IEICE TRANS. INF. & SYST., VOL.E100–D, NO.9 SEPTEMBER 2017

LETTER *Special Section on Picture Coding and Image Media Processing*

A Study on Video Generation Based on High-Density Temporal Sampling

Yukihiro BANDOH[†a)], **Seishi TAKAMURA**[†], ***and*** **Atsushi SHIMIZU**[†], ***Senior Members***

SUMMARY In current video encoding systems, the acquisition process is independent from the video encoding process. In order to compensate for the independence, pre-filters prior to the encoder are used. However, conventional pre-filters are designed under constraints on the temporal resolution, so they are not optimized enough in terms of coding efficiency. By relaxing the restriction on the temporal resolution of current video encoding systems, there is a good possibility to generate a video signal suitable for the video encoding process. This paper proposes a video generation method with an adaptive temporal filter that utilizes a temporally over-sampled signal. The filter is designed based on dynamic-programming. Experimental results show that the proposed method can reduce encoding rate on average by 3.01 [%] compared to the constant mean filter.

key words: *high temporal resolution, adaptive temporal filter, dynamic programming*

1. Introduction

The improvement of the acquisition rate of video equipment has been provided by progress in semiconductor technology. The increase in the temporal resolution of a video signal can improve subjective video quality and analysis accuracy for motion. The former aims at representing smooth movement on a display by approaching the perceptual limit of the temporal resolution. It assumes that the captured video is played at the display frame-rate. The latter aims at high accuracy analysis for fast moving objects by using a video signal captured at the temporal resolution beyond the perceptual limit. Note that it assumes that the captured video is replayed in slow motion for analysing fast motion.

Current video coding systems are designed on the premise that all captured frames are played on a display. Thus, video signals are captured at the display frame-rate. Additionally, the above premise is applied to pre-filters which are an adjustment process prior to the encoder. For example, conventional pre-filters for a non-scalable encoder [1] are designed on the assumption that its input signal and its output signal keep the same frame-rate. Conventional pre-filters for a temporal scalable encoder [2] supports variable frame-rate partially, but these do not assume filter design based on sampling with higher time resolution than display frame-rate.

By relaxing the restriction on the temporal resolution of current video encoding systems, there is a good possibility to generate a video signal suitable for the video encoding process. So, we focus on the usage of a temporally over-sampled video signal. An example of this is the case to generate a video signal for playing at normal speed (e.g. 30Hz) using the temporally over-sampled frames (e.g. 1000Hz) as an input signal. This is because it is possible to control the generation process with high temporal resolution so as to reduce the coded bits for the generated video signal. In [3], [4] and [5], it is demonstrated that temporal filters designed by the above-mentioned concept can generate a video signal suitable for the encoding process.

In this paper, as an extension of our previous studies, we propose a video generation method with an adaptive temporal filter that utilizes temporally over-sampled frames. The proposed method extends a temporal filter design [4] by introducing scheme to select sampling position [3] and a metric to evaluate encoding rate and subjective quality of generated video for filter design.

2. Design of Temporal Filter

2.1 Notations on Design of Temporal Filter

For preparation, we provide notations and terminologies for describing the proposed algorithm of the filter design. Here, for simplicity, and with no loss of generality, the case of a one-dimensional signal is given. The i-th frame generated by the temporal filter with $(2\Delta + 1)$-tap is given as follows:

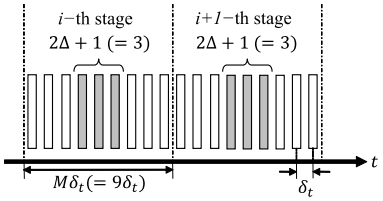
$$\hat{f}(x, iM\delta_t, \boldsymbol{w}_i, p_i) = \sum_{j=-\Delta}^{\Delta} w_i[j] f(x, (iM + \lfloor \frac{M}{2} \rfloor + p_i + j)\delta_t) \quad (1)$$

where $f(x,t)$ $(x = 0, \cdots, X-1)$ is a one-dimensional signal sampled at position x in the t-th frame as a reference frame of the temporal filter. Each frame is sampled at $t = j\delta_t$ $(j = 0, 1, \cdots)$, letting δ_t be the frame interval. $\boldsymbol{w}_i = (w_i[-\Delta], \cdots, w_i[\Delta])$ are the filter coefficients of the reference frames and satisfy $\sum_{j=-\Delta}^{\Delta} w_i[j] = 1$. p_i plays a role in shifting the position where the temporal filter applies and its value range is $0, \cdots, \pm P$. M is a down-conversion ratio that determines the frame rate of the sequence generated by the temporal filter; Eq. (1) says the frame-rate of the generated sequence is $\frac{1}{M\delta_t}$ in the case of $p_i = 0$. Suppose that $2\Delta + 2P + 1 \leq M$. As a special case of the above-mentioned filter of Eq. (1), a filter with constant a coefficient $\frac{1}{2\Delta+1}$ is called mean filter. Figure 1 illustrates the reference

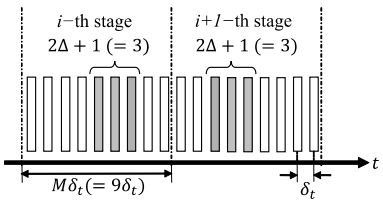
Manuscript received December 18, 2016.
Manuscript revised April 18, 2017.
Manuscript publicized June 14, 2017.

[†]The authors are with NTT Media Intelligence laboratories, NTT Corporation, Yokosuka-shi, 239–0847 Japan.
a) E-mail: bandou.yukihiro@lab.ntt.co.jp

DOI: 10.1587/transinf.2016PCL0009



(a) In the case of $p_{i-1} = p_i = p_{i+1} = 0$



(b) In the case of $p_{i-1} = 0, p_i = 1, p_{i+1} = -1$

Fig. 1 Reference frames of temporal filter denoted by Eq. (1) in the case of $\Delta = 1$, $M = 9$.

frames used by the temporal filter as gray rectangles in the case of $\Delta = 1$, $M = 9$. Figure 1 (a) and (b) show an example of $\Delta = 1$, $M = 9$, $p_{i-1} = p_i = p_{i+1} = 0$ and that of $p_{i-1} = 0, p_i = 1, p_{i+1} = -1$, respectively.

As the candidates of the filter coefficient $\boldsymbol{w}_i$, we prepare N sets of coefficient vectors (abbreviated to CV henceforth); $\boldsymbol{\gamma}_n = (\gamma_n[-\Delta], \cdots, \gamma_n[\Delta])$ where $n = 0, \cdots, N-1$. In parallel, we prepare $2P + 1$ candidates of p_i which is referred to as shifted position (abbreviated to SP henceforth). We find the best CV and SP for each generated frame among these $N \times (2P + 1)$ candidates. In what follows, the set of all candidates of CV is called *dictionary*, and let $\boldsymbol{\Gamma}_N = (\boldsymbol{\gamma}_0, \cdots, \boldsymbol{\gamma}_{N-1})$ dictionary. Each element in the dictionary is called *atom*, which is represented by $\boldsymbol{\gamma}_n$. Each atom is identified using an index $(n = 0, \cdots, N-1)$ which is called *atom index*.

2.2 Criterion for Optimization

As a measure for designing the temporal filter, we use the amount of coded bits for frames generated with the temporal filter. Let us suppose that the coded bits are obtained as the output of a lossless encoder with motion-compensated prediction (MC prediction). Let $B[k]$ $(k = 0, \cdots, K-1)$ be the k-th segment that is a one-dimensional sub-region with A pixels in frame $\hat{f}(x, iM\delta_t, \boldsymbol{w}_i, p_i)$. When segment $B[k]$ in the frame $\hat{f}(x, iM\delta_t, \boldsymbol{w}_i, p_i)$ is predicted from the reference frame $\hat{f}(x, (i-1)M\delta_t, \boldsymbol{w}_{i-1}, p_{i-1})$ by using estimated displacement $d[k]$ for the segment $B[k]$, MC prediction error becomes as follows.

$$e_i(x, \boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}) = \hat{f}(x, iM\delta_t, \boldsymbol{w}_i, p_i) - \hat{f}(x - d_i[k], (i-1)M\delta_t, \boldsymbol{w}_{i-1}, p_{i-1}) \quad (2)$$

where $\boldsymbol{d}_i = (d_i[0], \cdots, d_i[K-1])$. We evaluate the amount of coded bits generated by an encoder that supports MC prediction as follows:

$$\Psi(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}) = R_e(\boldsymbol{e}_i(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1})) + R_d(d_i[0], \cdots, d_i[K-1]) + R_h \quad (3)$$

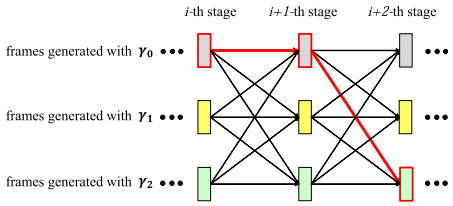


Fig. 2 Dependency among generated frames in neighbouring stages.

where $\boldsymbol{e}_i(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1})$ is the X-dimensional vector whose x-th element is $e_i(x, \boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1})$, and $R_e(\boldsymbol{e}_i(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}))$ is the amount of coded bits for MC prediction error. $R_d(d_i[0], \cdots, d_i[K-1])$ is the amount of coded bits for estimated displacements $d_i[0], \cdots, d_i[K-1]$, and R_h is the amount of coded bits for header syntax.

Since we use a lossless encoder as mentioned above, the MC prediction error depends on only the current frame (the generated i-th frame) and the reference frame (the generated $i-1$-th frame). Thus, Eq. (3) says the amount of coded bits $\Psi()$ depends on both $\boldsymbol{w}_i$, p_i for the generated i-th frame and $\boldsymbol{w}_{i-1}$, p_{i-1} for the generated $i-1$-th frame.

Additionally, in order to evaluate the subjective quality of the generated frames, we introduce the following:

$$\Phi[\boldsymbol{w}_i, p_i] = \sum_{k=0}^{M-1} \sum_{x=0}^{X-1} \{f(x, (iM+k)\delta_t) - \hat{f}(x, iM\delta_t, \boldsymbol{w}_i, p_i)\}^2$$

The internal sum in the above equation is the sum of square difference between a generated frame and each original frame in the time interval $iM\delta \le t < iM\delta + M\delta$ where the i-th generated frame is representative of all frames.

In order to reduce the amount of coded bits while keeping the subjective quality of the generated frames, as the criterion for optimizing the temporal filter, we use the following evaluation metric for the i-th frame:

$$\Xi[(\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}] = \Psi[\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}] + \lambda\Phi[\boldsymbol{w}_i, p_i] \quad (4)$$

2.3 Optimization of Filter Coefficients and Shifted Positions

In order to minimize the evaluation metric over all generated frames, it is necessary to find the J/M sets of CVs and SPs that satisfy the following equation:

$$(\boldsymbol{w}_0^*, \cdots, \boldsymbol{w}_{J/M-1}^*, p_0^*, \cdots, p_{J/M-1}^*) = \underset{\substack{\boldsymbol{w}_0, \cdots, \boldsymbol{w}_{J/M-1} \in \boldsymbol{\Gamma}_N \\ p_0, \cdots, p_{J/M-1}}}{\arg\min} \sum_{i=1}^{J/M-1} \Xi[\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}] \quad (5)$$

Their relationship with regard to inter-frame prediction is shown using a directed diagram as shown in Fig. 2. Figure 2 illustrates the case of $N = 3$ and $P = 0$ which corresponds to fixed SP. In this figure, each stage has three candidate frames that are generated with three kinds of atoms

γ_0,γ_2,γ_2. Gray rectangles, yellow ones and green ones indicate frames generated with atom γ_0, atom γ_1 and atom γ_2, respectively. In the diagram, a predicted frame and its reference frame are connected by an arrow. A predicted frame is depicted as the frame at the right end of each arrow, and its reference frame is depicted as the frame at the left end of each arrow. There are $N^{J/M}$ possible paths if the diagram has J/M stages. Assuming that the heavy red lines represent the optimum path, we have to find the optimum path among all paths. The brute force method to find the optimum solution of the above solution takes exponential time $O((N \times (2P+1))^{J/M})$ and is not realistic in terms of complexity.

Considering that $\Xi[\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}]$ depends on $\boldsymbol{w}_i$, $\boldsymbol{w}_{i-1}$, p_i and p_{i-1}, the optimization problem of Eq. (5) can be solved in polynomial time by recourse to dynamic programming. We define the following $S_i(\boldsymbol{w}_i, p_i)$ for each $\boldsymbol{w}_i$ and p_i $(i = 1, \cdots, J/M-1)$:

$$S_i(\boldsymbol{w}_i, p_i) = \min_{\substack{\boldsymbol{w}_0,\cdots,\boldsymbol{w}_{i-1}\in\boldsymbol{\Gamma}_N \\ p_0,\cdots,p_{i-1}}} \sum_{j=1}^{i} \Xi[\boldsymbol{w}_j, \boldsymbol{w}_{j-1}, p_j, p_{j-1}] \tag{6}$$

Since $\Xi[\boldsymbol{w}_j, \boldsymbol{w}_{j-1}, p_j, p_{j-1}]$ depends on just $\boldsymbol{w}_{j-1}$ and p_{j-1} if $\boldsymbol{w}_j$ and p_j are fixed, $S_i(\boldsymbol{w}_i, p_i)$ can be expressed using the following recursive equation:

$$\begin{aligned} &S_i(\boldsymbol{w}_i, p_i) \\ &\quad = \min_{\substack{\boldsymbol{w}_{i-1}\in\boldsymbol{\Gamma}_N \\ p_{i-1}}} \{\Xi[\boldsymbol{w}_i, \boldsymbol{w}_{i-1}, p_i, p_{i-1}] + S_{i-1}(\boldsymbol{w}_{i-1}, p_{i-1})\} \end{aligned} \tag{7}$$

$S_{i-1}(\boldsymbol{w}_{i-1}, p_{i-1})$ can be computed using a similar recursive equation and the computed value is reused in computing $S_i(\boldsymbol{w}_i, p_i)$. Using recursive Eq. (7), the computation of $S_i(\boldsymbol{w}_i, p_i)$ results in the selection of the best CV in the directory $\boldsymbol{\Gamma}_N$ and the best SP.

Based on the definition of Eq. (6), the minimization problem of Eq. (5) becomes as follows:

$$\min_{\substack{\boldsymbol{w}_{J/M-1}\in\boldsymbol{\Gamma}_N \\ p_{J/M-1}}} S_{J/M-1}(\boldsymbol{w}_{J/M-1}, p_{J/M-1}) \tag{8}$$

Using recursive Eq. (7), the minimization problem of Eq. (8) is to find the optimum solution $(\boldsymbol{w}_0^*, \cdots, \boldsymbol{w}_{J/M-1}^*, p_0^*, \cdots, p_{J/M-1}^*)$ from among $\{N\times(2P+1)\}^2 J/M$ candidates. This reduces the time complexity from exponential to polynomial.

3. Experiments

We performed the following experiments in order to evaluate the adaptive temporal filter (called *proposed filter*) designed based on the algorithm described in **2.3** from the viewpoint of coding efficiency. We captured original sequences with a high speed camera in 24 bit RGB format at 1000 Hz. Then, the color format of the sequences was converted from RGB data to YCbCr data, and the Y-data in 8 bit gray scale was used in these evaluation experiments. Each sequence consisted of 900 frames at 1000 [Hz]. The contents were a scene of two kinds of skyscrapers ("Building A", "Building B") and a cruise ship ("Ship") which were taken in panning photography. As the amount of coded bits defined in Eq. (3), we evaluated the H.264/AVC lossless stream created by x264 encoder with the lossless mode. As the GOP structure of the encoder, the head frame in a sequence was set to I-picture and the all subsequent frames were set to P-picture.

Table 1 Atoms in dictionary used in the experiments

Atom index	coefficinet vector
0	(1/3,1/3,1/3)
1	(29/96,19/48,29/96)
2	(13/48,11/24,13/48)
3	(35/96,13/48,35/96)
4	(19/48,5/24,19/48)

Table 2 The amount of coded bits

	mean filter [bits/pixel]	proposed filter [bits/pixel]	reduction ratio [%]
Building A	2.54	2.49	2.04
Building B	2.80	2.77	1.23
Ship	3.69	3.48	5.77

We compared proposed filter with mean filter in terms of coding efficiency. Table 2 shows the amount of coded bits for sequences generated by proposed filter and that by mean filter. As a common configuration of both filters, we set $M = 32$ and $\Delta = 1$ which corresponds to a three-tap filter. As the atoms of the proposed filter, we prepared five kinds of CVs shown in Table 1. We used five candidates $(0, \pm1, \pm2)$ of SPs. The results confirm that the proposed filter can achieve a reduction of coded bits on average by 3.01 [%] compared to the mean filter. This is because the proposed filter was designed considering the structure of inter-frame prediction. We confirmed that it was possible to keep an indetectable difference between the sequences yielded by the proposed filter and those yielded by the mean filter in terms of subjective quality. In the above-mentioned configuration where candidates of SPs are $0, \pm1, \pm2$, the shifted position of the proposed filter causes the inconsistency in the sampled temporal position at maximum 4 msec length. From the biochemical viewpoint [6], it is shown that the minimum detectable temporal resolution of the human visual system is 7 to 5 msec length. So, the above results on indetectable difference between both filter agree with the findings in the biochemical research.

4. Conclusion

This paper proposes the design of an adaptive temporal filter for the purpose of optimizing the video generation process in terms of coding efficiency. The filter design is formulated as a minimization problem that evaluates the weighted sum of the amount of coded bits and the metric of subjective quality for sequences generated by the filter. The minimization problem is resolved with a dynamic programming based approach. The proposed method can reduce the amount of

coded bits on average by 3.01 [%] compared to the constant mean filter, subject to keeping the subjective video quality. The experimental results show the bit-rate savings brought by increasing the temporal resolution of the video signal and designing the temporal filter for video encoding.

References

[1] L.J. Kerofsky, R. Vanam, and Y.A. Reznik, "Improved adaptive video delivery system using a perceptual pre-processing filter," Proc. IEEE Global Conf. Signal & Inf. Process., pp.1058–1062, 2014.

[2] A. Golwelkar and J.W. Woods, "Motion-compensated temporal filtering and motion vector coding using biorthogonal filter," IEEE Trans. Circuits Syst. Video, vol.CSVT-17, no.4, pp.417–428, 2007.

[3] Y. Bandoh, S. Takamura, K. Kamikura, and Y. Yashima, "Temporal down-sampling algorithm of high frame-rate video for reducing inter-frame prediction error," Proc. IEEE Int. Conf. Image Process., pp.1017–1020, 2009.

[4] Y. Bandoh, S. Takamura, and A. Shimizu, "Video generation algorithm based on high temporal-resolution imaging," Proc. IEEE Int. Conf. Image Process., pp.2172–2176, 2016.

[5] Y. Bandoh, S. Takamura, and A. Shimizu, "Encoding-oriented video generation algorithm based on control with high temporal resolution (in Japanese)," FIT, vol.3, pp.5–9, 2016.

[6] L. Spillmann and J.S. Werner, Visual perception the neurophysiological foundations, Academic Press, 1990.