SPATIO-TEMPORAL MARKOV RANDOM FIELD-BASED PACKET VIDEO ERROR CONCEALMENT

Daniel Persson and Thomas Eriksson

Chalmers University of Technology
Department of Signals and Systems
412 96 Göteborg
Sweden

ABSTRACT

In this paper, a *spatio-temporal Markov random field method* is proposed for block-based packet video error concealment. We suggest the combined usage of two estimators, one for lost pixels, and one for lost motion vectors. The estimator for the lost pixel field takes surrounding pixels in the same frame where the loss occurred and motion-compensated pixels from a previous frame based on a motion field estimate into account, while the optimal estimator of the motion field takes surrounding pixels in the same frame where the loss occurred, pixels from a previous frame, and *the estimator function for the pixel field* into account. Our method increases performance in peak signal-to-noise ratio as well as subjective visual performance compared to several other previous error concealment algorithms.

Index Terms— Error concealment, block-based packet video, estimation, Markov random field.

1. INTRODUCTION

The state-of-the-art video-coding scheme H.264/MPEG-4 part 10 is block-based, i.e. block-based motion-compensated inter-frame prediction, transformation, and quantization is employed in the scheme [1]. While such an encoder achieves high compression efficiency, the resulting bit stream is vulnerable to communication channel impairments. Packet errors occur in video transmission over a packet network such as the Internet, and may be characterized by a simultaneous loss of a bigger amount of data locally in the video stream.

Error concealment is the name for the category of techniques that repair errors without auxiliary information from the encoder [2]. Block-based packet video error concealment methods are usually categorized into spatial approaches such as [3], that use only spatially surrounding pixels for estimation of lost blocks, and temporal approaches such as [4] and [5], that use motion information and pixels from previous frames. A third group of strategies such as [6], [7], [8] and [9] combines spatial and temporal information for error concealment.

In this paper, we introduce mathematical notation for analyzing potential and existing solutions to the block-based packet video error concealment problem in terms of which information that is used by the schemes. In the reasoning that follows, we suggest a *spatio-temporal strategy* that combines two estimators. An estimator for the lost pixel field takes surrounding pixels in the same frame where the loss occurred and motion-compensated pixels from a previous frame based on a motion field estimate into account, while the optimal estimator of the motion field takes surrounding pixels in the same frame where the loss occurred, pixels from a previous frame, and *the estimator function for the pixel field* into account.

Section 2 analyzes solutions to the block-based packet video error concealment problem in terms of which information that is used, and spatio-temporal Markov random field (MRF)-based packet video error concealment is proposed. In Section 3, our method is compared to previous efforts. The paper is concluded in Section 4.

2. SPATIO-TEMPORAL MRF-BASED PACKET VIDEO ERROR CONCEALMENT

In this section, spatio-temporal MRF-based packet video error concealment is introduced. First in Section 2.1, the block-based packet video error concealment problem is analyzed, and the assumptions used in spatial, temporal, and spatio-temporal error concealment are quantified. Thereafter, in Section 2.2, our methodology, that is based on observations in Section 2.1, is proposed.

2.1. Spatial, temporal and spatio-temporal error concealment

Assume that a loss of a group of neighboring pixels represented by the stochastic vector X, occurs in frame t. Suppose more specifically that the loss is such that the motion vectors (MV) belonging to X, and represented by the stochastic vector V_X, as well as the displaced frame difference for X in the case of inter-frame coding, are lost. Pixels in frames t and $t-1$ surrounding the lost area, here represented by the vectors S^t_{SUR} and S^{t-1}_{SUR} respectively, as well as MV information surrounding the lost area, here represented by the vector V_{SUR}, are available for forming a replacement of X. From an information theoretic perspective, a spatio-temporal optimal estimate

$$\hat{X} = g_1(S^t_{\mathrm{SUR}}, S^{t-1}_{\mathrm{SUR}}, V_{\mathrm{SUR}}) \tag{1}$$

where all available information is considered, is desirable. However, because objects move between the frames, the number of pixels included in S^{t-1}_{SUR} has to be very big in order to include all pixels in frame $t-1$ that may be of interest. Therefore, subsets of the information S^{t-1}_{SUR}, S^t_{SUR}, V_{SUR}, or subsequent usage of parts of the information S^{t-1}_{SUR}, S^t_{SUR}, V_{SUR} are considered for error concealment in previous approaches.

In what is traditionally known as spatial error concealment, information in frame t surrounding the lost area is used for replacement of X, i.e. the optimal estimate of the lost area may be written

$$\hat{X} = g_2(S^t_{\mathrm{SUR}}). \tag{2}$$

The method [3] mentioned in Section (1) belongs to this category.

However, temporal error concealment methods may restore details better inside the lost blocks. Traditional temporal error concealment is based on the thought that by estimating MVs for the lost area X, pixel information from frame $t-1$ that has the highest correlation with X is first sorted out, and may then be used for error concealment of X. This implies that error concealment is carried out by subsequent usage of two estimators. An estimator *function* for the lost pixel area takes an estimate $\hat{V}_X$ of MVs together with a vector of pixels in the previous frame $S_{\mathrm{SUR}}^{t-1}(\hat{V}_X) \subset S_{\mathrm{SUR}}^{t-1}$ as arguments

$$\hat{X} = g_3\Big(S_{\mathrm{SUR}}^{t-1}(\hat{V}_X)\Big) = S_{\mathrm{SUR}}^{t-1}(\hat{V}_X). \tag{3}$$

For providing the optimal MV estimate $\hat{V}_X^*$, the information S_{SUR}^t, S_{SUR}^{t-1}, V_{SUR}, and the *trivial function* g_3 are employed. This optimal estimate may be written

$$\hat{V}_X^* = h_3(S_{\mathrm{SUR}}^t, S_{\mathrm{SUR}}^{t-1}, V_{\mathrm{SUR}}, g_3). \tag{4}$$

The methods [4] and [5] discussed in Section (1) belong to this category. It is well known that motion-compensated frame estimation achieves good results with low computational complexity in state-of-the-art video coding [1]. However, in a scene with fast motion, or following a scene change, spatial methods may work better.

A third group of methods that is traditionally considered as spatio-temporal error concealment provides an estimate of the lost area from pixels S_{SUR}^t in frame t, and a vector of pixels $S_{\mathrm{SUR}}^{t-1}(V_{\mathrm{SUR}}) \subset S_{\mathrm{SUR}}^{t-1}$ and $S_{\mathrm{SUR}}^{t-1}(\hat{V}_X^*)$ in frame $t-1$

$$\hat{X} = g_4\Big(S_{\mathrm{SUR}}^t, S_{\mathrm{SUR}}^{t-1}(V_{\mathrm{SUR}}), S_{\mathrm{SUR}}^{t-1}(\hat{V}_X^*)\Big). \tag{5}$$

MV estimates $\hat{V}_X^*$ that these methods rely on are however retrieved without consideration of the estimator function g_4 of the pixel field. The methods [6], [7], and [8] discussed in Section (1) belong to this category.

In [9], the estimator function for the pixel field may be written

$$\hat{X} = g_5\Big(S_{\mathrm{SUR}}^t, S_{\mathrm{SUR}}^{t-1}(V_{\mathrm{SUR}}), S_{\mathrm{SUR}}^{t-1}(\hat{V}_X)\Big). \tag{6}$$

Optimal MV estimates are in turn retrieved considering the estimator function g_5 of the pixel field

$$\hat{V}_X^* = h_5(S_{\mathrm{SUR}}^t, S_{\mathrm{SUR}}^{t-1}, V_{\mathrm{SUR}}, g_5). \tag{7}$$

Such an approach should, in terms of information, be superior to the strategies (2), (4) followed by (3), as well as (5). The reason for this is that in (6), as much information as in (5) is taken into account, while at the same time, (7) takes an estimate of the pixel field (6) into account.

In the following, we will propose a spatio-temporal error concealment scheme that has the form in (6) and (7), and that is based on MRF modeling.

2.2. Spatio-temporal MRF-based error concealment

In this section, we adopt to the error concealment formulation incorporating the estimator pair (6) and (7) from Section 2.1. A MRF-based maximum a posteriori (MAP)-optimal estimator of lost MVs having the form (7), i.e. considering the estimator function g_5, is derived in Section 2.2.1. Thereafter, estimation of lost pixels as in (6) is treated in Section 2.2.2.

2.2.1. Estimation of lost motion vectors

According to the Hammersley-Clifford theorem, a MRF is equivalent to a Gibbs random field, that has an associated Gibbs distribution [10]. The joint probability density function (pdf) for the pixel and MV fields is modeled as a MRF with a Gibbs distribution

$$p(s,v) = \frac{1}{Z} e^{-\frac{1}{T}U(s,v)} \tag{8}$$

where Z is a normalizing constant called the partition function, T is a constant referred to as temperature, and $U(s,v)$ is a potential function. Following [5] we choose to further specify

$$\begin{aligned} U(s,v) &= U_{\mathrm{S}}(s) + U_{\mathrm{V}}(v) &(9)\\ &= \sum_{\gamma} U_{\mathrm{S}}^{\gamma}(s) + \sum_{\gamma} U_{\mathrm{V}}^{\gamma}(v) &(10)\end{aligned}$$

where U_{S} and U_{V} are potential functions constituted by clique potentials U_{S}^{γ} and U_{V}^{γ} in the neighborhood system. Details of U_{S} and U_{V}, that are the same as in [5], will be included in a longer journal version of this paper, but are left out here because of insufficient space. In order to form a pdf

$$Z = \sum_{s}\sum_{v} e^{-\frac{1}{T}U(s,v)}. \tag{11}$$

If $s = \{s_{\mathrm{SUR}}^t, \hat{v}_X\}$ and $v = \{\hat{v}_X, v_{\mathrm{SUR}}\}$, we may formulate (7) as the constrained MAP optimization problem

$$\hat{v}_X^* = \arg\max_{\hat{v}_X} p(\hat{x}, s_{\mathrm{SUR}}^t, \hat{v}_X, v_{\mathrm{SUR}}) \tag{12}$$

where the pdf is given by (8) and the constraint is given by (6). Taking the logarithm of (12), it is finally possible to write

$$\begin{aligned}\hat{v}_X^* = &\\ &\arg\min_{\hat{v}_X}\Bigg\{U_{\mathrm{S}}\Big(\hat{x}\Big(s_{\mathrm{SUR}}^t, s_{\mathrm{SUR}}^{t-1}(\hat{v}_X), s_{\mathrm{SUR}}^{t-1}(v_{\mathrm{SUR}})\Big), s_{\mathrm{SUR}}^t\Big)\\ &+U_{\mathrm{V}}(\hat{v}_X, v_{\mathrm{SUR}})\Bigg\}\end{aligned} \tag{13}$$

where neither the constant partition function Z nor the temperature T appear. The optimization in (13) is efficiently solved by the iterated conditional modes (ICM) algorithm [5]. To summarize, in (13), the information S_{SUR}^t, S_{SUR}^{t-1}, V_{SUR}^t, and the estimator function g_5 are taken into account, which means that we have achieved an expression on the form (7) for our estimator.

2.2.2. Estimation of lost pixels given estimates of lost motion vectors

Here a method for achieving (6), i.e. an estimate of the lost pixel field using surrounding pixels S_{SUR}^t, $S_{\mathrm{SUR}}^t(V_{\mathrm{SUR}})$, and $S_{\mathrm{SUR}}^{t-1}(\hat{V}_X)$ is presented. One 8×8-block of pixels that forms a vector $X_{\mathrm{BLOCK}} \subset X$ is estimated at a time. Inspired by the estimators [6], [7], and [8], that have provided good results in peak signal-to-noise ratio (PSNR) in comparison with other methods, we choose an estimator with the

form

$$\begin{aligned}\hat{X}_{\mathrm{BLOCK}} &= g_5\Big(S^t_{\mathrm{SUR}}, S^{t-1}_{\mathrm{SUR}}(V_{\mathrm{SUR}}), S^{t-1}_{\mathrm{SUR}}(\hat{V}_X)\Big) \\ &= wA(S^t_{\mathrm{SUR}}) \\ &\quad +(1-w)IS^{t-1}_{\mathrm{SUR}}(\hat{V}_X) \qquad (14)\\ w &= w\Big(S^t_{\mathrm{SUR}}, S^{t-1}_{\mathrm{SUR}}(V_{\mathrm{SUR}})\Big) \qquad (15)\end{aligned}$$

where A is a linear function of the border pixels to X_{BLOCK} that belong to S^t_{SUR} and I is a matrix that chooses the motion-compensated pixels for $\hat{X}_{\mathrm{BLOCK}}$. Further w is a scalar function that depends on the local video statistics and that regulates the influence of the pixels S^t_{SUR} in the same frame where the loss occurred and the motion-compensated pixels $S^{t-1}_X(\hat{V}_X)$ from the previous frame on the final estimate. The function w is only evaluated once for each block in X, prior to applying (14) and (13) for finding a replacement.

Instead of (14), a form of (6) could have been chosen to resemble the schemes in [6], [7], or [8] more closely. However, (14) has the attractive feature that the first part of the estimator that uses pixels from frame t may be separated from the computationally inexpensive second part of the estimator that depends on the MV estimate $\hat{V}_X$. In this way, the computationally expensive part of the estimator $wA(S^t_{\mathrm{SUR}})$ needs to be calculated only once, while only the computationally inexpensive part $(1-w)IS^{t-1}_{\mathrm{SUR}}(\hat{V}_X)$ of the estimator will vary with different candidate MVs when solving the optimization problem (13).

For determining the linear function A, a regularizing approach similar to the one in [3], and built on the assumption that the first derivative should be minimal in the lost area, is developed. While the method in [3] was iterative, we have for complexity reasons reformulated the method so that it may be used in a non-iterative manner. More specifically, the iterative estimator of [3] reuses previously estimated pixels of X when estimating X_{BLOCK}. This means that if the estimator in [3] would be applied without modification in our algorithm, the whole estimator (14) of the pixel field would have to be recalculated in each iteration of the iterative ICM algorithm when solving (13). This is avoided by reformulating the method in [3] so that it works in a non-iterative manner. For reasons of insufficient space, the details of the derivation of the linear function A are omitted here, but will be included in a longer journal version of the paper. For now we only state the result in the case when pixels are available on the upper (u) and lower (l) borders of the lost block. The other cases, when different borders to the lost block are available, give rise to similar expressions. Pixels from a realization s^t_{SUR} are included in two vectors b_{u} and b_{l} together with zero entries, and the linear function A may be written

$$A(S^t_{\mathrm{SUR}}) = \Big((A_{\mathrm{u}})^{\mathrm{T}}A_{\mathrm{u}}+(A_{\mathrm{l}})^{\mathrm{T}}A_{\mathrm{l}}\Big)^{-1}\Big((A_{\mathrm{u}})^{\mathrm{T}}b_{\mathrm{u}}+(A_{\mathrm{l}})^{\mathrm{T}}b_{\mathrm{l}}\Big) \quad (16)$$

where A_{u} and A_l impose that the vertical first derivative should be minimized between every pixel in the lost block.

We derive the scalar w in (14) in the minimum mean square error (MMSE) sense by solving

$$\begin{aligned}w &= \arg\min_{w'} \mathrm{E}\Big[\Big\|X_{\mathrm{BLOCK}} - w'A(S^t_{\mathrm{SUR}}) \\ &\quad -(1-w')IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)\Big\|^2_2\Big] \qquad (17)\end{aligned}$$

where the norm $\|\cdot\|_2$ is the Euclidean norm. This is equivalent to maximizing the PSNR. We achieve

$$\begin{aligned}w &= \frac{1}{\mathrm{E}\Big[\Big\|A(S^t_{\mathrm{SUR}}) - IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)\Big\|^2_2\Big]} \\ &\quad \times\mathrm{E}\Big[\Big(X_{\mathrm{BLOCK}} - IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)\Big)^{\mathrm{T}} \\ &\quad \times\Big(A(S^t_{\mathrm{SUR}}) - IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)\Big)\Big]. \qquad (16)\end{aligned}$$

The parameter w is calculated for each block prior to applying (13), which implies that we neither have access to X_{BLOCK} nor $IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)$ in (16). Therefore, we use neighboring blocks to X_{BLOCK} in order to achieve w, i.e. X_{BLOCK} is replaced by a function of S^t_{SUR}, and $IS^{t-1}_{\mathrm{SUR}}(\hat{V}^*_X)$ is replaced by a function of $S^{t-1}_{\mathrm{SUR}}(V_{\mathrm{SUR}})$. More specifically, blocks in the neighborhood of X_{BLOCK} are used for evaluating w. For evaluation, the expectation in (16) is replaced by a sample mean of several available blocks surrounding X_{BLOCK}.

To summarize, we have in Section 2.2 achieved an estimator of the lost pixel field stated in (14) and (15) that has the sought form (6) and an estimator of the lost MV field (13) that has the sought form (7).

3. EXPERIMENTS

In this section, the proposed method is compared to methods suggested by other authors. Simulation details, which are chosen to fit state-of-the-art block-based video coders, are given in Section 3.1. These conditions are impartial to all the compared schemes. Results of the experiments are presented in Section 3.2.

3.1. Simulation prerequisites

Video. We use randomly chosen clips with a mean number of 18 frames from 89 MPEG-1 movies from [11] that have a frame rate of 29.97 frames per second and an image size of 352×240 pixels. Only the luminance component is used, but it is straight-forward to apply the method on the chrominance components as well. MVs are calculated for 8×8-blocks. Calculation of MVs for 8×8-blocks is supported by H.264/MPEG-4 part 10 [1]. A search for a MV is performed by checking every integer displacement vector $(\Delta u, \Delta v)$ where $-8 \leq \Delta u, \Delta v \leq 8$.

Packet errors. The video frames are first decoded, and thereafter are lost contiguous areas comprising several blocks introduced in the frames as in [5]. We assign a slice of 8×8-blocks to each packet, and accordingly simulate packet loss by randomly distributing slices of lost 8×8-blocks in the test sequences with error probabilities ranging from 5 to 20%. Assigning information for closely situated 8×8-blocks to different packets, as is done here by putting neighboring slices of 8×8-blocks in different packets, has previously shown to increase effectiveness of spatio-temporal error concealment schemes [6]. Errors propagate temporally. It is further assumed that we know at the decoder side where the errors occurred in the frames.

Proposed estimator. Optimization of (13) is carried out in the multi-scale manner explained in [12] and [13]. The MVs for the lost blocks $\hat{v}_X$ in (13) were initialized by the median of the MVs of the surrounding available blocks [4]. If necessary, this strategy was applied repeatedly so that also blocks without decodable neighbor blocks were assigned initializing MVs. The search range of the ICM algorithm when solving (13) was not specified in [5]. Each component of the MVs was searched within the range between the minimum and

maximum of the corresponding components of the initialization MV of the upper, lower, left, and right blocks.

Benchmarking. The proposed estimator is compared to Zhu *et al*'s method [6], that is a spatio-temporal method of the form (5). Zhu *et al*'s method was inspired by [3], that also influenced our derivation of the linear function A in (14). We also compare our scheme to Zhang and Ma's method [5], that influenced our choice of a MRF-based strategy and that has previously shown good results in PSNR in comparison with other error concealment schemes. Moreover, a comparison is made with the boundary matching approach (BMA) [9], that has the same spatio-temporal form as our method stated in (6) and (7). Also, we compare our method to motion-compensated copying and replacement of a lost MV by the median of surrounding MVs [4]. A comparison with motion-compensated copying and replacement of a lost MV by the mean of surrounding MVs is also made, as this method was used for comparison in [5].

3.2. Results

Performance in PSNR of the proposed method is benchmarked against the methods described in 3.1. Slices of lost 8×8-blocks are distributed randomly in the test sequences with error probabilities varying from 5 to 20%. Results are seen in Figure 1.

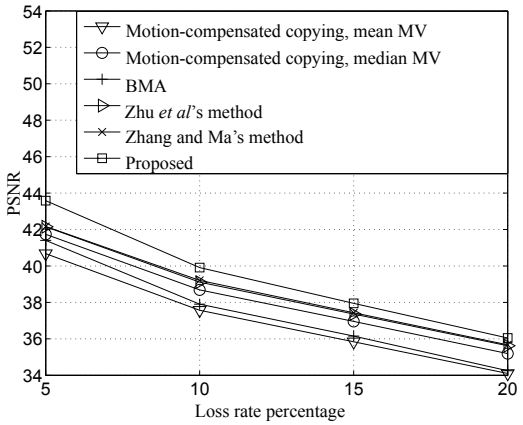

Fig. 1. Error concealment performance in PSNR. Slices of lost 8×8-blocks are distributed randomly in the test sequences with error probabilities varying from 5 to 20%.

The proposed method gives best performance in PSNR in all comparisons. Moreover, it is seen in the simulations that Zhu *et al*'s method that works merely by mixing pixel information from the same frame where the loss occurred and from a previous frame, as well as Zhang and Ma's method that works merely by providing refined MV estimates, both increase performance in PSNR compared to motion-compensated copying with median MV estimate. Our approach works in both these ways.

Images that show that the proposed method improves subjective visual quality will be included in a longer journal paper.

4. CONCLUSION

In this paper, spatio-temporal block-based packet video error concealment is addressed using a combination of two estimators. An estimator for the lost pixel field takes surrounding pixels in the same frame where the loss occurred and motion-compensated pixels from a previous frame based on a motion field estimate into account, while the MAP-optimal estimator of the motion field takes surrounding pixels in the same frame where the loss occurred, pixels from a previous frame, and *the estimator function for the pixel field* into account.

Our method increases performance in PSNR compared to several other previous error concealment algorithms. Moreover, it is seen in the simulations that a method that works merely by mixing pixel information from the same frame where the loss occurred and from a previous frame, as well as a method that works merely by providing refined MV estimates, both are effective in terms of PSNR in the same scenario. Our approach works in both these ways.

5. REFERENCES

[1] A. Tamhankar and K. R. Rao, "An overview of h.264/mpeg-4 part 10," in *Video/Image Processing and Multimedia Communications, 2003. 4th EURASIP Conference focused on*, July 2003, vol. 1, pp. 1 – 51.

[2] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: a review," *Proc. IEEE*, vol. 86, pp. 974–997, May 1998.

[3] Y. Wang, Q.-F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *IEEE Trans. Commun.*, vol. 41, pp. 1544 – 1551, Oct. 1993.

[4] P. Haskell and D. Messerschmitt, "Resynchronization of motion compensated video affected by atm cell loss," in *Proc. ICASSP*, Mar. 1992, pp. 545–548.

[5] Y. Zhang and K.-K. Ma, "Error concealment for video transmission with dual multiscale markov random field modeling," *IEEE Trans. Image Processing*, vol. 12, pp. 236–242, Feb. 2003.

[6] Q.-F. Zhu, Y. Wang, and L. Shaw, "Coding and cell-loss recovery in DCT-based packet video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, pp. 248–258, June 1993.

[7] D. Persson, T. Eriksson, and P. Hedelin, "Qualitative analysis of video packet loss concealment with gaussian mixtures," in *Proc. ICASSP*, May 2006, pp. II–961 – II–964.

[8] D. Persson and T. Eriksson, "A minimum mean square error estimation and mixture-based approach to packet video error concealment," in *Proc. ICASSP*, Apr. 2007.

[9] W. M. Lam, A. R. Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," in *Proc. ICASSP*, Apr. 1993, pp. 417–420.

[10] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. of Royal Stat. Soc. B*, vol. 36, pp. 192–226, 1974.

[11] "Prelinger archives," http://www.archive.org/details/prelinger, Online resource.

[12] J. Zhang and D. Ma, "Nonlinear prediction for Gaussian mixture image models," *IEEE Trans. Image Processing*, vol. 13, pp. 836–847, June 2004.

[13] F. Heitz, P. Perez, and P. Bouthemy, "Multiscale minimization of global energy functions in some visual recovery problems," in *CVGIP: Image Understanding archive*, Jan. 1994, pp. 125 – 134.