# Complexity Reduction Techniques For Long-Term Memory Motion Compensated Prediction Based On Spectral Distortion Analysis

*Waqar Zia*[1] and *Faisal Shafait*[2]

[1]BenQ Mobile GmbH & Co. OHG
Munich, Germany
waqar.zia@benq.com

[2]German Research Center for
Artificial Intelligence (DFKI GmbH)
Kaiserslautern, Germany
faisal.shafait@dfki.de

**Abstract.** Long-term memory motion compensated prediction and up to ¼-pel accurate motion compensation contribute a considerable portion of the compression gain provided by H.264/AVC over its predecessors. This paper investigates the factors contributing to the spectral distortions introduced in the digitized video signal. A quantitative analysis shows that fractional-pel interpolation is the main source of these spectral distortions. Using these results, two techniques are proposed for reducing computational complexity with negligible effects on the quality of the video. Simulation of the proposed techniques show up to 56% complexity reduction compared to the reference scheme without any significant decrease in signal-to-noise ratio.

***Index Terms***— Spectral distortion, Complexity reduction, Long-term memory motion compensated prediction, Fractional-pel interpolation

## 1. INTRODUCTION

Long-term memory motion compensated prediction (LTM-MCP) [1] and up to ¼-pel accurate motion compensation are two of the most important features introduced in H.264/AVC. LTM-MCP is considered to be a giant leap in the technology of hybrid video coding, similar to moving from intra-prediction to inter-prediction back in 1980s [1]. Compression and visual performance of inter-prediction process in earlier standards like ITU-T Rec. H.261 was considerably improved by adding fractional-pel interpolation in ISO/IEC MPEG-1 on the cost of some added complexity. Fractional-pel interpolation became an integral part of all the following standards like H.262 | MPEG-2 and H.263. However, all these standards allow only ½-pel interpolation. Further interpolation using the same simple interpolation filters did not show any further improvements. Since the time of introduction of MPEG-1, silicon technology has progressed significantly and at present more complexity can be afforded to achieve better compression performance.

To this end several new inter-prediction options were introduced in H.263+, including Annex N: Reference Picture Selection Mode (RPS), which was later improved as LTM-MCP [1], [2] and is incorporated in H.264/AVC and MPEG-4 Advanced Simple Profile.

However, despite the popularity of LTM-MCP and fractional-pel interpolation, their inter-dependence to achieve compression gain has been scantily addressed in literature. Wiegend et al. [2] presented an analysis for a few macroblocks in a single frame based on bilinear ½-pel interpolation. In [3], [4] analysis of the aliasing distortion introduced by fractional-pel interpolated motion compensation is presented. The dominant source of spectral distortion is reported to be the digitization of video. In this paper, the comparative impact of various phenomenon leading to the gain of LTM-MCP is given for practical systems. Also the comparison of various sources of spectral distortions, including video digitization is considered. The obtained results enable us to establish the relation between LTM-MCP and fractional-pel interpolation. This understanding is later used to introduce two complexity reduction techniques for real-time video processing systems.

The structure of this paper is as follows. In section 2 the system model for an H.264/AVC encoding system is given. In section 3 addresses the dependence of LTM-MCP on fractional-pel interpolation. Section 4 proposes two complexity reduction techniques based on results from section 3. Section 5 shows simulation results. Conclusion is presented in section 6.

## 2. SYTEM MODEL

Fig. 1 shows a simplified block diagram of an H.264/AVC encoding system. Analog video input $s(t)$ based on a format supported by ITU-R Rec. BT.601 is first sampled by an A/D converter at the input. Necessary pre- and post-processing is also performed. The resultant digitized data $s_{\Delta x,\Delta y}(x, y)$ is passed on to an optional pre-processor that can perform a variety of functions including noise filtering, format conversion and downscaling if required.

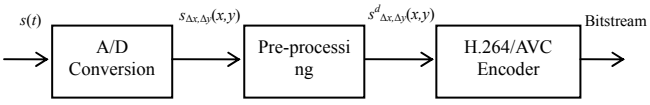


**Fig. 1.** Block Diagram of H.264/AVC Encoding System

Following this the data $s^{d}_{\Delta x,\Delta y}(x, y)$ is fed to the H.264/AVC video encoder for compression. Compressed bitstream is the output of the system.

## 3. Prediction gain of LTM-MCP

In [1], [2] several phenomena have been reported to contribute in the prediction gain achieved by using LTM-MCP. Here, we comment on the relative impact of the contributors in a practical encoding system.

*Repetition of Image scene content* can provide compression gain but it is limited in practical systems because of small amount of memory used for long-term prediction and the relative speed of motion with respect to search range used for motion estimation. *Scene Cuts* can be indicated to the encoding system by external means as suggested [2]. The gain is possible only in this specialized system. *Noise Triggered Random Matches* are possible for scene content with considerable input noise**.**

*Spectral Distortions* are introduced in the processed video by the following ways:

### 3.1 Analog to digital conversion of video:

To measure this distortion, we proceeding on similar lines as in [3], assuming the spectrum $\boldsymbol{S}(j\Omega)$ of the analog input is band-limited to $\Delta\Omega$, where

$$\Delta\Omega = 2\pi/\Delta t \qquad (1)$$

The base-band spectrum $\boldsymbol{S}^{b}_{\Delta t}(e^{j\omega})$ of the digitized signal $s_{\Delta t}(t)$ is given as

$$\boldsymbol{S}^{b}_{\Delta t}(e^{j\omega}) = \frac{1}{\Delta t}\left(\boldsymbol{S}(j\Omega) + \boldsymbol{S}\left(j\Omega\cdot(1-\frac{\Delta\Omega}{|\Omega|})\right)\right) \qquad (2)$$

where $\Delta t$ is the sampling interval. A practical A/D conversion system based on a configuration suggested in 5 consists of an analog video source encoder (for example ADV7314 6) and a video decoder (for example SAA7118 7). For this system, using equation 2, overall aliasing component at $\omega_H$ is -100dB. $\omega_H$ is here defined as $0.8\omega_N$, $\omega_N$ being Nyquist frequency ($\omega \rightarrow \pi$)

The digitized signal $s_{\Delta t}(t)$ is arranged in 2D images given as $s_{\Delta x,\Delta y}(x, y)$ with corresponding frequency spectrum $\boldsymbol{S}(e^{j\omega_x}, e^{j\omega_y})$.

### 3.2 Downscaling:

The above digitized image is downscaled to SIF, which involves 4:2:2 to 4:2:0 conversion, and horizontal and vertical downscaling by a factor of 2. For example horizontal downsampling by a factor of $M_x$ gives us the frequency spectrum

$$\boldsymbol{S}_{d}\left(e^{j\omega_x}, e^{j\omega_y}\right) = \frac{1}{M_x}\sum_{i=0}^{M_x-1}\boldsymbol{S}\left(e^{j(\omega_x/M_x - 2\pi i/M_x)}, e^{j\omega_y}\right) \qquad (3)$$

For $M_x$=2 aliasing is caused by $\boldsymbol{S}(e^{j\omega_x}, e^{j\omega_y})$ for $\omega_x \geq \pi/2$, and is reduced by prior low-pass filtering of the data. Fig. 3 shows filter characteristics of a down sampling filter used in some applications [8]. As a result there is an aliasing component of -6.3dB at horizontal high frequencies $\omega_{xH}$. Similar aliasing distortion is introduced for vertical downsampling.

### 3.3 Fractional-pel Interpolation:

The process of interpolated prediction for $M^{\text{th}}$ fraction of a pel is shown in Fig. 2 for horizontal direction. It involves upsampling, filtering, shifting and finally downsampling of the samples at Full-Pel (FP) positions. Aliasing is introduced in the downsampling as given by equation 3 and $h_i(t)$ must be suitably selected to attenuate the upsampled spectrum above $\omega_x = \pi/M_x$.

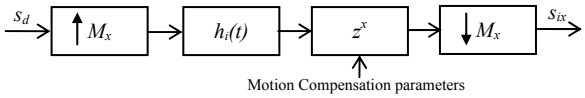


**Fig. 2.** Fractional-pel interpolation

The magnitude spectrum for the ½-pel (HP) and ¼-pel (QP) interpolation filters used in H.264/AVC for this purpose are shown in Fig. 3. The values of distortion for ½-pel and ¼-pel interpolation in either horizontal or vertical direction at $\omega_H$ are tabulated in Table 1. For ½-pel interpolation, simple bilinear filter is used and there is a pass-band attenuation of 1.8dB for high frequencies.

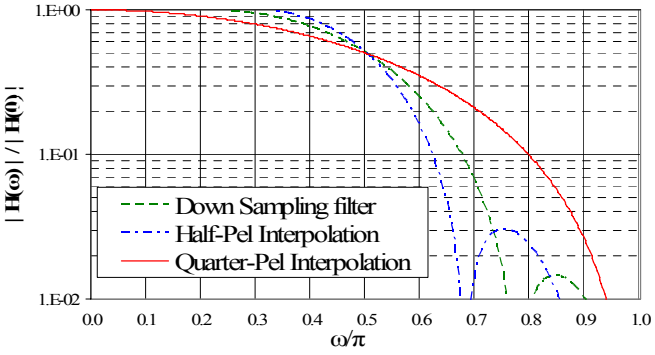


**Fig. 3.** Spectrum of downsampling and interpolation filters

Comparing the numerical values of the distortion for the three different phenomena tabulated in Table 1, we can see that fractional-pel interpolation process is a dominant contributor. Hence a considerable impairment is introduced for block matching. In Table 2 a relative Distortion Metric (DM) is assigned to various possible interpolation combinations for H.264/AVC. A lower number represents a relatively lower spectral distortion.

**Table 1.** Distortion Magnitude for various encoding processes

| Process | Distortion |
|---|---|
| Video Digitization | -100dB |
| Downscaling | -6.3dB |
| ½-pel interpolation | -8dB |
| ¼-pel interpolation | -4.7dB |

With LTM-MCP, there is possibility of finding a block match in one of the previous frames where the interpolation configuration introduces a lower distortion. This can occur in practice whenever there is camera or object motion, which is common in natural scene content. This fact results in the dominant portion of prediction gain for LTM-MCP compared to the other contributing phenomenon discussed in this section.

**Table 2.** Distortion Metric definitions for various interpolation configurations

| | Interpolation Configuration | | | | | |
|---|---|---|---|---|---|---|
| | FP-FP | FP-HP | FP-QP | HP-HP | QP-QP | HP-QP |
| Distort-ion Metric | 1 | 2 | 3 | 4 | 5 | 6 |

## 4. Complexity Reduction for LTM-MCP

As discussed in the preceding discussion, interpolated motion compensation introduces considerable distortion. LTM-MCP combats this by allowing inter-prediction using interpolation configurations with lower distortion. Also, compared to other contributing phenomenon, this can occur very frequently in natural scene content. Hence a large portion of block matches in long-term memory will have a lower DM compared to the match of the same block in the most recent image. This finding can be exploited to develop a complexity reduction technique, which will be verified later by simulations. The candidate motion estimators for this technique are the "Three Step Search" (3SS), "New Three Step Search" (N3SS) [9], "Four Step Search" (4SS) [10], "Diamond Search" (DS) and "Orthogonal Block Search" (OBS). We define two variant techniques with respect to the complexity reduction they offer and the performance loss.

The first technique is for medium level complexity reduction, and is described as follows. For a given block, let $S$ be the set of all the search points in a reference frame as specified by the motion estimation technique. Let $M$ be the set of all the block prediction modes defined in H.264/AVC.

For the most recent frame, perform motion estimation on each of the search points $k \in S$, for all the block prediction modes $i \in M$, and for all the blocks $j$ in the image for a given block mode $i$.

Store the values of the distortion metric $DM_{\min}(i, j)$ as given in Table 2 for each block $j$ belonging to block prediction mode $i$ given by

$$DM_{\min}(i, j) = DM(i, j, k)\big|_{E(i,j)=\min_{k\in S} E(i,j,k)} \quad (4)$$

where $E(i, j, k)$ is the cost or mean absolute difference (MAD) of inter-prediction for the current block at the given search position $k$. Hence the DM corresponding to the interpolation configuration of the best block match is stored in $DM_{\min}(i, j)$.

For each of the remaining frames, for any block $j$ belonging to block prediction mode $i$, perform motion estimation exclusively for $k \in S$ that satisfy the criteria

$$DM(i, j, k) \le DM_{\min}(i, j) \quad (5)$$

This technique skips motion estimation for any search position that corresponds to a higher DM compared to the DM of the best match in the most recent image. It will be shown in the following section that this corresponds to a large potion of the total search positions. Another similar technique that results a larger complexity reduction is given below

Start motion estimation from the most recent frame first, and proceed backward in temporal order. Let the cost of prediction at a search point $k \in S$ for a block $j$ with a block mode $i$ in a frame $f$ be $E_f(i, j, k)$. Let the minimum inter-prediction cost for that block for all frames up to the frame $f$ be $E_{\min}(i, j)$.

Initialize the parameter $DM_{\min}(i, j)$ with 6 and $E_{\min}(i, j)$ with the maximum perceivable inter-prediction cost. Update $DM_{\min}(i, j)$ using the following criteria

$$\begin{aligned}
&\text{if} : E_f(i, j, k) < E_{\min}(i, j) \\
&\text{then} : DM_{\min}(i, j) \leftarrow DM(i, j, k) \\
&E_{\min}(i, j) \leftarrow E_f(i, j, k) \\
&\text{fi}
\end{aligned}$$

For any block $j$ belonging to block prediction mode $i$, perform motion estimation exclusively for $k \in S$ that satisfies the criteria

$$DM(i, j, k) \le DM_{\min}(i, j) \quad (6)$$

This technique relies on cost estimate to make decision in step ii, and the loss of compression can be larger than the technique with medium complexity.

Let $s_i, i = 1, \ldots, 6$ be the number of search positions corresponding to the distortion metric $i$ in the reference scheme. Using above-mentioned criterion, the number of search positions at each DM are reduced to $r_i$. Then, the complexity reduction for LTM-MCP

achieved by the proposed scheme is given by:

$$\rho = 1 - \frac{\sum_{i=1}^{6} r_i c_i}{\sum_{i=1}^{6} s_i c_i} \qquad (7)$$

where $c_i$ is the number of cycles required to execute the search for the $i^{th}$ distortion metric.

## 5. Simulation Results

The techniques presented in the previous section work on the argument that a dominant contribution of prediction from long-term memory is to find matches with lower DM. Simulation results on test sequences depicting typical motion scenarios in natural scene content verify this argument. The selected test sequences are "Foreman", "Flower Garden," and "Mobile and Calendar". 10 previous frames are used for inter-prediction at 768 kbps. The compensated blocks are divided in two categories. Category I consists of blocks that had a better match in one of the previous frames with a larger interpolation distortion compared to the best block match in the most recent frame. Category II consists of the rest of the blocks that have a better match with a similar or lower interpolation distortion compared to the best block match in the most recent frame. Table 3 shows the portion of the total compression gain achieved by the first category. Also the amount of searches required at various interpolation configurations for this category are given as a fraction of total searches performed for the particular interpolation configuration.

**Table 3.** Relative compression gain and search effort for Category I blocks

| Test Sequence | Gain % | Search positions % | | | | |
|---|---|---|---|---|---|---|
| | | FP-HP | FP-QP | HP-HP | QP-QP | HP-QP |
| Foreman | 2 | 14 | 33 | 48 | 53 | 66 |
| Flower Garden | 1.5 | 24 | 42 | 70 | 71 | 77 |
| Mobile and Calendar | 4 | 11 | 28 | 49 | 53 | 67 |

We observe that for a small portion of the over-all compression gain, a large number of searches have to be performed for category I blocks. This verifies the discussion in section III. The complexity reduction techniques given in section IV focus on eliminating the searches tabulated in Table 3, which take up considerable portion of processing time without giving a significant prediction gain. To demonstrate this, these techniques are simulated in an environment that is close to a practical real-time processing system on an embedded DSP [8]. For this processor, operations required per pixel for various interpolation configurations are given in Table 4.

**Table 4.** Operations per pixel for various block interpolation configurations

| | Interpolation Configuration | | | | | |
|---|---|---|---|---|---|---|
| | FP-FP | FP-HP | FP-QP | HP-HP | QP-QP | HP-QP |
| Operations /pixel | 5 | 21.6 | 24.6 | 43.8 | 41.6 | 47.8 |

The reference motion estimation technique used is the 4SS [10]. The fractional-pel refinement is performed as discussed in [11]. The two proposed techniques are compared with the reference technique for complexity reduction and performance. The first technique is referred to as "Medium Complexity" technique, the second technique will be referred to as "Low Complexity" technique while the reference scheme will be referred to as "Full Complexity" technique. The performance curves are also compared to the case when only one past reference frame is used. Also, the searches performed for each of the techniques are given in Table 5 as a fraction of the full-pel search position.

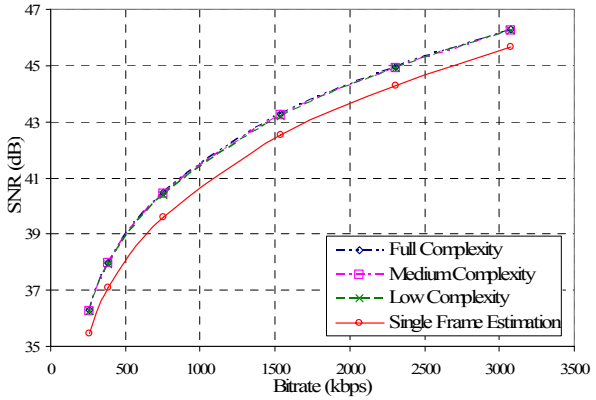


**Fig. 4.** Rate-distortion curves for test sequence "Forman"

Fig. 4 shows the rate distortion curves for the test sequence "Foreman." The SNR performance of the proposed techniques lies within 0.16% of the reference technique. A comparison with the case when a single frame is used as a reference shows that the techniques benefit from most of the advantage that is achievable by multi-frame motion compensation. At the same time there is a 48% complexity reduction as given in Table 6.

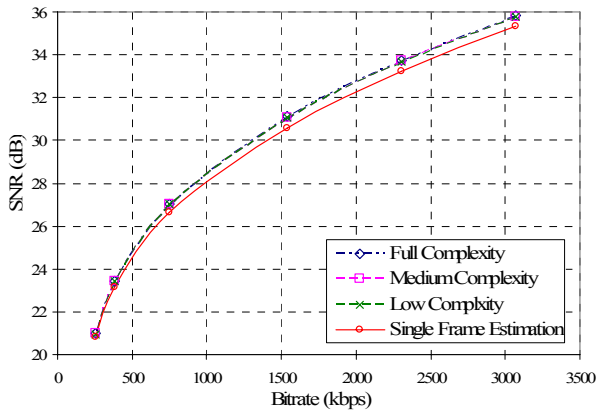


**Fig. 5.** Rate-distortion curves for "Flower Garden"

Fig. 5 shows the results for the test sequence "Flower Garden." With almost same compression performance as the reference scheme, there is a complexity reduction of up to 56%.

Fig. 6 shows the results for the test sequence "Mobile and Calendar." The performance difference of the techniques can be observed, and the medium complexity technique shows lower loss compared to the low complexity technique. This is contributed by high texture and complex motion of objects in the sequence.

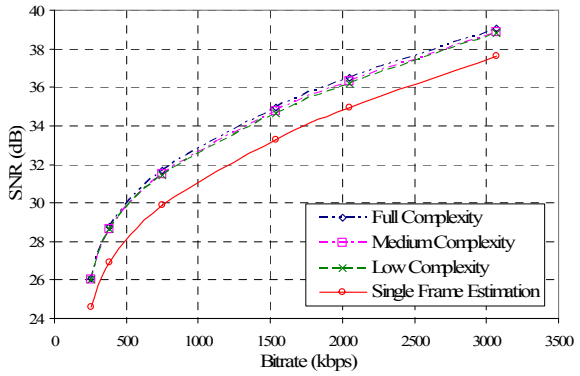


**Fig. 6.** Rate-distortion curves for "Mobile and Calendar"

**Table 5.** Searches performed at each interpolation configuration as a fraction of full-pel search positions

| Test Sequence | Technique (Complexity) | Searches performed (Normalized) | | | | |
|---|---|---|---|---|---|---|
| | | FP-HP | FP-QP | HP-HP | QP-QP | HP-QP |
| Foreman | Full | 0.2083 | 0.1152 | 0.2083 | 0.2083 | 0.0931 |
| | Medium | 0.1786 | 0.0911 | 0.1084 | 0.0969 | 0.0246 |
| | Low | 0.1606 | 0.0788 | 0.0815 | 0.074 | 0.0187 |
| Flower Garden | Full | 0.221 | 0.1332 | 0.221 | 0.221 | 0.0878 |
| | Medium | 0.1678 | 0.0933 | 0.0663 | 0.0637 | 0.0133 |
| | Low | 0.1593 | 0.0857 | 0.0566 | 0.0546 | 0.0113 |
| Mobile and Calendar | Full | 0.2105 | 0.1181 | 0.2105 | 0.2105 | 0.0924 |
| | Medium | 0.1869 | 0.0985 | 0.1072 | 0.0998 | 0.0249 |
| | Low | 0.1707 | 0.0861 | 0.073 | 0.0687 | 0.0166 |

However, comparing it with the curve for single frame motion estimation, the techniques make most out of the multi-frame motion compensation. The complexity reduction is approximately 49%.

**Table 6.** Percentage complexity reduction ($\rho$) and SNR loss

| Sequence | Low Complexity | | Medium Complexity | |
|---|---|---|---|---|
| | Complexity reduction % | Max. SNR loss % | Complexity reduction % | Max. SNR loss % |
| Foreman | 48.1 | 0.16 | 39.1 | 0.11 |
| Flower Garden | 56.1 | 0.15 | 52.6 | 0.11 |
| Mobile and Calendar | 49.3 | 0.75 | 38.3 | 0.51 |

The proposed techniques get conveniently integrated in the existing motion estimators. These techniques can be used on the top of other complexity reduction techniques, like the one discussed in [12].

## 6. Conclusion

The factors contributing to the compression gain for LTM-MCP are discussed. The relative impact of each of these is analyzed for real-time systems. Especially, the impact of spectral distortions in the form of aliasing on the inter-prediction process is studied in detail. It is shown that for practical systems distortions introduced by fractional-pel interpolation are dominant over other factors like aliasing distortion originating from digitization of images, and that the most important contribution of LTM-MCP is to avoid these spectral distortions. LTM-MCP achieves this by enabling block matches with lower distortion in one of the previous frames. By skipping motion estimation at search points with high distortion, two complexity reduction techniques are introduced. These techniques show a complexity reduction of up to 56% with a minimal loss of SNR performance. Simulations are performed to verify the theoretical results presented in the paper.

## REFERENCES

1. T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Transaction on Circuits and Systems for Video Technology (CSVT)*, Vol. 9 (February 1999) 70-84
2. T. Wiegand and B. Girod, Multi-Frame Motion-Compensated Prediction for Video Transmission. Norwell, MA: Kluwer (2001)
3. T. Wedi and H. G. Musmann "Motion- and Aliasing-Compensated Prediction for Hybrid Video Coding," *IEEE CSVT*, Vol. 13, No. 7

(July 2003)
4. T. Wedi, “Hybrid Video Coding Based on High-Resolution Displacement Vectors,” *Electronic Imaging 2001: VCIP*, San Jose, USA (January 2001)
5. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. 2nd edition, Prentice Hall (February 1999)
6. “Multiformat 216 MHz Video Encoder,” Analog Devices http://www.analog.com/UploadedFiles/Data_Sheets/276167652ADV7314_0.pdf (August 2003)
7. *Video Data Handbook*. Philips Semiconductors, CA (1995)
8. “TM-1300 Media Processor,” CA, Philips Semiconductors, http://www.semiconductors.philips.com/acrobat_download/datasheets/TM1300_T_2.pdf (May 2000)
9. R. Li, B. Zeng, and M. L. Liou, “A new three-step search algorithm for block motion estimation”, *IEEE CSVT*, Vol. 4 (August 1994) 438-443
10. L. M. Po and W. C. Ma, “A novel four-step search algorithm for fast block motion estimation”, *IEEE CSVT*, Vol. 6 (June 1996) 313-317
11. T. Wiegand, B. Lincoln, and B. Girod, “Fast Search for Long-Term Memory Motion-Compensated Prediction,” *Proceedings of the IEEE International Conference on Image Processing*, Chicago, USA (October 1998)
12. H. Chung, D. Romacho, and A. Ortega, “Fast long-term motion estimation for H.264 using multiresolution search,” *Proceedings ICIP* (2003)