A STAIRCASE TRANSFORM CODING SCHEME FOR SCREEN CONTENT VIDEO CODING

Cheng Chen, Jingning Han, Yaowu Xu, and James Bankoski

WebM Codec Team, Google Inc.
1600 Amphitheatre Parkway, Mountain View, CA 94043
Emails: {chengchen, jingning, yaowu, jimbankoski}@google.com

ABSTRACT

Demand for screen content videos that contain computer generated text and graphics is growing. They are very different from natural videos, because they include much sharper edge transitions and very repetitive patterns. On this type of material, the efficacy of the conventional discrete cosine transform (DCT) is questionable because it relies on the assumption that a Gauss-Markov model leads to a base-band signal. However, the assumption may not hold true for screen content material. This work exploits a class of staircase transforms. Unlike the DCT whose bases are samplings of sinusoidal functions, the staircase transforms have their bases sampled from staircase functions, which better approximate the sharp transitions often encountered in the context of screen content. The staircase transform is integrated into a hybrid transform coding scheme, in conjunction with DCT. It is experimentally shown that the proposed approach provides an average of 2.9% compression performance gains in terms of BD-rate reduction. A perceptual comparison further demonstrates that the use of staircase transform achieves substantial reduction in ringing artifact due to the Gibbs phenomenon.

Index Terms— Transform coding, screen content, ringing artifact, staircase transform

1. INTRODUCTION

Transform coding is widely adopted in image and video coding to reduce the spatial correlation and compact energy in the transform domain. The DCT has been proven to be a close approximation to the optimal Karhunen-Loeve transform (KLT) under the Gauss-Markov model and has been used by most modern codecs that are designed for compression of natural image and video contents, whose statistical properties are well captured by such model. Recent trends in online video services suggest greater demand for screen content videos. Screen content videos are typically composed of computer-generated text and graphics, with sharp edge transition between smooth regions. This makes their statistical characteristics distinctive from the general Gauss-Markov model, and hampers the efficacy of DCT in the context of screen content coding.

Consider a zero-mean auto-regressive process with unit variance

$$x(n) = \rho x(n-1) + e(n),$$

where $e(n)$ follows Gaussian distribution and ρ is the correlation coefficient. Its power spectral density function can be written as

$$S(w) = \frac{1-\rho^2}{1+\rho^2-2\rho cos(w)},$$

where $w \in (-\pi, \pi)$. The base-band signal can be described by a Gauss-Markov random process and therefore the near optimal transform should have its bases in the form of sinusoidal functions. The projection onto these bases will compact most energy in the low-frequency coefficients [1]. In practice, transform coding is incorporated with spatial or motion compensated prediction, and is applied to the prediction residual. The spatial prediction introduces certain boundary conditions to the auto-regressive process and affects the phase offset of the sinusoidal bases. It has been shown in [2, 3] that the optimal KLT is well approximated by a close relative of the sine transform. The motion compensated prediction residuals largely retain the above Gauss-Markov model, albeit with lower correlation coefficient values. Hence the DCT is commonly used for the inter frame prediction residuals for natural video content.

In the context of screen content coding, the prediction residuals often take the form of impulse or staircase signals due to the sharp transition across plain regions. This makes their power spectral density function rather flat across the spectrum, or it makes them contain substantially more power in the high frequency region. Naturally the sinusoidal transform bases would not be able to describe such signal in a sparse form. Current generation video codecs support variable transform block sizes. In principle, one can break down the region into smaller transform block sizes for a better fit. However, this adversely affects coding performance in smooth regions when compared to large transform block sizes [1]. Recent research efforts address this issue by allowing the codec to skip the transform in either vertical, horizontal, or both directions [4, 5] via a mode called transform skip mode (TSM). Substantial coding gains are achieved for screen content coding [6]. Other notable approaches to improve screen content coding efficiency include intra frame motion compensated prediction (IntraBC) [7] and an adaptive color table and index map coding scheme [8]. A comprehensive survey of the related techniques can be found in [9].

This work is inspired by the observation that a staircase-shape residual signal can be compactly captured by a set of staircase functions. It exploits a class of orthonormal transforms whose bases are samplings of staircase functions, as an alternative to the DCT and TSM. It is noteworthy that skipping transform is essentially applying an identity matrix, which itself is a special case of the staircase transform. The Walsh-Hadamard transform (WHT) is a classic staircase transform that has long been used for fast approximation of the DCT [10] and for lossless compression [11]. Another variant, the Haar transform, has been extensively studied in literature of wavelet transform [12]. This work leverages the staircase transform in combination with an adaptive 1D/2D transform coding scheme to improve the compression performance for screen content. It is incorporated in a rate-distortion optimization framework, along with a conventional DCT, the TSM, and a variable transform block size. Experimental results show that it provides solid additional coding gains.

A well-known shortcoming in sinusoidal transform coding for a sharp edge transition is the ringing artifact, commonly known as the Gibbs phenomenon. This phenomenon is caused by the truncation of high frequency coefficients in the quantization process [13]. It is typically addressed by post-filtering to remove the artifact, which may result in blurring actual edges. The proposed staircase transform has the added benefit of reducing this problem. The basis functions are capable of describing the staircase transition directly, instead of through a superposition of a series of sinusoidal functions. The perceptual quality improvement is shown in the experiments.

2. STAIRCASE TRANSFORM

We review the Walsh-Hadamard and Haar transforms here and evaluate their theoretical coding gains with respect to the DCT.

2.1. Walsh-Hadamard Transform

The Walsh-Hadamard transform can be represented by an orthonormal and symmetric matrix . Let W_m be a $2^m \times 2^m$ Walsh-Hadamard transform matrix, where m is a positive integer. The Walsh-Hadamard transform can be defined recursively by:

$$W_m = \frac{1}{\sqrt{2}} \begin{bmatrix} W_{m-1} & W_{m-1} \\ W_{m-1} & -W_{m-1} \end{bmatrix} = W_1 \otimes W_{m-1} \qquad (1)$$

where

$$W_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad (2)$$

and $\otimes$ a Kronecker product. In practice, we rearrange the matrix elements in descending order of frequency to help improve the coefficient entropy coding.

2.2. Haar Transform

A $2^n \times 2^n$ Haar matrix H_n can be written as:

$$H_n = \begin{bmatrix} H_{n-1} \otimes \begin{bmatrix} 1 & 1 \end{bmatrix} \\ 2^{\frac{n-1}{2}} I_2 \otimes I_{2^{n-2}} \otimes \begin{bmatrix} 1 & -1 \end{bmatrix} \end{bmatrix} \qquad (3)$$

where $H_1 = W_1$ and I_m is the $m \times m$ identity matrix. Simple and fast computation algorithms exist for both transforms.

2.3. Quantitative Analysis

We consider the theoretical coding performance of the staircase transforms in comparison to that of DCT. We model the sharp transition across two plain regions with a simplified scalar zero-mean random vector

$$\bar{x} = \{x_1, \cdots, x_M, x_{M+1}, \cdots, x_N\}^T,$$

with autocorrelation matrix

$$R(i,k) = E\{x_i x_k\} = \begin{cases} \rho^{|i-k|}, & \text{if}\{i,k\} \leq M \;\; or \;\; \{i,k\} > M \\ 0, & \text{otherwise} \end{cases}$$

where ρ is the correlation coefficient with a value close to 1, and M is the offset of the transition from the block boundary. Apply an $N \times N$ unitary matrix Q to $\bar{x}$

$$\bar{z} = Q\bar{x} = [z_1, z_2, \cdots, z_N]^T.$$

The objective of an encoder is to distribute a fixed number of bits to the elements in $\bar{z}$ such that the average distortion is minimized. This bit-allocation problem is addressed by the water-filling algorithm [14]. Under the high resolution quantization assumption with a non-integer bit-allocation allowed, the minimum obtainable mean-squared distortion (D_T) is approximately proportional to the geometric mean of the transform coefficient variance $\sigma^2_{z_i}$, i.e.,

$$D_T \propto \Big(\prod_{i=1}^{N} \sigma^2_{z_i} \Big)^{1/N}$$

These variances can be obtained as the diagonal elements of the autocorrelation matrix of $\bar{z}$, i.e.,

$$R_T = E\{\bar{z}\bar{z}^T\} = Q * R * Q^T.$$

The coding gain G_Q in decibel of a given transform Q is hence defined by

$$G_Q = 10log_{10}(D_I / D_Q),$$

where I is the $N \times N$ identity matrix.

We compare the theoretical coding gains of the Walsh-Hadamard transform and the Haar transform with that of DCT. When the transition is half the distance from the block boundary, i.e., $M = N/2$, both the Walsh-Hadamard transform and the Haar transform provide superior coding gains to the DCT as shown in Fig. 1. When the transition is quarter the distance from the block boundary, the Haar transform noticeably outperforms the Walsh-Hadamard transform, due to the fact that the basis functions of Haar transform enable better adaptation to the transition offset (both still outperform the DCT). When it comes to the odd offset case (Fig. 3), however, the coding gains for the Walsh-Hadamard transform is substantially below that of the DCT and the Haar transform, as the Walsh-Hadamard transform requires superposition of more bases to handle an odd offset transition. In practice, the rate-distortion optimization algorithm will avoid this situation by either selecting a motion vector that satisfy the transition condition for the staircase transform or it will select the DCT.

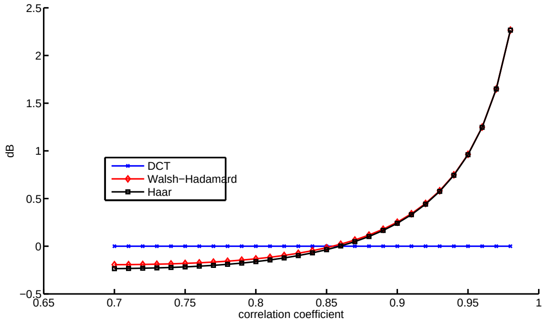


Fig. 1. Relative transform coding gains of staircase transforms with respect to the DCT. The transform dimension is 16×16. The transition offset M is 8.

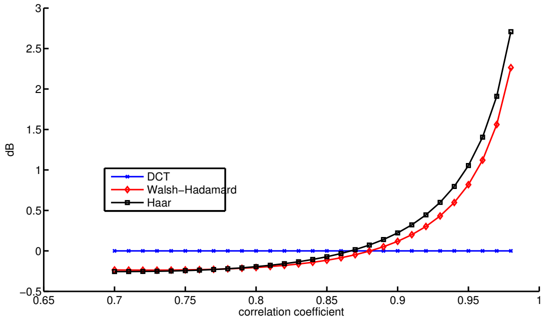


Fig. 2. Relative transform coding gains of staircase transforms with respect to the DCT. The transform dimension is 16×16. The transition offset M is 4.

The statistical model of a motion compensated prediction residual block can be considered as a superposition of the above model with similar relative coding gains. Its coding performance will be experimentally validated next.

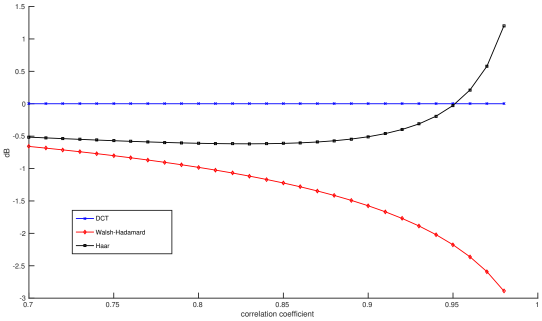


Fig. 3. Relative transform coding gains of staircase transforms with respect to the DCT. The transform dimension is 16×16. The transition offset M is 5.

3. HYBRID TRANSFORM CODING SCHEME FOR SCREEN CONTENT

Inspired by the adaptive 1-D/2-D DCT approach [4], we propose a hybrid transform coding system where the staircase transform can be applied to either vertical, horizontal, or both directions, in addition to the 2D-DCT and the TSM. Let Q denote a given staircase transform matrix. Given an $N \times N$ residual pixel block X, the scheme allows the following three transform types:

- *Type I*. Vertical transform: $X^* = QX$
- *Type II*. Horizontal transform: $X^* = XQ^T$
- *Type III*. 2D transform: $X^* = QXQ^T$

The encoding decision among the proposed staircase transform, the DCT, and the TSM, is made in a rate-distortion optimization framework, where the encoder examines each individual transform option for the given residual block, and selects the one with minimum rate-distortion cost. To keep the additional encoding complexity under check, the cross combination between the staircase transform and the DCT is temporarily disabled in the experiment.

4. EXPERIMENTAL RESULTS

We implemented the proposed scheme in a modified VP9 experimental codebase that supports screen content coding extensions, including transform skipping, IntraBC, and palette coding. The allowed coding block sizes range from 4×4 to 64×64. The transform block size for the DCT goes from 4×4 up to 32×32. In the experiments we found that the majority of the gains from the staircase transform came from transform block sizes between 4×4 and 16×16 and as such we decided to limit the maximum transform block size for the staircase transform to 16×16. In addition, the staircase transform is limited to coding block sizes under 32×32. All the encoding decisions were determined in a rate-distortion optimization framework.

Each individual staircase transform was plugged into the framework to evaluate its compression performance as compared to the modified VP9 codec for screen content coding. The rate-distortion performance is shown in Fig. 4. The average coding performance gains in terms of BD-rate reduction

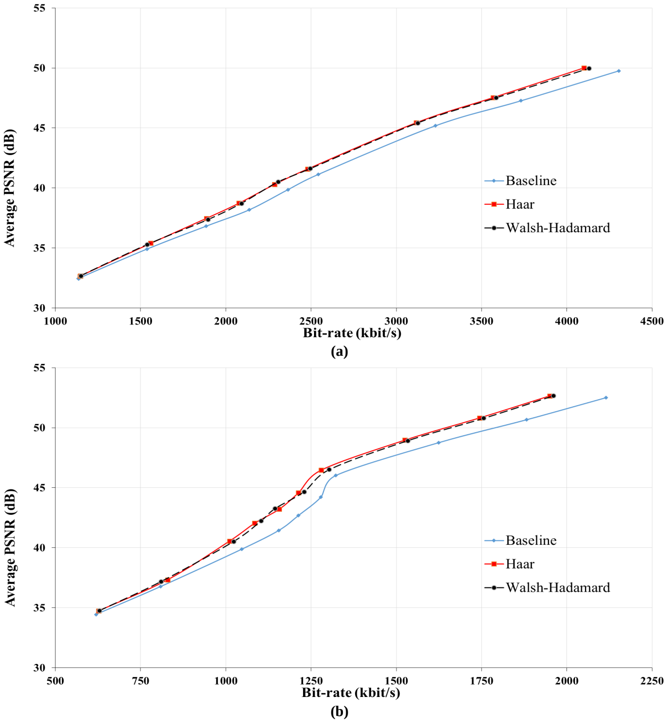


Fig. 4. Compression performance comparison between the Walsh-Hadamard transform, the Haar transform, and the DCT. Test sequences: (a). Console. (b). Desktop.

are shown in Table 1. The compression performance gains are more pronounced in the medium to high bit-rate range. In the lower bit-rate range, most coding blocks are quantized to all zero coefficients and the benefits of introducing alternative transform kernels diminish. The Haar transform provides better coding performance than the Walsh-Hadamard transform as expected in Sec. 2.3, since its basis functions can adapt to the transition offset more efficiently.

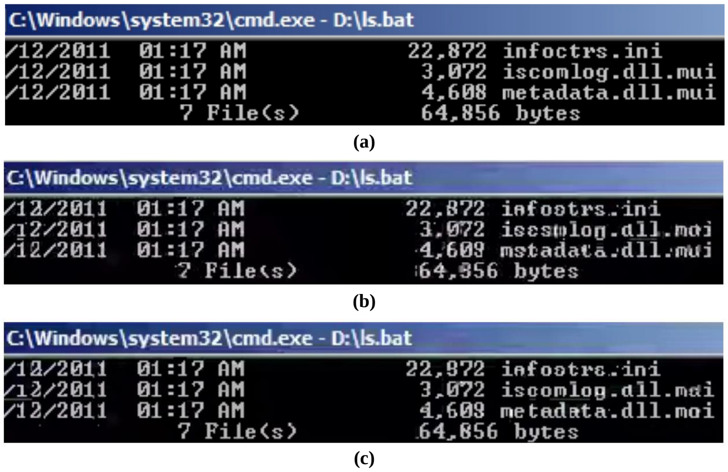
We then evaluate the perceptual quality at low bit-rate. The test clip is coded with the same target bit-rate by the reference codec and the proposed scheme using a Walsh-Hadamard transform. As shown in Fig. 5, the staircase transform substantially reduces the ringing artifacts caused by truncating the DCT coefficients, particularly around the numbers and top bar.

5. CONCLUSIONS

A class of staircase transforms are exploited in this work, in conjunction with an adaptive 1D/2D transform. The proposed scheme provides an alternative transform kernel to the conventional DCT and TSM. It improves the compression performance for screen content videos. A substantial reduction in ringing artifacts is achieved.

Table 1. Coding performance gains due to the staircase transform in terms of BD-rate reduction. A positive number means better compression.

	WHT-DCT	Haar-DCT
FlyingGraphics	3.37	4.00
SocialNetworkMap	1.65	1.90
Programming	1.91	2.20
Map	1.24	1.62
SlideShow	1.14	1.21
Web_browsing	1.16	1.82
Console	4.97	5.38
Desktop	6.48	7.01
Robot	0.80	0.95
Average	2.52	2.90



```
C:\Windows\system32\cmd.exe - D:\ls.bat
/12/2011  01:17 AM            22,872 infoctrs.ini
/12/2011  01:17 AM             3,072 iscomlog.dll.mui
/12/2011  01:17 AM             4,608 metadata.dll.mui
               7 File(s)       64,856 bytes
```

(a)

```
C:\Windows\system32\cmd.exe - D:\ls.bat
/12/2011  01:17 AM            22,872 infostrs.ini
/12/2011  01:17 AM             3,072 isesmlog.dll.mui
/12/2011  01:17 AM             4,609 mstadata.dll.mui
               7 File(s)       64,856 bytes
```

(b)

```
C:\Windows\system32\cmd.exe - D:\ls.bat
/12/2011  01:17 AM            22,872 infostrs.ini
/12/2011  01:17 AM             3,072 iscomlog.dll.mui
/12/2011  01:17 AM             4,609 metadata.dll.mui
               7 File(s)       64,856 bytes
```

(c)

Fig. 5. Perceptual quality comparison: (a). Original video *desktop*. (b). DCT and TSM. (c). Walsh-Hadamard, DCT, and TSM.

6. REFERENCES

[1] A. K. Jain, "A fast Karhunen-Loeve transform for a class of random processes," *IEEE Transactions on Communications*, vol. 24, no. 9, pp. 1023–1029, 1976.

[2] J. Han, A. Saxena, and K. Rose, "Towards jointly optimal spatial prediction and adaptive transform in video/image coding," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 726–729, 2010.

[3] J. Han, A. Saxena, V. Melkote, and K. Rose, "Jointly optimized spatial prediction and block transform for video and image coding," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1874–1884, 2011.

[4] F. Kamisli and J. S. Lim, "1-D transforms for the motion

compensation residual," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 1036–1046, 2011.

[5] A. Gabriellini, M. Naccari, M. Mrak, and D. Flynn, "Spatial transform skip in the emerging high efficiency video coding standard," *IEEE International Conference on Image Processing*, pp. 185–188, 2012.

[6] M. Mrak and J. Xu, "Improving screen content coding in hevc by transform skipping," *European Signal Processing Conference*, pp. 1209–1213, 2012.

[7] D.-K. Kwon and M. Budagavi, "Fast intra block copy (intrabc) search for hevc screen content coding," *IEEE International Symposium on Circuits and Systems*, pp. 9–12, 2014.

[8] Z. Ma, W. Wang, M. Xu, and H. Yu, "Advanced screen content coding using color table and index map," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4399–4412, 2014.

[9] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging hevc screen content coding extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 50–62, 2016.

[10] Y.-J. Chen, S. Oraintara, and T. Nguyen, "Video compression using integer dct," *IEEE International Conference on Image Processing*, pp. 844–12, 2000.

[11] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A technical overview of vp9 - the latest open-source video codec," *SMPTE*, vol. 2013, no. 10, pp. 1–17, 2013.

[12] G. Walter and X. Shen, *Wavelets and Other Orthogonal Systems*, Boca Raton: Chapman, 2001.

[13] S. H. Oğuz, Y.-H. Hu, and T.-Q. Nguyen, "Image coding ringing artifact reduction using morphological postfiltering," *IEEE Second Workshop on Multimedia Signal Processing*, pp. 628–633, 1998.

[14] A. Gersho and R. Gary, *Vector Quantization and Signal Compression*, Springer, 1991.