DISCRIMINATIVE STRUCTURED SET PREDICTION MODELING WITH MAX-MARGIN MARKOV NETWORK FOR OPTIMAL LOSSLESS IMAGE CODING

Wenrui Dai, Hongkai Xiong

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ABSTRACT

In this paper, we investigate and propose a novel prediction model for lossless image coding in which the optimal correlated prediction for block of pixels are simultaneously obtained in the sense of the least code length. It not only utilizes the spatial statistical correlation for the optimal prediction directly based on 2-D contexts, but also formulates the data-driven structural interdependencies to make the prediction error coherent with the underlying probability distribution for coding. Besides the discriminative adaptive pixel-wise prediction, the Markov network is adaptively derived to maintain the coherence of prediction in the blocks and seek the concurrent optimization of set of prediction by relating the loss function to actual code length. The prediction error is shown to be asymptotically upper bounded by the training error under the decomposable loss function. For validation, we apply the proposed model into lossless image coding and experimental results show that the proposed scheme outperforms the best prediction scheme reported.

Index Terms— Structured set prediction model, max-margin Markov network, lossless image coding

1. INTRODUCTION

Advances in lossless image coding can be achieved through either 1-D sequential data compression or 2-D context predictive coding, since the concept of context is constructed for universal sequential prediction by Rissanen [1]. Recently, context-based adaptive linear predictors have witnessed significant improvements. Among which, the state-of-the-art include gradient adjusted predictor (GAP) utilized in the context-based adaptive lossless image coder (CALIC [2]) and median edge detector (MED) adopted in the low-complexity lossless compression for images (LOCO-I [3]). GAP determines the active predictor for the current pixel based on its neighboring pixels' gradients. While MED adaptively chooses the median of the neighboring encoded pixels for the current pixel. However, the stationary linear predictors are not eligible in practice, as most natural images are far from being stationary.

Recognizing the nonstationarity of natural images, a class of least-square (LS) autoregression based predictors are developed. First proposed in [4], LS-based adaptation adapt itself in a pixel-by-pixel basis and improved the predictive performance by seeking the adaptive optimization of the prediction coefficients based on the sequentialized context. As a significant alternative, the edge-directed prediction (EDP) in [5] figured out the edge-directed property of LS-based adaptation in the edge areas and proposed to initiate the LS-based adaptation only when the prediction error exceeds the predetermined threshold. Inspired by the edge-directed property, a switching predictor structure for lossless coding was proposed in [6] for areas with distinct local statistics: run-length coding for piecewise smooth areas and LS-based adaptation for the edge areas. The further improvements of the LS-based adaptation are also considered, which mainly lie on two aspects: weighting for obtained contexts by readjusting the eigenvalues of the convariance matrix [7] and adaptive sequentialization with minimum description length (MDL) [8]. However, the sequentialization destructs the morphology of 2-D context region and obscures inherent statistical correlation among the pixels in the correlated region. Such that spatial structures cannot be fully exploited to regularize the predictions.

As an alternative, spatial structures have been considered in the probabilistic modeling of the encoded pixels to compensate the pixel-wise prediction [9]. Furthermore, two-pass prediction schemes were proposed to enable mixture distribution and global image analysis beyond one-pass prediction. Typically, TMW used a blending of multiple probability distributions and a correlation-based segmentation to achieve higher coding performance for most natural images [10]. Furthermore, Matsuda *et al.* [11] proposed a generalized Gaussian model for image prediction error and categorized image blocks with variable size in terms of their variance, and achieves the best compression performance with relatively high computational complexity. However, such generative methods are restrictive because the smooth of prediction error is isolated by considering the variance of each error rather than directly consider the interdependencies among sets of prediction error. Moreover the two-pass predictive schemes require to transmit side information about the predictor.

The contribution of this paper is twofold. First, the discriminative learning-based model is proposed to directly measure and evaluate the inherent statistical correlation between 2-D contexts and their corresponding prediction for optimal individual pixel-wise prediction. Contrary to LS-based adaptation, the proposed model maintains the inherent statistical correlation by avoiding sequentilization and directly construct the mapping between the contexts and obtained prediction. The model parameters are learned off-line by discriminating the actual pixel value from the others to the maximal margin. Second, the data-driven structural interdependencies are formulated to regulate the individual pixel-wise prediction in a correlated region. Distinguished from the restrictive generative methods concerning the variance of prediction error, the formulation is proposed to maintain the coherence of prediction in the region and adaptively derived from the varying local statistics. The series of estimates are constrained with the structural interdependencies derived from the current region of pixels and optimized to minimize the joint code length.

To validate the efficacy of the predictor with the proposed model, we apply it into lossless image coding. The practical coder is block-based and the optimal prediction of each block is obtained with the sequential minimal optimization over the generated junction tree. It should be stressed that only one-pass coding is needed as the proposed predictor is causal and performs only based on past encoded samples. The proposed lossless technique commits, an average 1.05 percent shorter code length than the optimal predictor in [11].

The rest of the paper is organized as follows. In Section 2, we formulate the structured set prediction model for lossless image coding by deriving decomposable loss function and upper bound for prediction errors, and find the optimal solution over the generated Markov network. The practical coder based on the structured set prediction model is proposed in Section 3, where the general framework is also described. Experimental results for both oscillatory patterns from natural images and common grayscale test images are given in Section 4. Finally, we draw the conclusion in Section 5.

2. FORMULATION OF STRUCTURED SET PREDICTION MODEL

2.1. Problem Formulation

The structured set prediction model simultaneously considers the inherent statistical correlation for individual context-based prediction and the structural interdependencies for set of prediction in local regions. Consequently, the constrained concurrent training and prediction are performed for the proposed model, where the enforced constraints are derived from the local structural interdependencies and utilized to regulate the set of prediction. For the application into lossless image coding, the set prediction is defined as the joint prediction for block of pixels with fixed size. In view of the fact that actual code length is based on the underlying probability distribution for prediction error, the loss function is designed to relate such practical measurement. Denote $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ the collected set of training data, where $\mathbf{y}_i$ is the ith labeled block of pixels for predicting and $\mathbf{x}_i$ is the ith observed contexts for $\mathbf{y}_i$. The max-margin individual prediction is conducted under the class of feature functions $\{\mathbf{f}_i\}$, each of which is constructed as a basis function to distinguish the various context-based spatial statistics: $\mathbf{f}_i(\mathbf{y}) = P(\mathbf{y}|\mathbf{x}_i)$ Consequently, the prediction are made by combining the corresponding feature function representing various local statistics. In training, the concurrent optimization for set of prediction in the local region is sought to minimize the practical measurements derived from the well-defined loss function and the weighting vector $\mathbf{w}$ is tuned to fit. In subsequent prediction, the joint prediction is made by combining a series of feature functions $\{\mathbf{f}_i\}$ in the spanned space $\mathcal{F}$ with the learned normal vector $\mathbf{w}$. The min-max formulation of the max-margin Markov network is proposed by considering the constraints derived from the structural interdependencies among the pixels for predicting.

$$\begin{cases} \min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{s.t.} \quad \mathbf{w}^T\mathbf{f}_i(\mathbf{y}_i) + \xi_i \geq \max_{\mathbf{y}} \left(\mathbf{w}^T\mathbf{f}_i(\mathbf{y}) + \mathcal{L}(\mathbf{y}_i, \mathbf{y})\right) \quad \forall i \end{cases} \tag{1}$$

In Eq. (1), the weighting vector $\mathbf{w}$ is the normal vector perpendicular to the hyperplane of the feature functions, and $\{\xi_i\}$ is the slack vector which allows for the violations of the constraints at a cost proportional to $\{\xi_i\}$.

The pixels for predicting are naturally correlated with their spatial relationship, as shown in Fig. 1. When denote $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^{M}$ the collection of M pixels for predicting, the 2-D Markov random field is constructed accordingly where each edge clique contains the two neighboring pixels connecting by the edge.

2.2. Loss Function

Since there exists strong connection between the loss-scaled margin and the expected risk of the learned model, we are to make a study for the loss function in the loss-augmented inference. For the M-ary estimated output $\hat{\mathbf{y}}$, its approximation error is supposed to be measured by the loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$.

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_i \ell_i\left(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\right) \tag{2}$$

where $\ell_i(\cdot)$ is the loss function for the ith component $\hat{\mathbf{y}}^{(i)}$ of estimation $\hat{\mathbf{y}}$. Let $\epsilon_i = \hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}$ be the ith error led by the estimate $\hat{\mathbf{y}}^{(i)}$, and denote σ^2 the variance derived by the M errors $\{\epsilon_i\}_{i=1}^{M}$. Commonly, the probability density function for prediction errors is modeled by the Gaussian function with a variance of σ^2. The proposed loss function is required not only to indicate the number of pixels predicted incorrectly but to measure the exact code length led by the prediction errors.

$$\ell_i(\epsilon_i) = \log_2 \sqrt{2\pi\sigma^2} + \frac{\epsilon_i^2}{2\sigma^2}\log_2 e \tag{3}$$

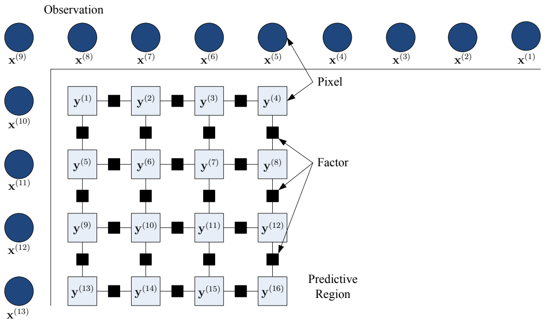


Fig. 1. Graphical model for the structured prediction model

where e is the base of the natural logarithms. Consequently, the solutions to the loss-augmented optimization problem will minimize the practical code length as defined by the loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$. According to the Markovian property of the grid-like Markov network shown in Fig. 1, the cumulative probability distribution of the correlated pixels is the production of all node and edge cliques.

According to D-separation theorem, we obtain

$$p(\mathbf{y}) = \prod_{1 \le i \le M} \prod_{j > i, j \in \text{ne}\{i\}} p\left(\mathbf{y}^{(j)} | \mathbf{y}^{(i)}\right) = p\left(\mathbf{y}^{(1)}\right) \cdot \prod_{C} \psi_C(\mathbf{y}_C)$$

where $\{\mathbf{y}^{\text{ne}(i)}\} = \{\mathbf{y}^{(j)} | j > i, j \in \text{ne}(i)\}$ includes all neighboring nodes $\mathbf{y}^{(j)}$ locating downside or rightside to $\mathbf{y}$, and $\psi_C(\cdot)$ is the potential function for the edge clique C. Referring to Eq. (3), the distribution over the states $\{\mathbf{y}_i\}$ is hence decomposable over the edges in the graphical model. Such that the log-Gaussian loss function under the probabilistic distribution is decomposable over the cliques in the graphical model.

2.3. Upper Bound for the Prediction Errors

In this section, we show that the upper bound for prediction error is asymptotically equivalent to the training error. Such upper bound allows us to relate the error on the training data to the prediction error. Consequently, when the weighting vector $\mathbf{w}$ has been well-tuned to fit the training data, the prediction error will not diverge owing to the consistency between training and prediction.

Extending the average error $\mathbf{L}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ for the blocks of M pixels to measure with the γ-margin hypersphere, we define a γ-margin per-label loss.

$$\mathbf{L}^{\gamma}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y}) = \sup_{\mathbf{y}' : \|\mathbf{w} \cdot \mathbf{f}(\mathbf{y}) - \mathbf{w} \cdot \mathbf{f}(\mathbf{y}')\| \le \gamma \mathcal{L}(\mathbf{y}, \mathbf{y}')} \frac{1}{M} \mathcal{L}(\mathbf{y}, \mathbf{y}').$$

The γ-margin per-label loss $\mathbf{L}^{\gamma}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ picks from all proper $\mathbf{y}'$ (satisfies $\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}) \le \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}')$) that maximizes the log-Gaussian measure from $\mathbf{y}$ in a $\gamma \mathcal{L}(\mathbf{y}, \mathbf{y}')$ wider hypersphere. It is actually closed to the loss in the proposed min-max formulation.

We can now show that the prediction and training is asymptotically consistent, which means that the upper bound for prediction error will converge to the training error with sufficient sampling.

Proposition 1. *For the trained normal vector* $\mathbf{w}$ *and arbitrary constant* $\eta > 0$*, the predictive error is asymptotically equivalent to the one obtained over the training data with probability at least* $1 - e^{-\eta}$.

The mean error of prediction and training is related as

$$\mathbb{E}_{\mathcal{X}} \mathbf{L}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y}) \le \mathbb{E}_{\mathcal{S}} \mathbf{L}^{\gamma}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y}) + o\left(\frac{\log N}{N}\right) \tag{4}$$

In Eq. (4), the first term bounds the training error based on $\mathbf{w}$. The low training error $\mathbb{E}_{\mathcal{S}} \mathbf{L}^{\gamma}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ can be achieved with the well-tuned weighting vector $\mathbf{w}$. Such that the performance of the prediction model can be assured with the low error $\mathbb{E}_{\mathcal{S}} \mathbf{L}^{\gamma}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ and high margin γ. The second term is the excess loss corresponding to the complexity of the predictor, which vanishes with the growth of sample size N. Thus the expected predictive per-label error is asymptotically equivalent to the γ-margin per-label error. Proposition 1 ensures the predictive performance by relating the theoretical upper bound for prediction to the tunable one for training. Since the loss derived by the log-Gaussian loss function is equivalent to the actual coding length led by the predictive error the coding cost led by the structured conditional prediction asymptotically approaches the training results with the sample size growth.

2.4. Solving Structured Set Prediction Model

Since the standard quadratic programming (QP) for Eq. (1) is often prohibitive in the structured set prediction model even for small training sets. To solve Eq. (1), we obtain its dual to use the coordinate dual ascent method analogous to the sequential minimal optimization (SMO). SMO breaks the optimization problem into a series of small QP problems and takes an ascent step that modifies the least number of variables.

$$\begin{cases} \max \left[v_i(\mathbf{y}') - v_i(\mathbf{y}'')\right] \delta - \dfrac{1}{2} C \|\mathbf{f}_i(\mathbf{y}') - \mathbf{f}_i(\mathbf{y}'')\|^2 \delta^2 \\ \text{s.t.} \alpha_i(\mathbf{y}) + \delta \ge 0, \alpha_i(\mathbf{y}'') - \delta \ge 0 \end{cases} \tag{5}$$

where $v_i(\mathbf{y}) = \mathbf{w} \cdot \mathbf{f}_i(\mathbf{y}) + \mathcal{L}(\mathbf{y}_i, \mathbf{y})$. and $\mathbf{f}_i(\mathbf{y}) = \left\{\sum_{k=1}^{K} \beta_j \mathbf{x}_{ik}^{(j)}\right\}_{j=1}^{M}$. The minimization process chooses the SMO pairs with respect to the KKT conditions. The KKT conditions are the sufficient and necessary criteria for optimality of the dual solution, which commits the certain locality for each example.

To maintain the spatial structures, we build the Markov Random Field for the α_i and v_i in each block, and accordingly calculate the marginal for each component in the random field to decide the SMO pairs. Since the generated Markov Random Field is not a chordal graph, it is firstly triangulated into a corresponding junction tree for cliques which can be obtained. Since the junction tree is not unique for each graphical model, we choose the chain-like junction tree for simplicity in the training and inference, and denote $\{J_i\}$ the nodes in the junction tree. When selecting the SMO pair, the states of labels in certain junction J_i are fixed, based on which the states in the other junctions are inferred. The SMO pair is chosen by finding the pairs of the series of states which maximize the margin. For each junction J_i, its potential is obtained by cumulating the potentials of its cliques.

$$\psi\left(J_i\right) = \prod_{C \in J_i} \psi_C\left(\mathbf{x}_C, \mathbf{y}_C\right) \tag{6}$$

The inference of the states in the junctions is made by passing the messages between the neighboring junctions. For each junction to predict in the grid, its most probable state and corresponding largest marginal probability are sought with the max-sum algorithm. The maximum marginal probability and the most probable state for junction J_p can be calculated. Consequently, the potential for each junction can be maximized. The maximization process is required to traverse over all the $\|\mathcal{A}\|^{\|J_p\|}$ states of the junction J_p for the alleged alphabet $\mathcal{A}$, which is too large even for the grayscale natural images. Since all the cliques in the graphical model are derived from the probability distribution condition on the states $\mathbf{y}_i$, the product $\prod$ and the maximization $\max$ can be exchanged. Therefore, the maximization process in each junction can be implemented by combining the maximized results of all its cliques. With the max-sum algorithm, the most probable states for cliques in all the junctions can be obtained as the candidate for sequential minimal optimization. The detailed The SMO process and KKT conditions for SMO pairs can refer to [12].

3. STRUCTURED SET PREDICTION MODEL BASED CODER

As shown in Fig. 2, we describe the practical lossless image coder where the estimates of the blocks of pixels are obtained from the structured set prediction model. The coding process is composed of two steps: training and prediction. The training step is based on the labeled data sampled from the natural images and is off-line. The mapping on the directions of blocks is built among the observation blocks and the label block for predicting. The direction of each block is associated with its pixels, which can be in terms of the statistics such as gradient orientation. It is viewed as the feature for fitting local statistics since the prediction based on the observations depends on the local motion directions. After the sampled data are categorized according to the orientation, it is easy to suppose that prediction in each class is characterized with one linear predictor. The parameters for the predictor are learned based on the classified sample data by building the max-margin Markov network. Both the mapping for classification and the learned parameters for prediction obtained from the off-line training are served as the priors for the next-step prediction. The constructed *a priori* distribution for prediction is the maximum posterior estimation for the distribution which the training data imply.

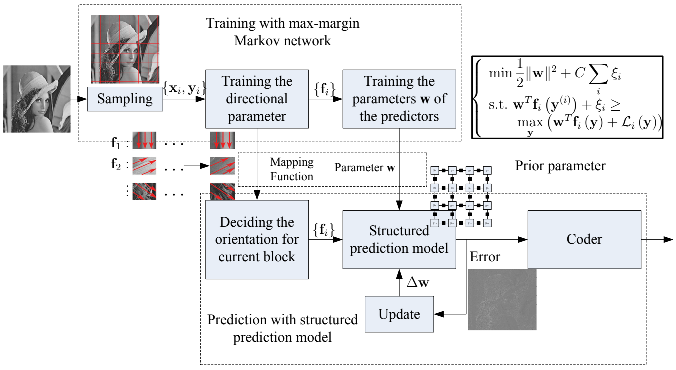

Training with max-margin Markov network
Sampling
$\{\mathbf{x}_i, \mathbf{y}_i\}$
Training the directional parameter
$\{\mathbf{f}_i\}$
Training the parameters $\mathbf{w}$ of the predictors
$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$
s.t. $\mathbf{w}^T\mathbf{f}_i\left(\mathbf{y}^{(i)}\right) + \xi_i \geq \max_{\mathbf{y}}\left(\mathbf{w}^T\mathbf{f}_i\left(\mathbf{y}\right) + \mathcal{L}_i\left(\mathbf{y}\right)\right)$
$\mathbf{f}_1$: . . .
$\mathbf{f}_2$: . . .
. . .
Mapping Function
Parameter $\mathbf{w}$
Prior parameter
Deciding the orientation for current block
$\{\mathbf{f}_i\}$
Structured prediction model
Coder
Error
$\Delta\mathbf{w}$
Update
Prediction with structured prediction model

Fig. 2. Illustrative framework for lossless image coder based on structured set prediction model

In the prediction step, we take the blocks that have been predicted as observations and classify the blocks for predicting according to their orientations. Then the estimation of all the pixels in each block is made by finding the joint max-margin estimation evaluated with the loss function. Here the minimization under the loss function will obtain the least code length under the Gaussian distributed entropy coder. Furthermore, to suppress the approximation errors led by the overfitting and the inconsistency with the training data, the online updates for predictors are proposed based on the errors obtained after the predictions.

The encoding engine has been developed in the range coder adopted by MRP. The prediction errors obtained by the structured set prediction model are combined with the minimum rate predictor, serving as an enhancement of the first loop prediction errors. In turn, the achieved prediction errors from the proposed predictor are used for the subsequent iterative optimization of the second loop. With the suppress of the errors around the oscillatory and texture regions, more efficient prediction is made.

3.1. Orientation for Blocks of Images

In the training step, we find the mapping function for the orientations of the blocks to predict and its neighbors. We sample the orientations from the blocks of observations along the boundary of the current predictive region at a pixel step. The orientation is measured by the Gaussian PSF G based gradient energy d of the blocks: $d = \|d_x\| + \|d_y\| = \|\frac{\partial G}{\partial x} \cdot \mathbf{I}\|^2 + \|\frac{\partial G}{\partial y} \cdot \mathbf{I}\|^2$ The orientation θ is obtained and categorized into 16 class at a gap of $\pi/16$, and we determine an orientation for the block by the variance with neural network with sigmoid function $\delta_i\left(\theta_i\right)=\frac{1}{1+e^{-\theta_i}}$ For each training data, the weight wd_i is updated by $\Delta wd_i = 2\eta(\theta-\theta_i)\delta_i\left(1-\delta_i\right)$ with the learning rate η.

3.2. Structured Set Prediction Model

The max-margin Markov network is established according to the graphical model in Fig. 1. The observation $\{\mathbf{x}, \mathbf{y}\}$ is obtained from the neighboring blocks that have been predicted, where $\mathbf{x}$ is the context for prediction in the training data and $\mathbf{y}$ the known prediction. The training process is conducted according to Eq. (1), where $\{\alpha_i\left(\cdot\right)\}$ and $\{v_i\left(\cdot\right)\}$ are obtained conditioning on the training data $\{\mathbf{x}, \mathbf{y}\}$. The learning parameter is set to 50 (equivalent to learning rate as 0.02). The normal vector $\mathbf{w}$ is iteratively learned with the sequential maximization optimization. In the prediction step, the prediction is obtained with the normal vector $\mathbf{w}$. To achieve the most probable prediction, the max-sum algorithm is conducted over the graphical model. The structured set prediction model infers the values of all the pixels simultaneously with the constraints characterized with the normal vector $\mathbf{w}$. Given L cliques with alphabet size $\|\mathbf{y}\|$ in each block, the computational complexity of the max-sum algorithm is $O\left(L\|\mathbf{y}\|^2\right)$, which means that the complexity is linear with the number of cliques.

3.3. Online Update

When the error ϵ_i for pixel x_i is obtained, we can update the weight vector by combining the observations in prediction. The reason for online update is that the prediction based on the training data would lead to the statistics minimizing the sample error in the data, and could not be optimized to suppress the approximation error led by the specific local feature of individual block. To asymptotically achieve the optimal prediction, we perform the online update after each prediction. In this step, the update for weight w_i of observation x_i is derived: $\Delta w_i = \frac{d\ell_i(\epsilon_i)}{dw_i}\cdot\epsilon_i$. Such that the normal vector $\mathbf{w}$ is readjusted along the direction that minimizing the error in the sense of the log-Gaussian loss function.

4. EXPERIMENTAL RESULTS

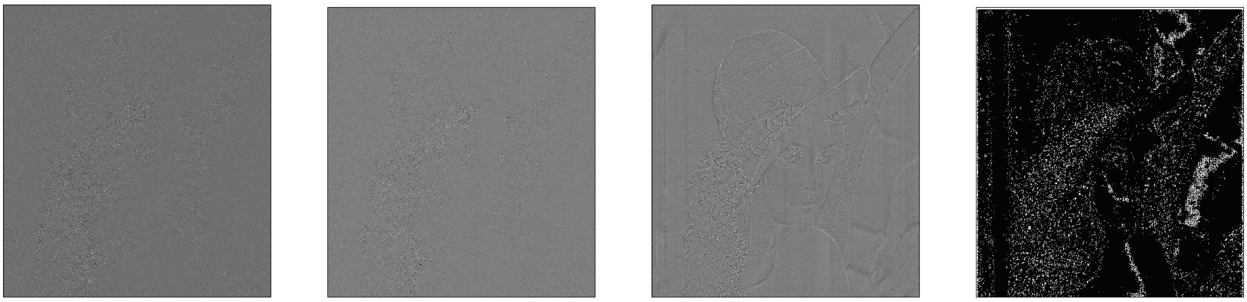
In this section, we present the experimental results and make comparison with the other predictors for validation. The size of pixels for simultaneous prediction is fixed to the 4×4 block with $M=16$. Corresponding to the boundary pixels neighboring leftside or upside to the blocks for prediction, the structural feature functions $\{\mathbf{f}_i\}$ serving as the contexts are set to 13 groups with each feature function indicating the relation of one context and one pixel for predicting in each group. A set of nine common grayscale test images are validated by comparing with the state-of-the-art lossless image coder MRP, TMW, and etc.

For validation, we compare the proposed structured set prediction-based-coder with the existing benchmark in lossless image coding. Their performance for lossless image coding is presented in Table 1. For generality of the validation, the selected test images span a wide rage in bit rates. It is witnessed that the proposed model is competitive as it achieves shorter code length than all the other coders. In a note of practical interest, the proposed method outperforms MRP, the best predictor reported, and witnesses an approximately 10% and 18% enhancement in bit rates over the JPEG 2000 lossless mode and JPEG-LS standard on average. This work improves the predictive performance of natural images with the structured set prediction model. Fig. 3(a)-(c) show the residual maps obtained from the proposed structured set prediction model, minimum rate predictor and edge directed predictor. The fact implies that the proposed model help to improve the predictive performance around oscillatory regions, which are often hard to deal in the individual context based adaptive methods. The binary map indicates where the proposed model is initiated when compared with MRP. It also validates that the proposed model is more efficient around the texture and oscillatory regions.

According to the implementation described in Section 3, the proposed coder based on the structured set prediction model commits the overheads indicating the model is initiated. For the sake of the fixed block size 4×4, the proposed coder has to consume more bits for identifying the block especially when the structural interdependencies is coherent for the blocks with larger size. However, the latest version of MRP optimizes for the supporting regions with variable block size ranging from 32×32 to 4×4, which helps reducing the cost in the description of the class of pixels. The proposed coder noticeably outperforms MRP when the block size in MRP is also fixed to 4×4. As shown in Table 2, the margin between the proposed coder and MRP with fixed block size is 3%. Such that there is still space for exploring the code cost of natural images when the proposed coder is designed to adapt the varying block size.

5. CONCLUSION

In this paper, the structured prediction model is proposed for the lossless image coding. The proposed model makes the prediction with multiple max margin estimation for each pixel in a correlated region, and subsequently, exploits the decomposition and combinatorial structure of the local prediction task. Evaluated with the well-defined loss function, the max margin Markov network is proposed for the pixel-wise prediction and the relevant parameters are trained. The prediction is bound to be asymptotically consistent with the training results with the decomposable loss function and the sufficient samples. For the practical coder, the training process gives the prior parameters for the prediction and the parameters are updated simultaneously with the progress of prediction. Ex-



(a) Proposed (b) MRP (c) EDP (d) Pixels predicted with the proposed model

Fig. 3. Prediction error maps for test image ”Lena” respectively obtained the proposed algorithm, the minimum rate predictor (MRP) and the edge directed predictor (EDP).

Table 1. Comparison with Existing Lossless Image Coders (bpp) for test image set 1

Image(size)	Proposed	MRP	BMF	TMW	Glicbawls	CALICa	JPEG-LS	JPEG 2000
Airplane(512×512)	3.539	3.591	3.602	3.601	3.668	3.743	3.817	4.013
Baboon(512×512)	5.648	5.663	5.714	5.738	5.666	5.875	6.037	6.107
Balloon(720×576)	2.526	2.579	2.649	2.649	2.640	2.825	2.904	3.031
Barb(720×576)	3.797	3.815	3.959	4.084	3.916	4.413	4.691	4.600
Barb2(720×576)	4.193	4.216	4.276	4.378	4.318	4.530	4.686	4.789
Couple(256×256)	3.349	3.388	3.448	3.446	3.543	3.609	3.699	3.915
Goldhill(720×576)	4.192	4.207	4.238	4.266	4.276	4.394	4.477	4.603
Lena(512×512)	3.871	3.889	3.929	3.908	3.901	4.102	4.238	4.303
Peppers(512×512)	4.166	4.199	4.241	4.251	4.246	4.421	4.513	4.629

perimental results show that its performance on oscillatory patterns is superior, where the regular feature can be caught by the training process.

6. REFERENCES

[1] J. Rissanen, “A universal data compression system,” *IEEE Trans. Inf. Theory*, vol. 29, no. 5, pp. 656–664, Sept. 1983.

[2] X. Wu and N. Memon, “Context-based adaptive lossless image coding,” *IEEE Trans. Commun.*, vol. 45, no. 4, pp. 437–444, Apr. 1997.

[3] M. Weinberger, G. Seroussi, and G. Sapiro, “The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS,” *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1309–1324, Aug. 2000.

[4] X. Wu and K. Barthel, “Piecewise 2D autoregression for predictive image coding,” in *Proc. Int. Conf. Image Process.*, Oct. 1998, vol. 3, pp. 901–905.

[5] X. Li and M. Orchard, “Edge-directed prediction for lossless compression of natural images,” *IEEE Trans. Image Process.*, vol. 10, no. 6, pp. 813–817, June 2001.

[6] L.-J. Kau and Y.-P. Lin, “Least-squares-based switching structure for lossless image coding,” *IEEE Trans. Circuits Syst. I*, vol. 54, no. 7, pp. 1529–1541, July 2007.

[7] H. Ye, G. Deng, and J. C. Devlin, “A weighted least-squares method for adaptive prediction in lossless image compression,” in *Proc. Picture Coding Symp.*, Sept. 2003, pp. 489–493.

[8] X. Wu, G. Zhai, X. Yang, and W. Zhang, “Adaptive sequential prediction of multidimensional signals with applications to lossless image coding,” *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 36–42, Jan. 2011.

[9] X. Zhao and Z. He, “Lossless image compression using super-spatial structure prediction,” *IEEE Signal Process. Lett.*, vol. 17, no. 4, pp. 383–386, Apr. 2010.

[10] B. Meyer and P. E. Tischer, “TMW - a new method for lossless image compression,” in *Proc. Int. Picture Coding Symp.*, Oct. 1997, pp. 533–538.

[11] I. Matsuda, N. Ozaki, Y. Umezu, and S. Itoh, “Lossless coding using variable block-size adaptive prediction optimized for each image,” in *Proc. 13th European Signal Process. Conf.*, Sept. 2005.

[12] B. Taskar, *Learning structured prediction models: A large margin approach*, Ph.D. thesis, Stanford Univ., CA, Dec. 2004.

Table 2. Comparison with the minimum rate predictor with fixed and variable block size (bpp)

	Airplane	Baboon	Balloon	Barb	Barb2	Couple	Goldhill	Lena	Peppers
Proposed	3.539	5.648	2.526	3.797	4.193	3.349	4.192	3.871	4.166
MRP with VBS	3.591	5.663	2.579	3.815	4.216	3.388	4.207	3.889	4.199
MRP with FBS	3.439	5.747	2.691	4.003	4.365	3.520	4.318	3.987	4.286