Robust Learning of 2-D Separable Transforms for Next-Generation Video Coding

Sezer, O.G.; Cohen, R.; Vetro, A.

TR2011-021 March 2011

Abstract

With the simplicity of its application together with compression efficiency, the Discrete Cosine Transform(DCT) plays a vital role in the development of video compression standards. For next-generation video coding, a new set of 2-D separable transforms has emerged as a candidate to replace the DCT. These separable transforms are learned from residuals of each intra prediction mode; hence termed as Mode dependent- directional transforms (MDDT). MDDT uses the Karhunen-Loeve Transform (KLT) to create sets of separable transforms from training data. Since the residuals after intra prediction have some structural similarities, transforms utilizing these correlations improve coding efficiency. However, the KLT is the optimal approach only if the data has a Gaussian distribution without outliers. Due to the nature of the least-square norm, outliers can arbitrarily affect the directions of the KLT components. In this paper, we will address robust learning of separable transforms by enforcing sparsity on the coefficients of the representations. With this new approach, it is possible to improve upon the video coding performance of H.264/AVC by up to 10.2% BD-rate for intra coding. At no additional cost, the proposed techniques can also provide up to 3.9% improvement in BD-rate for intra coding compared to existing MDDT schemes.

Data Compression Conference (DCC)

ROBUST LEARNING OF 2-D SEPARABLE TRANSFORMS FOR NEXT-GENERATION VIDEO CODING

Osman G. Sezer[†*], *Robert Cohen*[‡] *and Anthony Vetro*[‡]

[†]Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA 30308, USA

[‡]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA

osman@ece.gatech.edu, {cohen, avetro}@merl.com

ABSTRACT

With the simplicity of its application together with compression efficiency, the Discrete Cosine Transform (DCT) plays a vital role in the development of video compression standards. For next-generation video coding, a new set of 2-D separable transforms has emerged as a candidate to replace the DCT. These separable transforms are learned from residuals of each intra prediction mode; hence termed as Mode dependent- directional transforms (MDDT). MDDT uses the Karhunen-Loeve Transform (KLT) to create sets of separable transforms from training data. Since the residuals after intra prediction have some structural similarities, transforms utilizing these correlations improve coding efficiency. However, the KLT is the optimal approach only if the data has a Gaussian distribution without outliers. Due to the nature of the least-square norm, outliers can arbitrarily affect the directions of the KLT components. In this paper, we will address robust learning of separable transforms by enforcing sparsity on the coefficients of the representations. With this new approach, it is possible to improve upon the video coding performance of H.264/AVC by up to 10.2% BD-rate for intra coding. At no additional cost, the proposed techniques can also provide up to 3.9% improvement in BD-rate for intra coding compared to existing MDDT schemes.

1. INTRODUCTION

This paper describes a novel mode dependent design of 2-D separable transforms to be used in video coding. Compared to the current state of the art methods, this method enforces sparsity on the transform coefficients for given data fidelity. Iterative optimization updates each separable transform after finding the optimal coefficients in a mode-dependent framework. The mode dependent aspect of the transform design also deviates from prior work [1, 2] in that the rate-distortion-optimal selection of transforms is abandoned. Hence, no extra bits are required to signal the transform selection, which makes our approach compatible with current video coding architectures.

In video coding, frames are typically encoded in two ways: i) intra coding, ii) inter coding. In intra coding the correlation of blocks within a frame is utilized to generate prediction residuals, which will have significantly less energy than the corresponding original image

[*] Part of this work was performed while O.G. Sezer was an intern at MERL.

blocks. The prediction residual is the difference between an original block and its prediction. Hence, fewer bits are required to encode the blocks at a given level of fidelity. For inter coding, motion-compensated prediction residuals are generated using blocks within a temporal neighborhood.

In state-of-the-art video codecs such as H.264/AVC, the prediction for an intra coded block is computed from previously coded neighboring blocks. Several directional predictions are generated, and a fitness measure such as sum of absolute differences (SAD), sum of squared error (SSE), or sum of absolute transformed differences (SATD) is computed for each direction. In H.264/AVC, the best prediction direction or "mode" is selected, and the corresponding prediction residual is transformed via the conventional integer Discrete Cosine Transform (DCT) prior to quantization. Since the residuals of the same mode possess common patterns of correlation, one can design transforms that will further exploit these patterns to reduce the bit rate. One such set of transforms are the Mode Dependent Directional Transforms (MDDT) proposed in [3]. While MDDT utilizes the KLT or Singular Value Decomposition (SVD) to learn 2-D separable transforms for residuals of each intra prediction mode, this paper describes shortcomings of KLT in the presence of outliers in the training data. Next, a new $\mathcal{L}_0$-norm regularized optimization method is proposed as a more robust way to learn 2-D separable transforms for video coding. By employing new transforms, which are termed as Mode Dependent Sparse Transforms (MDST), into H.264/AVC-based video codec (JM11.0KTA2.6r1), the compression efficiency is improved by up to 3.9% BD-rate compared to MDDT, while the coding architecture is kept the same.

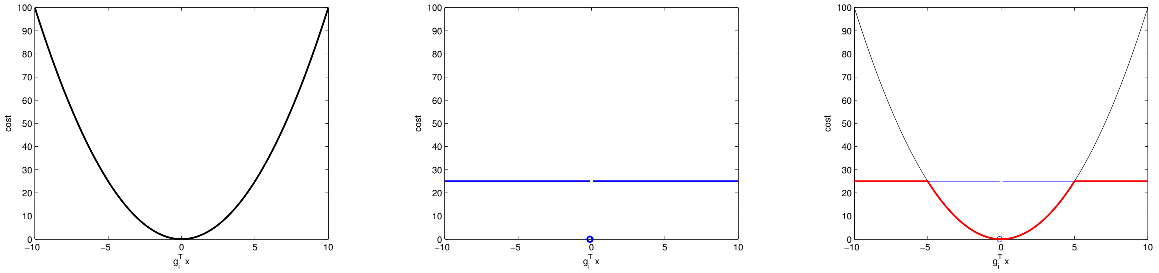
The outline of the paper is as follows. In the next section, we point out why KLT is vulnerable to outliers in the data, and show how $\mathcal{L}_0$-norm regularization can bring robustness to the transform learning process. Section 3 outlines the proposed iterative optimization method used to generate 2-D separable transforms for video coding, which is followed by Section 4 where we introduce a new ordering method for the locations of the coefficients to improve coding efficiency of the entropy coder. In Section 5, experiments to validate the proposed method are provided. Finally, we make some concluding remarks in Section 6.

2. LEARNING TRANSFORMS FROM DATA

Given a set of random signals, KLT is the standard procedure to extract transforms that will decorrelate the data to a smaller number of variables. With the KLT, the signal energy is concentrated mostly to the first few coefficients of this linear orthogonal decomposition, such that a reduced dimensional representation is achieved within certain fidelity. KLT solves the following minimization to find the principal component $\mathbf{g}_1$

$$\min_{\mathbf{g}_1} \sum_{j \in \mathcal{S}} \|\mathbf{x}^j - \mathbf{g}_1 c_1^j\|_2^2 \;\; s.t \;\; \mathbf{g}_1^T \mathbf{g}_1 = 1, \tag{1}$$

where $\mathbf{x}^j$ is the j-th vector of size $n \times 1$ in the dataset $\mathcal{S}$, and c_1^j is the coefficient of the principal component. The principal vector aligns itself to the direction of maximum variation, and the solution can be found by using singular value decomposition (SVD). Similarly, the subsequent k-th components can be found from the residual data after the subtraction of the first $k-1$ principal components. Another way to express the KLT formulation is as



(a)

(b)

(c)

Fig. 1. Cost functions of (a) $\mathcal{L}_2$ norm, (b) $\mathcal{L}_0$ norm, and (c) $\rho(.)$ as a function of $\mathbf{g}_i^T\mathbf{x}$ for fixed $\lambda = 25$ in (5).

follows:

$$\min_{\mathbf{G}} \sum_{j\in\mathcal{S}} \|\mathbf{x}^j - \mathbf{G}\mathbf{c}^j\|_2^2 \;\; s.t \;\; \mathbf{G}^T\mathbf{G} = \mathbf{I}, \tag{2}$$

where $\mathbf{g}_1$ is the first column of matrix $\mathbf{G}$. One of the problems with KLT-based learning arises from its noise intolerance. The least square norm in (1) is prone to outliers, especially to the ones with large energy. These outliers can arbitrarily skew the direction of the principal component. In cascade, the subsequent components and the overall performance of this representation will be affected. In computer vision and statistics literature there are several methods proposed to overcome this challenge, such as outlier rejection [4], weighted least squares [5], and utilizing robust error norms to learn subspaces [6].

In compression, recent learning-based designs have been shown to provide superior performance compared to standard methods such as the DCT or wavelets. Ye and Karczewicz [3] proposed to use KLT to learn 2-D separable transforms for video coding. Sezer et al. [1, 7] used sparsity enforced transform designs. Apart from iterative update of clusters and corresponding transforms, the Sparse Orthonormal Transforms (SOT) of [1] provide a learning algorithm that is more robust than the KLT, by regularizing the cost in (1) with the sparsity of the coefficients.

To be more specific, let $\mathbf{G}$ be of size $N \times N$. A robust estimation of the principal components can be achieved when the following cost is minimized

$$\min_{\mathbf{G}} \sum_{j\in\mathcal{S}} \min_{\mathbf{c}^j} \{\|\mathbf{x}^j - \mathbf{G}\mathbf{c}^j\|_2^2 + \lambda\|\mathbf{c}^j\|_0\} \;\; s.t \;\; \mathbf{G}^T\mathbf{G} = \mathbf{I}, \tag{3}$$

where $\mathbf{c}^j$ is the coefficient of $\mathbf{G}$ for data vector $\mathbf{x}^j$, λ is Lagrange multiplier, and $\|.\|_0$ is the $\mathcal{L}_0$ norm, which is equivalent to the the number of nonzero elements. Next, (3) can be written as

$$\min_{\mathbf{G}} \sum_{j\in\mathcal{S}} \sum_{i} \min_{\mathbf{c}_i^j} \{\|\mathbf{x}^j - \mathbf{g}_i c_i^j\|_2^2 + \lambda\|c_i^j\|_0\} \;\; s.t \;\; \mathbf{G}^T\mathbf{G} = \mathbf{I}, \tag{4}$$

where $\mathbf{g}_i$ is the i-th column of $\mathbf{G}$, and c_i^j is the corresponding coefficient. The cost defined in (4) penalizes nonzero c_i's; thus enforcing a sparse representation for component $\mathbf{g}_i$. For further analysis, let us examine the first minimization term. This expression can be expressed as,

$$\rho(\mathbf{g}_i^T\mathbf{x}^j, \lambda) = \min_{c_i^j} \{\|\mathbf{g}_i^T\mathbf{x}^j - c_i^j\|_2^2 + \lambda\|c_i^j\|_0\}, \tag{5}$$

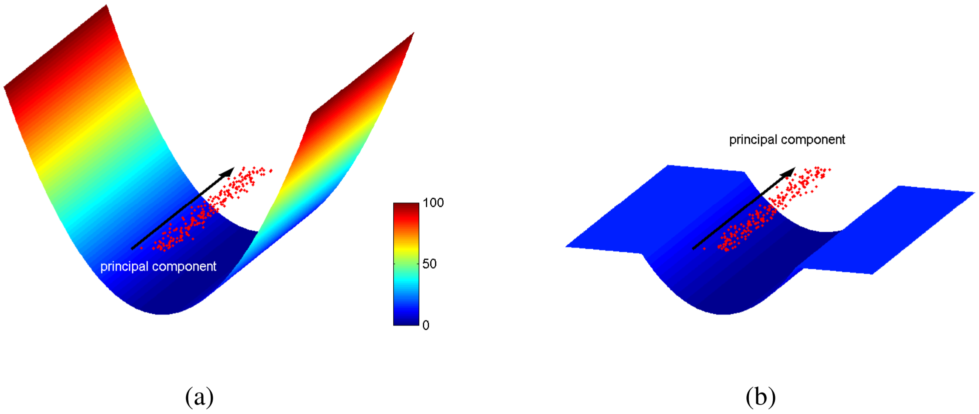


Fig. 2. Cost function of KLT (a), $\mathcal{L}_0$-norm regularized solution (b), and their corresponding principal components.

since $\mathbf{g}_i^T(\mathbf{x}^j - \mathbf{g}_i c_i^j) = \mathbf{g}_i^T \mathbf{x}^j - c_i^j$. The cost defined by ρ is a union of $\mathcal{L}_2$ and $\mathcal{L}_0$ norms. For small values of c_i, the $\mathcal{L}_2$ norm takes precedence, whereas the $\mathcal{L}_0$ norm has a greater priority for larger values of c_i. The transition between two norms is defined by λ. Figures 1(a) and 1(b) show the $\mathcal{L}_2$ and $\mathcal{L}_0$ norms as a function of $\mathbf{g}_i^T\mathbf{x}$. Fig. 1(c) plots $\rho(\mathbf{g}_i^T\mathbf{x}, \lambda)$, which picks the minimum of these norms for given $\mathbf{g}_i^T\mathbf{x}$ and λ values.

The minimization over the coefficients can be substituted by the M-estimator $\rho(.)$ as follows:

$$\min_{\mathbf{G}} \sum_{j \in \mathcal{S}} \sum_i \rho(\mathbf{g}_i^T \mathbf{x}^j, \lambda) \quad s.t \;\; \mathbf{G}^T\mathbf{G} = \mathbf{I}. \tag{6}$$

Due to orthonormality conditions imposed on the solution, this expression essentially searches for the axis rotations that will minimize the cost function $\rho(.)$ over a set of observations. Sparsity imposed on a component helps robust estimation of components orthogonal to that. To visualize this, Fig. 2 gives a $3D$ perspective of the $\mathcal{L}_2$ and $\mathcal{L}_0$-regularized cost functions used in (1) and (6) for 2-D data. If $\mathbf{g}_1$ assumed to be the principal component, the cost function in Fig. 2(b) is attained by imposing sparsity on the coefficients of $\mathbf{g}_2$, where $\mathbf{g}_1 \perp \mathbf{g}_2$. Here we show how the principal components should be aligned with respect to given data (dots in 2-D) to minimize the costs. Note that even a single large outlier would arbitrarily change the direction of KLT-solution shown in Fig. 2, due to rapid increase of the cost function. On the other hand, the proposed approach limits this influence. The robustness of $\mathcal{L}_0$-norm regularized SOT solution comes from this reality. Also in order to avoid local minima, annealing λ is a common approach in robust statistics literature [8, 9]. A linear regression experiment with outliers is provided in Section 5 to compare the robustness of standard KLT and $\mathcal{L}_0$-norm regularized solution.

3. MODE DEPENDENT SPARSE TRANSFORMS (MDST)

There are two standard approaches for block-based 2-D data transforms: i) separable, and ii) non-separable transforms. In the separable case, each column and row of the block is considered as a 1-D signal, and 1-D transforms are used to map the block of data to a set of coefficients. The 1-D transforms used in each direction could be the same, but may also be different. For non-separable transforms, the block is generally ordered as a 1-D vector by lexicographically ordering columns or rows of the block. The disadvantage of this is that non-separable transforms would require more memory to hold the entries of the transform matrix. Also, large matrix multiplications are generally too complex for hardware implementations. Therefore, separable transforms are appealing. However, there is a cost for separable transforms, since they only utilize the correlation with a column or row; hence the compression performance of the separable transforms is lower around directional edges as compared to non-separable transforms.

Intra coding of H.264/AVC has been shown to provide higher coding efficiency compared to standard block based image compression methods such as JPEG, and it has competitive performance with, if not better than, wavelet based JPEG2000 [10, 11]. The success is largely due to intra prediction methods employed prior to transform coding. In general, the residual data generated by intra prediction has less energy than the original image block, hence requires fewer bits to represent coefficients after transform coding. Nevertheless, even after the intra prediction, residuals are observed to possess directional structures often aligned with the direction of prediction. Therefore for each directional prediction mode a new transform is trained in [3] to further utilize the inherent structure of that prediction mode to reduce the bitrate. We will improve upon that transform design process with a new iterative optimization method to learn 2-D separable transforms for each intra prediction mode.

We define the number of prediction modes as M, where $M = 9$ for intra prediction of 4×4 and 8×8 block sizes, and $M = 4$ for 16×16 blocks. For each mode, two separable transforms are needed. The vertical and horizontal transform for mode i is denoted as $\mathbf{V}_i$ and $\mathbf{H}_i$, respectively. Let the $N \times N$ block $\mathbf{X}_i^j$ be the j-th residual block encoded using intra mode i, and $\mathbf{C}_i^j$ be the corresponding coefficient matrix of the residual signal. The sparsity-distortion cost function can be written as follows:

$$\begin{aligned} & i \in \{1, ..., M\}: \\ & \min_{\mathbf{V}_i, \mathbf{H}_i} \left(\sum_{j \in S_i} \min_{\mathbf{C}_i^j} \|\mathbf{X}_i^j - \mathbf{V}_i \mathbf{C}_i^j {\mathbf{H}_i}^T\|_2^2 + \lambda \|\mathbf{C}_i^j\|_0 \right) \\ & \textit{s.t}\ \ {\mathbf{V}_i}^T \mathbf{V}_i = \mathbf{I},\ \ {\mathbf{H}_i}^T \mathbf{H}_i = \mathbf{I}\,. \end{aligned} \tag{7}$$

To learn the transforms for mode i, we have formed a training dataset S_i, over which the cost function will be minimized. The given cost models distortion as the reconstruction error (first term in the summation), and an approximation to rate is given by $\mathcal{L}_0$ norm term, which is the number of nonzero coefficients. In Section 2 we have also pointed out how $\mathcal{L}_0$-norm regularization relates to robust estimation. The proposed method iteratively finds optimal coefficients and updates one of the separable transform at each iteration. Let us assume vertical and horizontal transforms are initialized with the DCT, then for the i-th mode we apply the following steps:

I. *Optimal coefficients for a given transform:* The sparsest representation for a given transform can be found by solving

$$\mathbf{C}_i^{j*} = \arg\min_{\mathbf{D}} (\|\mathbf{X}_i^j - \mathbf{V}_i \mathbf{D}_i^j {\mathbf{H}_i}^T\|_2^2 + \lambda \|\mathbf{D}_i^j\|_0). \tag{8}$$

Note that since both $\mathbf{V}_i$ and $\mathbf{H}_i$ are orthonormal, the optimal solution is obtained by hard-thresholding the components of $\mathbf{D} = {\mathbf{V}_i}^T \mathbf{X}_i^j \mathbf{H}_i$ with $\sqrt{\lambda}$.

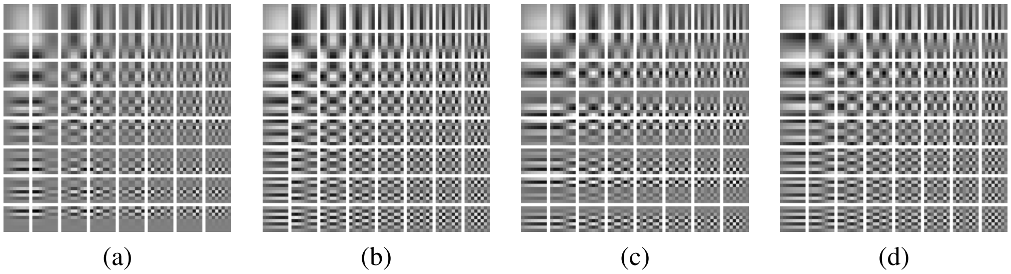


Fig. 3. Comparison of separable transforms of MDST and MDDT. MDST of vertical prediction (mode 0) (a), MDDT of vertical prediction (mode 0) (b), MDST of horizontal prediction (mode 1) (c), MDDT of horizontal prediction (mode 1) (d).

II. *Optimal vertical transforms for given coefficients:* The optimal vertical separable orthonormal transform for given coefficient vectors from previous step can be found by solving

$$\begin{gathered} \mathbf{V}_i^* = \arg\min_{\mathbf{A}} \left\{ \sum_{\mathbf{X}_i^j \in S_i} \|\mathbf{X}_i^j - \mathbf{A}\mathbf{C}_i^{j^*}\mathbf{H}_i{}^T\|_2^2 \right\} \\ s.t. \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{gathered} \tag{9}$$

Note here that the horizontal separable transform, $\mathbf{H}_i$, is assumed to be fixed. Let $\mathbf{Y} = \sum_{\mathbf{X}_i^j \in S_i} \mathbf{X}_i^{j^T} \mathbf{H}_i{}^T \mathbf{C}_i^j$, and its SVD be $\mathbf{Y} = \mathbf{U}\Lambda^{1/2}\mathbf{W}^T$. The solution for the optimal orthonormal transform can be found by $\mathbf{V}_i^* = \mathbf{W}\mathbf{U}^T$. For details of the optimization please refer to [1].

III. *Optimal coefficients with updated vertical transform:* This time optimal coefficients are found with optimized transform, $\mathbf{V}_i^*$, from the previous step,

$$\mathbf{C}_i^{j^*} = \arg\min_{\mathbf{D}} (\|\mathbf{X}_i^j - \mathbf{V}_i^* \mathbf{D}_i^j \mathbf{H}_i{}^T\|_2^2 + \lambda \|\mathbf{D}_i^j\|_0). \tag{10}$$

Note that since both $\mathbf{V}_i$ and $\mathbf{H}_i$ are orthonormal, the optimal solution is obtained by hard-thresholding the components of $\mathbf{D} = \mathbf{V}_i^{*T}\mathbf{X}_i^j\mathbf{H}_i$ with $\sqrt{\lambda}$.

IV. *Optimal horizontal transforms for given coefficients:* Similarly, the optimal horizontal separable orthonormal transform can be calculated with updated coefficients and the vertical transform found in previous steps;

$$\begin{gathered} \mathbf{H}_i^* = \arg\min_{\mathbf{A}} \left\{ \sum_{\mathbf{X}_i^j \in S_i} \|\mathbf{X}_i^j - \mathbf{V}_i^* \mathbf{C}_i^{j^*} \mathbf{A}^T\|_2^2 \right\} \\ s.t. \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}. \end{gathered} \tag{11}$$

Note this time vertical separable transform $\mathbf{V}_i$ is assumed to be fixed. Next, let $\mathbf{Y} = \sum_{\mathbf{X}_i^j \in S_i} \mathbf{C}_i^{j^T} \mathbf{V}_i{}^T \mathbf{X}_i^j$, and its SVD be $\mathbf{Y} = \mathbf{U}\Lambda^{1/2}\mathbf{W}^T$. The solution for the optimal orthonormal transform is $\mathbf{H}_i^* = \mathbf{W}\mathbf{U}^T$.

Return to Step I and repeat the process till the cost function converges to a steady state value. Sample transforms are shown in Fig. 3 together with their MDDT counterparts. This optimization method differs from those used in [1] and [2]. In [1], the proposed transform design method reduces the sparsity-distortion cost of a set of data extracted from natural images via iterative clustering and transform optimization for the nonseparable case. In this paper, the data is residual blocks extracted from a video coder, and the corresponding clusters are defined by the intra prediction mode. Hence, the data clusters are fixed, so relabeling after the transform optimization is not needed. Thus, the mode-dependent term is coined for the transforms in the current design.

The 2-D separable transform design provided in [2], which is based on the optimization given in [1], lacks the mode-dependent characteristic, and its iterative optimization has two shortcomings. The first problem is the update step for vertical transforms in Equation (9) of [2], whereby the vertical and horizontal separable components converges to the same transforms. For mode-dependent transforms, it is expected that the vertical and horizontal transforms will be different due to the directional characteristics of the residual data. Correction is provided in Step II of our iterative optimization procedure. The second problem stems from the iterative update of vertical and horizontal transforms without updating coefficients. When vertical or horizontal transforms are updated, the coefficients do not belong to new transform anymore. Therefore, in the iterative optimization described above, the transform update is always followed by a coefficient update.

4. REORDERING TRANSFORMS

Entropy coders in current video codecs are optimized to work with the DCT. Although the optimization described in this paper initializes transforms with DCT, the resulting transform coefficients may compact energy in a different order than with the DCT. Therefore, the columns of the vertical and horizontal 2-D separable transform are reordered depending on the energy of the coefficient values of the residual data set of the corresponding mode.

Let $\mathbf{Q}$ be an $N \times N$ matrix whose entries are defined as follows:

$$\mathbf{Q}(m,n) = \sum_{j \in \mathcal{S}} \mathbf{C}^j(m,n)^2. \tag{12}$$

where $\mathbf{C}^j$ is the coefficient matrix of the j-th block in the training set of mode $\mathcal{S}$. Then sum of the energies along the rows and columns can be defined respectively as,

$$q_r(m) = \sum_n \mathbf{Q}(m,n) \ \ \forall m, \quad q_c(n) = \sum_m \mathbf{Q}(m,n) \ \ \forall n. \tag{13}$$

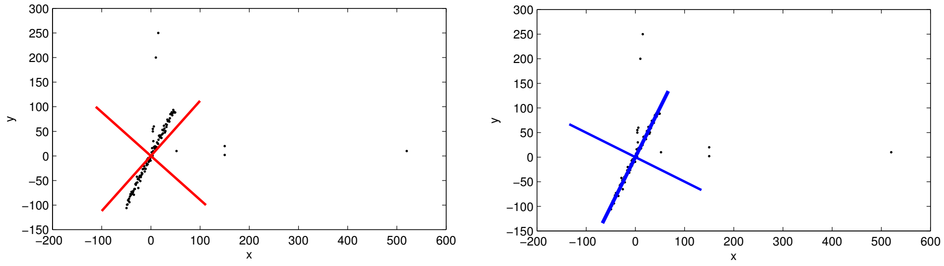
To rank these energies, the x and y variables that will satisfy

$$q_r(x_1) \geq q_r(x_2) \geq \ldots \geq q_r(x_N), \quad q_c(y_1) \geq q_c(y_2) \geq \ldots \geq q_c(y_N) \tag{14}$$

can be found. Next the columns of the optimized vertical and horizontal separable transforms are reordered as

$$\mathbf{V}^o(m,n) = \mathbf{V}(m,x_n), \ \ \mathbf{H}^o(m,n) = \mathbf{H}(m,y_n) \quad \forall m,n \tag{15}$$

where $\mathbf{H}$ and $\mathbf{V}$ become $\mathbf{H}^o$ and $\mathbf{V}^o$ after reordering. The new order statistically ensures that the coefficients with higher energy appear closer to the top-left corner of the coefficient matrix similar to DCT. The transforms for each mode and block size are ordered in same fashion. Later, they are scaled up and rounded off to have integer values.



(a) (b)

Fig. 4. Crosses show axises of components found by KLT (a), and $\mathcal{L}_0$-norm regularized solution (b).

5. RESULTS

Two sets of experiments are provided in this section. First, the robustness of the $\mathcal{L}_0$-norm regularized solution is compared with KLT for a linear regression problem. The second set of experiments show the video coding performance of the proposed MDST method with respect to MDDT, which is already implemented in the JM11.0KTA2.6r1 (or KTA) codec. In addition, a set of KLT-based 2-D separable transforms is trained by using the same method as MDDT but with our data set. This enables us to analyze the effect of training data for the performance improvement that we achieved.

5.1. Model-based Experiment on Robust Regression

In this part a simple linear regression application of KLT and $\mathcal{L}_0$-norm regularized solution is given. A 2-D set of vector are generated by the following model

$$y = 2x + 5w \tag{16}$$

where w is a zero-mean and unit-variance Gaussian random variable. One-hundred Gaussian noise samples are generated and added to x values from -50 to 50. Both the KLT and the proposed $\mathcal{L}_0$-norm regularized solutions recover the correct principal direction. However, when random sparse outliers are included in the data set, the KLT fails to capture the correct direction, as shown in Fig. 4(a). The $\mathcal{L}_0$-norm regularized solution, however, almost perfectly aligns with the direction of correlation set in (16), as shown in Fig. 4(b). The only disadvantage of $\mathcal{L}_0$-norm regularized solution over KLT is its complexity, which is in general less of a concern for off-line training. Nevertheless, initializing the algorithm with the components of KLT and annealing λ improves the convergence speed. For this experiment, the cost function converged in 15 iterations with a fixed $\lambda = 50^2$ when the components were initialized with KLT.

5.2. Video Coding with MDST

The transforms generated by the proposed algorithm is used to replace the MDDT transforms currently implemented in the KTA software. As mentioned before, the transforms are trained by extracting intra prediction residuals for 4×4, 8×8, and 16×16 block sizes. For each block size, a set of 2-D separable transforms is trained with the proposed iterative optimization scheme described in Section 3. The training data contains High-Definition (HD) and CIF (352×288) sequences. The video frames used for training are not used for testing. The sequences are encoded as all intra pictures using four QP values 25, 29, 33, and 37. A different set of values are used for HD sequences: 25, 28, 31, and 34. These QP values are identical to those used in [12]. The CABAC entropy coder is used, and the anchors used for comparison are generated using the DCT-enabled KTA encoder.

Table 1 shows experimental results for several sequences. To understand how much of the performance improvement comes from the data that is used to learn transforms, a set of controlled experiments are performed with KLT as the learning method, using the same set of training data. One could expect similar coding results between our controlled experiment and MDDT, provided that the training data of both are similar. For performance comparisons, BD metrics are used [13]. On the second and third columns of Table 1, although BD-rate improvements of individual sequences differ, the averages are very close (our implementation is only 0.2% better). The fourth column shows performance improvement of the proposed method, MDST. Overall a BD-rate improvement of 1.29% is achieved over MDDT at *no extra cost*, and the improvement goes up to 2.97% in HD sequences.

Table 1. Coding performance, reference is JM-KTA 2.6r1

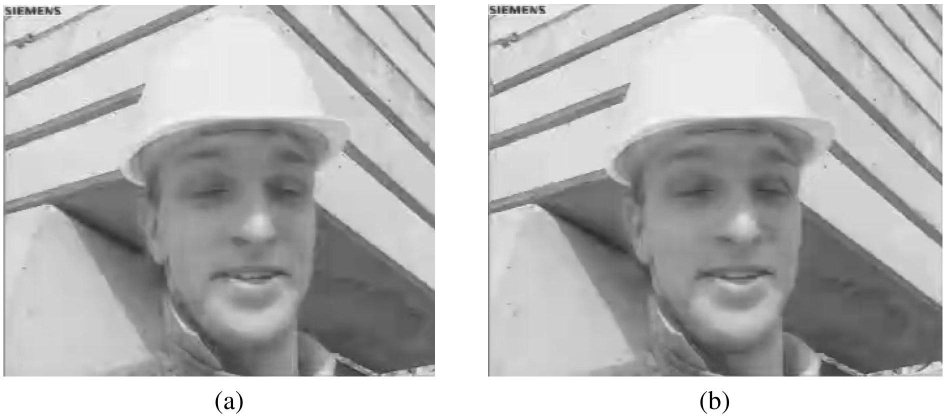
		MDDT			**KLT**			**MDST**			**KLT HD**			**MDST HD**		
Sequences	***number of frames***	**BD-Rate (%)**	**BD-PSNR (dB)**	**Avg BD-Rate**	**BD-Rate (%)**	**BD-PSNR (dB)**	**Avg BD-Rate**	**BD-Rate (%)**	**BD-PSNR (dB)**	**Avg BD-Rate**	**BD-Rate (%)**	**BD-PSNR (dB)**	**Avg BD-Rate**	**BD-Rate (%)**	**BD-PSNR (dB)**	**Avg BD-Rate**
352x288				**-5.07**			**-5.07**			**-5.73**			**-5.20**			**-5.52**
Foreman	100	-5.93	0.322		-5.84	0.318		-8.10	0.446		-6.16	0.335		-7.35	0.402	
Mobile	100	-4.27	0.444		-4.74	0.495		-4.77	0.500		-4.52	0.471		-4.77	0.499	
Coastguard	100	-5.39	0.335		-5.00	0.310		-5.09	0.315		-5.39	0.337		-4.86	0.301	
Container	100	-4.67	0.301		-4.71	0.305		-4.94	0.320		-4.74	0.307		-5.08	0.329	
832x480				**-5.07**			**-5.18**			**-6.26**			**-5.48**			**-6.01**
BasketballDrill	30	-5.87	0.288		-6.14	0.304		-6.91	0.344		-6.51	0.322		-6.90	0.342	
PartyScene	30	-3.85	0.294		-4.09	0.313		-5.31	0.408		-4.14	0.317		-4.88	0.376	
BQMall	30	-5.59	0.362		-5.42	0.353		-7.21	0.474		-5.96	0.389		-6.78	0.445	
RaceHorses	30	-4.69	0.306		-5.08	0.333		-5.61	0.370		-5.30	0.348		-5.48	0.361	
1920x1080				**-5.72**			**-6.09**			**-7.54**			**-7.22**			**-8.14**
Kimono1	10	-6.58	0.245		-6.65	0.249		-8.90	0.341		-9.68	0.372		-10.25	0.397	
ParkScene	10	-6.38	0.298		-6.96	0.327		-7.12	0.334		-6.99	0.329		-7.07	0.332	
Cactus	10	-6.39	0.257		-6.96	0.281		-7.70	0.312		-7.49	0.304		-8.05	0.326	
BasketballDrive	10	-4.47	0.114		-4.73	0.121		-7.44	0.192		-6.10	0.157		-8.38	0.217	
Tennis	10	-4.80	0.154		-5.16	0.167		-6.52	0.211		-5.86	0.191		-6.94	0.226	
Average				**-5.30**			**-5.50**			**-6.59**			**-6.07**			**-6.68**

For visual quality comparison, frames coded at the same rate using MDDT and MDST are also provide in Fig. 5. MDST result on the right have slightly better reconstruction of facial features and edges compared to MDDT.

Due to increased importance of efficiently coding HD sequences, one last set of experiments is done by changing training data set to all HD sequences. Both KLT of controlled experiment and MDST are learned from this new data. Columns five and six of Table 1 shows these results. Surprisingly, the KLT in this case has significant performance improvement not just on HD but for all the sequences. This can be attributed to the statistics of the residuals extracted from HD sequences. Compared to previous training data, it is likely that these residuals have fewer outliers, hence the components of KLT align better with the data. On the other hand, the training method used for MDST outperforms KLT-based learning in all these settings.

6. CONCLUSIONS

This paper presents the Mode-Dependent Sparse Transform (MDST), a new 2-D separable transform design for video coding. The implicit relation between sparsity-enforced opti-



(a) (b)

Fig. 5. Reconstructed foreman image (a) with MDDT 32.93dB and 0.196bpp, and (b) with MDST at 33.04dB and 0.194bpp

mization of transforms and robust learning is revealed. When the training data has outliers, the proposed training method is more robust than the conventional KLT-based training. Utilizing this approach, a new set of 2-D separable transforms are trained using residual data from each intra prediction mode in the KTA codec. Compared to DCT and MDDT-based video coding, bit-rate reductions of up to 10.2% and 3.9% are achieved, respectively.

7. REFERENCES

[1] O. G. Sezer, O. Harmanci, and O. G. Guleryuz, “Sparse orthonormal transforms for image compression,” *In Proc. of 15th IEEE Int. Conf. on Image Processing, San Diego, CA*, pp. 149–152, Oct 2008.

[2] J. Sole, P. Yin, Y. Zheng, and C. Gomila, “Joint sparsity-based optimization of a set of orthonormal 2D separable block transforms,” *In Proc. of 16th IEEE Int. Conf. on Image Processing*, Nov 2009.

[3] Yan Ye and Marta Karczewicz, “Improved H.264 intra coding based on bi-directional intra prediction, direction transform, and adaptive coefficient scanning,” *In Proc. of 15th IEEE Int. Conf. on Image Processing*, Oct 2008.

[4] L. Xu and A. Yuille, “Robust principal component analysis by self-organizing rules based on statistical physics approach,” *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 131–143, 1995.

[5] H. Shum, K. Ikeuchi, and R. Reddy, “Principal component analysis with missing data and its application to polyhedral object modeling,” *IEEE Pattern Analysis and Machine Intelligence*, vol. 17, no. 9, pp. 855–867, 1995.

[6] F. De la Torre and M. J. Black, “A framework for robust subspace learning,” *International Journal of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003.

[7] O. G. Sezer, Y. Altunbasak, and O. G. Guleryuz, “A sparsity-distortion optimized multiscale representation of geometry,” *In Proc. of 17th IEEE Int. Conf. on Image Processing, Hong Kong*, Oct 2010.

[8] M. J. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields,” *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, March 1996.

[9] M. J. Black and A. D. Jepson, “Eigentracking: Robust matching and tracking of objects using view-based representation,” *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[10] F. De Simone, M. Ouaret, F. Dufaux, A.G. Tescher, and T Ebrahimi, “A comparative study of JPEG 2000, AVC/H.264, and HD photo,” *SPIE Optics and Photonics, Applications of Digital Image Processing XXX, San Diego, CA USA*, 2007.

[11] M. Ouaret, F. Dufaux, and T Ebrahimi, “On comparing JPEG2000 and intraframe AVC,” *SPIE Optics and Photonics, Applications of Digital Image Processing XXIX, San Diego, CA USA*, vol. 6312, 2006.

[12] ISO/IEC JTC1/SC29/WG11, “Call for evidence on high-performance video coding (HVC),” ISO/IEC JTC1/SC29/WG11 N10553, April 2009.

[13] G. Bjontegaard, “Calculation of average PSNR differences between RD curves,” *ITU-T SG16/Q6, 13th VCEG meeting, Austin, Texas, USA*, pp. Doc. VCEG–M33, April 2001.