Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# Content-adaptive robust error concealment for packet-lossy H.264 video streaming*

LIAO Ning†, YAN Dan, QUAN Zi-yi, MEN Ai-dong

(*Multimedia Center, School of Telecommunication Engineering, Beijing University of Posts & Telecom, Beijing 100876, China*)

†E-mail: ning-777@tom.com

Received Dec. 5, 2005; revision accepted Feb. 18, 2006

**Abstract:** In this paper, we present a spatio-temporal post-processing error concealment (EC) algorithm designed initially for a H.264 video-streaming scheme over packet-lossy networks. It aims at optimizing the subjective quality of the restored video under the constraints of low delay and computational complexity, which are critical to real-time applications and portable devices having limited resources. Specifically, it takes into consideration the physical property of motion field in order to achieve more meaningful perceptual video quality, in addition to the improved objective PSNR. Further, a simple bilinear spatial interpolation approach is combined with the improved boundary-match (B-M) based temporal EC approach according to texture and motion activity analysis. Finally, we propose a low complexity temporal EC method based on motion vector interpolation as a replacement of the B-M based approach in the scheme under low-computation requirement, or as a complement to further improve the scheme's performance in applications having enough computation resources. Extensive experiments demonstrated that the proposal features not only better reconstruction, objectively and subjectively, than JM benchmark, but also robustness to different video sequences.

**Key words:** Error concealment, Error control, H.264/AVC, Video streaming, Lossy transmission
**doi:**10.1631/jzus.2006.AS0041 **Document code:** A **CLC number:** TN919.8

## INTRODUCTION

With the ubiquitous usage of Internet and the deployment of next generation of networks, video communication is increasingly becoming more demanded. Unlike data transmission, video communication is essentially time-sensitive but allows for some transmission errors, because human visual observation system can make up for a certain degree of errors. Further, in order to compress the large volume of video data, the spatial and temporal correlations within and between pictures are generally exploited in video coding standards. Therefore, video transmitted over packet-switched networks suffers varied quality degradation from packet loss and thus spatio-temporal error propagation. To deal with these new problems arising from lossy transmission, and still achieve higher compression gain, the new video compression standard, H.264/JVT (JVT of ISO/IEC MPEG & ITU-T VCEG, 2003) is designed. H.264 includes many error-resilient tools to facilitate the decoder's error concealment (EC) procedure that is a necessity in order to provide an intelligible reconstructed video because packet loss is inevitable. After an in-depth investigation of the error-resilient tools of H.264, we chose to employ the flexible macroblock order (FMO) tool shown in Fig.1, which is simple and efficient in improving the decoder's EC performance (Wenger, 2003).

The goal of EC is to estimate missing macroblocks (MBs) in a compressed video stream that arise from bit-erasure or packet loss, in order to achieve a minimum degree of perceptual quality degradation. Methods that have been developed roughly fall into two categories: spatial approach and temporal

* Project supported by the Teaching and Research Award Program for Outstanding Young Professor in High Education Institute, Ministration of Education, China

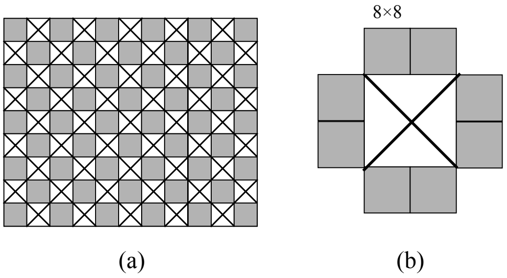


**Fig.1 (a) MB-interleaving FMO mode. Gray and white MBs are assembled into different slice packets. If the white MB is lost; (b) Its neighboring 8×8 gray blocks, if received, can be used to reconstruct it**

approach. In the former class, spatial correlation between local pixels is exploited. This includes various interpolation techniques in pixel domain such as bilinear (Salama *et al*., 1995), multidirectional edge-based methods (Kwok and sun, 1993; Sun and Kwok, 1995), maximum a posteriori (MAP) estimation algorithms (Shirani *et al*., 2002; Salama *et al*., 2000), and DCT coefficients recovery techniques (Alkachouch and Bellanger, 2000). Generally, spatial approach provides blurred estimates of the missing MB. After investigating several approaches among them, we found that the bilinear technique is, simple, fast, and in practice, a good tradeoff between performance and computational efficiency, although other algorithms may be more effective for recovering MBs having certain type of edges but at an expensive cost of complexity.

In temporal class, both the coherence of motion field and the spatial smoothness of pixels along edges across the block boundary, are exploited to estimate motion vector (MV) of a lost MB. Learning from experiments, we found that motion field interpolation methods (Mualla *et al*., 2000; Zheng and Chau, 2003), and block-match (B-M) based schemes (Chen *et al*., 1997; Tsekeridou and Pitas, 2000) are among the best conventional temporal techniques. For H.26L (previous version of H.264), Wang *et al*.(2002b) incorporated a B-M based temporal algorithm for corrupted P-frame in reference model (Available at http://iphome.hhi.de/suehting/tml/download/old_jm/) as benchmark for evaluation of any future proposal. It selects the optimum MV of a lost MB from a limited candidate MV set comprised of neighboring correctly received MVs. Obviously, however, success of this algorithm critically depends on the availability of the MVs in the neighborhood. Generally, temporal concealment yields better results if the MVs can be estimated with a satisfactory degree of accuracy, as the replaced MB is sharp and fairly similar to the original. Inaccurate MVs, on the other hand, may introduce unacceptable discontinuities, to which a blurred version of the MB estimated by spatial approach may be preferential.

Notice that the feasibility of spatial or temporal approach is not necessarily bonded to certain slice type or MB type. Conversely, complemented combination of the spatial and temporal methods is naturally an important approach from the above analysis. Thus, we propose that the bilinear pixel interpolation and the temporal approach work complementally based on local texture and motion activity analysis.

The paper is structured as follows. Section 2 gives theoretical insight into motion field estimation and an improved B-M based temporal EC algorithm. In Section 3, the bilinear pixel interpolation approach is described first, and then combined with the improved temporal EC algorithm based on texture and motion activity analysis, and B-M criterion. In Section 4, we propose a 4×4 block based bilinear MV interpolation approach to cope with complex local motion, where B-M based method may fail. Finally, simulation results and conclusion are given in Section 5.

## MOTION SMOOTHNESS CONSTRAINED TEMPORAL EC

In H.264, the popular block-based motion estimation approach specifies MV at every block. It assumes that, as long as each block is small enough, the motion variation within each block can be characterized well by a simple translational motion model, and the motion parameter (i.e., MV) for each block can be estimated independently. As a result, in practice, the resulting motion may be discontinuous across block boundaries, where the real motion field is changing smoothly from block to block. Fig.2b illustrates a typical block-wise motion representation. The prediction motion field that is quite chaotic on the left part of the picture should actually be small translational motion in the original sequence. This is due to the fact that encoder generally does not impose any constraint on the motion transition between adjacent blocks at the motion estimation stage, and thus the MVs for blocks such as homogeneous blocks, are indeterminate and, often, physically incorrect. For a complete mathematical explanation refer to (Wang *et al*., 2002a). Although the physically incorrect MVs do not affect the encoder's predictive video coding performance, we found they often result in false B-M based temporal replacement (Unlike in some literatures, the term "temporal replacement" here refers to the block recovered by motion estimation) at the EC stage of the decoder, as shown in Fig.2c.

We propose to exploit the physical constraint that the MVs should vary smoothly spatially, to eliminate incorrect candidate MVs for BM-based EC of (Wang *et al*., 2002b), by measuring the motion variation over a small area surrounding the lost MB as follows.

Denote the available MVs of surrounding 8×8 blocks in Fig.1 as $\boldsymbol{v}_i=(v_{i,x},v_{i,y})^{\mathrm{T}}$, $i$=1, …, $M$, $M$≤8, and their set as $V_{\mathrm{nbr}}$. Considering the large likelihood of practical motions like zooming and rotation, in addition to translation, in the neighborhood, we choose average motion magnitude defined as

$$v_{\mathrm{avg}} = (1/M)\sum_{i=1}^{M}[|v_{i,x}|+|v_{i,y}|], \tag{1}$$

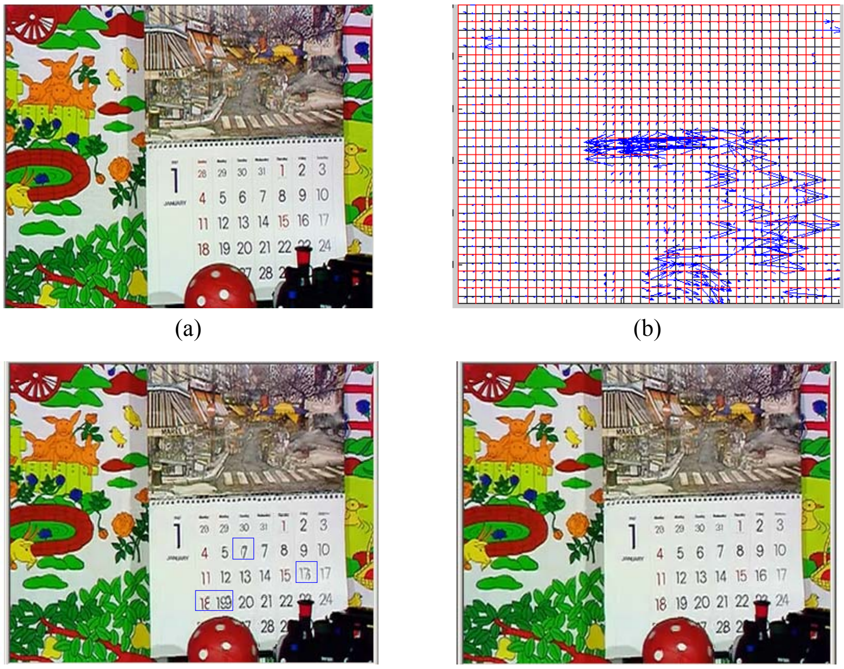
To restrict the candidate MVs set as:

$$V_{\mathrm{cdt}} = \{\boldsymbol{v}_i, i\in[1,M]\,|\,|v_{i,x}|+|v_{i,y}|< w\cdot v_{\mathrm{avg}}\}. \tag{2}$$

The weight factor $w$ determines the spatial coherence degree of the local motion field. A larger value of $w$ allows for larger motion discontinuity between neighboring blocks while a smaller value requires more smooth motion variation in the neighborhood. In practice, a value of 2 proves appropriate.

## CONTENT-ADAPTIVE SPATIO-TEMPORAL EC

There are many cases where bilinear spatial interpolation fails but temporal replacement does not, and vice versa, as shown in Figs.3b and 3c. This inspired us to conceive a wise combination of the two approaches that can provide better robust performance.

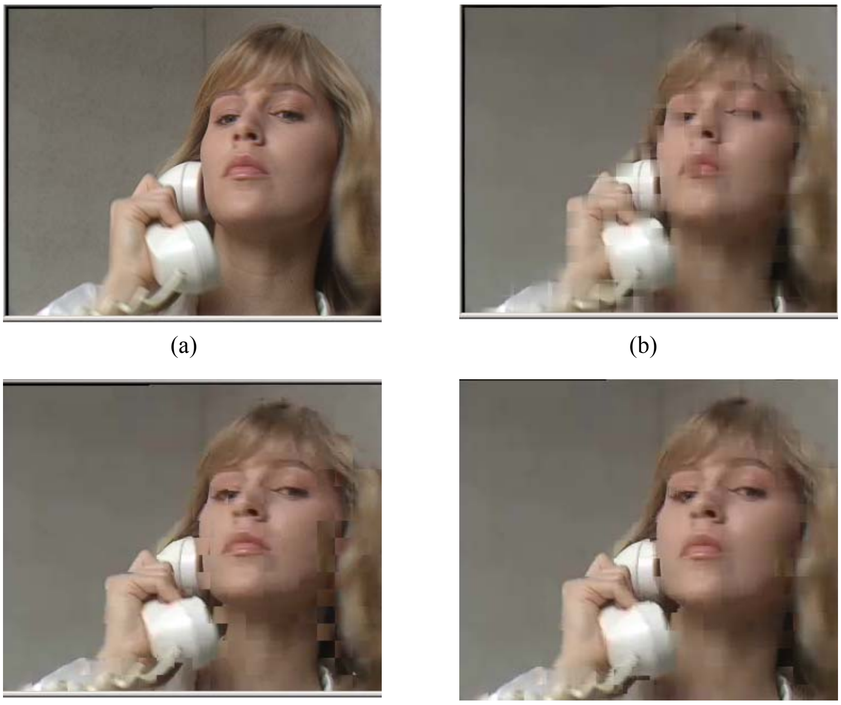

(a)

(b)

(c)

(d)

**Fig.2 Mobile CIF (frame P1). (a) Original image; (b) Chaotic motion field; (c) Restored by JM86; (d) Restored by Proposal 1**



(a) (b) (c) (d)

**Fig.3 Miss CIF (frame P0). (a) Original image; (b) Restored by bilinear interpolation; (c) Restored by JM86; (d) Restored by Proposal 2**

In bilinear pixel interpolation approach, each pixel $f^{\mathrm{r}}_{i,j}$ of the lost MB is reconstructed by spatially averaging its four closest decoded neighbors $f$ as

$$f^{\mathrm{r}}_{i,j}=\lambda[\mu_1 f_{i,-1}+(1-\mu_1)f_{i,N}]+(1-\lambda)[\mu_2 f_{-1,j}+(1-\mu_2)f_{N,j}],\ i,j\in[0,\ 15], \tag{3}$$

where, $N$=16 for an MB. Weights $\mu_1$ and $\mu_2$ are used to weigh the contributions from the neighboring vertical and horizontal pixels, respectively. $\mu_1$ is a function of the distances between the lost pixel and its closest vertical neighbor. $\mu_2$ is of similar form. Contributions from the MBs on either side are weighted by 1−$\lambda$, and those from above and below by $\lambda$. For simplicity, $\lambda$ is set to 1/2. Thus, rewrite $f^{\mathrm{r}}_{i,j}$ as

$$f^{\mathrm{r}}_{i,j}=\frac{(d_{\mathrm{N}}f_{i,-1}+d_{\mathrm{S}}f_{i,N}+d_{\mathrm{W}}f_{-1,j}+d_{\mathrm{E}}f_{N,j})}{d_{\mathrm{N}}+d_{\mathrm{S}}+d_{\mathrm{W}}+d_{\mathrm{E}}}, \tag{4}$$

where

$$d_{\mathrm{N}}=\begin{cases}N-j, & \text{if} f_{i,-1} \text{ exists, correctly received,}\\ 0, & \text{otherwise.}\end{cases}$$

and $d_{\mathrm{S}}$, $d_{\mathrm{W}}$, $d_{\mathrm{E}}$ are derived similarly.

As can be seen from Eq.(3), bilinear pixel interpolation recovers reasonably well the MBs that are smooth, or have merely horizontal or vertical edge(s), but fails to recover other complex textures. On the other hand, the improved B-M based temporal EC algorithm can preserve details of the lost MB if the MV is estimated correctly, but may bring undesirable edges if the wrong MV is used. Therefore, we propose an adaptive EC scheme as illustrated in Fig.4.

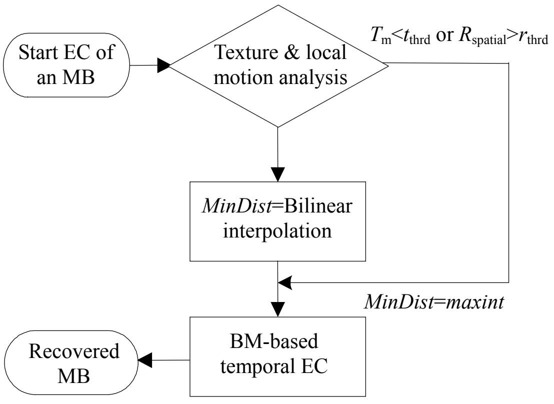

Start EC of an MB
Texture & local motion analysis
$T_{\mathrm{m}}<t_{\mathrm{thrd}}$ or $R_{\mathrm{spatial}}>r_{\mathrm{thrd}}$
*MinDist*=Bilinear interpolation
*MinDist*=*maxint*
BM-based temporal EC
Recovered MB

**Fig.4 Adaptive spatio-temporal EC of a lost MB**

After decoding all the correctly received MBs in the current frame, the proposed EC procedure is invoked if any missing MB is detected. Both the MVs and the pixel value of the adjacent correct 8×8 blocks are exploited to recover the lost MB (Fig.1). First, the temporal activity in the neighborhood of the damaged MB is quantified as follows:

$$T_{\mathrm{m}} = (1/C_{\mathrm{M}}^2)\sum_{k=1}^{M-1}\sum_{j=k+1}^{M} |\boldsymbol{v}_j - \boldsymbol{v}_k|,\ \boldsymbol{v}_j,\ \boldsymbol{v}_k \in V_{\mathrm{nbr}}.$$

And the measure of local spatial regularity around the missing area is derived from the histogram of luma differences in its surrounding 8-pixel wide areas by:

$$R_{\mathrm{spatial}} = histogram(|\Delta f_{i,j}| > e_{\mathrm{thrd0}},\ \forall \Delta f_{i,j}),$$

where $e_{\mathrm{thrd0}}$ is a constant, $|\Delta f_{i,j}|$ takes different formulation for the four neighbor areas. That is

$$|\Delta f_{i,j}| = \begin{cases} |f_{i,j} - f_{i-1,j}|,\ j \in [0,15],\ i \in [-1,-7],\ \text{above}, \\ |f_{i,j} - f_{i+1,j}|,\ j \in [0,15],\ i \in [16,22],\ \text{below}, \\ |f_{i,j} - f_{i,j-1}|,\ i \in [0,15],\ j \in [-1,-7],\ \text{left}, \\ |f_{i,j} - f_{i,j+1}|,\ i \in [0,15],\ j \in [16,22],\ \text{right}. \end{cases}$$

If $T_m$ does not exceed a threshold $t_{\mathrm{thrd}}$, then the missing MB, together with its surrounding areas, can be assumed to undergo relatively slow motion or translational motion. If $R_{\mathrm{spatial}}$ exceeds a threshold $r_{\mathrm{thrd}}$, then the damaged MB is less likely to be homogeneous, or vertically or horizontally edged region. In both above cases, bilinear spatial interpolation is not performed. Otherwise, bilinear spatial interpolation is performed. Then the best reconstruction is selected from all temporal replacements obtained from candidate MVs, and the spatial interpolation version if it exists, by minimizing the following boundary error cost function:

$$\varepsilon_{\mathrm{BM}} = \sum_{i=0}^{15}(f_{i,-1} - f_{i,0}^{\mathrm{r}})^2 + \sum_{i=0}^{15}(f_{i,16} - f_{i,15}^{\mathrm{r}})^2 + \sum_{j=0}^{15}(f_{-1,j} - f_{0,j}^{\mathrm{r}})^2 + \sum_{j=0}^{15}(f_{16,j} - f_{15,j}^{\mathrm{r}})^2.$$

The EC procedure in Fig.4 runs repeatedly until all the missing MBs in a frame are concealed. Then proceed to decoding the received packets of the next frame. This scheme is called "Proposal 1" in Table 1.

**Table 1 *PSNR*$_{\mathrm{avg}}$ of JM86, MFIEC, (Zheng and Chau, 2003), and the proposals**

| Sequence | Miss | Foreman | Mobile | Tempete | Mother-Doctor |
|---|---|---|---|---|---|
| JM86 | 27.99 | 30.57 | 21.66 | 26.12 | 36.43 |
| MFIEC | 26.61 | 31.80 | 22.57 | 27.07 | 37.25 |
| (Zheng and Chau, 2003) | 26.16 | 31.39 | 20.77 | 25.62 | 37.14 |
| Proposal 1 | 29.18 | 31.23 | 24.15 | 28.26 | 37.30 |
| Proposal 2 | 29.63 | 32.30 | 24.19 | 28.45 | 37.41 |

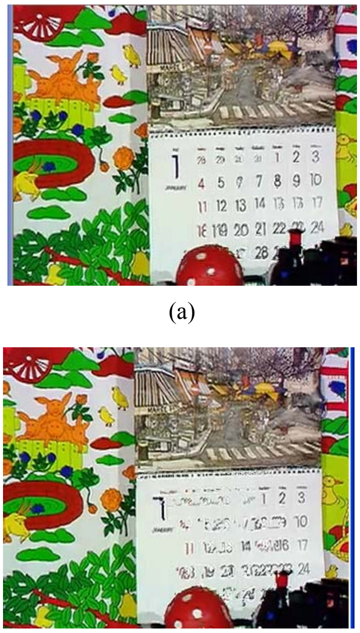
## MV INTERPOLATION BASED TEMPORAL EC

Mualla *et al.*(2000) proposed to interpolate MV for each missing pixel from the adjacent correctly received MVs, and reported satisfactory performance. Zheng and Chau (2003) applied the same idea to each missing 4×4 block by Lagrangian interpolation. We investigated them and employed bilinear interpolation for MV of each lost 4×4 block, by substituting $\boldsymbol{v}_{i,j}^{\mathrm{r}}$, $i$, $j\in$ [0,3] for pixel $f_{i,j}^{\mathrm{r}}$ in Eq.(4), correctly received MV, $\boldsymbol{v}_{i,-1}$, $\boldsymbol{v}_{i,N}$, $\boldsymbol{v}_{-1,j}$, $\boldsymbol{v}_{N,j}$ of surrounding 4×4 blocks for pixels $f_{i,-1}, f_{i,N}, f_{-1,j}, f_{N,j}$, and letting $N$=4.

Our approach is justified by its quite similar PSNR performance to (Zheng and Chau, 2003) (Table 1) and largely reduced computations. More importantly, in areas having smooth/global motion, all these motion field interpolation EC (MFIEC) approaches yielded poor visual reconstruction as compared with B-M approach (BMEC), even if their PSNR results outperform the later, as depicted in Fig.5. Actually, MFIEC works well with MBs having complex local motions, while BMEC is superior with MBs having smooth motion. Therefore, we further incorporate the bilinear MV interpolation based version of restored MB into the BM-based selection of best reconstruction as depicted in Fig.4. We call this "Proposal 2".

## SIMULATIONS AND CONCLUSION

In order to evaluate the proposed approach's



(a)

(b)

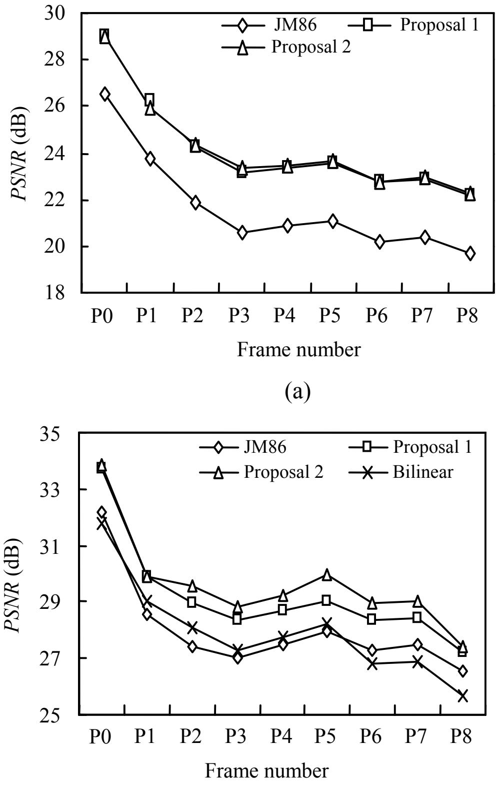
**Fig.5 Mobile CIF (frame P4). (a) Restored by JM86, *PSNR*=20.84 dB; (b) Restored by MFIEC, *PSNR*=22.26 dB**

performance, we incorporated it in JM86 (Available at http://iphome.hhi.de/suehting/tml/download/old_jm/), which is the benchmark for our comparison. The thresholds $t_{\rm thrd}$, $e_{\rm thrd0}$, and $r_{\rm thrd}$ are set to 8 (corresponding to 1 pixel in 1/4-pixel unit), 10, and 16, respectively by training. In order to assess the robustness of our proposal to different video contents, we experimented with extensive sequences including Miss, Foreman, Mobile, Tempete, and Mother-Doctor. Sequence is structured as one I-frame followed by nine P-frames, and coded by H.264 encoder with fixed *QP*=28 and MB-interleaving FMO mode. Finally, the channel models given in (VCEG, 1999) were used to simulate the backbone Internet. A packet loss rate of 20% was applied.

Since our aim is to evaluate the performance of post-processing EC techniques, not the error-free compression efficiency, error-free decoded video, instead of original video, was used as reference in our calculation of PSNR. It indicated exactly the objective video distortion resulting from inevitable lossy transmission and the EC technique employed.

More importantly, subjective assessment is necessary in addition to PSNR, since human eyes are the ultimate consumer of video and PSNR has been widely recognized for its inconsistence with perceived video quality by human visual system (Wang *et al*., 2003), particularly at low bit rate or in the case that compressed video undergoes severe channel distortion.

Typical objective performance is given in Table 1. For all tested sequences, our proposals outperform the conventional BMEC in JM86, more or less, depending on the sequences themselves and their interaction with the packet loss pattern. For Mobile, the per-frame PSNR improvement is significant, seen from Fig.6a. In particular, the visual quality of the reconstructed video is more physically meaningful as illustrated in Fig.2. This mainly benefits from the motion field smoothing filter that limits the set of candidate MVs. For Miss, the PSNR improvement in Fig.6b results mainly from the content-adaptive combination of the bilinear interpolation approach and the temporal approaches. As can be seen from Fig.3, good temporal replacement provides sharper and



(a)

(b)

**Fig.6 Per-frame PSNR performance. (a) Traditional BMEC in JM86 and the proposals; (b) Bilinear pixel interpolation approach, the JM86, and the proposals**

hence more pleasant reconstruction; however, inaccurate motion estimation introduces blocky effects, to which spatial interpolation version of the MB is preferable.

Also note that in Table 1 the PSNR of MFIEC emulates or even outperforms that of the BMEC in JM86 depending on sequences. This partially justifies MFIEC as a promising temporal EC method in applications with low complexity requirement, because it entails much less computation than BMEC. On the other hand, however, PSNR metric is often inconsistent with subjective evaluation in EC task. Fig.5 illustrates such an example. There is a global motion of the calendar in Mobile sequence, where MFIEC renders less intelligible picture but with a larger *PSNR*. This shows that MFIEC excels at recovering MB having complex motions and that BMEC works better with MB having translational motion.

In conclusion, we present a content-adaptive spatio-temporal EC scheme for H.264 video communication over packet-lossy networks. It imposes a motion field smoothing filter in the B-M based temporal EC approach, thus enabling more perceptually meaningful reconstruction. Further, the bilinear spatial interpolation approach, the improved BMEC and MFIEC temporal approaches are combined in a complemented way, based on texture and motion activity analysis and the boundary match criterion. Extensive experiments and detailed analysis demonstrated that our proposal provides not only better subjective and objective reconstruction than the JM benchmark, but also maintains robustness to different sequences.

## References

Alkachouch, Z., Bellanger, M.G., 2000. DCT-based spatial domain interpolation of blocks in images. *IEEE Trans. on Image Processing*, **9**(4):729-732. [doi:10.1109/83.841948]

Chen, M.J., Chen, L.G., Weng, R.M., 1997. Error concealment of lost motion vectors with overlapped motion compensation. *IEEE Trans. Circuits and Systems for Video Technology*, **7**(3):560-563. [doi:10.1109/76.585936]

Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 2003. Draft ITU-T Recomm. and Final Draft Int'l. Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), Doc.G050r1.

Kwok, W., Sun, H., 1993. Multidirectional interpolation for spatial error concealment. *IEEE Trans. on Consumer Electronics*, **39**(3):455-460. [doi:10.1109/30.234620]

Mualla, M.A., Canagarajah, C.N., Bull, D.R., 2000. Motion field interpolation for temporal error concealment. *IEEE Proc. Visual Image Signal Processing*, **147**(5):445-453.

Salama, P., Shroff, N.B., Coyle, E.J., Delp, E.J., 1995. Error Concealment Techniques for Encoded Video Streams. Pro. Int. Conf. Image Processing, p.9-12.

Salama, P., Schroff, N.B., Delp, E.J., 2000. Error concealment in MPEG video streams over ATM networks. *IEEE J. on Select. Areas in Communication*, **18**(6):1129-1144. [doi:10.1109/49.848263]

Shirani, S., Kossentini, F., Ward, R., 2000. A concealment method for video communication in an error-prone environment. *IEEE J. Select. Areas Commun.*, **18**(6): 1122-1128. [doi:10.1109/49.848261]

Sun, H., Kwok, W., 1995. Concealment of damaged block transform coded images using projections onto convex sets. *IEEE Trans. on Image Processing*, **4**(4):470-479. [doi:10.1109/83.370675]

Tsekeridou, S., Pitas, I., 2000. MPEG-2 error concealment based on block-matching principles. *IEEE Trans. Circuits and Systems for Video Tech.*, **10**(4):646-658. [doi:10.1109/76.845010]

VCEG, 1999. Internet Error Patterns VCEG-O38r1.doc. ftp://ftp.imtc-files.org/jvt-experts/9910_Red/Q15-I16r1.zip.

Wang, Y., Ostermann, J., Zhang, Y.Q., 2002a. Two-dimensional Motion Models. Video Processing and Communications, Section 5.5. Prentice Hall.

Wang, Y.K., Hannuksela, M.M., Varsa, V., Hourunranta, A., Gabbouj, M., 2002b. The error concealment feature in the H.26L test model. *Proc. Int. Conf. Image Processing*, **2**:729-732.

Wang, Z., Sheikh, H.R., Bovik, A.C., 2003. Objective Video Quality Assessment. *In*: Furht, B., Marqure, O. (Eds.), Handbook of Video Databases: Design and Applications. CRC Press, p.1041-1078.

Wenger, S., 2003. H.264/AVC over IP. *IEEE Trans. on Circuits and Systems for Video Tech.*, **13**(7):645-656. [doi:10.1109/TCSVT.2003.814966]

Zheng, J.H., Chau, L.P., 2003. A motion vector recovery algorithm for digital video using Lagrangian interpolation. *IEEE Trans. Broadcasting*, **49**(4):383-389. [doi: 10.1109/TBC.2003.819050]