations of the video signal *at the same time*. The algorithm was applied to the extrapolation problem concealment in video communications and showed excellent extrapolation performance. In order to finally evaluate the performance, the method has to be integrated in a video coder like H.264/AVC. Current research focusses further on applying the extrapolation principle to prediction having the advantage that no motion vectors have to be transmitted.

REFERENCES

[1] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: a review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–977, May 1998.

[2] W.-M. Lam, A. R. Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, April 1993, pp. V417–V420.

[3] Y.-K. Wang, M. M. Hannuksela, V. Varsa, A. Hourunranta, and M. Gabbouj, "The error concealment feature in the H.26L test model," in *Proc. Int. Conf. on Image Processing (ICIP)*, Sept. 2002, vol. 2, pp. 729–732.

[4] A. Kaup, K. Meisinger, and T. Aach, "Frequency selective signal extrapolation with applications to error concealment in image communication," *Int. J. Electron. Commun. (AEÜ)*, vol. 59, pp. 147–156, June 2005.

[5] K. Meisinger and A. Kaup, "Minimizing a weighted error criterion for spatial error concealment of missing image data," in *Proc. Int. Conf. on Image Processing (ICIP)*, Singapore, Oct. 2004, pp. 813–816.

[6] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *IEEE Trans. Image Process.*, vol. 4, no. 4, pp. 470–477, April 1995.

[7] L. Xin and M. T. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 10, pp. 857–864, Oct. 2002.

[8] K. Meisinger, T. Troeger, M. Zeller, and A. Kaup, "Automatic TV logo removal using statistical based logo detection and frequency selective inpainting," in *Proc. European Signal Processing Conference (EUSIPCO)*, Turkey, Sept. 2005.

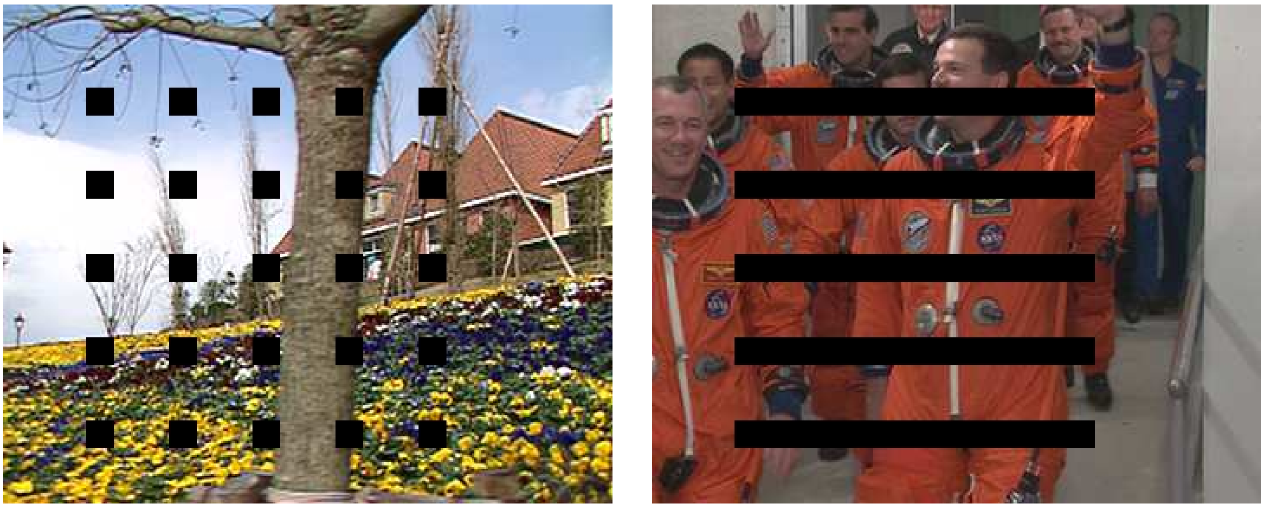

Figure 5: MB losses. Left: Isolated losses. Right: Consecutive losses.

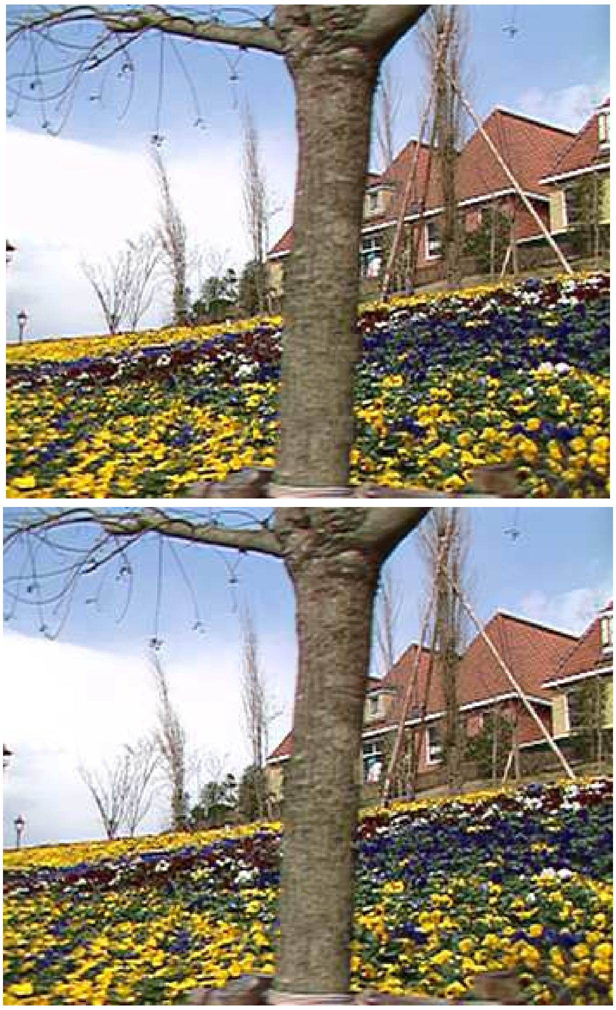

Figure 6: Concealed isolated MB losses. Top: EBMA. Bottom: 3-D extrapolation.

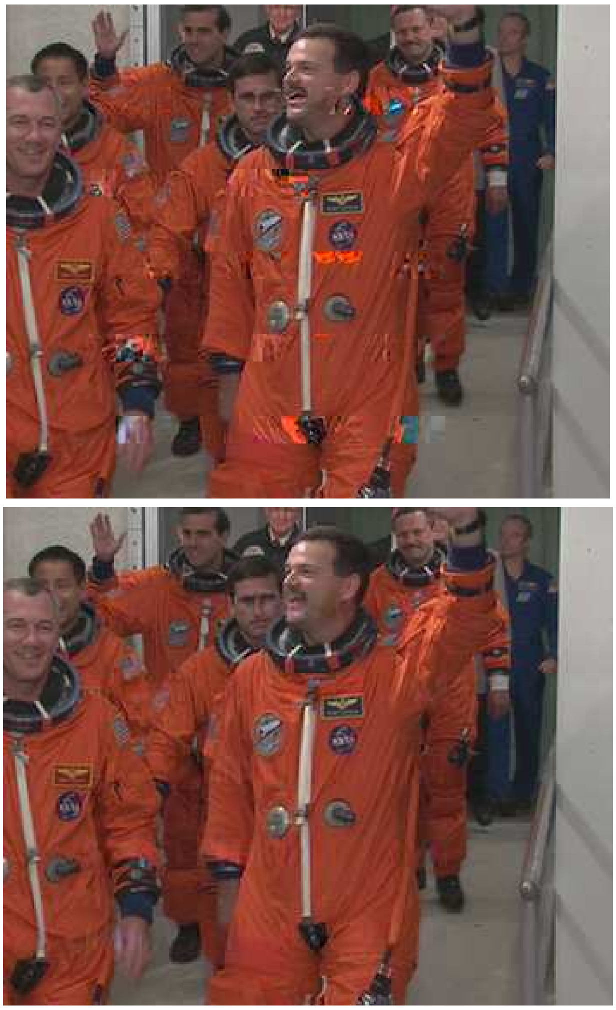

Figure 7: Concealed consecutive MB losses. Top: BMA. Bottom: 3-D extrapolation.