Large Discriminative Structured Set Prediction Modeling With Max-Margin Markov Network for Lossless Image Coding

Wenrui Dai, Hongkai Xiong, *Senior Member, IEEE*, Jia Wang, and Yuan F. Zheng, *Fellow, IEEE*

***Abstract*— Inherent statistical correlation for context-based prediction and structural interdependencies for local coherence is not fully exploited in existing lossless image coding schemes. This paper proposes a novel prediction model where the optimal correlated prediction for a set of pixels is obtained in the sense of the least code length. It not only exploits the spatial statistical correlations for the optimal prediction directly based on 2D contexts, but also formulates the data-driven structural interdependencies to make the prediction error coherent with the underlying probability distribution for coding. Under the joint constraints for local coherence, max-margin Markov networks are incorporated to combine support vector machines structurally to make max-margin estimation for a correlated region. Specifically, it aims to produce multiple predictions in the blocks with the model parameters learned in such a way that the distinction between the actual pixel and all possible estimations is maximized. It is proved that, with the growth of sample size, the prediction error is asymptotically upper bounded by the training error under the decomposable loss function. Incorporated into the lossless image coding framework, the proposed model outperforms most prediction schemes reported.**

***Index Terms*— Structured set prediction, max-margin Markov networks, lossless image coding, discriminative model.**

I. Introduction

ADVANCES in lossless image coding can be achieved through either 1-D sequential data compression or 2D context predictive coding, since the concept of context is constructed for universal sequential prediction by Rissanen [1]. A seminal work on sequential coding can be traced back to the Lempel-Ziv (LZ) dictionary compression algorithm in a raster-scanning order [2], which scans the input string till it finds one substring that is not in the dictionary, e.g. GIF, TIFF, and PNG coding. Specifically, wavelet-based branch [3]–[8] has been developed to achieve lossless or near-lossless compression and progressive reconstruction with the lifting structure. However, it is still inferior to pixel-domain predictive coding in lossless image coding.

Recently, context-based adaptive linear predictors [9] have achieved significant improvements through the fitting of the varying statistics. Among which, the state-of-the-art includes gradient adjusted predictor (GAP) in the context-based adaptive lossless image coder (CALIC [10]) and median edge detector (MED) in the low-complexity lossless compression for images (LOCO-I [11]). 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. Later, MED was proven to achieve the minimum entropy when the median of the mixture of symmetric distributed unimodels coincides [12]. However, these stationary linear predictors are not eligible in practice, as most natural images are far from being stationary.

Recognizing the nonstationary property, least-square (LS) autoregression based predictors have been considered as a significant alternative. In [13], the LS-based adaptation hypothesized that image signals were piecewise autoregressive (PAR) and improved the predictive performance by optimizing prediction coefficients. Traditionally, the contexts in LS-based predictors are obtained with the sequentialization of predicted pixels and the model parameters are estimated with the varying contexts featuring local statistics. Such predictors have been proven to be the maximum likelihood estimators for stationary Gaussian random processes. Furthermore, edge-directed prediction (EDP) in [14] figured out the edge-directed property of LS-based adaptation which inspired the LS optimization exactly in the edge area [15], [16]. The further improvements of LS-based adaptation involve weighting for the contexts and sequentializing for multidimensional signals [17], [18]. Although it favors individual prediction, the morphology of 2D context region would be destructed accordingly and inherent statistical correlation among the correlated region gets obscure. As an alternative, spatial structure has been considered to compensate the pixel-wise prediction. Inspired by the success of prediction by partial matching (PPM [19]) in sequential compression, [20] introduced the probabilistic modeling of the encoding symbol based on its previous context occurrences. In [21], super-spatial structure prediction aims to find an optimal prediction of the structure components, e.g. edges, patterns, and textures, within the previously encoded image regions instead of the spatial causal neighborhood.

Manuscript received March 26, 2013; revised November 16, 2013; accepted November 17, 2013. Date of publication December 3, 2013; date of current version December 17, 2013. The work was supported by the National Science Foundation of China under Grant U1201255, Grant 61271218, and Grant 61228101. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pascal Frossard.

W. Dai, H. Xiong, and J. Wang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: daiwenrui@sjtu.edu.cn; xionghongkai@sjtu.edu.cn; jiawang@sjtu.edu.cn).

Y. F. Zheng is with the Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH 43210 USA (e-mail: zheng@ece.osu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2293429

To further enhance the coding efficiency, two-pass prediction schemes were proposed to enable mixture distribution and global image analysis beyond one-pass prediction. Remarkably, TMW [22], [23] adopted a blending of multiple probability distributions and a correlation-based segmentation to achieve higher coding performance. In [24], an adaptive prediction was achieved by choosing the predictor that minimizes the energy of prediction error in a specified cluster of causal pixels and updating its coefficients with the gradient descent rule. To date, Matsuda *et al.* [25], [26] attained the minimum rate predictor (MRP) with the best compression performance by a generalized Gaussian model and block categorization with variable size in terms of the variance. These generative methods are limited by utilizing local statistics to smooth the prediction error in a local region because the smoothing is isolated without structural interdependencies. Moreover, two-pass prediction requires to transmit side information.

In this paper, we propose a lossless image coding scheme with discriminative structured set prediction (SSP) model which incorporates max-margin Markov networks (M3Ns). Contrary to LS-based adaptation, the proposed (SSP) model maintains the inherent statistical correlations by avoiding sequentialization and directly makes conditional prediction based on the observed 2D contexts by discriminative learning. Furthermore, the data-driven structural interdependencies are formulated to regulate the set prediction in a correlated region. This formulation is adaptively derived from the varying local statistics to maintain the coherence of the set prediction. Unlike the generative methods, the proposed model is optimized to minimize the joint code length under the constraints for local coherence.

Max-margin Markov networks leverage Markov networks to combine support vector machines (SVMs) structurally in order to make max-margin estimation for a set of pixels. Based on the obtained contexts, multi-class SVMs are trained to distinguish the actual value from other possible estimations. They are desirable for single prediction task, but ignore the structural interdependencies for local coherence of the set prediction. On the other hand, Markov networks can enforce local coherence by estimating the most likely joint probability assignment to the set of pixels. However, such methods do not usually achieve the prediction accuracy that is comparable to the max-margin estimation [30]. Combining the advantages of both, M3Ns unify Markov networks to enforce local coherence for the set prediction with the max-margin estimation to exploit statistical correlations for context-based prediction.

Therefore, the proposed model can jointly take into account the context-based prediction and the data-driven structural interdependencies in a local region. Concretely, the conditional prediction for the SSP model is achieved for most probable estimations based on the trained model parameters. The proposed model incorporates the max-margin Markov network to distinguish the actual values of the set of pixels from the other possible estimations. Based on the randomly collected training data, the model parameters (i.e. weighting vector) are iteratively optimized in the terms of log-Gaussian loss function. The optimized solution to the max-margin Markov network is obtained with the sequential minimal optimization (SMO) over the generated junction trees. Under the decomposable loss function, the clique-based estimation is able to be in parallel for the reduction of computational complexity.

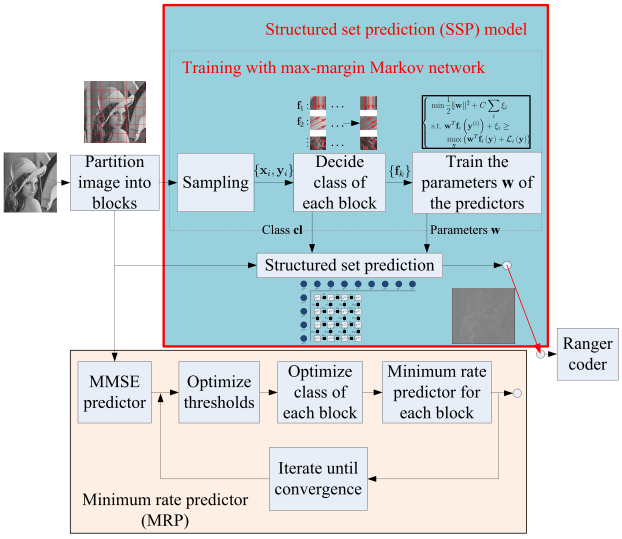


Fig. 1. The lossless image coder diagram with structured set prediction model.

The proposed model is well-established in the view of performance guarantee. The theoretical upper bound of its prediction errors is developed, which is demonstrated to asymptotically approach the training error with decomposable loss function and sufficient sampling. To validate the efficacy of the lossless coder, the prediction residual of the proposed model is coded by the available MRP engine [40] to generate the practical bits. It is worth mentioning that the proposed model is causal so that encoding and decoding can be synchronously achieved in one-pass coding. For an extensive range of natural images, the proposed lossless image coding scheme achieves a maximum 0.1 bpp (bit per pixel) gain in code length in comparison to MRP.

The rest of the paper is organized as follows. Section II describes the lossless image coder scheme with the proposed model. In Section III, the log-Gaussian loss function is designed and the upper bound of the prediction error is shown. Section IV provides the solution to the structured set prediction model based on the max-margin Markov network. Extensive experimental results are evaluated in Section V on natural images and common grayscale test images. Section VI concludes this paper.

II. General Framework

This section presents a general framework of lossless image coder with the SSP model, where the model parameters are trained to optimize the loss function by jointly considering the context-based spatial correlation and the structural interdependencies for local coherence.

A. The Proposed Coding Scheme

As shown in Fig. 1, each block of pixels will choose an optimal predictor that achieves least code length by comparing the proposed model with MRP. The set prediction of each block is characterized by the parameter set

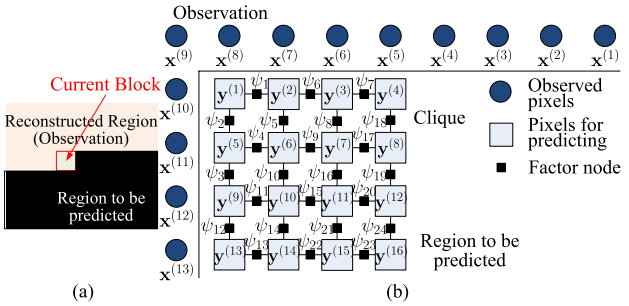


Fig. 2. The graphical model for the structured set prediction model, where $\{\mathbf{y}^{(i)}\}$ is the set of pixels to be predicted, $\{\mathbf{x}^{(i)}\}$ is the set of observed pixels serving as contexts.

PARAM $= \{\text{BLK\_SIZE}, \mathbf{cl}, \mathbf{w}\}$, where $\mathbf{cl} = \{cl\}$ is the set of class identifications for blocks and $\mathbf{w}$ is the trained weighting vector as model parameters. Note that, in the remainder of this paper, we reserve bold face symbols to vector variables and such symbols with superscripts to their components. The proposed model predicts blocks of pixels with $\mathbf{w}$, namely, the predictor is subtracted from the currently encoding block to generate the residual which is subsequently sent to the ranger coder. Each class identification cl corresponds to one group of model parameters $\mathbf{w}(cl)$. The coding cost of the proposed model comes from coding its prediction residual and the class identification of blocks.

$$J_{STRUCT} = C\,(\text{residual}) + C\,(\mathbf{cl})\,, \tag{1}$$

where $C\,(\cdot)$ is the cost function indicating the assigned code length. The coding cost of each block is evaluated and compared with the MRP to decide the active predictor.

$$\begin{aligned} J_{MRP} = {} & C\,(\text{residual}) + C\,(\text{Class}) \\ & + C\,(\text{Threshold}) + C\,(\text{Pred\_Coeff}) \end{aligned} \tag{2}$$

The structured set prediction model is formulated in Section II-B and II-C, where the set prediction for block of pixels and the training of model parameters are addressed, respectively. Its solution is composed of derivation of prime-dual formulation, generation of junction tree, message passing, and max-margin prediction, as described in Section IV.

B. The Structured Set Prediction Model

The class of feature functions $\mathcal{F} = \{\mathbf{f}_k\,(\mathbf{x}, \mathbf{y})\}$ is combined through the linear model with trained weighting vector $\mathbf{w}$ to make the joint prediction for block of pixels. The conditional probabilistic model for prediction is constructed over various contexts with structural interdependencies. Consider that $\mathbf{f}_k(\mathbf{x}, \mathbf{y})$ is conditioned on the k-th pixel in context $\mathbf{x} = \{\mathbf{x}^{(k)}\}_{k=1}^{K}$.

$$\mathbf{f}_k\,(\mathbf{x}, \mathbf{y}) = P\left(\mathbf{y}|\mathbf{x}^{(k)}\right)$$

In the graphical model of Fig. 2, $\mathbf{y}$ denotes the set of pixels being predicted and $\mathbf{x}$ is the causal context. Within the block of M pixels $\mathbf{y} = \left\{\mathbf{y}^{(j)}\right\}_{j=1}^{M}$ for predicting, each clique ψ of the Markov network represents two neighboring pixels linked with an edge. The set prediction is derived in a concurrent form of the linear combination of feature functions.

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \mathbf{w}^T \mathbf{f}\,(\mathbf{x}, \mathbf{y}), \tag{3}$$

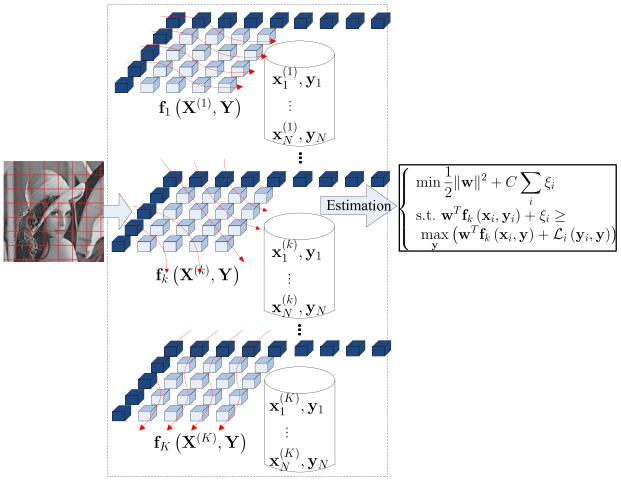


Fig. 3. Training of weighting vector $\mathbf{w}$ in structured set prediction model. The training data $\{\mathbf{x}_i, \mathbf{y}_i\}$ are sampled under the hypothesized structural constraints $\{\mathbf{f}_k\}$. The max-margin Markov network is trained by combining the class of feature functions $\{\mathbf{f}_k\}$ with the weighting vector $\mathbf{w}$.

where $\mathbf{f}$ is the collection of feature functions $\mathbf{f}_k$. As shown in Section II-C, the training of the weighting vector $\mathbf{w}$ is modeled as an optimization problem which considers both context-based spatial correlation and the interdependencies among the pixels for predicting.

C. Training Based Prediction Model

Fig. 3 illustrates training of the weighting vector $\mathbf{w}$ in the SSP model. Denote $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ the training set with sample size N, where $\mathbf{y}_i$ is the ith block of pixels for predicting and $\mathbf{x}_i$ is the observed context for $\mathbf{y}_i$. For optimal prediction of a single pixel y_i, multi-class SVMs [29] with soft margin can be utilized to train model parameters to relate y_i with its context $\mathbf{x}_i$, $1 \le i \le N$ and $1 \le j \le M$.

$$\begin{cases} \min \dfrac{1}{2}\|\mathbf{w}\|^2 + C\displaystyle\sum_i \xi_i \\ \text{s.t.} \quad \mathbf{w}^T \mathbf{f}\,(\mathbf{x}_i, y_i) + \xi_i \ge \max\limits_{y} \left[\mathbf{w}^T \mathbf{f}\,(\mathbf{x}_i, y) + \ell\,(y_i, y)\right] \quad \forall i \end{cases}$$

where ξ_i is the slack vector which allows for the violations of the constraints at a cost proportional to ξ_i and C is a constant. Evaluating with single loss function $\ell(\cdot)$, the actual value of pixel y_i is distinguished from the others to the maximum margin. However, such a formulation is isolated for predicting a set of pixels. Adopting joint loss function and feature functions for set prediction, M3Ns make max-margin estimation for a set of pixels regulated with local coherence for structural interdependencies. Hence, the min-max formulation based on the max-margin Markov network [27], [28], [41] is attained.

$$\begin{cases} \min \dfrac{1}{2}\|\mathbf{w}\|^2 + C\displaystyle\sum_i \xi_i \\ \text{s.t.} \quad \mathbf{w}^T \mathbf{f}\,(\mathbf{x}_i, \mathbf{y}_i) + \xi_i \ge \max\limits_{\mathbf{y}} \left[\mathbf{w}^T \mathbf{f}\,(\mathbf{x}_i, \mathbf{y}) + \mathcal{L}\,(\mathbf{y}_i, \mathbf{y})\right] \quad \forall i \end{cases} \tag{4}$$

where $\mathcal{L}(\mathbf{y}_i, \mathbf{y})$ is the joint loss function that measures the distance between $\mathbf{y}_i$ and $\mathbf{y}$. The weighting vector $\mathbf{w}$ is the

normal vector perpendicular to the hyperplane spanned by the class of feature functions $\{\mathbf{f}_k\}$. To adjust $\mathbf{w}$, the collection of training data $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ is iteratively arranged from 1 to N. C is related to the learning rate in the training based model. A large C will lead to the fine adjustment $\Delta\mathbf{w}$ of parameter $\mathbf{w}$ but with a slow convergence rate.

According to Eq. (4), the adjustment of $\mathbf{w}$ is based on the loss function $\mathcal{L}(\mathbf{y}_i, \mathbf{y})$. For practical lossless image coding, the log-Gaussian loss function is designed to relate the structured set prediction with the actual code length, as shown in Section III-A. With the class of feature functions $\mathcal{F}$ and the log-Gaussian loss function $\mathcal{L}(\mathbf{y}_i, \mathbf{y})$, weighting vector $\mathbf{w}$ is trained to achieve minimized code length for a set of pixels.

III. Formulation of the Structured Set Prediction Model

A. Loss Function

Since there exists strong connection between the loss-scaled margin and the expected loss of the training-based model, we study the loss function for the loss-augmented inference. Given the M-ary estimated output $\hat{\mathbf{y}}$ of the set of pixels $\mathbf{y}$, the prediction error is supposed to be measured by the loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{1 \le j \le M} \ell_j \left(\hat{\mathbf{y}}^{(j)}, \mathbf{y}^{(j)}\right), \tag{5}$$

where $\ell_j(\cdot)$ is the loss function for the prediction $\hat{\mathbf{y}}^{(j)}$ of the j-th pixel $\mathbf{y}^{(j)}$. Let $\epsilon_j = \hat{\mathbf{y}}^{(j)} - \mathbf{y}^{(j)}$ be the j-th error in the set of pixels, and σ^2 the variance derived by the M errors $\{\epsilon_j\}_{j=1}^{M}$. The log-Gaussian function with a variance of σ^2 is considered to measure prediction error.

$$\ell_j \left(\hat{\mathbf{y}}^{(j)}, \mathbf{y}^{(j)}\right) = \ell_j(\epsilon_j) = -\log_2 \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\epsilon_j^2}{2\sigma^2} \log_2 e \tag{6}$$

where e is the base of natural logarithms. Contrary to the $0/1$ loss, squared error loss function, and the deduced Hamming distance function, the underlying loss function should be designed to represent the exact code length from the structured set prediction. In the proposed model, the log-Gaussian loss function is adopted to be compatible with the coder in MRP [40], which estimates the Gaussian-like distributions for prediction error from various classes. Hence, the log-Gaussian loss function is related to the exact code length which can be represented by the total amount of information on blocks of prediction errors [32]. Later, the logarithm of generalized Gaussian function has ever been introduced to adapt a series of distributions including Laplacian and Gaussian [26]. However, a mixture of distributions derived by various shape and normalization parameters might not guarantee the decomposable property of the loss function, which would affect its theoretical upper bound. Furthermore, an upper bound for prediction error shall be provided under the loss function, as shown in Section III-B.

As follows, we define the decomposable loss function for the derivation of upper bound.

Definition 1. *The loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ is of decomposable, if it holds over the cliques in the graphical model $\mathcal{G}$*

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{c \in \mathcal{G}} \ell\left(\hat{\mathbf{y}}_c, \mathbf{y}\right), \tag{7}$$

where $\hat{\mathbf{y}}_c$ is the estimations of pixels in clique c.

According to Definition III-A, the designed log-Gaussian loss function is decomposable.

Proposition 1. *The loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ is decomposable.*
Proof: Please refer to Appendix A. ■

Since the loss function is decomposable, the loss in prediction and training can be viewed as a combination of loss from all individual pixels. Hence, the upper bound for the prediction error can be developed.

B. Upper Bound for the Prediction Error

Hereinafter, the upper bound for prediction error is shown to be asymptotically equal to the training error. In turn, the prediction error shall not diverge with the well-tuned weighting vector $\mathbf{w}$. As mentioned, the structured set prediction model minimizes the cumulative coding length of a correlated region in terms of the log-Gaussian loss function. Inspired by [27], we define the average prediction error $\mathbf{L}(\mathbf{w}^T\mathbf{f}, \mathbf{y})$ for the set of M pixels.

$$\mathbf{L}(\mathbf{w}^T\mathbf{f}, \mathbf{y}) = \frac{1}{M}\mathcal{L}\left(\arg\max_{\hat{\mathbf{y}}} \mathbf{w}^T\mathbf{f}(\mathbf{x}, \hat{\mathbf{y}}), \mathbf{y}\right)$$

To estimate the extreme case of the log-Gaussian loss function, its tight upper bound $\overline{\mathbf{L}}(\mathbf{w}^T\mathbf{f}, \mathbf{y})$ is obtained.

$$\overline{\mathbf{L}}(\mathbf{w}^T\mathbf{f}, \mathbf{y}) = \max_{\hat{\mathbf{y}}: \mathbf{w}^T\mathbf{f}(\mathbf{x},\mathbf{y}) \le \mathbf{w}^T\mathbf{f}(\mathbf{x},\hat{\mathbf{y}})} \frac{1}{M}\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) \tag{8}$$

Note that the upper bound is derived by picking all proper $\hat{\mathbf{y}}$ (satisfies $\mathbf{w}^T\mathbf{f}(\mathbf{x}, \mathbf{y}) \le \mathbf{w}^T\mathbf{f}(\mathbf{x}, \hat{\mathbf{y}})$) that can maximize the log-Gaussian loss function from $\mathbf{y}$. It is called tight because only when $\mathbf{y} = \arg\max_{\hat{\mathbf{y}}} \mathbf{w}^T\mathbf{f}(\mathbf{x}, \hat{\mathbf{y}})$, $\mathbf{L}(\mathbf{w}^T\mathbf{f}, \mathbf{y}) = \overline{\mathbf{L}}(\mathbf{w}^T\mathbf{f}, \mathbf{y})$ holds. Relating the average prediction error to the margin of the predictors, the upper bound $\overline{\mathbf{L}}(\mathbf{w}^T\mathbf{f}, \mathbf{y})$ is extended with the γ-margin hypersphere. The γ-margin relaxed loss is defined as:

$$\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f}, \mathbf{y}) = \sup_{\hat{\mathbf{y}}: \|\mathbf{w}^T\mathbf{f}(\mathbf{x},\mathbf{y}) - \mathbf{w}^T\mathbf{f}(\mathbf{x},\hat{\mathbf{y}})\| \le \gamma\mathcal{L}(\hat{\mathbf{y}},\mathbf{y})} \frac{1}{M}\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}). \tag{9}$$

The γ-margin relaxed loss $\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f}, \mathbf{y})$ similarly picks $\hat{\mathbf{y}}$ in a $\gamma\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ wider hypersphere, which is closed to the loss in the max-margin formulation Eq. (4).

Now, we show the consistency between prediction and training. In Proposition 2, we prove that the prediction and training are asymptotically consistent, which means that the prediction error is upper-bounded by the empirical γ-margin relaxed loss in training with the exception of an inversely growing additional term.

Proposition 2 *Given the trained weighting vector $\mathbf{w}$ and arbitrary constant $\eta > 0$, with sufficient sampling, there exists $\varepsilon(\mathcal{L}, \gamma, N, \eta) arrow 0$ satisfying*

$$P[\sup[\mathbb{E}_{\mathcal{X}}\mathbf{L}(\mathbf{w}^T\mathbf{f}, \mathbf{y}) - \mathbb{E}_{\mathcal{S}}\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f}, \mathbf{y})] \le \varepsilon] > 1 - \eta, \tag{10}$$

where $\mathbb{E}_{\mathcal{X}}\mathbf{L}(\mathbf{w}^T\mathbf{f},\mathbf{y})$ *and* $\mathbb{E}_{\mathcal{S}}\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f},\mathbf{y})$ *are the average prediction error and the average relaxed training error derived by the* γ*-margin relaxed loss, respectively.*

Proof: Please refer to Appendix B. ∎

In view of the average prediction error, Proposition 2 can be translated as: given the trained normal vector $\mathbf{w}$ and arbitrary $\eta > 0$, with probability of at least $1-\eta$, the prediction error satisfies

$$\mathbb{E}_{\mathcal{X}}\mathbf{L}(\mathbf{w}^T\mathbf{f},\mathbf{y}) \le \mathbb{E}_{\mathcal{S}}\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f},\mathbf{y}) + \varepsilon(\mathcal{L},\gamma,N,\eta). \quad (11)$$

In Eq. (11), the first term on the right side indicates the training error based on $\mathbf{w}$. The average training error $\mathbb{E}_{\mathcal{S}}\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f},\mathbf{y})$ can be reduced with the well-tuned weighting vector $\mathbf{w}$, such that the performance of prediction can be upper-bounded by the low training error $\mathbb{E}_{\mathcal{S}}\mathbf{L}^{\gamma}(\mathbf{w}^T\mathbf{f},\mathbf{y})$ and high margin γ. The second term is the excess loss corresponding to the complexity of the predictor. Eq. (24) in Appendix B shows that the excess loss vanishes with the growth of sample size N. Thus, the expected per-pixel prediction error is asymptotically equivalent to the γ-margin relaxed loss in training.

Proposition 2 ensures the predictive performance by relating the theoretical upper bound for prediction to the tunable one for training. Actually, since the loss derived by the log-Gaussian loss function meets with the actual coding length [32] by the prediction error, the average loss can be viewed in the sense of practical coding. It implies that, with sufficient sampling, the max-margin Markov network can asymptotically minimize the average coding length to the well-tuned loss over training data.

In learning-based methods, prediction tends to be efficient for the regions with regular features. From Proposition 2, the prediction with better performance is prospected by catching the local regular features in the training process. Applying Proposition 2 to the proposed model with log-Gaussian loss function, its prediction error is upper-bounded in Corollary 1.

Corollary 1. *Given the log-Gaussian loss function and the well-tuned parameter* $\mathbf{w}$ *from training, the average prediction error asymptotically approaches zero.*

Proof: Please refer to Appendix C. ∎

Corollary 1 implies that the prediction error derived from the proposed model approaches zero with the well-tuned training parameter $\mathbf{w}$ and sufficient sampling. Thus, the predictive performance of the proposed model is assured.

IV. Solving Structured Set Prediction Model

The standard quadratic programming (QP) for Eq. (4) is often prohibitive in the structured set prediction model with large state space. We obtain its dual and solve Eq. (4) by the coordinate dual ascent method similar to the sequential minimal optimization (SMO [33]). It extends SMO to the structured set prediction.

$$\begin{cases} \max \Big[\sum\limits_{i,\mathbf{y}} \alpha_i(\mathbf{y})\,\mathcal{L}(\mathbf{y}_i,\mathbf{y}) \\ \qquad -\frac{1}{2}\| \sum\limits_{i,\mathbf{y}} \alpha_i(\mathbf{y})\left(\mathbf{f}_i(\mathbf{x}_i,\mathbf{y}_i) - \mathbf{f}_i(\mathbf{x}_i,\mathbf{y})\right)\|^2 \Big] \\ \text{s.t.} \quad \sum\limits_{\mathbf{y}} \alpha_i(\mathbf{y}) = C, \alpha_i(\mathbf{y}) \ge 0, \quad \forall i \end{cases} \quad (12)$$

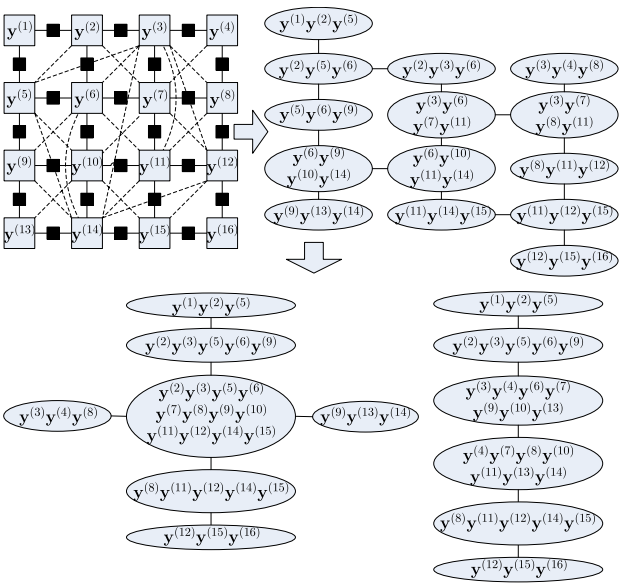


Fig. 4. Junction tree for the generated Markov network. (a, up-left) The triangulation of the graphical model by adding some dashed edges; (b, up-right) The intermediate result of construction of junction tree; (c, bottom) Two proper junction trees derived from the graphical model.

Eq. (12) can be solved by SMO, which breaks it into a series of small QP problems and takes an ascent step to update a least number of variables.

$$\begin{cases} \max \left[v_i(\hat{\mathbf{y}}) - v_i(\hat{\mathbf{y}}') \right]\delta - \frac{1}{2}C\|\mathbf{f}(\mathbf{x},\hat{\mathbf{y}}) - \mathbf{f}(\mathbf{x},\hat{\mathbf{y}}')\|^2\delta^2 \\ \text{s.t.} \quad \alpha_i(\mathbf{y}) + \delta \ge 0, \alpha_i(\hat{\mathbf{y}}') - \delta \ge 0 \end{cases} \quad (13)$$

where $v_i(\mathbf{y}) = \mathbf{w}^T\mathbf{f}(\mathbf{x},\mathbf{y}) + \mathcal{L}(\mathbf{y}_i,\mathbf{y})$. The minimization process chooses the SMO pairs with respect to the KKT conditions [34]. The KKT conditions are the sufficient and necessary criteria for optimality of the dual solution, which allow certain locality with respect to each example for repeatedly searching the optimal solution. The max-margin Markov network is built for each block with variable α_i and v_i. Their marginals are calculated to validate KKT conditions for deciding SMO pairs.

A. Junction Tree Based Solution

Since the generated Markov netwrok is not a chordal graph, it should be triangulated into a corresponding junction tree of all available cliques. The construction of junction tree is illustrated in Fig. 4, where the junction tree is not unique for the given graphical model. Henceforth, we choose the chain-like junction tree for simplicity in the training and prediction, and denote $\{J_i\}$ the nodes in the junction tree.

The SMO pairs are the pairs of possible estimations that maximize the margin. For each junction J_i, its potential is obtained by cumulating the potentials of its cliques.

$$\psi(J_i) = \prod_{C \in J_i} \psi_C(\mathbf{x}_C,\mathbf{y}_C) \quad (14)$$

When selecting the SMO pairs, the estimations of cliques in certain junction J_i are fixed, and the others are inferred based on them. These cliques are inferred by passing the messages between the neighboring junctions. For each junction, its most probable estimation and corresponding largest marginal probability are calculated with the max-sum algorithm. Fig. 5 shows

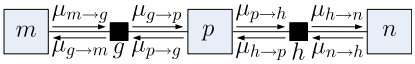


Fig. 5. Message passing in the generated junction tree.

the propagation of the local messages. For junction J_p, the messages that it sends and receives are

$$\mu_{g\to p}\left(J_p\right) = \max_{ne(g)\backslash p}\left[\ln g\left(J_p, J_m\right) + \sum_{m\in ne(g)\backslash p}\mu_{m\to g}\left(J_m\right)\right] \tag{15}$$

$$\mu_{p\to g}\left(J_p\right) = \sum_{h\in ne(p)\backslash g}\mu_{h\to p}\left(J_p\right) \tag{16}$$

where the neighboring factors g and h of junction p are associated with the definition of $\{\alpha_i(\cdot)\}$ or $\{\upsilon_i(\cdot)\}$. The maximum marginal probability and the most probable estimation for junction J_p can be calculated as:

$$p^{max} = \max_{J_p}\sum_{g\in ne(p)}\mu_{g\to p}\left(J_p\right) \tag{17}$$

$$J_p^{max} = \arg\max_{J_p}\sum_{g\in ne(p)}\mu_{g\to p}\left(J_p\right) \tag{18}$$

The potential for each junction is maximized.

$$\max\psi\left(J_p\right) = \max\prod_{C\in J_p}\psi_C\left(\mathbf{x}_C, \mathbf{y}_C\right) \tag{19}$$

The maximization process is required to traverse over all the $\|\mathcal{A}\|^{\|J_p\|}$ possible estimations in J_p for the alleged alphabet $\mathcal{A}$, which is too large even for the grayscale natural images. To simplify, we exchange the product and the maximization for the parallel solution to reduce the volume of possible estimations.

$$\max\psi\left(J_p\right) = \prod_{C\in J_p}\max\psi_C\left(\mathbf{x}_C, \mathbf{y}_C\right) \tag{20}$$

According to the Holder inequality, we can obtain

$$\max\prod_{C\in J_p}\psi_C\left(\mathbf{x}_C, \mathbf{y}_C\right) \le \prod_{C\in J_p}\max\psi_C\left(\mathbf{x}_C, \mathbf{y}_C\right).$$

Such that $\{\mathbf{x}_C, \mathbf{y}_C\}_{C\in J_p}$ that achieve maximization in Eq. (19) can satisfy Eq. (20). On the other hand, the neighboring cliques in the same junction are conditional independent according to the D-separation property [35]. Consequently, the multiplication and the maximization in Eq. (19) can be exchanged. According to Eq. (20), the maximization of each junction is implemented by combining the maximized results of all its cliques. With the max-sum algorithm, the most probable estimations for all the junctions can be obtained as the candidate for sequential minimal optimization. The SMO process is briefly described in Algorithm 1. The detailed KKT conditions for SMO pairs can be referred to [27].

Algorithm 1 Implementation with SMO

```
obj_old = 0
repeat
    obj_new = obj_old
    for all i do
        Initialize {υ_i(·)}, {α_i(·)} and violation=0
        Find violation ŷ and ŷ′ with KKT conditions
        if violation>0 then
            a = υ_i(ŷ) − υ_i(ŷ′)
            b = C‖f_i(ŷ) − f_i(ŷ′)‖²
            c = −α_i(ŷ)   d = α_i(ŷ′)
            δ = max(c, min(d, a/b))
            obj_new = obj_new − (1/2) a · δ
            update w and α_i with δ
        end if
    end for
until ‖1 − obj_new/obj_old‖ < 0.5
```

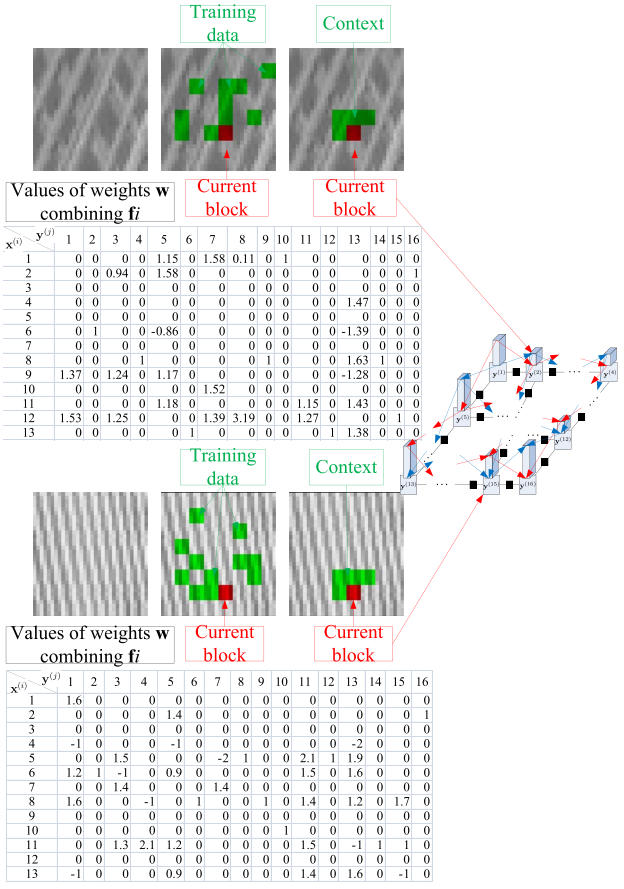
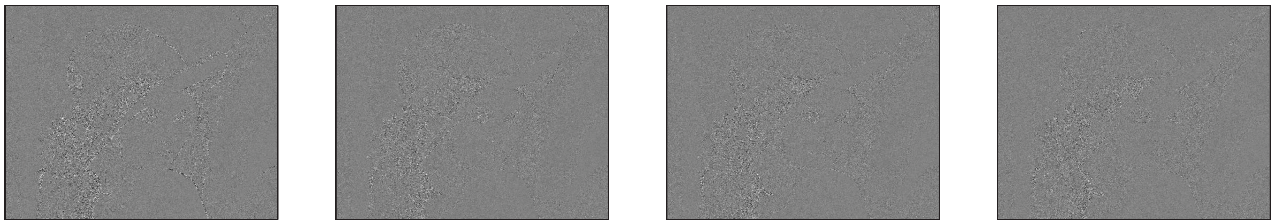


Fig. 6. Two examples of the training and prediction of the structured set prediction model. In each example, the selection of training data and contexts are indicated. For the specific block of pixels, its weights for combining the feature functions $\{\mathbf{f}_i\}$ indicating structural information are learned over the training data while such pixels are constrained with the Markov network in the form of the pixel states transition.

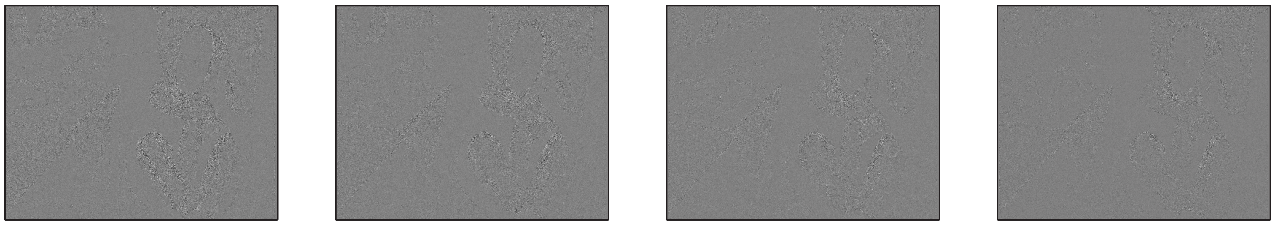
B. Discussion on the Structured Set Prediction Model

To clarify the philosophy of the proposed model, Fig. 6 shows the contexts and training data selection of the proposed model for a block of pixels. It shows that anisotropic distributions are constructed for weighting vector $\mathbf{w}$ with various class identifications. The potential for clique ψ_j is



(a) (b) (c) (d)

Fig. 7. The residual images for "Lena" with iterations 1, 2, 5 and 10. The first order entropy of the residual images is 3.977, 3.951, 3.928 and 3.904 bpp, respectively. (a) Iteration 1. (b) Iteration 5. (c) Iteration 10. (d) Iteration 50.



(a) (b) (c) (d)

Fig. 8. The residual images for "Barb" with iterations 1, 2, 5 and 10. The first order entropy of the residual images is 3.916, 3.901, 3.890 and 3.884 bpp, respectively. (a) Iteration 1. (b) Iteration 5. (c) Iteration 10. (d) Iteration 50.

(a) (b)

Fig. 9. Two sets of test images used in our experiments. (a) Test image set 1. From left-top to right-bottom: "Airplane", "Baboon", "Lena", "Peppers", "Balloon", "Barb", "Barb2", "Goldhill", "Couple", and "Cameraman". (b) Test image set 2. From left-top to right-bottom: "Boat1", "Boat2", "House", "Man", "Sailboat", "Cafe", "Monarch", "Beacon", "Clown", and "Milk".

obtained by

$$\psi_j\left(\mathbf{y}_{\psi_j}, \mathbf{x}\right) = \sum_k \mathbf{w}_k \mathbf{f}_k\left(\mathbf{y}_{\psi_j}, \mathbf{x}\right)$$

where weights $\mathbf{w}_k$ depend on the class identification of the block. The feature function $\mathbf{f}_k\left(\mathbf{y}_{\psi_j}|\mathbf{x}^{(i)}\right)$ implies the spatial statistics over the observed contexts.

$$\mathbf{f}_k\left(\mathbf{y}_{\psi_j}, \mathbf{x}\right) = P\left(\mathbf{y}_{\psi_j}|\mathbf{x}^{(k)}\right)$$

The set of pixels $\mathbf{y}_{\psi_j}$ that achieves the minimum loss for the linear combination is selected as the prediction of the proposed model. The feature function for different cliques ψ_j and $\psi_{j'}$ can be merged by

$$\mathbf{f}_k\left(\mathbf{y}_{\psi_j}, \mathbf{y}_{\psi_{j'}}, \mathbf{x}\right) = \mathbf{f}_k\left(\mathbf{y}_{\psi_j}, \mathbf{x}\right) \mathbf{f}_k\left(\mathbf{y}_{\psi_{j'}}, \mathbf{x}\right) \mathbb{I}\left(\mathbf{y}_{\psi_j}, \mathbf{y}_{\psi_{j'}}\right),$$

where $\mathbb{I}(\cdot)$ is the ising function. Accordingly, the feature function for the block of M pixels can be obtained.

Furthermore, we consider the consistency between training and prediction of the proposed model. Figs. 7 and 8 show the predictive results for "Lena" and "Barb" with the weighting vector $\mathbf{w}$ trained under 1, 5, 10, 50 iterations, respectively. In the iterative process, the weighting vector $\mathbf{w}$ is tuned with step $\Delta\mathbf{w} = 0.00625$. Obviously, the predictive performance becomes increasingly well with the growth of iterations. According to Proposition 2, the prediction performance will approach an upper bound. The residual in edge structure of the

TABLE I

COMPARISON WITH THE MINIMUM RATE PREDICTOR WITH FIXED AND VARIABLE BLOCK SIZE (bpp)

	Airplane	Baboon	Balloon	Barb	Barb2	Camera	Couple	Goldhill	Lena	Peppers	Average
Proposed	3.536	5.635	2.548	3.764	4.175	3.901	3.323	4.173	3.877	4.163	3.910
MRP with VBS	3.591	5.663	2.579	3.815	4.216	3.949	3.388	4.207	3.889	4.199	3.950
MRP with FBS	3.658	5.747	2.691	4.003	4.365	4.518	3.520	4.318	3.987	4.286	4.109
	Beacon	Boat1	Boat2	Cafe	Clown	House	Man	Milk	Monarch	Sailboat	Average
Proposed	4.089	4.591	4.355	4.775	3.880	3.732	4.425	3.353	4.023	3.557	4.078
MRP with VBS	4.165	4.639	4.401	4.839	3.992	3.753	4.469	3.412	4.159	3.624	4.145
MRP with FBS	4.225	4.734	4.482	5.031	4.105	3.865	4.555	3.466	4.226	3.695	4.238

TABLE II

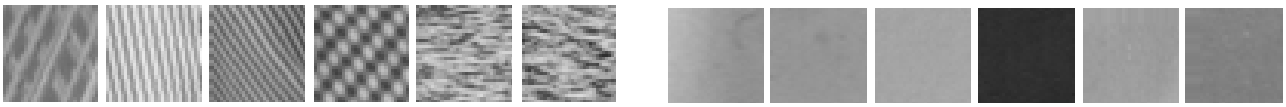
CODING PERFORMANCE (bpp) FOR THE PROPOSED (SSP) MODEL, EDP AND MRP IN OSCILLATORY REGIONS

	block 1	block 2	block 3	block 4	block 5	block 6	Average
SSP	**5.328**	**5.773**	**6.078**	**6.180**	**5.586**	**5.844**	**5.798**
EDP	5.742	6.531	6.609	6.547	7.125	7.195	6.625
MRP	6.834	6.467	6.271	6.693	8.521	8.576	7.227

TABLE III

CODING PERFORMANCE (bpp) FOR THE PROPOSED (SSP) MODEL, EDP AND MRP IN FLAT REGIONS

	block 1	block 2	block 3	block 4	block 5	block 6	Average
SSP	3.686	3.412	3.583	3.712	3.490	3.709	3.599
EDP	3.734	3.539	3.742	3.992	3.555	3.914	3.746
MRP	3.514	3.334	3.350	3.537	3.436	3.592	3.461



(a) (b)

Fig. 10. Selected 32×32 blocks in oscillatory and flat regions, respectively. (a) Selected 32×32 blocks in oscillatory regions for lossless coding. The left-up coordinates of the selected blocks are $(353, 6)$ and $(225, 511)$ in "Barb2", $(391, 415)$ and $(43, 366)$ in "Barb", and $(435, 64)$ and $(41, 63)$ in "Baboon", respectively. (b) Selected 32×32 blocks in flat regions for lossless coding. The left-up coordinates of the selected blocks are $(33, 33)$ and $(577, 481)$ in "Barb2", $(321, 193)$ and $(289, 33)$ in "Barb", and $(1, 1)$ and $(257, 1)$ in "Lena", respectively.

Algorithm 2 Proposed Scheme for Lossless Image Coding

```
/* Structured set prediction model */
Structured set prediction
Store class cl and residual, and calculate the coding cost J_STRUCT
/* Minimum rate predictor */
Initialization and utilize MMSE predictor as initial predictor
while Coding cost J_MRP does not converge do
    Optimize thresholds of each class
    Optimize class of each block
    Design MRP predictor by optimizing coefficients w.r.t. class and thresholds
    Calculate the coding cost J_MRP
end while
/* Compare the coding cost */
if J_STRUCT < J_MRP then
    Activate the structured set prediction model
else
    Activate the minimum rate predictor
end if
```

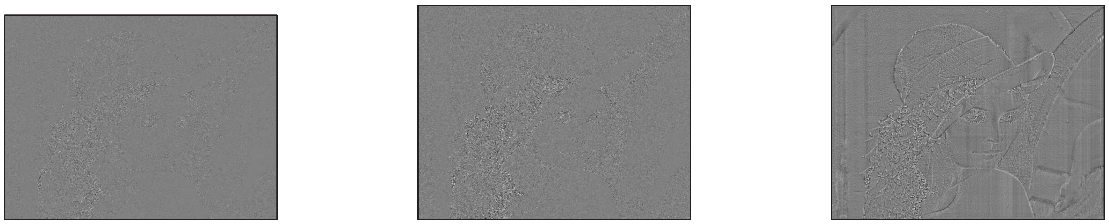
selected block is suppressed and the first order entropy also decreases, which provides sufficient evidence that the weighting vector tends to represent the anisotropic local statistics.

V. EXPERIMENTAL RESULTS

A. Implementation

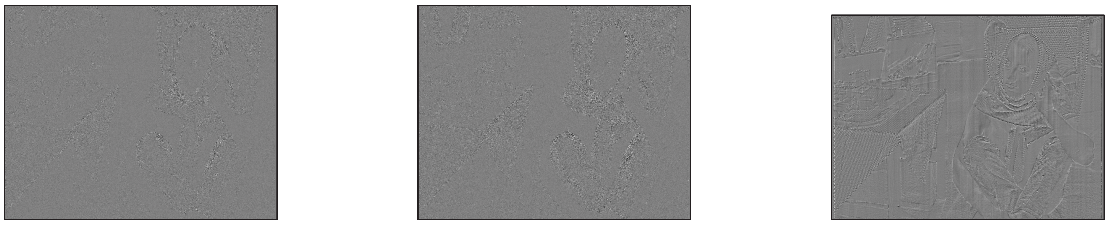
The size of each block for set prediction is 4×4 ($M = 16$). Corresponding to the boundary pixels neighboring leftside or upside to the blocks for prediction, the feature functions $\{\mathbf{f}_k\}_1^K$ are utilized to imply the spatial correlations directly conditioned on $K = 13$ contexts. For each block of pixels, $\mathbf{w}$ is selected according to its class identification. The class identifications of the proposed model are partly associated with MRP. Blocks of pixels are classified with 64 class identifications, which consider 8 variance intervals with 8 directions. The 8 variance intervals depend on the classes in MRP. Given a maximum D classes in MRP, the d-th variance interval in SSP is $[\lceil (d-1) \times D/8 \rceil, \lfloor d \times D/8 \rfloor]$. The training set $\mathcal{S}$ randomly collects 100 samples for each class, which means $N = 6400$ in this paper.

In the training process, the learning rate C is 50 to fine-tune the weighting vectors $\mathbf{w}$ which are obtained by optimizing loss function constrained by the local structures $\mathcal{F}$ over training data. Hence, the parameters are not required to be encoded and transmitted to the decoder. The procedure of the proposed scheme can be referred to Algorithm 2, where prediction residual from the SSP model is coded by ranger coder [36]. Two sets of grayscale test images (20 test images shown in Fig. 9)



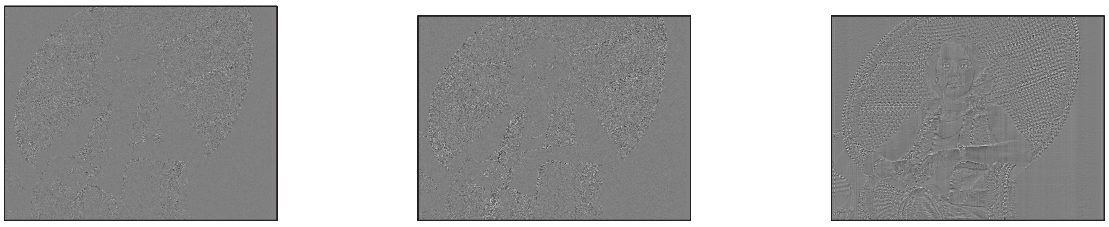
(a) (b) (c)

Fig. 11. Prediction error maps for test image "Lena" obtained by the proposed scheme, MRP and EDP. Their first order entropy from left to right is 3.901, 4.025, and 4.237 bpp, respectively. (a) Proposed. (b) MRP. (c) EDP.



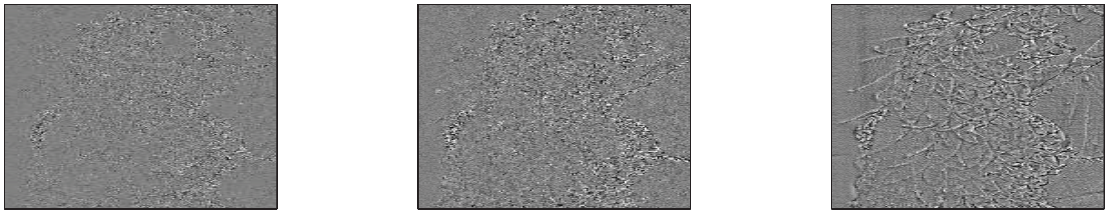
(a) (b) (c)

Fig. 12. Prediction error maps for test image "Barb" obtained by the proposed scheme, MRP, and EDP. Their first order entropy from left to right is 3.881, 3.918, and 4.352 bpp, respectively. (a) Proposed. (b) MRP. (c) EDP.



(a) (b) (c)

Fig. 13. Prediction error maps for test image "Barb2" obtained by the proposed scheme, MRP, and EDP. Their first order entropy from left to right is 4.447, 4.471, and 4.790 bpp, respectively. (a) Proposed. (b) MRP. (c) EDP.



(a) (b) (c)

Fig. 14. The zoomed detail of "hair" regions in "Lena" obtained by the proposed scheme, MRP, and EDP, respectively. (a) Proposed. (b) MRP. (c) EDP.

are: The first set is evaluated by MRP and TMW whose coding performance is directly taken from the benchmark [40]; The second set is natural images selected from the standard test sets, such as KODAK, USC SIPI, and etc.

B. Predictive Performance Compared to MRP With Fixed and Variable Block Size

Since the block size in the proposed model is 4×4, more bits are consumed for class identifications of blocks, especially in the smooth blocks with a large scale. Table I provides the coding gain of 4.2% on average over MRP with 4×4 block size, and up to 3% over MRP with variable block size ranging from 32×32 to 4×4. Therefore, more improvements can be achieved by designing the proposed model with variable block size.

C. Predictive Performance for Oscillatory and Flat Regions

The proposed SSP model is evaluated by coding 32×32 blocks, and Table II and III provide the coding performance in oscillatory and flat regions compared to MRP and EDP. Table II shows that the proposed model (SSP) is obviously more effective in oscillatory regions. In flat regions, it is slightly inferior to MRP because the prediction accuracy of the training-based model is interfered with the slight differences of pixel values. Thus, it is promising to combine the proposed SSP model with MRP to improve the universal coding performance.

D. Predictive Performance for Natural Images

To evaluate the performance of the proposed model, we compare the first order entropy of its residuals with other

TABLE IV

FIRST ORDER ENTROPY (bpp) OF THE NATURAL TEST IMAGES OBTAINED BY THE PROPOSED SCHEME, MRP AND EDP, RESPECTIVELY

	Airplane	Baboon	Balloon	Barb	Barb2	Camera	Couple	Goldhill	Lena	Peppers	Average
Proposed	**3.809**	**5.744**	**2.562**	**3.881**	**4.447**	**4.522**	**3.586**	**4.330**	**3.901**	**4.211**	**4.099**
MRP	3.855	5.879	2.606	3.918	4.471	4.635	3.665	4.371	4.025	4.270	4.170
EDP	4.133	6.012	3.101	4.352	4.790	5.122	4.077	4.597	4.237	4.512	4.493

	Beacon	Boat1	Boat2	Cafe	Clown	House	Man	Milk	Monarch	Sailboat	Average
Proposed	**4.315**	**4.782**	**4.401**	**4.860**	**4.104**	**4.109**	**4.589**	**3.430**	**4.123**	**3.745**	**4.246**
MRP	4.412	4.889	4.454	4.921	4.111	4.184	4.665	3.498	4.254	3.771	4.316
EDP	4.690	5.138	4.759	5.522	4.494	4.523	4.943	3.755	4.585	4.126	4.654

TABLE V

COMPARISON WITH EXISTING LOSSLESS IMAGE CODERS (bpp) FOR TEST IMAGE SET 1

Image(size)	Proposed	MRP	BMF [37]	TMW	Glicbawls [38]	CALIC	JPEG-LS	JPEG 2000	HD Photo
Airplane(512×512)	**3.536**	3.591	3.602	3.601	3.668	3.743	3.817	4.013	4.247
Baboon(512×512)	**5.635**	5.663	5.714	5.738	5.666	5.875	6.037	6.107	6.149
Balloon(720×576)	**2.548**	2.579	2.649	2.649	2.640	2.825	2.904	3.031	3.320
Barb(720×576)	**3.764**	3.815	3.959	4.084	3.916	4.413	4.691	4.600	4.836
Barb2(720×576)	**4.175**	4.216	4.276	4.378	4.318	4.530	4.686	4.789	5.024
Camera(256×256)	**3.901**	3.949	4.060	4.098	4.208	4.190	4.314	4.535	4.959
Couple(256×256)	**3.323**	3.388	3.448	3.446	3.543	3.609	3.699	3.915	4.318
Goldhill(720×576)	**4.173**	4.207	4.238	4.266	4.276	4.394	4.477	4.603	4.746
Lena(512×512)	**3.877**	3.889	3.929	3.908	3.901	4.102	4.238	4.303	4.477
Peppers(512×512)	**4.163**	4.199	4.241	4.251	4.246	4.421	4.513	4.629	4.850
Average	**3.910**	3.950	4.012	4.042	4.038	4.210	4.338	4.453	4.693

TABLE VI

COMPARISON WITH EXISTING LOSSLESS IMAGE CODERS (bpp) FOR TEST IMAGE SET 2

Image(size)	Proposed	MRP	BMF	CALIC	EDP	JPEG-LS
Beacon(768×512)	**4.089**	4.165	4.129	4.345	4.540	4.494
Boat1(512×512)	**4.591**	4.639	4.595	4.875	5.066	5.001
Boat2(512×512)	**4.355**	4.401	4.376	4.634	4.775	4.790
Cafe(768×512)	**4.775**	4.839	4.879	5.056	5.349	5.298
Clown(512×512)	**3.880**	3.992	3.815	3.941	4.386	4.255
House(512×512)	**3.732**	3.753	3.718	3.970	4.197	4.103
Man(512×512)	**4.425**	4.469	4.440	4.601	4.851	4.762
Milk(512×512)	**3.353**	3.412	3.343	3.532	3.753	3.636
Monarch(768×512)	**4.023**	4.159	2.883	4.023	4.497	4.324
Sailboat(768×512)	**3.557**	3.624	3.583	3.796	4.037	3.960
Average	**4.078**	4.145	3.976	4.277	4.545	4.462

predictors, including MRP, EDP, and etc. The orders of EDP and MRP are fixed at 12. Table IV shows the first order entropy of the residuals, where the proposed model achieves 0.07 bpp and 0.40 bpp on average less than MRP and EDP, respectively. It implies that the histogram of residual is more concentrated. To reflect the effect of capturing regular features in images, Figs. 11–13 illustrate the residual maps obtained by the proposed model, MRP and EDP. These residual maps are refined by magnifying the contrast of its prediction errors. It can be observed that the proposed model is apt to characterize the varying local statistics, which is even evident around the oscillatory and texture regions. Fig. 14 shows three zoom-in examples of prediction residue where the proposed model achieves most illegible effect.

Table V and VI provide the lossless code lengths of the natural images achieved by the proposed scheme in Section II-A. For validation, it is compared with benchmarks in existing lossless image coders and spans a wide rage in bit rates over the test images. It can be seen that the proposed model is competitive as it achieves the shortest code length, and outperforms MRP, the best predictor reported, by an average margin of 0.054 bpp and the coding gain of up to 3%. In a note of practical interest, it achieves an approximately 10% and 14% improvement in bit rates over the JPEG 2000 lossless mode and JPEG-LS standard on average.

E. Computational Complexity

Once the weighting vector $\mathbf{w}$ is trained, the computational complexity of the proposed model is equivalent to the complexity of the max-sum algorithm. Given L cliques with alphabet size $\|\mathbf{y}\|$ in each block, its complexity is $O\left(L\|\mathbf{y}\|^2\right)$, which means that it is linear with the number of cliques. For the graphical model of $M_x \times M_y$ block, there are $L = (M_x - 1)\, M_y + M_x \left(M_y - 1\right)$ cliques in total.

In practice, the encoder and decoder operate on a PC with a 3.2GHz Intel Core i7 processor and complied with VC++ 9.0 with same configuration (“DEBUG” mode). Given a 4×4 block, L is set to 24 and $\|\mathbf{y}\|$ is 256 for the 8 bits grayscale images. In the encoder side, the proposed model

TABLE VII

COMPUTATIONAL COMPLEXITY OF THE PROPOSED MODEL AND THE MINIMUM RATE PREDICTOR (MRP). RUN-TIME RATIOS (%) ARE ASSESSED AS: RT_RATIO $= t_{PROP}/t_{MRP} \times 100\%$

	Airplane	Baboon	Balloon	Barb	Barb2	Camera	Couple	Goldhill	Lena	Peppers	Average
Proposed	347.22	425.01	865.35	936.97	1078.08	85.59	95.09	1004.75	365.21	410.93	-
MRP with FBS	28.15	34.71	35.69	78.02	89.13	13.97	7.58	36.63	29.20	44.42	-
Run-time ratio	1233.46	1224.46	2424.63	1200.94	1209.56	612.67	1254.49	2742.97	1250.72	925.10	1407.90
MRP with VBS	97.92	144.07	217.50	345.20	284.93	30.26	16.07	340.24	102.45	152.55	-
Run-time ratio	354.60	295.00	397.86	371.43	378.34	282.85	591.72	295.31	356.48	269.37	359.30
	Beacon	Boat1	Boat2	Cafe	Clown	House	Man	Milk	Monarch	Sailboat	Average
Proposed	1123.11	358.75	392.05	763.05	387.11	330.70	397.26	361.09	839.49	859.01	-
MRP with FBS	86.51	26.96	37.09	47.14	31.77	19.34	42.23	35.08	33.11	83.38	-
Run-time ratio	1298.24	1330.68	1057.02	1618.69	1218.48	1709.93	940.71	1029.33	2535.46	1030.24	1376.88
MRP with VBS	311.70	99.78	124.00	226.49	94.15	72.76	108.25	64.01	477.14	331.92	-
Run-time ratio	360.32	359.54	316.17	336.90	411.16	454.51	366.98	564.11	175.94	258.80	360.44

TABLE VIII

SELECTION RATIO (%) OF THE SSP MODEL UNDER VARIOUS TRAINING ITERATIONS IN TEST IMAGES "LENA", "BARB", AND "BARB2"

	iteration=1	iteration=2	iteration=5	iteration=10	iteration=50
Lena	8.66	8.97	10.79	13.01	14.58
Barb	5.27	5.82	8.55	12.86	14.55
Barb2	3.46	3.98	7.83	12.52	14.81

(a) (b) (c)

Fig. 15. Binary maps indicating where the proposed structured set prediction model performs better when compared with MRP in practical coding. In detail, the counts of pixels for "Lena", "Barb", and "Barb2" are 38224 (14.58%), 60336 (14.55%), and 61424 (14.81%), respectively. (a) Lena. (b) Barb. (c) Barb2.

would be selected in the sense of least code length. Table VII shows the decoding speed of the proposed scheme and the minimum rate predictor (MRP), where run-time ratio over MRP with both VBS and FBS is used to assess computational complexity. Depending on the selection ratio of the proposed SSP model, the run-time ratio ranges from 175% to 600%. The encoding speed is approximately 600 pixels per second. In fact, the complexity of the max-sum algorithm can be improved by a stochastic gradient decent algorithm [39] which solves large scale prediction problems with a decomposable and differentiable loss function.

F. Selection Ratio

Fig. 15 shows the selection map of the proposed model when compared with MRP in test images, e.g., "Lena", "Barb", and "Barb2". It reveals that the proposed model is mainly distributed in the regions of edges or textures. Table VIII shows the selection ratio varies with the growth of iterations in training. With the growth of iterations, the performance of the proposed model is improved and the selection ratio of the structured set prediction model increases. It demonstrates that the predictive performance approaches an upper bound derived by the training error.

VI. CONCLUSION

In this paper, a structured set prediction model with max-margin Markov networks is proposed for lossless image coding. It exploits the decomposition and combinatorial structure of the local prediction task, and makes the conditional prediction with multiple max-margin estimation in a correlated region. With the well-defined decomposable loss function relevant to actual code length, the max-margin Markov network combines support vector machines structurally and obtain the model parameters to make a maximized distinction between the actual pixel and all possible estimations. The prediction error is demonstrated to be asymptotically upper bounded by the training error with sufficient sampling, and the theoretical upper bound of prediction errors is also provided. For the practical coder, the data are arranged in form of blocks and classified with their variance and orientation. The structured set model is optimized to minimize the joint code length for the prediction residual, and outperforms most prediction schemes in literature.

APPENDIX A
PROOF OF PROPOSITION 1

When formulated with the grid-like Markov network, the cumulative probability distribution of the correlated pixels is the production of all node and edge cliques. As shown in Fig. 2, each edge in the graphical model corresponds to one clique. According to the Markovian property, we can find

$$p(\mathbf{y}) = \prod_{1 \le i \le M} p\left(\left\{\mathbf{y}^{ne(i)}\right\} | \mathbf{y}^{(i)}\right) \tag{21}$$

where $\{\mathbf{y}^{\mathrm{ne}(i)}\}$ includes all neighboring nodes $\mathbf{y}^{(j)}$ locating downside or rightside to $\mathbf{y}$.

$$\left\{\mathbf{y}^{\mathrm{ne}(i)}\right\} = \left\{\mathbf{y}^{(j)} | j > i,\, j \in \mathrm{ne}(i)\right\}$$

According to D-separation theorem, $\mathbf{y}^{(i)}$ and its set of neighboring nodes $\left\{\mathbf{y}^{\mathrm{ne}(i)}\right\}$ are conditional interdependent. Thus, it holds

$$p\left(\left\{\mathbf{y}^{\mathrm{ne}(i)}\right\} | \mathbf{y}^{(i)}\right) = \prod_{j>i,\, j \in \mathrm{ne}(i)} p\left(\mathbf{y}^{(j)} | \mathbf{y}^{(i)}\right)$$

Consequently, we obtain

$$\begin{aligned} p(\mathbf{y}) &= \prod_{1 \le i \le M} \prod_{j>i,\, j \in \mathrm{ne}\{i\}} p\left(\mathbf{y}^{(j)} | \mathbf{y}^{(i)}\right) = p(\mathbf{y}_1) \cdot \prod_C \psi_C(\mathbf{y}_C) \\ &= \frac{1}{Z} \prod_C \psi_C(\mathbf{y}_C). \end{aligned} \tag{22}$$

In Eq. (22), $\psi_C(\cdot)$ is the potential function for the clique C, which corresponds to the edge in the graphical model, and Z is the normalized parameter. Referring to Eq. (6), the distribution over the states $\{\mathbf{y}_i\}$ is hence decomposable over the edges in the graphical model. Since the log-Gaussian function for errors $\{\epsilon_i\}$ is decomposable over these errors and the errors is linear with the variation of estimation $\left\{\hat{\mathbf{y}}_i\right\}$, the log-Gaussian loss function is decomposable over the cliques in the graphical model.

APPENDIX B
PROOF OF PROPOSITION 2

Since the loss function $\mathcal{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)$ is proven to be decomposable in Proposition I, the results in [31] can be extended for the decomposable loss function here. Given arbitrary constant $\varepsilon > 0$ and the decomposable loss function $\mathcal{L} : \mathbb{R}^M \to [0, 1]$, the prediction error can be asymptotically upper-bounded by the γ-margin relaxed loss over training data.

$$\begin{aligned} P[\sup[\mathbb{E}_{\mathcal{X}} \mathbf{L}(\mathbf{w}^T \mathbf{f}, \mathbf{y}) - \mathbb{E}_{\mathcal{S}} \mathbf{L}^{\gamma}(\mathbf{w}^T \mathbf{f}, \mathbf{y})] \le \varepsilon] \\ > 1 - 4\mathbb{E}[\mathcal{N}_{\infty}(\mathcal{L}, \gamma, X_1^N)] e^{-\frac{N\varepsilon^2}{32}} \end{aligned} \tag{23}$$

where $\mathcal{N}_{\infty}(\mathcal{L}, \gamma, \mathcal{S})$ is the infinity covering number of the sample $\mathcal{S}$. Denote $\mathcal{N}_{\infty}(\mathcal{L}, \gamma, N) = \sup_{X_1^N} \mathcal{N}_{\infty}\left(\mathcal{L}, \gamma, X_1^N\right)$ the supremum of the infinity covering number. Defining

$$\eta = 4\mathcal{N}_{\infty}(\mathcal{L}, \gamma, N) e^{-\frac{N\varepsilon^2}{32}},$$

$\varepsilon = \varepsilon(\mathcal{L}, \gamma, N, \eta)$ can be formulated as

$$\varepsilon = \sqrt{\frac{32}{N}\left(\ln 4\mathcal{N}_{\infty}(\mathcal{L}, \gamma, N) + \ln \frac{1}{\eta}\right)}. \tag{24}$$

For the M-ary decomposable loss function, the upper bound of its infinity covering number is derived in [28]

$$\ln \mathcal{N}_{\infty}(\mathcal{L}, \gamma, N) \le 36(p-1) \frac{a^2 b^2}{\gamma^2} \ln(2\lceil 4ab/\gamma \rceil N + 1).$$

where $\|\mathbf{x}\|_p \le b$, $\|\mathbf{w}\|_q \le a$ and $1/p + 1/q = 1$, such that we can draw the conclusion that with sufficient sampling,

$$\frac{32}{N} \cdot \ln 4\mathcal{N}_{\infty}(\mathcal{L}, \gamma_i, \mathcal{S}) \sim o\left(\frac{\log N}{N}\right) \to 0. \tag{25}$$

Consequently, given arbitrary η, ε in Eq. (24) vanishes when $N \to \infty$. As a conclusion, since

$$\mathbb{E}\left[\mathcal{N}_{\infty}\left(\mathcal{L}, \gamma, X_1^N\right)\right] \le \sup_{X_1^n} \mathcal{N}_{\infty}\left(\mathcal{L}, \gamma, X_1^N\right) = \mathcal{N}_{\infty}(\mathcal{L}, \gamma, N), \tag{26}$$

Proposition 2 is drawn from Eq. (23) and (25).

APPENDIX C
PROOF OF COROLLARY 1

At first, we recall the γ-margin relaxed loss function $\mathbf{L}^{\gamma}\left(\mathbf{w}^T \mathbf{f}, \mathbf{y}\right)$:

$$\mathbf{L}^{\gamma}\left(\mathbf{w}^T \mathbf{f}, \mathbf{y}\right) = \max_{\hat{\mathbf{y}}: \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}) \le \mathbf{w}^T \mathbf{f}(\mathbf{x}, \hat{\mathbf{y}}) + \gamma \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})} \frac{1}{M} \mathcal{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right).$$

$\mathbf{L}^{\gamma}\left(\mathbf{w}^T \mathbf{f}, \mathbf{y}\right)$ is derived from the loss function $\mathcal{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right)$, which is defined as the logarithms of multivariate Gaussian with zero mean and variance σ^2.

$$\mathcal{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = \mathcal{L}(\epsilon) = \sum_i \ell_i(\epsilon_i) = \sum_i \left(-\log_2 \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\epsilon_i^2}{2\sigma_i^2}}\right)$$

where $\epsilon = \mathbf{y} - \hat{\mathbf{y}}$. Since the loss function is decomposable, each of its component can be considered individually. Without loss of generality, we consider the ith component $\ell_i(\epsilon_i)$.

$$\ell_i(\epsilon) = -\log_2 \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{\epsilon_i^2}{2\sigma_i^2}}.$$

For $\mathbf{y}^{(i)}$, it is relaxed by $\gamma \ell_i\left(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}\right)$.

$$\mathbf{w}_i^T \mathbf{f}\left(\mathbf{x}, \mathbf{y}^{(i)}\right) \le \mathbf{w}^T \mathbf{f}\left(\mathbf{x}, \hat{\mathbf{y}}^{(i)}\right) + \gamma \ell_i\left(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}\right)$$

The γ-margin relaxed loss function is tight when $\mathbf{y}^{(i)} = \arg\max_{\hat{\mathbf{y}}^{(i)}} \left[\mathbf{w}_i^T \mathbf{f}\left(\mathbf{x}, \hat{\mathbf{y}}^{(i)}\right) + \gamma \ell_i\left(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}\right)\right]$, which is also the boundary of the γ-margin relaxed region of $\mathbf{y}$. Such that the loss function ℓ_i is relaxed by $\epsilon_{\gamma}^{(i)}$:

$$\begin{aligned} \epsilon_i^{\gamma} = &\arg\max_{\hat{\mathbf{y}}^{(i)}} \left[\mathbf{w}_i^T \mathbf{f}\left(\mathbf{x}, \hat{\mathbf{y}}^{(i)}\right) + \gamma \ell_i\left(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}\right)\right] \\ &- \arg\max_{\hat{\mathbf{y}}^{(i)}} \mathbf{w}_i^T \mathbf{f}\left(\mathbf{x}, \hat{\mathbf{y}}^{(i)}\right) \end{aligned}$$

According to the definition of the γ-margin relaxed loss function in Eq. 9, we can obtain that

$$\ell_i^{\gamma}(\epsilon_i) = \ell_i^{\gamma}\left(\mathbf{y}^{(i)}, \hat{\mathbf{y}}^{(i)}\right) = \begin{cases} \ell_i\left(\epsilon_i - \epsilon_i^{\gamma}\right) & \epsilon_i \le \epsilon_i^{\gamma} \\ \ell_i(\epsilon_i) & \epsilon_i > \epsilon_i^{\gamma}/2 \end{cases} \tag{27}$$

In Eq. (27), the mean of $\ell_i^{\gamma}(\epsilon_i)$ is $\epsilon_i^{\gamma}/2$. Such that we can derive the mean vector ϵ^{γ} of the γ-margin relaxed loss function $\mathbf{L}^{\gamma}\left(\mathbf{w}^T\mathbf{f}, \mathbf{y}\right)$. It is obvious that ϵ^{γ} is zero when

$$\mathbf{y} = \arg\max_{\hat{\mathbf{y}}} \mathbf{w}^T \mathbf{f}\left(\mathbf{x}, \hat{\mathbf{y}}\right) + \gamma \mathcal{L}\left(\hat{\mathbf{y}}, \mathbf{y}\right) \tag{28}$$

Eq. (28) can be achieved or approximately achieved by tuning the parameters $\mathbf{w}$. When given margin γ and sampled data $\mathcal{S}$ in training, the mean can be closed to zero with the well-tune parameter $\mathbf{w}$. Moreover, since the excess term $\varepsilon(\mathcal{L}, N, \gamma, \eta)$ is shown to vanishes with the growth of N in Proposition 2, the mean of the prediction error asymptotically equals to the one of training error. As a result, the average prediction error can be zero with the well-tuned parameters $\mathbf{w}$ and the log-Gaussian loss function.

REFERENCES

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

[2] J. Ziv, "A universal prediction Lemma and applications to universal data compression and prediction," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1528–1532, May 2001.

[3] D. Taubman and M. W. Marcellin, *JPEG2000, Image Compression, Fundamentals, Standards and Practice*. Norwell, MA, USA: Kluwer, 2002.

[4] J. Reichel, G. Menegaz, M. J. Nadenau, and M. Kunt, "Integer wavelet transform for embedded lossy to lossless image compression," *IEEE Trans. Image Process.*, vol. 10, no. 3, pp. 383–392, Mar. 2001.

[5] N. V. Boulgouris, D. Tzovaras, and M. G. Strintzis, "Lossless image compression based on optimal prediction, adaptive lifting, and conditional arithmetic coding," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 1–14, Jan. 2001.

[6] M. Grangetto, E. Magli, M. Martina, and G. Olmo, "Optimization and implementation of the integer wavelet transform for image coding," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 594–604, Jun. 2002.

[7] A. T. Deever and S. S. Hemami, "Lossless image compression with projection-based and adaptive reversible integer wavelet transforms," *IEEE Trans. Image Process.*, vol. 12, no. 5, pp. 489–499, May 2003.

[8] X. Wu and T. Qiu, "Wavelet coding of volumetric medical images for high throughput and operability," *IEEE Trans. Med. Imag.*, vol. 24, no. 6, pp. 719–727, Jun. 2005.

[9] I. Avcibas, N. Memon, B. Sankur, and K. Sayood, "A successively refinable lossless image coding algorithm," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 445–452, Mar. 2005.

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

[11] 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.

[12] T.-L. Chen and S. Geman, "On the minimum entropy of a mixture of unimodal and symmetric distributions," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3166–3174, Jul. 2008.

[13] X. Wu and K. Barthel, "Piecewise 2D autoregression for predictive image coding," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 1998, pp. 901–905.

[14] 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, Jun. 2001.

[15] L.-J. Kau and Y.-P. Lin, "Adaptive lossless image coding using least squares optimization with edge-look-ahead," *IEEE Trans. Circuits Syst.*, vol. 52, no. 11, pp. 751–755, Nov. 2005.

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

[17] H. Ye, G. Deng, and J. C. Devlin, "A weighted least squares method for adaptive prediction in lossless image coding," in *Proc. IEEE Picture Coding Symp.*, Sep. 2003, pp. 489–493.

[18] 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.

[19] J. G. Cleary and I. H. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Trans. Commun.*, vol. 32, no. 4, pp. 396–402, Apr. 1984.

[20] Y. Zhang and D. A. Adjeroh, "Prediction by partial approximate matching for lossless image compression," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 924–935, Jun. 2008.

[21] 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.

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

[23] B. Aiazzi, L. Alparone, and S. Baronti, "Fuzzy logic-based matching Pursuits for lossless predictive coding of still images," *IEEE Trans. Fuzzy Syst.*, vol. 10, no. 4, pp. 473–483, Aug. 2002.

[24] G. Motta, A. Storer, and B. Carpentieri, "Lossless image coding via adaptive linear prediction and classification," *Proc. IEEE*, vol. 88, no. 11, pp. 1790–1796, Nov. 2000.

[25] I. Matsuda, H. Mori, and S. Itoh, "Lossless coding of still images using minimum-rate predictors," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2000, pp. 132–135.

[26] I. Matsuda, N. Ozaki, Y. Umezu, and S. Itoh, "Lossless coding using variable block-size adaptive prediction optimized for each image," in *Proc. Eur. Signal Process. Conf.*, Sep. 2005, pp. 1–4.

[27] B. Taskar, "Learning structured prediction models: A large margin approach," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Dec. 2004.

[28] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 25–32.

[29] J. Weston and C. Watkins, "Multi-class support vector machines," Dept. Comput. Sci., Royal Holloway, Univ. London, Egham, Surrey, U.K., Tech. Rep. CSD-TR-98-04, May 1998.

[30] S. Zhu, "Statistical modeling and conceptualization of visual patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 691–712, Jun. 2003.

[31] T. Zhang, "Covering number bounds of certain regularized linear function classes," *J. Mach. Learn. Res.*, vol. 2, pp. 527–550, Mar. 2002.

[32] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[33] J. Platt, "Using analytic QP and sparseness to speed training of support vector machine," in *Proc. Adv. NIPS*, 1999, pp. 557–563.

[34] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[35] M. I. Jordan, "Graphical models," *Statist. Sci.*, vol. 19, no. 1, pp. 140–155, 2004.

[36] D. Subbotin. (1979). *Carrless Rangecoder* [Online]. Available: http://cpansearch.perl.org/src/SALVA/Compress-PPMd-0.10/Coder.hpp

[37] D. Shkarin. (2009). *A Special Lossless Compressor BMF, Version 2.0* [Online]. Available: http://compression.ru/ds/

[38] B. Meyer and P. E. Tischer, "Glicbawls—Grey level image compression by adaptive weighted least squares," in *Proc. Data Compress. Conf.*, Mar. 2001, pp. 1–503.

[39] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, pp. 116–123.

[40] I. Matsuda. (2005, Jan.). *Lossless Image Coding Using Minimum-Rate Predictors* [Online]. Available: http://itohws03.ee.noda.sut.ac.jp/matsuda/mrp/

[41] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proc. 22nd ICML*, Aug. 2005, pp. 896–903.

Wenrui Dai received the M.S. and B.S. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2005 and 2008, respectively, where he is currently pursuing the Ph.D. degree with the Department of Electronic Engineering.

His current research interests include learning-based video coding and signal processing.

Hongkai Xiong (M'01–SM'10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU) in 2003. He was with the Department of Electrical Engineering, SJTU, Shanghai, where he is currently a professor. From 2007 to 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, as a Research Scholar.

His research interests include source coding/network information theory, signal processing, computer vision and graphics, and statistical machine learning. In 2009, he was awarded a recipient of the New Century Excellent Talents in University. In 2008, he received the Young Scholar Award of SJTU. He has published over 70 international journal/conference papers. In SJTU, he directs Image, Video, and Multimedia Communications Laboratory and Multimedia Communication area in the Key Lab of Ministry of Education of China – Intelligent Computing and Intelligent System which is also co-granted by Microsoft Research. He has served for various the IEEE Conferences as a Technical Program Committee Member. He acts as a member of Technical Committee on Signal Processing of Shanghai Institute of Electronics.

Jia Wang received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, China, in 2002. He is currently an Associate Professor with the Institute of Image Communication and Network Engineering, Department of Electronic Engineering, Shanghai Jiao Tong University, and a member of the Shanghai Key Laboratory of Digital Media Processing and Transmission. His research interests include multiuser information theory and its application in image and video coding.

Yuan F. Zheng (F'97) received the M.S. and Ph.D. degrees in electrical engineering from Ohio State University, Columbus, OH, USA, in 1980 and 1984, respectively. He received the Undergraduate degree from Tsinghua University, Beijing, China, in 1970. From 1984 to 1989, he was with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA. Since 1989, he has been with Ohio State University, where he is currently Professor and was the Chairman of the Department of Electrical and Computer Engineering from 1993 to 2004. From 2004 to 2005, he was with the Shanghai Jiao Tong University, Shanghai, China, and the Dean of the School of Electronic, Information and Electrical Engineering in 2008.

His research interests include wavelet transform for image and video, and object classification and tracking and robotics which includes robotics for life science applications, multiple robots coordination, legged walking robots, and service robots. He was on the editorial board of five international journals. He received the Presidential Young Investigator Award from Ronald Reagan in 1986, and the Research Awards from the College of Engineering, Ohio State University, in 1993, 1997, and 2007. He received the Best Conference and Best Student Paper Award a few times in 2000, 2002, and 2006, and received the Fred Diamond for Best Technical Paper Award from the Air Force Research Laboratory, Rome, NY, USA, in 2006. In 2004, he was appointed to the International Robotics Assessment Panel by the NSF, NASA, and NIH to assess the robotics technologies worldwide in 2004 and 2005.