

Published in final edited form as:
*IEEE Trans Circuits Syst Video Technol*. 2013 November ; 23(11): 1941–1956. doi:10.1109/TCSVT.2013.2269776.

# Structured Set Intra Prediction With Discriminative Learning in a Max-Margin Markov Network for High Efficiency Video Coding

**Wenrui Dai**,
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Hongkai Xiong [Senior Member, IEEE]**,
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Xiaoqian Jiang**, and
Division of Biomedical Informatics, School of Medicine, University of California, San Diego, CA 92093 USA

**Chang Wen Chen [Fellow, IEEE]**
Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY 14260 USA

Wenrui Dai: daiwen-rui@sjtu.edu.cn; Hongkai Xiong: xionghongkai@sjtu.edu.cn; Xiaoqian Jiang: x1jiang@ucsd.edu; Chang Wen Chen: chencw@buffalo.edu

## Abstract

This paper proposes a novel model on intra coding for High Efficiency Video Coding (HEVC), which simultaneously predicts blocks of pixels with optimal rate distortion. It utilizes the spatial statistical correlation for the optimal prediction based on 2-D contexts, in addition to formulating the data-driven structural interdependences to make the prediction error coherent with the probability distribution, which is desirable for successful transform and coding. The structured set prediction model incorporates a max-margin Markov network (M3N) to regulate and optimize multiple block predictions. The model parameters are learned by discriminating the actual pixel value from other possible estimates to maximize the margin (i.e., decision boundary bandwidth). Compared to existing methods that focus on minimizing prediction error, the M3N-based model adaptively maintains the coherence for a set of predictions. Specifically, the proposed model concurrently optimizes a set of predictions by associating the loss for individual blocks to the joint distribution of succeeding discrete cosine transform coefficients. When the sample size grows, the prediction error is asymptotically upper bounded by the training error under the decomposable loss function. As an internal step, we optimize the underlying Markov network structure to find states that achieve the maximal energy using expectation propagation. For validation, we integrate the proposed model into HEVC for optimal mode selection on rate-distortion optimization. The proposed prediction model obtains up to 2.85% bit rate reduction and achieves better visual quality in comparison to the HEVC intra coding.

**Index Terms**

Discriminative learning; expectation propagation (EP); intra coding; max-margin Markov network; structured set prediction

## I. Introduction

The state-of-the-art video coding schemes developed jointly by ITU-T and ISO/IEC, e.g., H.264/AVC [1] and the new High Efficiency Video Coding (HEVC) standard [2], have achieved a vital efficiency by exploring statistical redundancy among pixels through intra and inter prediction. Over the past decade, more prediction methods have been suggested to achieve better performance [3], [4]. Within the development of HEVC, the macroblock up to 64 × 64 was modeled as the hierarchical partition tree. Besides bidirectional intraprediction and separable directional transforms [5], the angular prediction with up to 34 prediction modes [6] and mode-dependent directional transforms [7] have been advanced to exploit the remaining significant directional residual energy beyond existing intraprediction modes. Recently, the combined intra prediction (CIP) [8] was proposed to exploit the spatial redundancies with both open-loop prediction and the closed-loop prediction. Taking into consideration the prediction residual that is transformed within the region of correlated pixels, this paper is devoted to investigating the set prediction for coherence with the underlying probability distribution for transform rather than minimizing individual prediction error. It is worth mentioning that the proposed set prediction for a correlated region of pixels can approximate the optimal performance with an upper bound under the joint constraints for mutual structural interdependences in max-margin Markov networks. Max-margin Markov networks [26] leverage Markov networks to combine support vector machines structurally in order to make max-margin estimation for a set of pixels. Based on the obtained contexts, multiclass support vector machines [40] are trained to distinguish the actual value from other possible estimations. Consequently, conditional prediction is made to find the most probable estimation based on the training results. Because Markov networks can represent structural interdependences among a set of pixels, the max-margin Markov networks can enforce local coherence.

Revisiting the traditional video coding trajectory, directionality in discrete cosine transform (DCT) has been considered to catch textures, lines, and edges. Zeng and Fu [9] proposed a block-independent directional DCT from the shape-adaptive DCT. Furthermore, Xu *et al.* [10] suggested primary directional operations for lifting-based DCT to exploit the interblock correlations. Later, Chang *et al.* developed the direction-adaptive partitioned block transform by incorporating directional DCT into H.264 intra coding [11]. Recently, the discrete filtering transform was proposed to exploit correlations among pixels in H.264/AVC intra coding [12]. Unfortunately, existing improvements over the transform depend on the prediction residual from the traditional prediction modes and they do not use all the available information. Other efforts have also been made to improve the intra coding under the traditional prediction by reducing the bit rate of coding unit (CU) syntax elements or enhancing the efficiency of intraprediction algorithm. A typical mode decision strategy for intra prediction is to estimate the most probable prediction mode by extracting the

directional features [13]. Laroche *et al.* [14] improved the intra coding by reducing the bit rate of the intrapredictor indexes along with the distance-based metrics in the DCT domain. Commonly, neighboring spatial information is utilized in a block-based progressive manner [15]–[17]. However, such methods only consider spatial statistical correlation for the context-based prediction, where each prediction is isolated without considering the coherence in a local region.

To further improve prediction performance, texture synthesis and hallucination (for upsampling-based reconstruction) with good perceptual quality have been proposed. A related approach [18] using the texture analysis-synthesis scheme was developed, which reduces the entropy of source information by clustering the homogeneous area into a small patch that contains the epitome content of associated regions. Those patches, which are close to uniform, can be handled under the framework of Markov random fields (MRFs), and its optimization algorithms, e.g., belief propagation (BP), have been developed to solve it. Various attempts to restore the missing information have involved in various side information, e.g., edge [19] and auxiliary parameters [4]. To maintain a temporal consistency of video, a space–time completion has been proposed in a global optimization framework [20], [21]. It has been recognized that those methods fail to ensure pixelwise fidelity.

Recently, learning-based methods for intraframe prediction have drawn more attention. Assuming the weights for pixels with same coordinates in the block for predicting are fixed [44], least-square-based methods for calculating prediction weights were proposed in [45]. They aim to achieve the optimal prediction under the Gaussian assumption. However, they do not consider the local coherences in the blocks for predicting. Xiong *et al.* [3] proposed the structured priority BP-based inpainting prediction model to exploit the intrinsic nonlocal and geometric regularity in H.264/AVC. The geometric regularity in block-based prediction was also considered in [22], which explores interdependences among blocks with the BP approach to estimate probability mass function for existing nine intraprediction modes of H. 264/AVC and obtain a reduced set of intraprediction modes for low complexity. It has been shown to commit bit rate increment and peak signal-to-noise ratio (PSNR) loss in comparison to existing H.264/AVC intra coding. BP is a message passing algorithm to make inference on graphical models [41]. It calculates the marginal distribution for each unobserved node conditioned on observed nodes in the tree-like graphical model. For graphs containing cycles or loops, BP was extended to loopy BP [42], which finds the maximum a posteriori (MAP) inference by iteratively solving a finite set of equations till convergence. However, the precise condition for the convergence in BP is not yet available. Expectation propagation (EP) unifies and extends the Kalman filter and loopy BP to make MAP inference with a simpler distribution [43]. The divergence measure function of the messages in EP is close in terms of Kullback–Leibler (KL) divergence. To fit a wider scope of messages, EP measures the difference between messages with the expectation, e.g., means and variances, rather than exact values. However, such learning-based methods are recognized as generative learning models, which might not produce the best discrimination for the actual distribution when only partial knowledge of observations is available in prediction of video data [23].

We propose a novel model for intra coding in HEVC, which can simultaneously predict a set of pixels with optimal rate-distortion performance. The structured set prediction model can take into account both the context-based prediction (i.e., with respect to the probabilistic distribution of transform) and the data-driven structural interdependences (i.e., in a local region). Unlike traditional prediction models, a discriminative learning approach is adopted to exploit the inherent statistical correlation, which is directly conditioned on the 2-D contexts of correlated regions. The obtained prediction residual is in conformity with the underlying probabilistic distribution for succeeding transform, i.e., the transformed coefficients from the proposed model tend to concentrate on low-frequency domain in the correlated region. Specifically, the max-margin Markov network models structural dependences, where pixels are correlated with their neighbors, to regulate predictions with joint constraints. It optimizes the predictions in a correlated region so that the prediction error fits the probability distribution for transform-based coding. In the max-margin estimation, model parameters are learned to jointly consider local features that characterize varying statistics to discriminate the actual pixel values from the other possible estimates to satisfy the maximal margin criteria. The loss function is designed to fit the probability distribution of DCT coefficients to reduce the coding rate. Therefore, coefficients derived from the loss-augmented inference are optimal for the coding engine. Furthermore, the prediction error from both the trained model parameter and the decomposable loss function is asymptotically upper bounded by the training error with sufficient samples. Finally, the structured set prediction (i.e., find the states achieving the maximal probability without explicit posterior distribution) can be solved by the EP algorithm.

To validate the efficacy of the proposed model, we integrate it with the unified directional intra prediction in the HM reference software of HEVC for the optimal mode selection on rate-distortion optimization (RDO). In the training process, the loss function is iteratively optimized and the model parameters (such as weighting vector) are learned from randomly selected training data. During prediction, the model parameters are utilized to combine the class of feature functions, and the discriminative learning model can make the joint max-margin prediction directly conditioned on the predicted data. To suppress the approximation errors due to overfitting, an online update is used in intraframe prediction. It is worth mentioning that the proposed predictor is causal so that both encoding and decoding can simultaneously operate on the learned parameters in one pass.

The rest of this paper is organized as follows. In Section II, we describe the overall intracoding framework and the structured set prediction model. In Section III, we design the Laplacian loss function and develop the upper bound of the prediction error for the proposed model. Section IV provides the solution to the structured set prediction model with the EP algorithm. Experimental results are evaluated in Section V on both objective and visual performance. Finally, we conclude this paper and discuss the future work in Section VI.

## II. Proposed Framework

### A. Proposed Codec

The generic video coding framework with the proposed structured set prediction model is depicted in Fig. 1. In addition to the existing intra and inter modes, each CU is designed to

choose the optimal mode to minimize rate-distortion cost. As marked in Fig. 1, the proposed structured prediction model is blended with the traditional angular intra prediction to serve as an alternative mode. Therefore, the value MODE_STRUCT is added to syntax element PRED_MODE

```
PRED_MODE ∈ {MODE_SKIP, MODE_INTER,
MODE_INTRA, MODE_STRUCT}.
```

The STRUCT_MODE is initiated in intra(I) frame. The CU is iteratively predicted from the maximal possible size to the minimal one for the decision of prediction unit (PU) in intra prediction. After calculating the Lagrangian rate-distortion cost for all the possible intraprediction modes, the PUs achieving the least overall cost is selected for the intra prediction of current CU. For each PU with its PRED_MODE as either MODE_INTRA or MODE_STRUCT, the syntax elements INTRA_PRED_MODE and PU_SIZE for intra coding are

```
INTRA_PRED_MODE ∈ {0, …, 34}
PU_SIZE ∈ {PU_4 × 4, PU_8 × 8, PU_16 × 16,
PU_32 × 32, PU_64 × 64}.
```

If $P_n$ is the current PU and $P_n^{(r)}$ is the reconstructed PUs in the current CU, the Lagrangian cost $J_{P_n}$ of the predicted PU with parameter set PARAM = {PRED_MODE, PU_SIZE, INTRA_PRED_MODE} is

$$J_{P_n}(P_n, \mathrm{PARAM} | P_n^{(r)}, Qp, \lambda) = D(P_n, \mathrm{PARAM} | P_n^{(r)}, Qp) + \lambda \cdot R(P_n, \mathrm{PARAM} | P_n^{(r)}, Qp) \quad (1)$$

where $Qp$ is the quantization parameter and $\lambda$ is the Lagrange parameter associated with $Qp$. The Lagrange parameter is empirically set: $\lambda = 0.85 \cdot 2^{(Qp-12)/3}$. Similar to intramodes and intermodes, the predictor of STRUCT_MODE is subtracted from the current PU to generate the residual, which is subsequently transformed, quantized, and encoded to obtain the compressed bitstream. There is no additional side information to be required in the compressed bitstream. Therefore, the proposed mode's bit rate $R$ only involves the header information (e.g., STRUCT_MODE flag) and the corresponding DCT residual blocks. A detailed description of the proposed intraprediction process can be found in Algorithm 3 in Section V-A, which considers iterative rate distortion cost for the structured set prediction model. The derivation process of intra prediction with the proposed structured set prediction model is formulated in Sections II-B and II-C. Through (3), the weights are trained in the structured set prediction model. In training, the PU for prediction can be obtained and the observed contexts are reconstructed video signals to ensure the synchronism between encoder and decoder. The candidate mode for INTRA_PRED_MODE is derived analogously to the default HEVC intra coding. The prediction based on the specified weights is to find the block of most probable prediction simultaneously, which is described in (2). Therefore, the EP-based message passing algorithm is proposed for the prediction.

Both the encoder and decoder start from the weights trained offline and update such weights according to the derived INTRA_PRED_MODE. Section IV describes a detailed solution process, including the derivation of prime-dual formulation, the generation of junction tree, and subsequent message passing and max-margin prediction.

### B. Structured Set Prediction Model

In this section, we describe the proposed structured set prediction model. The state-of-the-art adaptive prediction methods commonly adopt the sequential prediction on each individual pixel using adaptive rules for varying spatial dependences. Remarkably, the proposed model takes into consideration the correlation among the pixels for predicting, such that the prediction errors can be customized for various contexts. Consequently, the structured set prediction model makes context-based inference and derive constraints over sets of pixels for predicting, as shown in Fig. 2.

Fig. 2 illustrates the training process with the class of feature functions in the proposed model. In the model, concurrent training and prediction are performed for blocks of pixels with the fixed size. When obtaining the training data $\mathcal{S}=\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^{N}$, the class of feature functions $\{\mathbf{f}_j\}_{j=1}^{K}$ serves as the enforced constraints indicating the structural interdependences. In the proposed model, $\mathbf{f}_i$ describes the conditional distribution based on the *i*th context $\mathbf{x}^{(i)}$ and the pixels for predicting **y**. To be consistent in representation, we denote $\mathbf{y}^{(i)}$ = **y**. The conditional discriminative learning model is established for training of model parameters. **X** and **Y** indicate the set of samples $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$. To be concrete, the spatial statistics are characterized by the linear combination of the class of feature functions $\mathcal{F}$ = $\{\mathbf{f}_i(\mathbf{x},\mathbf{y})\}$ which establish the conditional probabilistic model for prediction over the various contexts with the structural interdependences

$$\mathbf{f}_i(\mathbf{x},\mathbf{y})=P(\mathbf{y}|\mathbf{x}).$$

Fig. 3(a) presents a graphical model of the proposed model. Denoting **y** the set of pixels to be predicted, given the reconstructed pixels **x** as contexts, their prediction $\hat{\mathbf{y}}$ is derived in a concurrent form

$$\hat{\mathbf{y}}=\arg\max_{\mathbf{y}}\mathbf{w}^T\mathbf{f}(\mathbf{x},\mathbf{y}) \quad (2)$$

where **f** is the collection of feature functions indicating the probability distribution conditioned on the various spatial structural interdependences and **w** is the trained weighting vector of the linear model that combines the class of feature functions. The training process of the weighting vector **w** is modeled as an optimization problem that simultaneously considers context-based spatial correlation and the interdependences among pixels for predicting.

### C. Intra Prediction As Optimized Problem

The predictive performance is based on the training of the weighting vector **w**. Denote $\mathcal{S}=\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^{N}$ the collected set of training data, where $\mathbf{y}_i$ is the *i*th labeled block of pixels for predicting and $\mathbf{x}_i$ is the *i*th observed contexts for $\mathbf{y}_i$. Consequently, the training of weighting vector **w** is treated as an optimization problem over the training set $\mathcal{S}$. The min–max formulation of the max-margin Markov network is formulated for the training process with constraints representing the structural interdependences among the pixels being predicted.

Since the pixels being predicted are naturally correlated in local regions, the min–max formulation is constructed according to the graphical model in Fig. 3(b). Assuming that block of *M* pixels for predicting $\mathbf{y}=\left\{\mathbf{y}^{(j)}\right\}_{j=1}^{M}$ is correlated, an edge clique of the 2-D Markov network represents two neighboring pixels. As a result, the max-margin estimation for each pixel can be obtained by jointly optimizing all pixels. The margin of the actual value $\hat{\mathbf{y}}$ over the other possible estimation **y** is $\frac{1}{\|\mathbf{w}\|}\mathbf{w}^T\left[\mathbf{f}_i(\hat{\mathbf{y}})-\mathbf{f}_i(\mathbf{y})\right]$. In training, the lower bound of such margin is maximized so that the actual value can be discriminated from the other possible estimations. Constraining $\mathbf{w}^T\mathbf{f}_i(\mathbf{x},\hat{\mathbf{y}})-\mathbf{w}^T\mathbf{f}_i(\mathbf{x},\mathbf{y})\geq\mathcal{L}(\mathbf{y},\hat{\mathbf{y}})$, the training process is formulated as

$$\begin{cases}\min\frac{1}{2}\|\mathbf{w}\|^2+C\sum\limits_i\xi_i\\ \text{s.t.}\mathbf{w}^T\mathbf{f}_i(\mathbf{y}_i)+\xi_i\geq\max\limits_{\mathbf{y}}\left(\mathbf{w}^T\mathbf{f}_i(\mathbf{y})+\mathscr{L}(\mathbf{y}_i,\mathbf{y})\right)\forall i.\end{cases}\quad(3)$$

In (3), the weighting vector **w** is the normal vector perpendicular to the hyperplane spanned by the class of feature functions $\{\mathbf{f}_i\}$ and $\{\xi_i\}$ is the slack vector that allows violations to the constraints at a cost proportional to $\{\xi_i\}$. *C* is a constant related to the learning rate in the training-based model. A large *C* will lead to the fine adjustment of parameters **w** but with a slow convergence rate. In training, the collection of training data $\mathcal{S}=\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^{N}$ is available, such that it is feasible to deal with the training data from 1 to *N*. Since a practical coding is based on the probabilistic estimation of errors with the alleged distribution, the loss function $\mathcal{L}(\mathbf{y}_i,\mathbf{y})$ is defined to reflect the actual code length under such distribution. In consequence, the optimization problem is conducted under the class of feature functions $\mathcal{F}=\{\mathbf{f}_i(\mathbf{x},\mathbf{y})\}$ and the loss function $\mathcal{L}(\mathbf{y}_i,\mathbf{y})$ that measures the actual code length.

## III. Formulation of Structured Set Prediction Model

### A. Loss Function

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

In structured set prediction model, the loss function is proposed to consider the Laplacian errors derived for each node potential and the state transition of neighboring nodes for each

edge potential. Denoting ne (*i*) the set of nodes linking the node *i* with an edge in the graphical model in Fig. 3(b), the loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ is formulated on cliques of the generated graphical model

$$\mathscr{L}(\hat{\mathbf{y}}, \mathbf{y}) = \sum_i \ell_i(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}) + \sum_i \sum_{j \in \mathrm{ne}(i)} \mathbb{I}(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(i)}) \mathbb{I}(\hat{\mathbf{y}}^{(j)}, \mathbf{y}^{(j)}) \quad (4)$$

where $\ell_i(\cdot)$ is the Laplacian loss function for the *i*th node component based on the disparity between the *i*th label $\mathbf{y}^{(i)}$ and its estimation $\hat{\mathbf{y}}^{(i)}$. In (4), we denote $\varepsilon_i = \hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}$ the *i*th prediction error in set prediction and $\sigma^2$ the variance derived by the all *M* errors $\{\varepsilon_i\}_{i=1}^{M}$ in the predictive region.

Contrary to the 0/1 loss function, squared error loss function, and the deduced Hamming distance function, the proposed loss function is designed to indicate the disparities of the actual pixel values from the predicted ones and satisfy the 2-D DCT transform for a concentrated dc energy. The Laplacian loss function is adopted to meet the practical DCT transform-based coding [24], [39] with least empirical entropy

$$\ell_i(\varepsilon_i) = \begin{cases} \log_2\left(1 - e^{-\frac{1}{\sqrt{2}\sigma}}\right) & \varepsilon_i = 0 \\ \log_2\left(\frac{1}{2}\left(e^{-\frac{|\varepsilon_i|-0.5}{\sigma/\sqrt{2}}} - e^{-\frac{|\varepsilon_i|+0.5}{\sigma/\sqrt{2}}}\right)\right) & 0 < |\varepsilon_i| < 255 \\ \log_2\left(\frac{1}{2}e^{-\frac{|\varepsilon_i|-0.5}{\sigma/\sqrt{2}}}\right) & |\varepsilon_i| = 255 \end{cases} \quad (5)$$

where *e* is the base of the natural logarithm. The solution to the loss-augmented optimization problem will minimize the practical code length as measured by the loss function $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$.

## B. Upper Bound for the Prediction Errors

In this section, we show that the upper bound for prediction error is asymptotically consistent with the training error. Such upper bound allows us to relate the error on training data to the prediction error. Hopefully, the prediction error is assured to converge by such consistency between training and prediction as long as the weighting vector **w** is well tuned to fit the training data.

As mentioned previously, the proposed model aims to minimize the cumulative code length of a correlated region in terms of Laplacian measurement. Analogical to [25], we define the average error $\mathbf{L}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ for the blocks of *M* pixels

$$\mathbf{L}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y}) = \frac{1}{M} \mathscr{L}\left(\mathbf{y}, \arg\max_{\mathbf{y}'} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}')\right).$$

To relate the prediction error to the margin of the predictors, we consider the tight upper bound $\bar{\mathbf{L}}(\mathbf{w} \cdot \mathbf{f}, \mathbf{y})$ for the average error

$$\overline{\mathbf{L}}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})=\max_{\mathbf{y}':\mathbf{w}^T\mathbf{f}(\mathbf{y})\le\mathbf{w}^T\mathbf{f}(\mathbf{y}')}\frac{1}{M}\mathscr{L}(\mathbf{y},\mathbf{y}').$$

This upper bound is tight since $\mathbf{L}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})=\overline{\mathbf{L}}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})$ holds only when $\mathbf{y}=\arg\max_{\mathbf{y}'}[\mathbf{w}\cdot\mathbf{f}+\mathcal{L}(\mathbf{y},\mathbf{y}')]$. The upper bound 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 maximize the log-Gaussian measure from $\mathbf{y}$. Extending the upper bound with the $\gamma$-margin hypersphere, we define a $\gamma$-margin per-label loss

$$\mathbf{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})=\sup_{\mathbf{y}':\|\mathbf{w}\cdot\mathbf{f}(\mathbf{y})-\mathbf{w}\cdot\mathbf{f}(\mathbf{y}')\|\le\gamma\mathscr{L}(\mathbf{y},\mathbf{y}')}\frac{1}{M}\mathscr{L}\left(\mathbf{y},\mathbf{y}'\right). \quad (6)$$

The $\gamma$-margin per-label loss $\mathbf{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})$ similarly picks $\mathbf{y}'$ in a $\gamma\mathcal{L}(\mathbf{y},\mathbf{y}')$ wider hypersphere, which is closed to the loss in the previous max-margin formulation.

We can show that the prediction error asymptotically equals the training error, which is upper bounded by its empirical $\gamma$-margin per-label loss (with the exception of an inversely growing additional term in (13) of Appendix A). In the following proposition, we prove that the prediction and training are 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 prediction error asymptotically equals to the one obtained over the training data with probability at least $1-e^{-\eta}$.

**Proof:** Refer to Appendix A.

The prediction error is upper bounded by two additional terms. The first term bounds the training error based on $\mathbf{w}$. The low training error $\mathbb{E}_S\mathbf{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})$ is achieved with the well-tuned weighting vector $\mathbf{w}$ such that the performance of the prediction model can be assured with the low error $\mathbb{E}_S\mathbf{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})$ and high margin $\gamma$. The second term is the excess loss corresponding to the complexity of the predictor. The excess loss is shown to vanish with the growth of sample size $N$. Thus, the expected predictive per-label error asymptotically approaches the $\gamma$-margin per-label error.

Proposition III-B 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 Laplacian loss function meets the empirical distribution of DCT coefficients [24], the average individual loss reflects the practical coding of prediction residual. We note that the code length led by the structured set prediction asymptotically approaches the training results, as the excess loss between the prediction and training vanishes with the growth of the sample size. With sufficient sampling, the obtained residual based on the max-margin Markov network can minimize the coding cost to the well-tuned loss over the training data. As a result, it ensures the consistency between training and prediction and shows the excess bit cost vanishes for infinite length coding.

**Algorithm 1**

Implementation with SMO

1: obj_old = 0
2: **repeat**
3: obj_new = obj_old
4: **for** all *i* **do**
5: Initialize $\{v_i(\cdot)\}$, $\{a_i(\cdot)\}$ and violation=0
6: Find violation $\mathbf{y}'$ and $\mathbf{y}''$ with KKT conditions
7: **if** violation>0 **then**
8: $a = v_i(\mathbf{y}') - v_i(\mathbf{y}'')$
9: $b = C\|\mathbf{f}_i(\mathbf{y}') - \mathbf{f}_i(\mathbf{y}'')\|^2$
10: $c = -a_i(\mathbf{y}')$ $d = a_i(\mathbf{y}'')$
11: $\delta = \max(c, \min(d, a/b))$
12: $\text{obj\_new} = \text{obj\_new} - \frac{1}{2}a \cdot \delta$
13: update **w** and $a_i$ with $\delta$
14: **end if**
15: **end for**
16: **until** $\|1 - \text{obj\_new/obj\_old}\| < 0.5$

## IV. Solving Structured Set Prediction Model

Standard quadratic programming (QP) is an effective solution to the original optimization problem. However, it suffers a high computational cost, which makes it impractical for the problems with a large state space. As an alternative, the dual of (3) is obtained as

$$\begin{cases} \max \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \mathscr{L}(\mathbf{y}_i, \mathbf{y}) - \frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})(\mathbf{f}_i(\mathbf{y}_i) - \mathbf{f}_i(\mathbf{y}))\|^2 \\ \text{s.t.} \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C, \alpha_i(\mathbf{y}) \geq 0, \quad \forall i. \end{cases} \tag{7}$$

Equation (7) can be solved by sequential minimal optimization (SMO) [28], which breaks the dual problem into a series of small QP problems and takes an ascent step to update a least number of variables

$$\begin{cases} \max \left[ v_i(\mathbf{y}') - v_i(\mathbf{y}'') \right] \delta - \frac{1}{2}C\|\mathbf{f}_i(\mathbf{y}') - \mathbf{f}_i(\mathbf{y}'')\|^2 \delta^2 \\ \text{s.t.} \alpha_i(\mathbf{y}) + \delta \geq 0, \alpha_i(\mathbf{y}'') - \delta \geq 0 \end{cases} \tag{8}$$

where $v_i(\mathbf{y}) = \mathbf{w} \cdot \mathbf{f}_i(\mathbf{y}) + \mathcal{L}(\mathbf{y}_i, \mathbf{y})$ and $\mathbf{f}_i(\mathbf{y}) = \left\{ \sum_{k=1}^{K} \beta_j \mathbf{x}_{ik}^{(j)} \right\}_{j=1}^{M}$. Like Algorithm 1, the minimization process chooses the SMO pairs with respect to the Karush–Kuhn–Tucker (KKT) conditions [29]. The KKT conditions are the sufficient and necessary criteria for optimality of the dual solution. These conditions allow certain locality with respect to each example for repeatedly searching the optimal solution.

### A. Junction Tree

The key to solve the optimization problem with SMO is the selection of SMO pairs. To maintain the spatial structures in intraframe prediction, we generate the max-margin Markov network for the $a_i$ and $v_i$ in each 4 × 4 block. The SMO pairs are decided by calculating the conditional margin for each state over the generated graphical model. Since the grid-like Markov network is not a chordal graph, it should be triangulated into a tree-like structure for exact state inference. In triangulation, the junction tree is constructed over cliques by eliminating circles in the graphical model. As shown in Fig. 4, due to the order of elimination, the junction tree tends to be nonunique for a given graphical model. For the clarity and simplicity in training and inference, we choose the chain-like junction tree.

We denote $\{J_i\}$ the nodes in the junction tree and obtain its potential $\psi(J_i)$ by accumulating potentials of all its cliques

$$\psi(J_i) = \prod_{C \in J_i} \psi_C(\mathbf{x}_C, \mathbf{y}_C) \quad (9)$$

where $\psi_C(\mathbf{x}_C, \mathbf{y}_C)$ is the potential for the clique $C$. The selection of an SMO pair over the original graphical model is transferred to the clique-based junction tree. The SMO pair is chosen by finding the pairs of the series of states that maximize the margin.

### B. EP

Once the junction is determined, the next step is to apply the inference algorithm to find states achieving the maximal probability. It involves two choices: 1) the divergence measure and 2) the message-passing scheme. In this section, we develop the EP-based method to find the most probable states. EP is an extension of BP, which can solve the problems without explicit posterior distribution over a single variable (because it sends only expectations of features in message passing).

Initially, the behavior of message-passing algorithms depends directly on the behavior of divergence measures. The basic divergence measure for the distributions is the KL divergence

$$KL(p\|q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx + \int (q(x) - p(x))\, dx$$

where $q(\mathbf{x})$ is to approximate the complex probability distribution $p(\mathbf{x})$ in divergence measure. The formula is applicable to unnormalized distributions since it includes a correction factor. Power EP [30] could minimize the $\alpha$-divergence in the context of fractional BP [31] and the $\alpha$-divergence is a generalization of KL divergence

$$D_\alpha(p\|q) = \frac{\int_x \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1-\alpha)}.$$

This is actually a family of divergences, indexed by $a \in (-\infty, +\infty)$. Given the distribution $p$ and the functional family $\mathcal{F}$, EP estimates the distribution $q$ that is closest to $p$ in $\mathcal{F}$ in the predefined divergence measure.

For the generated Markov network $\mathcal{G}$, the probabilistic distribution is factorized by

$$p(\mathcal{G}) = \prod_{C \in \mathcal{G}} \psi_C(\mathbf{x}_C, \mathbf{y}_C) = \prod_{J \in \mathcal{G}} \psi(J).$$

The projection function $\mathcal{P}$ of the distributions $p$ on functional space $\mathcal{F}$ can be represented by the KL divergence

$$\mathcal{P}[p] = \arg\min_{q \in \mathcal{F}} KL(p \| q). \quad (10)$$

EP approximates the distributions of the grid-like graphical model in (9) by approximating the factors one by one. For each clique $C$, the marginal probability obtained by excluding itself is

$$\begin{aligned} q^{\backslash C}(\mathbf{x}) &= q(\mathbf{x}) \backslash \tilde{\psi}_C(\mathbf{x})^{1/n_C} \tilde{\psi}_C(\mathbf{x})^{\text{new}} \\ &= \mathcal{P}\left[\psi_C(\mathbf{x}) q^{\backslash \psi_C}(\mathbf{x})\right] / q^{\backslash \psi_C(\mathbf{x})}. \end{aligned}$$

Also, the potential of the clique $C$ is updated by

$$\psi_C(\mathbf{x}_C, \mathbf{y}_C) = \tilde{\psi}_C(\mathbf{x}_C, \mathbf{y}_C) = \frac{\prod_{(j,k) \in \mathcal{G}} \tilde{\psi}_C\left(x^{(j)}, x^{(k)}\right)}{\prod_{s \in \mathcal{S}} \tilde{\psi}_C\left(x^{(s)}\right)}.$$

**Algorithm 2**

Message passing with EP

1: Initialize $\tilde{\psi_C}(\mathbf{x})$ for all cliques and make $q(\mathbf{x})$ their product

2: **repeat**

3: **for** all $C$ **do**

4: $q^{\backslash \psi C}(\mathbf{x}) = q(\mathbf{x}) / \tilde{\psi_C}(\mathbf{x})$

5: $q'(\mathbf{x}) = \mathcal{P}[\psi_C(\mathbf{x}) q^{\backslash \psi C}(\mathbf{x})]$

6: $$\tilde{\psi}_C(\mathbf{x})^{\text{new}} = \tilde{\psi}_C(\mathbf{x})^{1-\gamma} \left(\frac{q'(\mathbf{x})}{q^{\backslash \psi_C}(\mathbf{x})_{n_C}}\right)^{\gamma} = \tilde{\psi}_C(\mathbf{x}) \left(\frac{q'(\mathbf{x})}{q(\mathbf{x})}\right)^{\gamma}$$

7: $$q(\mathbf{x})^{\text{new}} = q(\mathbf{x}) \left(\frac{\tilde{\psi}_C(\mathbf{x})^{\text{new}}}{\tilde{\psi}_C(\mathbf{x})}\right)^{n_C} = q(\mathbf{x}) \left(\frac{q'(\mathbf{x})}{q(\mathbf{x})}\right)^{\gamma n_C}$$

8: **end for**

9: **until** Convergence

Algorithm 2 shows how messages passed in and out of the clique C, where the estimation and update of the probabilistic distribution are propagated over the junction tree. Since it is prohibitive to traverse the whole state space of all junction trees, the local propagation in each junction is used in practice. The message-passing algorithm in EP is to seek the min–max formulation over the generated junction tree

$$\min_{\hat{p}_i}\max_{q}\sum_i n_i \int_x \hat{p}_i(x)\log\frac{\hat{p}_i(x)}{f_i(x)} + \left(1-\sum_i n_i\right)\int_x q(x)\log q(x)$$

such that $\int_x g_j(x)\, p_i\hat{}(x)\, dx = \int_x g_j(x)\, q(x)\, dx$ for normalized distributions $p_i\hat{}(x)$ and $q(x)$. For the unnormalized distributions $p$ and $q$, they are scaled by the normalization term $Z$

$$Z = \int_x q'(x)\, dx = \left(\int_x q(x)\, dx\right)\prod_a s_a.$$

**Proposition 2—**The EP for the MRF with Laplacian loss function is upper bounded.

**Proof:** Refer to Appendix B.

### C. Discussion on the Structured Set Prediction Model

In this section, the mechanism of the structured set prediction model is analyzed with relevant factors. In Fig. 5, each pixel in a block is predicted with linear combination of the feature functions based on observed contexts and the constraints with each other. The potential for a clique $\psi_j$ is obtained by

$$\psi_j = \sum_i \mathbf{w}_j^{(i)} P\left(\mathbf{y}^{(j)}|\mathbf{x}^{(i)}\right)$$

where the selection of weights $\mathbf{w}_j^{(i)}$ depends on the syntax element INTRA_PRED_MODE and the values of observed contexts. In the encoder side, the rate-distortion cost w.r.t. each intradirectional mode is calculated to select a corresponding set of weights. Like angular intra prediction, the value of INTRA_PRED_MODE indicates the directional mode and the selected set of weights. In this example, INTRA_PRED_MODE is 6, which means the predictive direction is diagonal. The weighting vector is isotropically initiated to a zero vector, and it is trained in an iterative manner. Corresponding to Fig. 5(c), Fig. 5(d) shows the trained weighting vector for each clique and observed contexts after 100 iterations. As noticed, the weighting vector shows the predictive tendency in an diagonal form. However, it is not the same with the angular intra prediction, since the interdependences of pixels for

predicting are considered and the weights are adjusted accordingly to minimize the loss function (4) in the training. Fig. 5(b) indicates the selection of SMO pairs with red and blue arrow lines. The feature function $P(\mathbf{y}^{(j)}|\mathbf{x}^{(i)})$ implies the spatial distribution over the observed contexts

$$P(\mathbf{y}^{(j)}|\mathbf{x}^{(i)})=\mathbb{I}(\mathbf{y}^{(j)},\mathbf{x}^{(i)}). \quad (11)$$

By traversing all possible values of pixel $\{\mathbf{y}^{(j)}\}$, the one achieving the minimum loss for the linear combination is chosen as the most probable estimate.

Further, we study the relationship between training and prediction of the proposed model. As proven in Proposition III-B, the prediction will be asymptotically consistent with the training. Fig. 6 shows the visual and PSNR performance of the predictive results by the weighting vector **w** trained under 1, 5, 10, and 50 iterations. The edge structure in the selected block becomes clear with more iterations, which provides additional evidence that the weighting vector tends to represent the anisotropic local statistics. Fig. 7 depicts the BD-rate reduction in % from predictive results by the weighting vector **w** trained with 1, 5, 10, 25, 50, 75, and 100 iterations. The iterative process forms the anisotropic distribution for the weighting vector **w** with the step $\Delta\mathbf{w} = 0.0625$. It shows that BD-rate reduction increases with the growth of iteration number and tends to approach an upper bound, which is also consistent with Proposition III-B.

## V. Experimental Results

### A. Implementation

Through integrating the proposed model (STRUCT mode) into the intra coding with hierarchical tree-structured intraprediction block sizes ranging from 64×64 to 4×4, the structured set intraprediction scheme is implemented on the HEVC test model HM 7.0 [32]. The STRUCT mode is enabled in I slices, and is compared with the hierarchical tree-structured modes for mode selection in a RDO. The proposed intraprediction model is initiated for luma samples, which can be referred to Algorithm 3. Algorithm 3 describes the detailed procedure of the proposed intra prediction, including integration of the proposed model, RDO, and derivation of intraprediction mode. The maximum CU size is 64 and the maximum partition depth is 4, such that the size of PU can range from 64 × 64 to 4 × 4. In the experiments, we evaluate the performance over test sequences with the YUV 4:2:0 format, various resolution including CIF (352 × 288), WQVGA (416 × 240), standard definition (832 × 480), 576p (720 × 576), and high-definition (1920 × 1080). Without loss of generality, those are coded with the same quantization parameters and conditions [36]. In the proposed model, the block size is set to 4 × 4, namely, $M = 16$ pixels are predicted simultaneously. The weighting vector is designed for 17 sets associated with the number of INTRA_PRED_MODE. In the training process, the learning rate $C$ is set to 0.05 to fine tune the weighting vector.

**Algorithm 3**

Procedure of the proposed intra prediction in HEVC

Input: luma location (xB,yB) of current coding block and the size of current luma coding block
Output: Residual block after subtracting prediction
**if** IntraSplitFlag = 0 **then**
  Derive optimal intraprediction mode for angular intra prediction and obtain BestPUCost.
  **for** All possible IntraPredMode[xB][yB] in luma location (xB,yB) **do**
    **Derivation process with the proposed prediction model (Referring to Section IV)**
    **Calculating R-D cost PUCost with** (1)
    **if** PUCost < BestPUCost **then**
      **Set PRED_MODE to MODE_STRUCT, luma intraprediction mode IntraPredMode[xB][yB], luma location (xB,yB) and luma block size PU_SIZE**
      BestPUCost = PUCost
    **end if**
  **end for**
  Obtain the residual block by subtracting the prediction of intraluma block
**Else**
  **for** blkIdx = 0 … 3 **do**
    Derive optimal angular intraprediction mode and obtain BestSubPUCost(blkIdx).
xBS = xB + ((1 ≪ log2CUSize) ≫ 1) * (blkIdx%2).
yBS = yB + ((1 ≪ log2CUSize) ≫ 1) * (blkIdx/2).
    **for** All possible IntraPredMode[xB][yB] in luma location (xB,yB) **do**
      **Derivation process with the proposed prediction model (referring to Section IV)**
      **Calculating R-D cost SubPUCost(blkIdx) in luma location (xBS,yBS) with** (1)
      **if** SubPUCost(blkIdx) < BestSubPUCost(blkIdx) **then**
        **Set PRED_MODE to MODE_STRUCT, luma intraprediction mode IntraPredMode[xBs][yBs], luma location (xBs,yBs), and luma block size PU_SIZE** ≫ 1
        SubBestPUCost(blkIdx) = SubPUCost(blkIdx)
      **end if**
    **end for**
  **end for**
  *PUCost* $\Sigma_{blkIdx=0\cdots3}$ SubBestPUCost(blkIdx)
  **if** PUCost < BestPUCost **then**
    Obtain the residual block by subtracting the prediction of intraluma block
  **end if**
**end if**

## B. Rate-Distortion Performance

Using both extended CU size and traditional CU size, the proposed scheme is compared with the CIP with HEVC and the HEVC intra prediction (angular intra prediction). As shown in Figs. 8 and 9 for test sequences of various resolutions, the rate-distortion points are obtained at various QP levels (30, 32, 34, 36, 38, 40). To be concrete, five QP levels (30–38 or 32–40) are chosen to set the bit rates within a moderate region. Remarkably, the PSNR

gain of the proposed model is up to 0.4 dB over the HEVC intraprediction (angular intraprediction, AIP). The coding gain is more obvious in video sequences with rich texture characterized by regular features.

For a complete validation, the BD-PSNR and BD-rate reduction [33] evaluations are also provided based on rate-distortion curve fitting. They can describe the average PSNR difference in dB over the entire range of bit rates and average bit rate difference in % over the whole spectrum of PSNR (between two RD plots under difference conditions). Tables I and II, respectively, show the BD PSNR and BD rate of the selected test video sequences. The four R-D points are obtained with QP levels set to 22, 27, 32, and 37 to reflect the curve in a wide range of rates or distortion. Table II shows that the proposed model obtains up to 2.85% bit rate reduction in intraluma prediction in comparison with the HEVC intracoding. Moreover, the BD-rate reduction over HEVC intracoding with full R-D optimization is provided. Since INTRA_PRED_MODE for the proposed model is still based on candidate modes for R-D optimization, the BD-rate reduction on luma samples decreases slightly by 0.06%.

As in Table III, adaptive hierarchical tree-structured mode decision degrades the performance of the proposed model. The degradation will deteriorate when the depth of the mode decision tree increases, e.g., an average 3.16% bit rate reduction in intraluma coding in comparison to HEVC with the traditional macroblock size (H.264/AVC).

### C. Visual Quality

Fig. 10 shows the reconstructed video frames with various resolutions by the proposed model, HEVC with CIP, and HEVC with AIP. Overall, the proposed scheme achieves best visual quality in regions with regular features. In Fig. 10(e), it can be seen that the contour of the person’s face and the region of nose reconstructed by the proposed model are clearer and more natural.

The corresponding prediction residuals from the proposed model are shown in Fig. 11. Compared to HEVC test model and HEVC with CIP, the proposed model obviously reduces the prediction residuals in edge regions and oscillatory regions. Furthermore, first-order entropies of the prediction residuals are provided to evaluate the distributions of prediction errors. As shown in Fig. 11, prediction residuals from the proposed model achieve the least first-order entropies, which implies the distribution of the prediction errors is most concentrated.

### D. Computational Complexity

Although the training process iterates over the collection of sample data, the complexity of the proposed prediction scheme is equivalent to the max-sum algorithm. For the proposed lattice-based graphical model of the $M_x \times M_y$ block, there are $L = (M_x - 1)\, M_y + M_x\, (M_y - 1)$ cliques in total. Given $L$ cliques with alphabet size $\|\mathbf{y}\|$ of prediction, the computational complexity of predicting each block is $O\,(L\|\mathbf{y}\|^2)$, which means that the complexity is linear with the number of cliques.

In practice, both the encoders and decoders operate on a PC with a 3.2-GHz Intel Core i7 processor and are compiled with VC++ 9.0 under the same configuration (DEBUG mode). The evaluation is tested on encoder_intra_main configuration when QP equals 24 and 36. Since the proposed model is initiated with 4 × 4 PU, $L$ is set to 24 and $\|\mathbf{y}\|$ is 256 for the 8 bits internal bit depth. In the encoder side, the total complexity of the proposed scheme is also affected by the number of INTRA_PRED_MODE for the sake of RDO. The R-D cost of the structured set prediction model with all 34 intraprediction modes should be counted. In detail, the encoding speed is approximately 118 pixels per second. Table IV shows the decoding speed of the proposed scheme, HEVC with CIP, and HEVC with AIP. Depending on the selection ratio of the proposed STRUCT mode, it ranges from 1700 to 6500 and 4000 to 14 000 pixels per second when QP equals 24 and 36, respectively. Because the prediction is noniterative, the decoding cost can be reduced by optimizing the solution process. Moreover, the run-time ratios for the proposed model over HEVC with AIP are also provided in Table IV for evaluation. Table IV shows that the run-time ratios of the proposed model are 60–178 times the ones of HEVC with AIP. When compared with the default HEVC with AIP, the proposed method is obviously more complex.

The additional complexity of the proposed model is derived from two aspects: one is that the proposed model shall make the full R-D optimization for all the candidate modes in the proposed intracoding scheme, as in Algorithm 3; the other is that all the possible estimates of block of pixels are traversed in order to find the optimal solution, as in Algorithm 1. It could be improved by two approaches under consideration. An attempt is to adopt parallelism techniques because the optimization problem is decomposable over the max-margin Markov network. It is possible to deal with all junctions in parallel and combine their results. The other solution is to introduce the stochastic gradient decent algorithm [46] to speed up the optimization process, which is an efficient and simple procedure with a decomposable and differentiable loss function. Its numeric implementation can converge very quickly.

### E. Selection Ratio

To make a study on the selection ratio of the proposed model under various QP levels and training iterations, Table V shows the average STRUCT mode selection ratio. The selection ratio increases with the lower QP levels, whereas the replacement ratio of the proposed mode over angular intra prediction decreases. The distribution map of STRUCT mode is displayed in Fig. 12, where the selected blocks are labeled by red rims. As shown in Fig. 12(b), it can be observed that the STRUCT mode tends to be initiated in the regions with regular features. As a result, the STRUCT mode can save bits of repeatable visual patterns beyond traditional intraprediction modes. Table VI shows that the selection ratio varies with the growth of iterations in training. With more consistency between training and prediction, the performance of the proposed model is expected to grow and the selection ratio of STRUCT mode will increase.

## VI. Conclusion

This paper proposed a learning-based structured set prediction model on intracoding, which simultaneously predicts a correlated region of pixels by considering the inherent statistical

correlation (i.e., 2-D context) and the coherence of set prediction (i.e., structural dependences). The prediction was optimized in alignment with two goals: 1) context-based prediction and 2) structure-based prediction in a global (set) manner. The training and prediction were formulated using the max-margin Markov network, where the optimization was achieved under the well-defined loss function conditioned on the obtained local observations. With the growing sample size, the loss-augmented inference was demonstrated to be asymptotically consistent with the fine-tuned training results. In turn, the distribution of DCT coefficients derived from the prediction residual was more concentrated. Since the proposed Laplacian loss function can be fully factorized, the proposed min–max formulation can be solved by combining optimized results of all individual cliques using EP with lower dimensional state spaces. In practice, the proposed model was integrated in the latest HEVC reference software to serve as an optional mode in RDO.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U1201255, 61271218, and 61228101. The work of X. Jiang was supported in part by the National Institutes of Health under Grants K99LM011392, UH2HL108785, U54HL108460, and UL1TR0001000, and AHRQ Grant R01HS019913.

## References

1. Wiegand T, Sullivan G, Bjontegaard G, Luthra A. Overview of the H.264/AVC video coding standard. IEEE Trans Circuits Syst Video Technol. Jul; 2003 13(7):560–576.
2. Ugur K, Andersson K, Fuldseth A, Bjontegaard G, Endresen LP, Lainema J, Hallapuro A, Ridge J, Rusanovskyy D, Zhang C, Norkin A, Priddle C, Rusert T, Samuelsson J, Sjoberg R, Wu Z. High performance, low complexity video coding and the emerging HEVC standard. IEEE Trans Circuits Syst Video Technol. Dec; 2010 20(12):1688–1697.
3. Xiong H, Xu Y, Zheng YF, Chen C. Priority belief propagation based inpainting prediction with tensor voting projected structure in video compression. IEEE Trans Circuits Syst Video Technol. Aug; 2011 21(8):1115–1129.
4. Xiong Z, Sun X, Wu F. Block-based image compression with parameter-assistant inpainting. IEEE Trans Image Process. Jun; 2010 19(6):1651–1657. [PubMed: 20215076]
5. Ye Y, Karczewicz M. Improved H.264 intracoding based on bidirectional intraprediction, directional transform, and adaptive coefficient scanning. Proc IEEE Int Conf Image Process. Oct. 2008 :2116–2119.
6. Bossen F, Drugeon V, Francois E, Jung J, Kanumuri S, Narroschke M, Sasai H, Sole J, Suzuki Y, Tan TK, Wedi T, Wittmann S, Yin P, Zheng Y. Video coding using a simplified block structure and advanced coding techniques. IEEE Trans Circuits Syst Video Technol. Dec; 2010 20(12):1667–1675.
7. Yeo C, Tan YH, Li Z, Rahardja S. Mode-dependent transforms for coding directional intraprediction residuals. IEEE Trans Circuits Syst Video Technol. Apr; 2012 22(4):545–554.
8. Gabriellini A, Flynn D, Mrak M, Davies T. Combined intraprediction for high-efficiency video coding. IEEE J Sel Topics Signal Process. Nov; 2011 5(7):1282–1289.
9. Zeng B, Fu JJ. Directional discrete cosine transforms—a new framework for image coding. IEEE Trans Circuits Syst Video Technol. Mar; 2008 17(3):305–313.
10. Xu H, Xu J, Wu F. Lifting-based directinal DCT-like transform for image coding. IEEE Trans Circuits Syst Video Technol. Oct; 2007 17(10):1325–1335.
11. Chang CL, Makar M, Tsai SS, Girod B. Direction-adaptive partitioned block transform for color image coding. IEEE Trans Image Process. Jul; 2010 19(7):1740–1755. [PubMed: 20215074]
12. Peng X, Xu J, Wu F. Directional filtering transform for image/intraframe compression. IEEE Trans Image Process. Nov; 2010 19(11):2935–2946. [PubMed: 20435540]

13. Pan F, Lin X, Rahardja S, Lim KP, Li ZG, Wu D, Wu S. Fast mode decision alogirthm for intraprediction in H.264/AVC video coding. IEEE Trans Circuits Syst Video Technol. Jul; 2005 15(7):813–822.
14. Laroche G, Jung J, Popescu BPP. Intracoding with prediction mode information inference. IEEE Trans Circuits Syst Video Technol. Dec; 2010 20(12):1786–1796.
15. Kim DY, Han KH, Lee YL. Adaptive single-multiple prediction for H.264/AVC intracoding. IEEE Trans Circuits Syst Video Technol. Apr; 2010 20(4):610–615.
16. Piao Y, Park H. Adaptive interpolation-based divide-and-predict intracoding for H.264/AVC. IEEE Trans Circuits Syst Video Technol. Dec; 2010 20(12):1915–1921.
17. Tao P, Wu W, Wang C, Xiao M, Wen J. Horizontal spatial prediction for high dimension intra coding. Proc IEEE Data Compression Conf. Mar.2010 :552.
18. Dumitras A, Haskell BG. An encoder-decoder texture replacement method with application to content-based movie coding. IEEE Trans Circuits Syst Video Technol. Jun; 2004 14(6):825–840.
19. Liu D, Sun X, Wu F, Zhang YQ. Edge-oriented uniform intra prediction. IEEE Trans Circuits Syst Video Technol. Oct; 2008 17(10):1827–1836.
20. Wexler Y, Shechtman E, Irani M. Space-time completion of video. IEEE Trans Pattern Anal Mach Intell. Mar; 2007 29(3):463–476. [PubMed: 17224616]
21. Yuan Z, Xiong H, Zheng YF. A generic video coding framework based on anisotropic diffusion and spatio-temporal completion. Proc IEEE Int Conf Acoust, Speech, Signal Process. Mar.2010 : 926–929.
22. Milani S. Fast H.264/AVC FRExt intra coding using belief propagation. IEEE Trans Image Process. Jan; 2011 20(1):121–131. [PubMed: 21172745]
23. Zhu SC. Statistical modeling and conceptualization of visual patterns. IEEE Trans Pattern Anal Mach Intell. Jun; 2003 25(6):1–22.
24. Lam EY, Goodman JW. A mathematical analysis of the DCT coefficient distributions for images. IEEE Trans Image Process. Oct; 2000 9(10):1661–1666. [PubMed: 18262905]
25. Taskar, B. PhD dissertation. Dept. Comp. Sci., Stanford Univ; Stanford, CA, USA: 2004 Dec. Learning structured prediction models: A large margin approach. [Online]. Available: http://robotics.stanford.edu/btaskar/pubs/thesis.pdf/
26. Taskar, B.; Guestrin, C.; Koller, D. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; Dec. 2003 Max-margin Markov networks; p. 25-32.
27. Zhang T. Covering number bounds of certain regularized linear function classes. J Mach Learn Res. Mar.2002 2:527–550.
28. Platt, J. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; Nov. 1999 Using analytic QP and sparseness to speed training of support vector machine; p. 557-563.
29. Boyd, S.; Vandenberghe, L. Convex Optimization. Cambridge, U.K: Cambridge Univ. Press; 2004.
30. Minka, TP. Power EP. 2004. [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/papers/ep/
31. Wiegerinck, W.; Heskes, T. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; 2003. Fractional belief propagation; p. 438-445.
32. High Efficiency Video Coding (HEVC). 2012 Jul. [Online]. Available: http://hevc.hhi.fraunhofer.de/
33. Bjontegaard, G. Calculation of Average PSNR Differences Between RD-Curves, document VCEG-M33, ITU-T SG16/Q6. 13th VCEG Meeting; Apr. 2001;
34. Bross, B.; Han, W-J.; Ohm, J-R.; Sullivan, GJ.; Wiegand, T. High efficiency video coding (HEVC) text specification draft 7. JCTVC-I1003; Geneva, Switzerland. Apr. 2012;
35. Saxena, A.; Fernandes, F. CE7: Mode-Dependent DCT/DST Without 4*4 Full Matrix Multiplication for Intra Prediction. JCTVC-E125; Geneva, Switzerland. Jul. 2010;
36. Bossen, F. Common HM Test Conditions and Software Reference Con-figurations. JCTVC-I1100; Geneva, Switzerland. Apr. 2012;
37. Karczewicz M, Ye Y, Chong I. Rate-Distortion Optimized Quantization, document VCEG-AH21, VCEG. ITU-T Q6/16. Jan.2008

38. Turkan M, Guillemot C. Online dictionaries for image prediction. Proc IEEE Int Conf Image Process. Sep.2011 :293–296.

39. Wu X, Zhai G, Yang X, Zhang W. Adaptive sequential prediction of multidimensional signals with applications to lossless image coding. IEEE Trans Image Process. Jan; 2011 20(1):36–42. [PubMed: 20679033]

40. Weston, J.; Watkins, C. Tech Rep CSD-TR-98-04. Dept. Comput. Sci., Royal Holloway, Univ. London; London, U.K: 1998. Multiclass support vector machines.

41. Pearl J. Reverend Bayes on inference engines: A distributed hierarchical approach. Proc 2nd Nat Conf Artif Intell. Aug.1982 :133–136.

42. Frey, BJ.; MacKay, DJ. Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; Nov. 1998 A revolution: Belief propagation in graphs with cycles; p. 479-485.

43. Minka TP. Expectation propagation for approximate Bayesian inference. Proc 17th Conf Uncertainty Artif Intell. 2001:362–369.

44. Chen J, Han W-J. Adaptive linear prediction for block-based lossy image coding. Proc 16th IEEE Int Conf Image Process. Sep.2009 :2833–2836.

45. Zhang Y, Zhang L, Ma S, Zhao D, Gao W. Context-adaptive pixel based prediction for intra frame encoding. Proc IEEE Int Conf Acoust Speech Signal Process. Mar.2010 :898–901.

46. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. Proc 21st Int Conf Mach Learn. 2004:116–123.

## Biographies

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

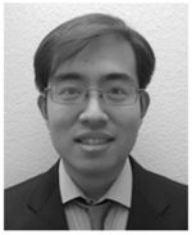
His research interests include learning-based video coding and signal processing.

**Hongkai Xiong** (M’01–SM’10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003.

Since 2003, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Professor. From December 2007 to December 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA,

as a Research Scholar. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics, University of California, San Diego. He has published more than 100 refereed journal/conference papers. At SJTU, he directs the Image, Video, and Multimedia Communications Laboratory and the multimedia communication area in the Key Laboratory of Ministry of Education of China—Intelligent Computing and Intelligent System, which is also co-granted by Microsoft Research. His research interests include source coding/network information theory, signal processing, computer vision and graphics, and statistical machine learning.

Dr. Xiong received the Top 10% Paper Award for Super-Resolution Reconstruction with Prior Manifold on Primitive Patches for Video Compression at the 2011 IEEE International Workshop on Multimedia Signal Processing. In 2011, he received the First Prize of the Shanghai Technological Innovation Award. In 2010, he received the SMC Excellent Young Faculty Award of SJTU. In 2009, he received the New Century Excellent Talents in University Award from the Ministry of Education of China. He is a technical program committee member or session chair for a number of international conferences.



**Xiaoqian Jiang** received the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, USA, in 2010.

He is a Post-Doctoral Scientist with the Division of Biomedical Informatics, School of Medicine, University of California, San Diego, CA, USA. His expertise is in data privacy and machine learning. He has researched imbalanced data analysis, predictive model calibration, and privacy-preserving data mining. His research interests include developing practical and scalable technologies for large data analysis.

Dr. Jiang received a Distinguished Paper Award from the American Medical Informatics Association Summits on transnational science in 2012 and served as the Tutorial Chair for the 2nd IEEE Conference on Health Informatics, Imaging, and System Biology.

**Chang Wen Chen** (F'04) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1983, the M.S.E.E. degree from the University of Southern California, Los Angeles, CA, USA, in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 1992.

Since 2008, he has been a Professor of computer science and engineering at the State University of New York at Buffalo, Buffalo, NY, USA. From 2003 to 2007, he was the Allen S. Henry Distinguished Professor with the Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL, USA. He was with the Faculty of Electrical and Computer Engineering, University of Missouri-Columbia, Columbia, MO, USA, from 1996 to 2003 and at the University of Rochester, New York, NY, USA, from 1992 to 1996. From 2000 to 2002, he was the Head of the Interactive Media Group, David Sarnoff Research Laboratories, Princeton, NJ, USA. He has also consulted with Kodak Research Laboratories; Microsoft Research, Beijing, China; Mitsubishi Electric Research Laboratories, Cambridge, MA, USA; NASA Goddard Space Flight Center, Greenbelt, MD, USA; and the U.S. Air Force Rome Laboratories, Rome, NY, USA.

Dr. Chen was the Editor-in-Chief for the IEEE Transactions on Circuits and Systems for Video Technology from 2006 to 2009. He has served as an Editor of Proceedings of the IEEE, IEEE Transactions on Multimedia, IEEE Journal on Selected Areas in Communications, IEEE Multimedia, *Journal of Wireless Communication and Mobile Computing*, *EURASIP Journal of Signal Processing: Image Communications*, and *Journal of Visual Communication and Image Representation*. He has also chaired and served on numerous technical program committees for the IEEE and other international conferences. He became a fellow of the International Society for Optical Engineers for his contributions in electronic imaging and visual communications.

## Appendix A. Proof of Proposition 1

At first, we use the corresponding $\gamma$-margin loss to serve as an empirical upper bound of the average loss

$$\mathscr{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})=\sup_{\mathbf{y}':\|\mathbf{w}\cdot\mathbf{f}(\mathbf{y})-\mathbf{w}\cdot\mathbf{f}(\mathbf{y}')\|\le 2\gamma}\frac{1}{M}\mathscr{L}(\mathbf{y},\mathbf{y}').$$

It equals to $L(\mathbf{w}\cdot\mathbf{f},\mathbf{y})$ when $\mathbf{y}=\arg\max_{\mathbf{y}'}[\mathbf{w}\cdot\mathbf{f}+\mathcal{L}(\mathbf{y},\mathbf{y}')]$. According to (4), the $M$-label loss function is decomposable over the cliques of labels. Given the decreasing sequence $\{\gamma_i\}$ and the positive sequence $\{p_i\}$ that satisfies $\Sigma_i p_i = 1$, [26] shows that for every constant $\eta > 0$ and at the probability $1-e^{-\eta}$, the mean for loss function on the sample space $\mathcal{X}$ is bounded by

$$\mathbb{E}_{\mathscr{X}}L(\mathbf{w}\cdot\mathbf{f},\mathbf{y})\le E_{\mathscr{S}}\mathscr{L}^{\gamma}(\mathbf{w}\cdot\mathbf{f},\mathbf{y})+\sqrt{\frac{32}{N}\left(\ln 4\mathscr{N}_{\infty}(\mathscr{L},\gamma_i,\mathscr{S})+\ln\frac{1}{p_i\eta}\right)} \quad (12)$$

where $\mathcal{S}$ is the sets of $N$ pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$ sampled from $\mathcal{X}$. When $\|\mathbf{x}\| \le b$, $\|\mathbf{w}\| \le a$ and $1/p + 1/q = 1$, [27] shows the numeric upper bound for $\mathcal{N}_\infty$ ( $\mathcal{L}$, $\varepsilon$, $n$) is

$$\log_2 \mathcal{N}_\infty(\mathcal{L}, \varepsilon, N) \le 36\,(p-1)\frac{a^2b^2}{\varepsilon^2}\log_2(2\lceil 4ab/\varepsilon \rceil N+1).$$

As a result, we can draw the conclusion that the excess term in (12) decays to 0 with the growth of $N$

$$\sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(\mathcal{L}, \gamma_i, \mathcal{S}) + \ln\frac{1}{p_i\eta}\right)} \sim o\left(\frac{\log N}{N}\right) \to 0. \quad (13)$$

Finally, the prediction errors for the optimal combination of the basis function are asymptotically equivalent to the results gained from the training data.

## Appendix B Proof of Proposition 2

As we have described, the proposed loss function is to regularize the distribution of the predictive errors to a Laplace distribution $p$. The integral of Laplacian distribution for predictive error $\varepsilon_i$ is

$$\ell_i(\varepsilon_i) = \begin{cases} \left(1 - e^{-\frac{1}{\sqrt{2}\sigma}}\right) & \varepsilon_i = 0 \\ \left(\frac{1}{2}\left(e^{-\frac{|\varepsilon_i|-0.5}{\sigma/\sqrt{2}}} - e^{-\frac{|\varepsilon_i|+0.5}{\sigma/\sqrt{2}}}\right)\right) & 0 < |\varepsilon_i| < 255 \\ \left(\frac{1}{2}e^{-\frac{|\varepsilon_i|-0.5}{\sigma/\sqrt{2}}}\right) & |\varepsilon_i| = 255. \end{cases}$$

Here, the actual distribution for the residual is approximated by the exponential family $\mathcal{F}$ with the Laplace exponent. Hence, we take a glance at $D_\alpha(p\|q)$ where the stationary point $q_0$ of the divergence is equivalent to the stationary point of the projection $\mathcal{P}$ $[p(\mathbf{x})^{1-\alpha} q(\mathbf{x})^\alpha]$ when considering the derivative of the $\alpha$-divergence with respect to its parameter $\theta$

$$\frac{dD_\alpha(p\|q)}{d\theta} = \frac{1}{\alpha}\left(\int_{\mathbf{x}} \frac{dq(\mathbf{x})}{\theta} d\mathbf{x} - \int_{\mathbf{x}} \frac{q'(\mathbf{x})}{q(\mathbf{x})}\frac{dq(\mathbf{x})}{d\theta} d\mathbf{x}\right)$$

where $q'(\mathbf{x}) = p(\mathbf{x})^\alpha q(\mathbf{x})^{1-\alpha}$. In consequence, the stationary point $\theta_0$ that achieves the optimality satisfies

$$\int_{\mathbf{x}} \left(1 - \left[\frac{p(\mathbf{x})}{q(\mathbf{x})}\right]^\alpha\right) \frac{dq(\mathbf{x})}{d\theta} d\mathbf{x} = 0. \quad (14)$$

From (14), it could be drawn that the ratio $p(\mathbf{x})/q(\mathbf{x})$ is bounded.

On the other hand, it holds $q = \mathcal{P}(p)$. Since the undirected graphical model generated from the MRF is fully factorized, the Laplacian loss function is decomposable. Thus, the projection onto the fully factorized distribution equals the matching of the two margins

$$q=\mathscr{P}(p) \iff \int_{\mathbf{x}\setminus x_i} q(\mathbf{x})d\mathbf{x}=\int_{\mathbf{x}\setminus x_i} p(\mathbf{x})d\mathbf{x} \quad \forall i \quad (15)$$

Since (15) holds for arbitrary $i$, it implies that the integrals of the two distribution are almost equivalent. In conclusion, the approximation of distribution $p$ with $q$ is upper bounded.

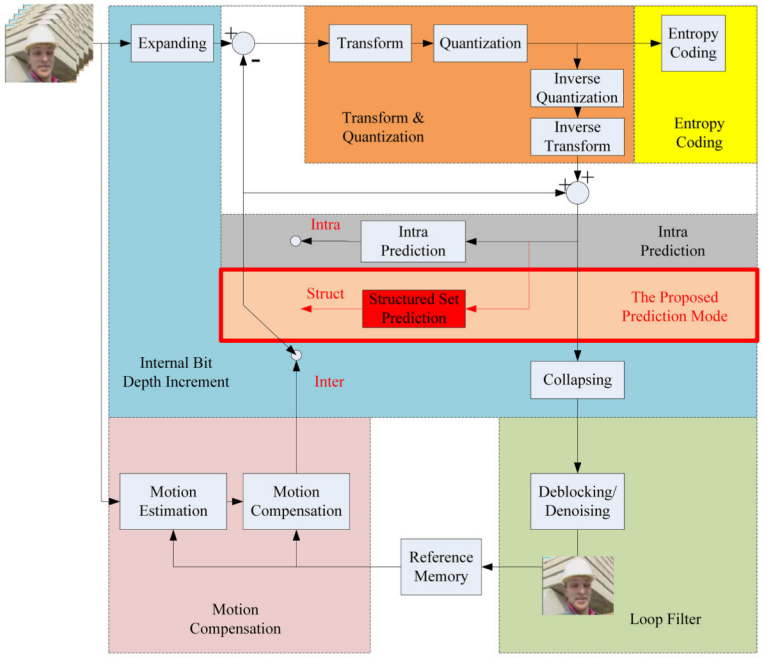



**Fig. 1.**

Proposed codec with structured set prediction model based on HM. The structured set prediction model is blended with the original angular intra prediction to serve as an alternative prediction mode. The proposed model is selected according to the rate-distortion cost.

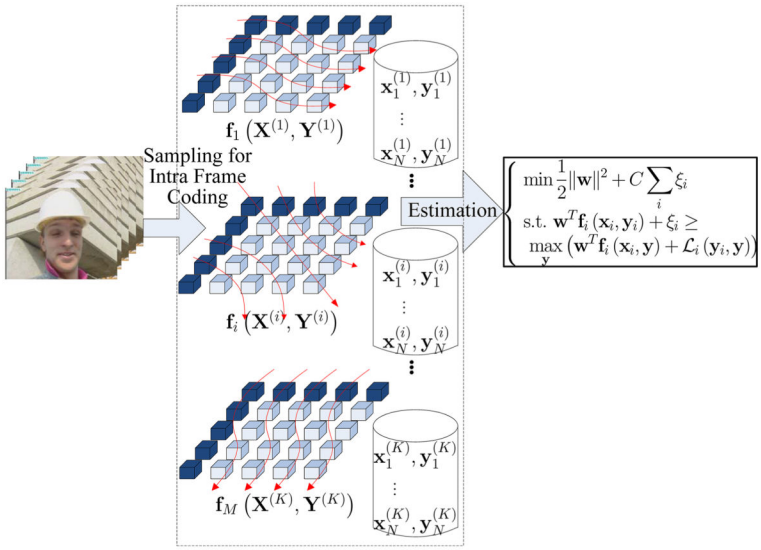



**Fig. 2.**

Diagram of the structured set prediction model, where the training data $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}$ for intra coding are conducted with the class of feature functions $\{\mathbf{f}_i\}$. The max-margin Markov network is generated over the training data by combining the class of feature functions $\{\mathbf{f}_i\}$ with the corresponding normal vector $\{\mathbf{w}_i\}$.

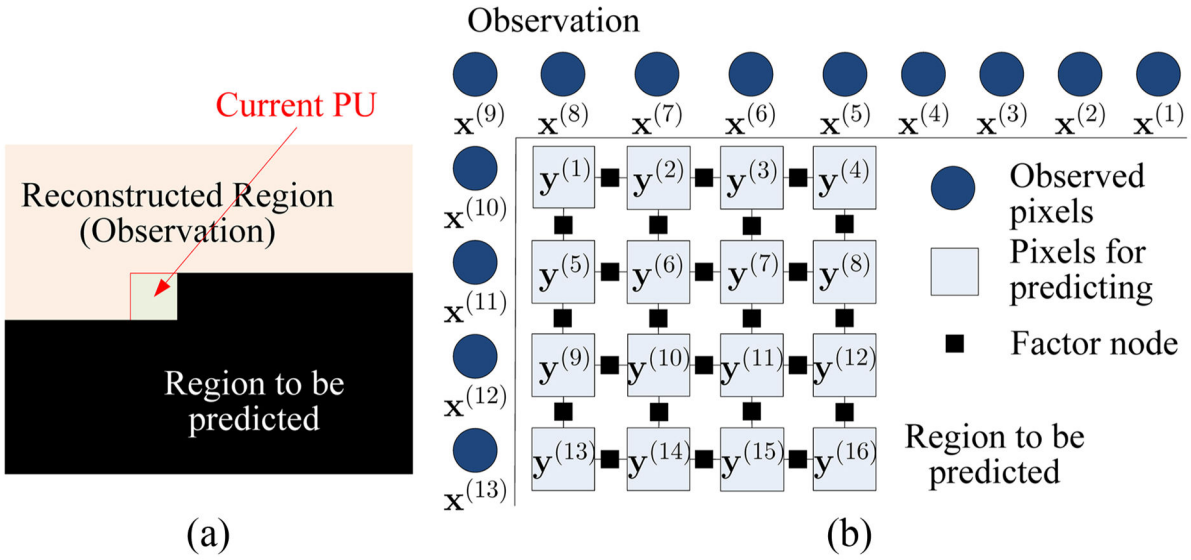



**Fig. 3.**

Graphical model for the structured set prediction model, where $\{\mathbf{y}^{(i)}\}$ is the set of pixels being predicted and $\{\mathbf{x}^{(i)}\}$is the set of observed pixels serving as contexts.

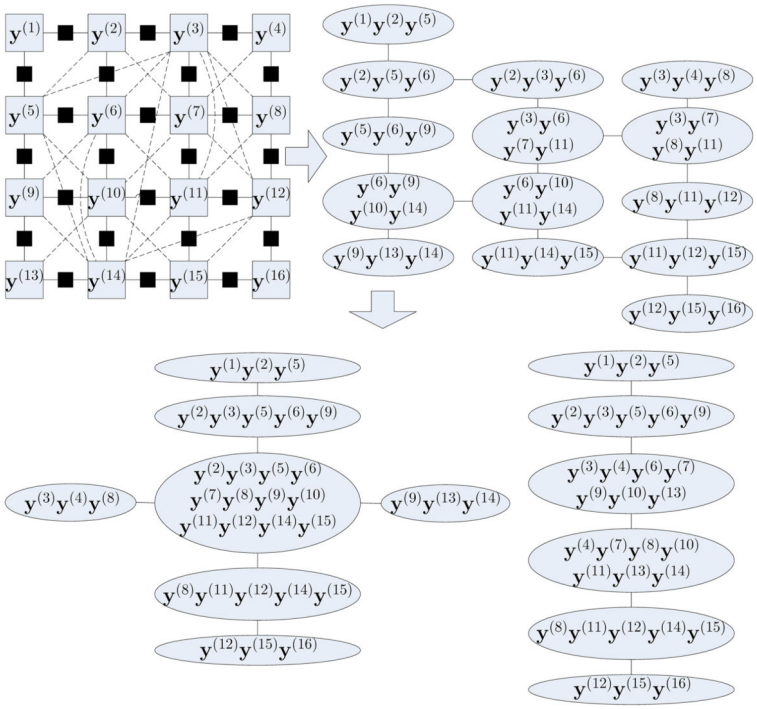
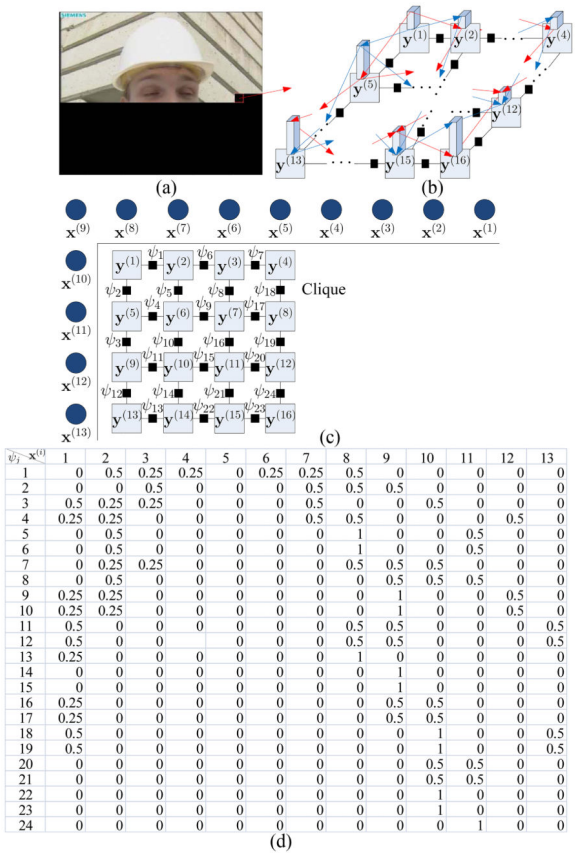


**Fig. 4.**

Junction tree for the generated MRF. (a) (Up left) Triangulation of the graphical model by adding some dashed edges. (b) (Up right) Intermediate result of construction of junction tree. (c) (Bottom) Two possible junction trees derived from the graphical model.




| $\psi_j$ \ $\mathbf{x}^{(i)}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.5 | 0.25 | 0.25 | 0 | 0.25 | 0.25 | 0.5 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0.5 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0 | 0 | 0 | 0 |
| 3 | 0.5 | 0.25 | 0.25 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0.5 | 0 | 0 | 0 |
| 4 | 0.25 | 0.25 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 | 0.5 | 0 |
| 5 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 | 0 | 0 |
| 6 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 | 0 | 0 |
| 7 | 0 | 0.25 | 0.25 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0 | 0 | 0 |
| 8 | 0 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0.5 | 0 | 0 |
| 9 | 0.25 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 | 0 |
| 10 | 0.25 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 | 0 |
| 11 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 | 0.5 |
| 12 | 0.5 | 0 | 0 | | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 | 0.5 |
| 13 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 16 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| 17 | 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| 18 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 |
| 19 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 |
| 21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 | 0.5 | 0 | 0 |
| 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

(d)

**Fig. 5.**

Analytic sample of the structured set prediction model. (a) Test 4 × 4 block in *Foreman* sequence. (b) Diagram for training in the max-margin Markov network. (c) Graphical model constructed for the 4 × 4 block. (d) Weighting vectors for all cliques over the observed contexts.

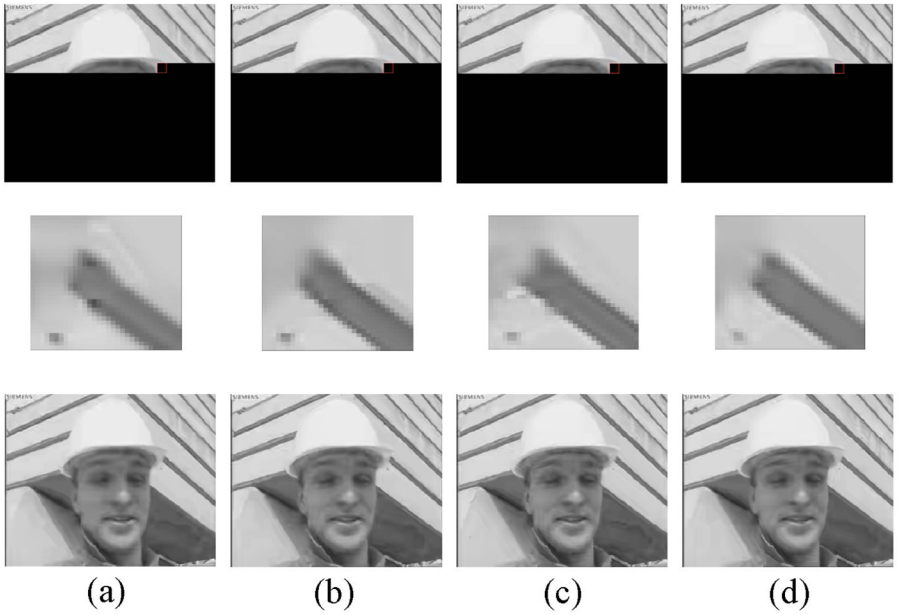



**Fig. 6.**

Prediction performance of visual quality and PSNR with training iterations 1, 5, 10, and 50. (a) Iteration = 1 PSNR = 32.17. (b) Iteration = 5 PSNR = 32.31. (c) Iteration = 10 PSNR = 32.41. (d) Iteration = 50 PSNR = 32.48.

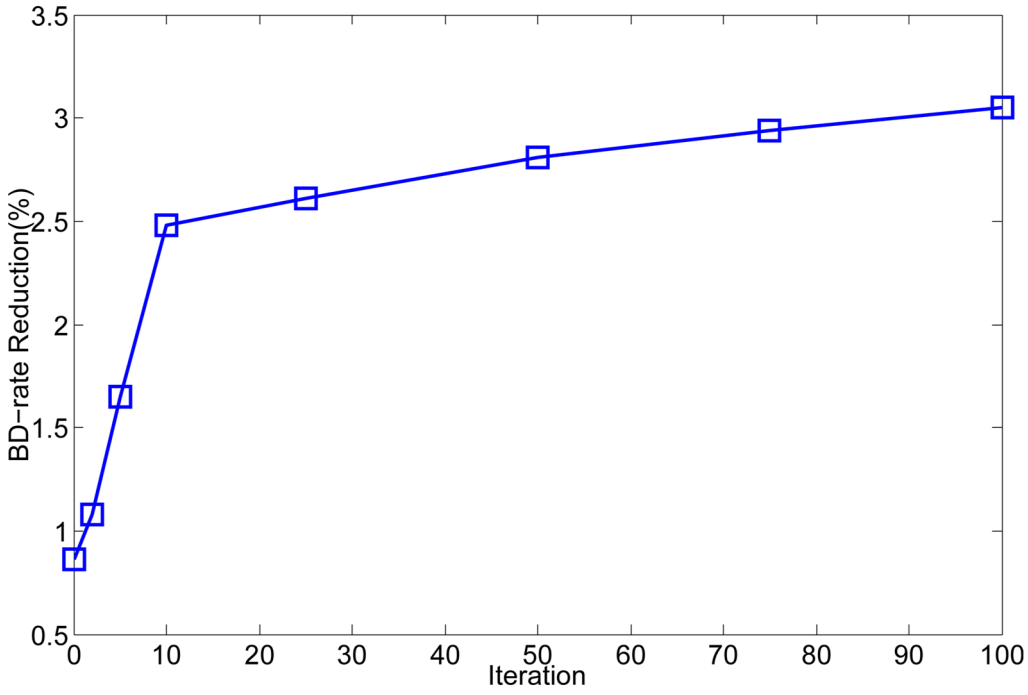



**Fig. 7.**

Prediction performance (BD-rate reduction in %.) under training process with various iterations 1, 2, 5, 10, 25, 50, 75, and 100.

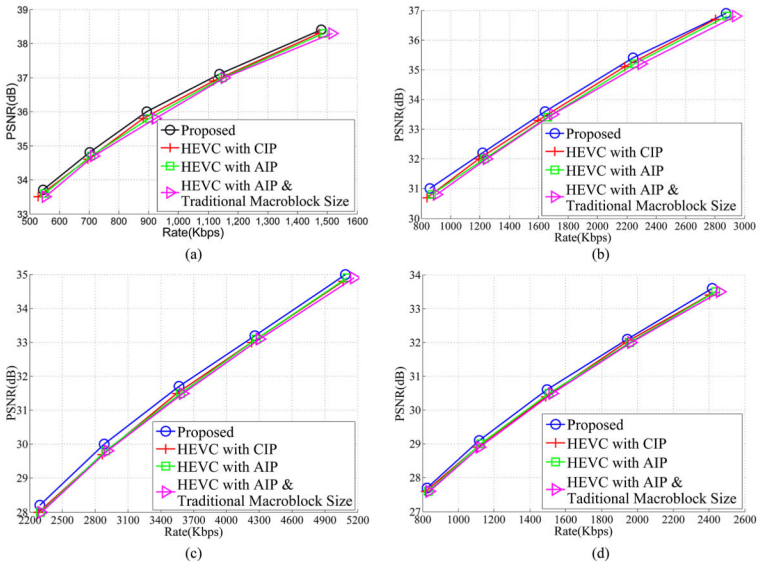



**Fig. 8.**

Rate-distortion curve for performance comparison of CIF sequences. The proposed model is compared with the CIP with HEVC, default HEVC intraprediction (angular intraprediction, AIP) with extended CU size (MAX_CU_SIZE=64), and traditional CU size (MAX_CU_SIZE=16), respectively. (a) Performance comparison of sequence *Foreman*_(352)×(288). (b) Performance comparison of sequence *Football*_(352)×(288). (c) Performance comparison of sequence *Mobile*_(352) × (288). (d) Performance comparison of sequence *Bus*_(352) × (288).

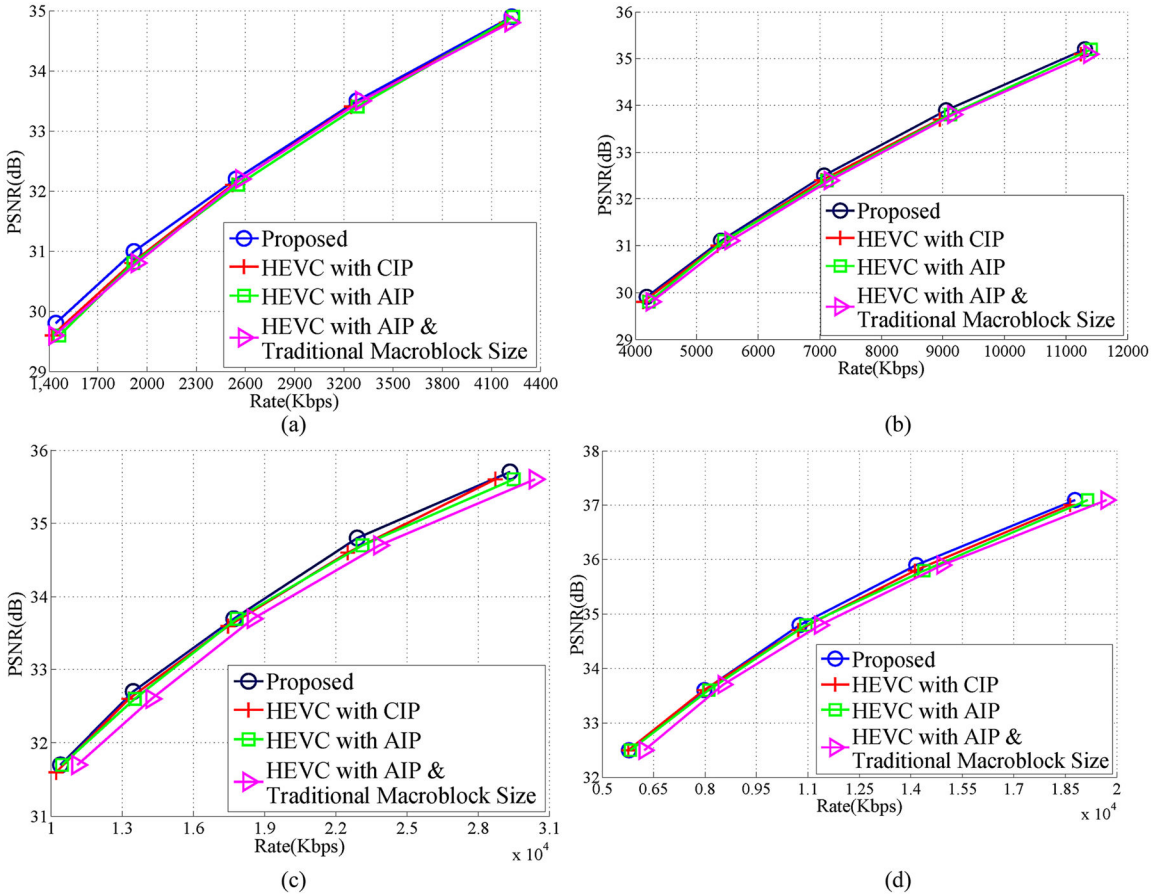



**Fig. 9.**

Rate-distortion curve for performance comparison of sequences with extensive resolutions. The proposed model is compared with the CIP with HEVC, default HEVC intra prediction (angular intraprediction, AIP) with extended CU size (MAX_CU_SIZE=64), and traditional CU size (MAX_CU_SIZE=16), respectively. (a) Performance comparison of sequence *BlowingBubbles*_416 × 240. (b) Performance comparison of sequence *BQMall*_832 × 480. (c) Performance comparison of sequence *Cactus*_1920 × 1080. (d) Performance comparison of sequence *ParkScene*_1920 × 1080

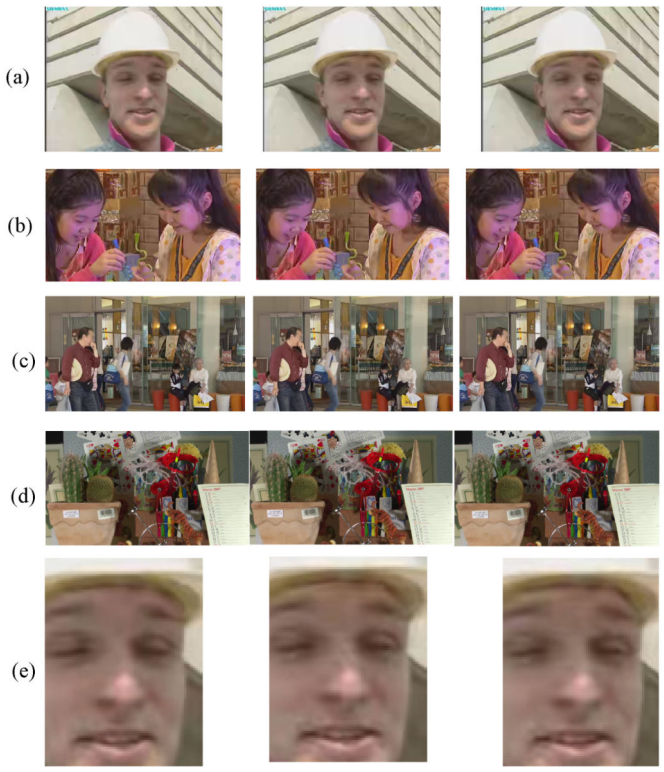



**Fig. 10.**

Visual results of reconstructed sequences. From left to right and from top to bottom, there are the reconstructed frames of *Foreman*, *BlowingBubbles*, *BQMall*, and *Cactus* by the proposed model, CIP with HEVC, and HEVC intraprediction, respectively. (a) Visual performance of *Foreman* sequence. From left to right, PSNR of the reconstructed frames are 33.82, 33.58, and 33.53 dB, respectively. (b) Visual performance of *BlowingBubbles* sequence. From left to right, PSNR of the reconstructed frames are 30.98, 30.80, and 30.77 dB, respectively. (c) Visual performance of *BQmall* sequence. From left to right, PSNR of the reconstructed frames are 32.59, 32.48, and 32.40 dB, respectively. (d) Visual performance of *Cactus* sequence. From left to right, PSNR of the reconstructed frames are 32.81, 32.68, and 33.61 dB, respectively.

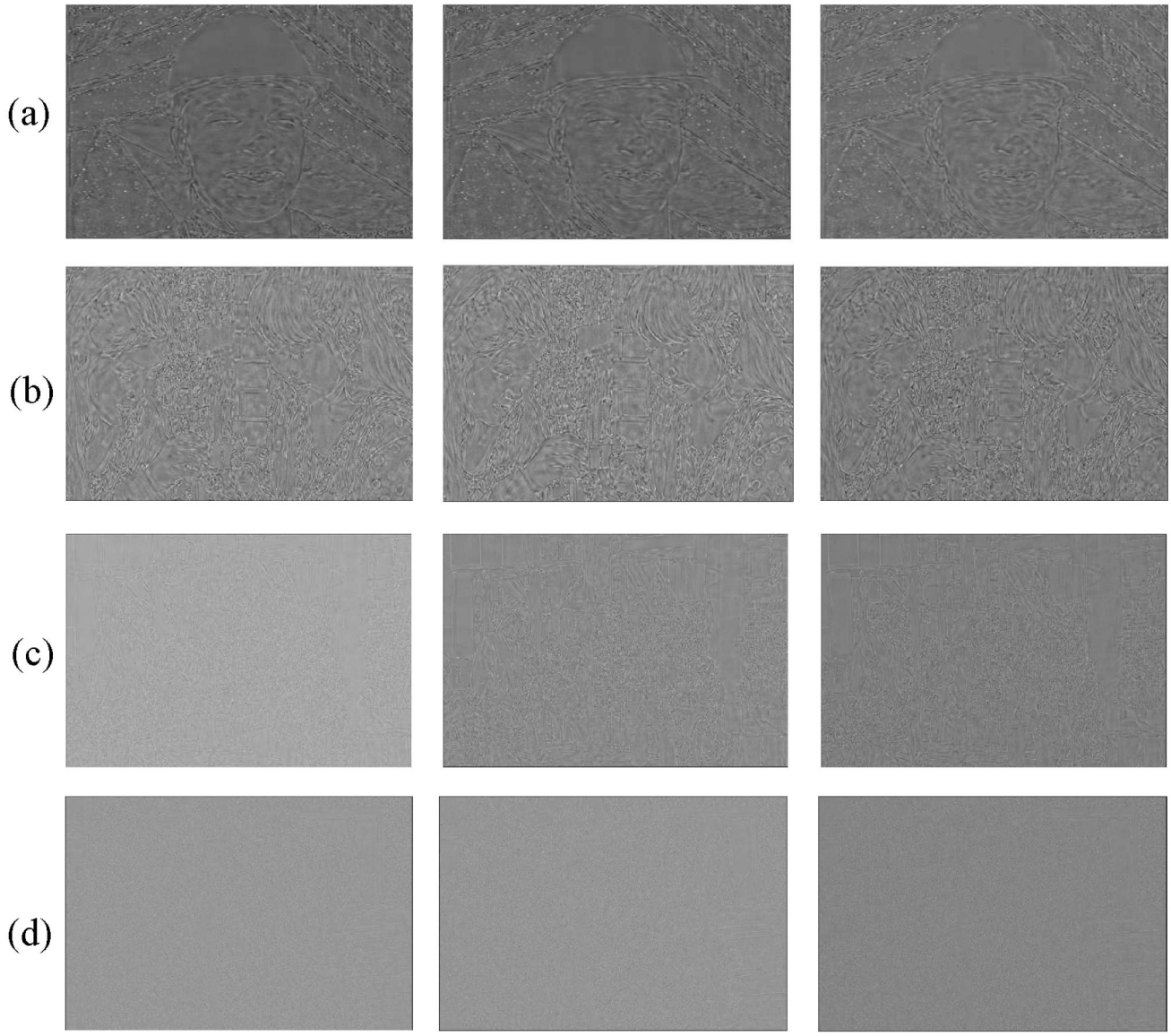



**Fig. 11.**

Prediction residuals obtained by subtracting reconstructed sequences from original sequences. From left to right and from top to bottom: prediction residuals of *Foreman*, *BlowingBubbles*, *BQMall*, and *Cactus* by the proposed model, CIP with HEVC, and HEVC intra prediction, respectively. (a) Prediction residuals of *Foreman* sequence. From left to right: first-order entropies of the prediction residuals are 4.2764, 4.2944, and 4.2842 b/p, respectively. (b) Prediction residuals of *BlowingBubbles* sequence. From left to right: first-order entropies of the prediction residuals are 4.8571, 4.8742, and 4.8647 b/p, respectively. (c) Prediction residuals of *BQMall* sequence. From left to right, first-order entropies of the prediction residuals are 4.4722, 4.5220, and 4.5150 b/p, respectively. (d) Prediction residuals of *Cactus* sequence. From left to right, first-order entropies of the prediction residuals are 4.3653, 4.3744, and 4.3709 b/p, respectively.

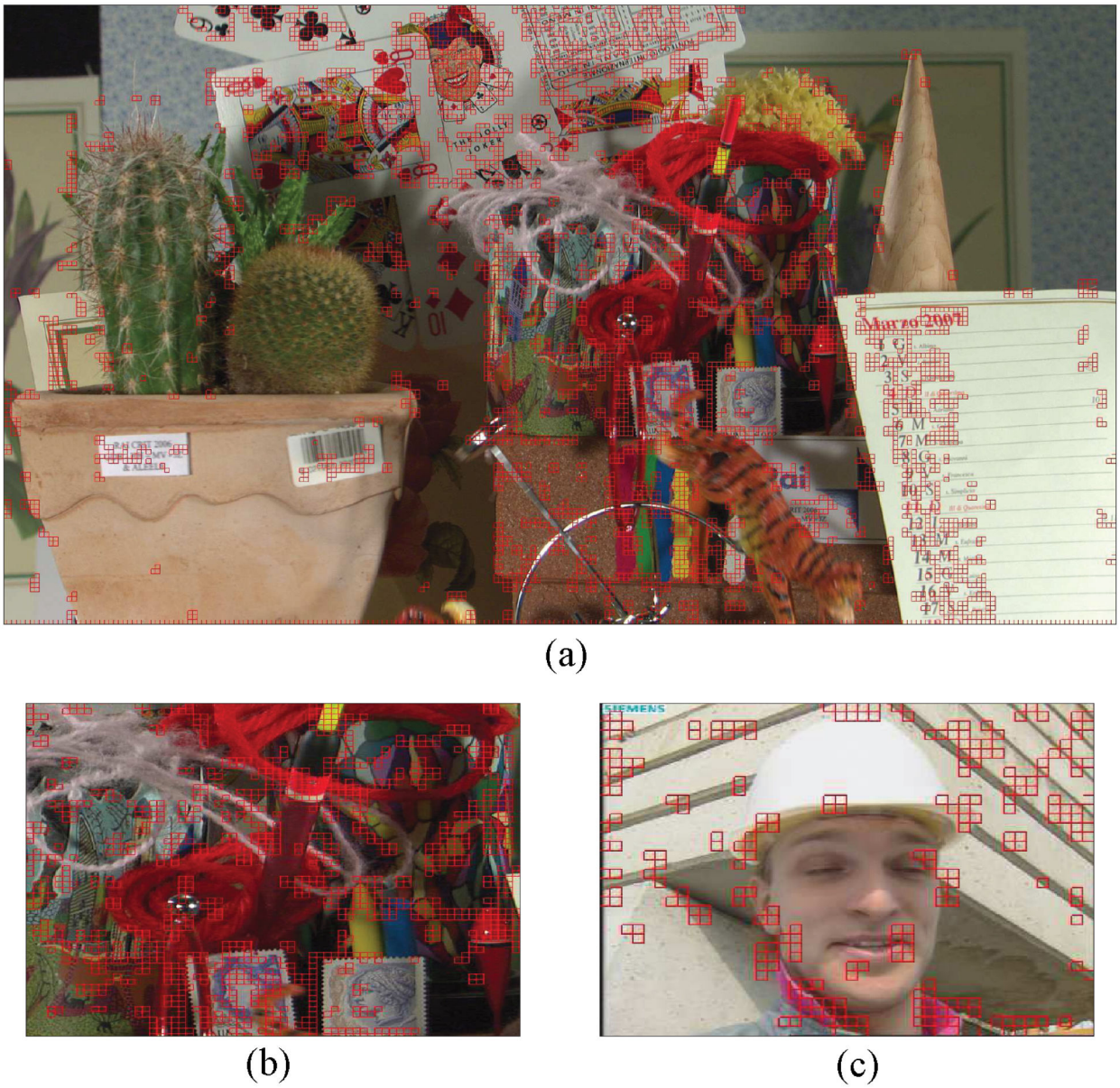



**Fig. 12.**

Block partition for video sequences by the proposed model. The proportions of the blocks predicted with the structured set prediction model are 11.3% and 7.6%, respectively. (a) Block partition for *Cactus* sequence. (b) Zoomed partition for *Cactus* sequence. (c) Block partition for *Foreman* sequence.

**TABLE I**

BD-PSNR (dB) Performance in Comparison to HEVC With CIP

| Sequence | Size | The proposed model | | | Combined intra prediction | | |
|---|---|---|---|---|---|---|---|
| | | Y BD-PSNR | U BD-PSNR | V BD-PSNR | Y BD-PSNR | U BD-PSNR | V BD-PSNR |
| Bus | 352 × 288 | 0.14 | 0.13 | 0.01 | −0.01 | 0.02 | 0.05 |
| Football | | 0.25 | 0.08 | 0.07 | 0.08 | 0.12 | 0.08 |
| Foreman | | 0.17 | 0.06 | 0.02 | 0.04 | 0.08 | 0.10 |
| Mobile | | 0.15 | 0.09 | 0.05 | 0.01 | 0.05 | 0.05 |
| Blowing Bubbles | 416 × 240 | 0.11 | 0.04 | 0.03 | 0.03 | 0.04 | 0.06 |
| Basketball | 720 × 576 | 0.12 | 0.04 | 0.01 | 0.07 | 0.04 | 0.05 |
| BQMall | 832 × 480 | 0.08 | 0.00 | 0.01 | 0.08 | 0.05 | 0.06 |
| Cactus | 1920 × 1080 | 0.12 | 0.05 | 0.05 | 0.01 | 0.05 | 0.12 |
| ParkScene | | 0.09 | 0.03 | 0.03 | 0.07 | 0.09 | 0.09 |

*IEEE Trans Circuits Syst Video Technol*. Author manuscript; available in PMC 2014 December 09.

**TABLE II**

BD-Rate (%) Change in Comparison to HEVC With CIP

| Sequence | Size | The proposed model | | | Combined intra prediction | | | The proposed (Full R-D) |
|---|---|---|---|---|---|---|---|---|
| | | Y BD-rate | U BD-rate | V BD-rate | Y BD-rate | U BD-rate | V BD-rate | Y BD-rate |
| Bus | 352 × 288 | −1.63 | −2.14 | −0.37 | 0.15 | −1.57 | −1.03 | −1.59 |
| Football | | −2.85 | −1.22 | −1.50 | −1.23 | −2.54 | −2.52 | −2.72 |
| Foreman | | −2.68 | −0.75 | −0.02 | −0.87 | −2.17 | −2.22 | −2.57 |
| Mobile | | −1.39 | −1.34 | −0.73 | −0.25 | −0.64 | −0.64 | −1.44 |
| BlowingBubbles | 416 × 240 | −1.87 | −0.27 | −0.08 | −0.28 | −0.97 | −1.08 | −1.86 |
| Basketball | 720 × 576 | −1.44 | −0.38 | −0.48 | −0.75 | −0.77 | −0.78 | −1.17 |
| BQMall | 832 × 480 | −1.08 | −0.05 | −0.13 | −1.05 | −1.07 | −1.20 | −1.18 |
| Cactus | 1920 × 1080 | −2.09 | −1.34 | −1.37 | −0.37 | −2.09 | −3.85 | −2.17 |
| ParkScene | | −1.70 | −0.99 | −1.08 | −1.31 | −2.86 | −5.08 | −1.46 |
| Average | | −1.86 | −0.94 | −0.64 | −0.66 | −1.63 | −2.04 | −1.80 |

*IEEE Trans Circuits Syst Video Technol*. Author manuscript; available in PMC 2014 December 09.

**TABLE III**

BD-Rate (%) Change With Various Maximum CU Size

| Sequence | Size | MAX_CU_SIZE=64 | | | MAX_CU_SIZE=16 | | |
|---|---|---|---|---|---|---|---|
| | | Y BD-rate | U BD-rate | V BD-rate | Y BD-rate | U BD-rate | V BD-rate |
| Bus | 352 × 288 | −1.55 | −2.08 | −0.47 | −2.30 | −3.54 | −2.56 |
| Football | | −2.81 | −1.26 | −1.51 | −4.57 | −5.19 | −5.98 |
| Foreman | | −2.69 | −0.79 | −0.02 | −4.80 | −6.08 | −6.16 |
| Mobile | | −1.42 | −1.29 | −0.67 | −2.30 | −3.45 | −2.74 |
| BlowingBubbles | 416 × 240 | −1.87 | −0.31 | −0.09 | −2.56 | −1.85 | −2.93 |
| Basketball | 720 × 576 | −1.39 | −0.38 | −0.50 | −2.22 | −2.16 | −1.31 |
| BQMall | 832 × 480 | −1.03 | −0.08 | −0.17 | −2.11 | −2.73 | −3.56 |
| Cactus | 1920 × 1080 | −2.18 | −1.31 | −1.35 | −4.45 | −5.55 | −8.43 |
| ParkScene | | −1.70 | −0.99 | −1.08 | −3.17 | −4.91 | −6.25 |
| Average | | −1.86 | −0.94 | −0.64 | −3.16 | −3.94 | −4.44 |

*IEEE Trans Circuits Syst Video Technol*. Author manuscript; available in PMC 2014 December 09.

## TABLE IV

Computational Complexity Comparison in Terms of Decoding Speed (s/frame) for the Proposed Model, HEVC With CIP, and HEVC With AIP, and Run-Time Ratio (%) Compared With HEVC With AIP

| Sequences | The proposed model | HEVC with CIP | HEVC with AIP | run-time ratio |
|---|---|---|---|---|
| Bus | 11.77/17.37 | 0.16/0.32 | 0.14/0.26 | 8407/6681 |
| Football | 8.12/15.66 | 0.14/0.23 | 0.13/0.23 | 6246/6809 |
| Foreman | 7.20/41.67 | 0.14/0.23 | 0.12/0.25 | 6000/16668 |
| Mobile | 25.86/30.82 | 0.21/0.37 | 0.20/0.36 | 12930/8561 |
| BlowingBubbles | 10.43/29.47 | 0.18/0.32 | 0.17/0.30 | 6135/9823 |
| Basketball | 45.13/87.60 | 0.69/1.34 | 0.60/1.29 | 7522/6791 |
| BQMall | 84.84/237.08 | 0.54/0.91 | 0.48/0.96 | 17675/24696 |
| Cactus | 134.60/433.84 | 2.04/4.09 | 1.87/3.71 | 7198/11694 |
| ParkScene | 100.78/279.55 | 1.66/4.16 | 1.63/4.11 | 6183/6801 |

**TABLE V**

Proposed STRUCT Mode Selection Ratio (%) Under Various QP Levels

| Sequences | Size | QP=30 | QP=32 | QP=34 | QP=36 | QP=38 |
|---|---|---|---|---|---|---|
| Bus | 352 × 288 | 26.45 | 24.12 | 20.33 | 18.56 | 17.87 |
| Football | | 10.04 | 7.07 | 6.12 | 5.05 | 3.66 |
| Foreman | | 18.94 | 15.78 | 12.12 | 7.64 | 4.92 |
| Mobile | | 45.61 | 45.27 | 45.14 | 43.37 | 43.24 |
| BlowingBubbles | 416 × 240 | 32.31 | 29.17 | 23.53 | 16.73 | 13.14 |
| Basketball | 720 × 576 | 26.39 | 22.84 | 21.23 | 17.47 | 14.54 |
| BQMall | 832 × 480 | 27.26 | 23.38 | 19.86 | 16.36 | 12.92 |
| Cactus | 1920 × 1080 | 11.38 | 9.52 | 7.90 | 6.15 | 4.87 |
| ParkScene | | 14.12 | 12.99 | 10.74 | 9.36 | 6.65 |

**TABLE VI**

Proposed STRUCT Mode Selection Ratio (%) Under Various Training Iterations When QP Level Is 32

| Sequences | Size | Iteration=1 | Iteration=2 | Iteration=5 | Iteration=10 | Iteration=50 |
|---|---|---|---|---|---|---|
| Bus | 352 × 288 | 2.17 | 3.09 | 9.65 | 17.84 | 22.94 |
| Football | | 1.49 | 1.71 | 3.18 | 5.76 | 6.81 |
| Foreman | | 1.26 | 1.79 | 5.29 | 13.15 | 14.63 |
| Mobile | | 1.55 | 2.45 | 12.27 | 32.66 | 41.58 |
| BlowingBubbles | 416 × 240 | 2.02 | 2.68 | 10.66 | 22.34 | 27.95 |
| Basketball | 720 × 576 | 1.37 | 2.01 | 8.56 | 18.16 | 20.67 |
| BQMall | 832 × 480 | 1.16 | 1.74 | 8.14 | 18.87 | 22.04 |
| Cactus | 1920 × 1080 | 0.91 | 1.44 | 3.22 | 7.10 | 9.07 |
| ParkScene | | 1.02 | 1.73 | 3.40 | 9.63 | 12.46 |