A Learning-based Video Compression on Low-Quality Data by Unscented Kalman Filters with Gaussian Process Regression

Hongkai Xiong, Zhe Yuan
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai 200240, China
Email: {xionghongkai, yuanzhe}@sjtu.edu.cn

Yuan F. Zheng
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210 USA
Email: zheng@ece.osu.edu

***Abstract*— With the ever increasing concern of vision-based video analysis and coding over resource-limited systems, this paper proposes a novel video coding scheme that exploits low-quality video data and formulates as an inverse learning based video reconstruction from online training by diverse stochastic processes. Given a sparsely sampled incomplete data, the intrinsic nonlocal and spatio-temporal geometric regularity related to online training examples in the key frames are considered as a state-dependent uncertainty estimation problem using Gaussian Process (GP) regression. Unlike non-parametric or exemplar-based sampling methods, we consider non-parametric system models for sequential state estimation by using the Unscented Kalman Filter (UKF) as the state estimator. It inherits the unscented transform for linearization to the transition function and the observation function. Once an approximate motion and observation model is available, it can naturally be incorporated to make a further performance improvement.**

I. INTRODUCTION

There are ever increasing number of resource-limited systems with a demand of vision-based video analysis and coding, e.g. mobile or wearable cameras. Video compression is endowed with a challenging task to exhibit a high-performance capability on a given low-quality video data. The mainstream video coding schemes, e.g. H.264/AVC, only focus on exploring statistical redundancy among pixels through intra and inter prediction. In 2005, High-performance Video Coding (HVC) was initialized to suit regions of different properties. In 2010, High Efficiency Video Coding (HEVC) joint project were expected to attain bit rate reduction of 50% at the same subjective image quality comparing to H.264/AVC. Hopefully, it provides an opportunity to revisit most compression techniques in a new paradigm allowing for pixel-wise difference while achieving acceptable subjective visual quality.

The vision based technologies have ever been envisaged to hallucinate missing image contents with good perceptual quality. Parallel with the vision based progress [1], image-based compression framework has been designed by either removing some smooth and flat blocks [2] or reducing the entropy of source by clustering the homogeneous area that contains the epitome content of all related regions [3]. Obviously, it is so difficult to reflect stochastic pixel intensity models (e.g., scaling, rotation, and local motion) and attributes (e.g. combination of different texture) without sufficient prior knowledge. Those have not yet been suitable for video compression because there often exists temporal inconsistence derived from independent reconstruction of each frame. The coding burden from the assistant information is also a critical issue for generic video coding. Another important effort is focusing on a correlation between a sparsely sampled low-resolution and high frequency contents [4]. To guarantee the sufficient prior and overcome the inaccurate estimated motion, a learned co-occurrence prior to predict the correspondence between low-resolution and high-resolution image patches has been investigated using a training set [5]. In order to achieve more patch patterns with a smaller database, manifold learning has been assumed similarity between the two manifolds in the high-resolution and the low-resolution patch spaces [6]. Such exemplar-based detail reconstruction would assume a linear relationship among high-resolution signals and can be precisely recovered from their low-dimensional projections. When the dataset is corrupted by noise or non-rigid transform, the restoration will always suffer from distortion.

In this paper, we proposes a novel video coding scheme that exploits low-quality video data and formulates as an inverse learning based video reconstruction from online training by diverse stochastic processes. It acts as a super-resolution video completion with non-parametric nonlinear system models for sequential state estimation where the state estimator is an Unscented Kalman Filter (UKF). It inherits the unscented transform for linearization to the transition function and the observation function. A subset of frames regularly spaced in the video sequence abstracts the co-occurrence prior in a sparser way, and the remaining key frames are treated as an online training set. Considering each sample of a frame as a system state, its estimated value is supposed to depend on both the manner of pre-filtering (the observation model) and the state transition (the process model). The nonlinear process and observation models in the reconstruction inference could be learned from the online training examples using Gaussian Process (GP) regression in a non-parametric manner. The

978-1-4244-9474-3/11/$26.00 ©2011 IEEE

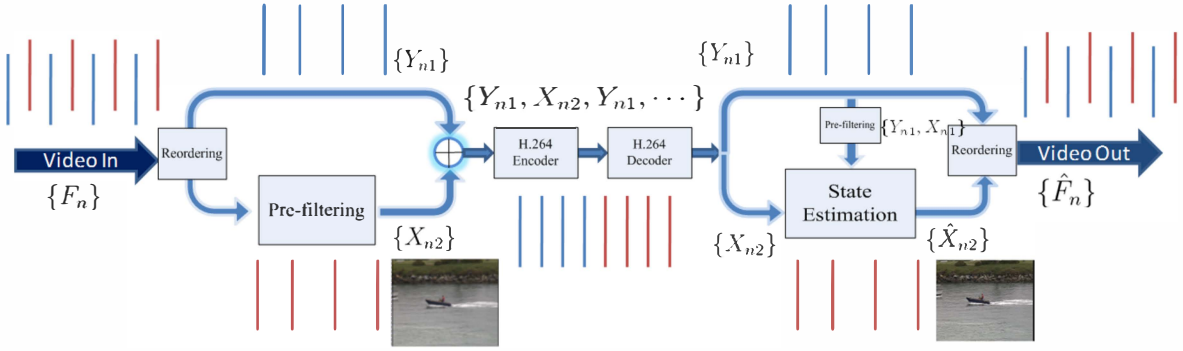


Fig. 1. The diagram of the proposed video coding framework where the red lines indicate the abstracted sample frames.

proposed approach relies on the state-dependent uncertainty estimates, which take into account both noise and regression uncertainty derived from the limited training examples of key frames. It is capable of estimating the state of arbitrary nonlinear systems on non-stationary video statistics. Once an approximate motion and observation model is available, it can naturally be incorporated to make a further performance improvement compared to existing texture synthesis methods.

II. Super-Resolution Video Completion with Nonlinear System Models

The proposed coding framework is depicted in Fig. 1, where a subset of video frames are pre-filtered to achieve a smoothed version as $\{X_{n2}\}$ in a sparse representation while remaining key frames are denoted as $\{Y_{n1}\}$. Hereinafter, X denotes image samples with smooth filtering (observations) and Y for image samples containing detail information (states). We adopt a prior filter with deterministic weights such that the decoder could reproduce the filter with an accurate observation model. The two types of frames will be arranged into groups of pictures (GOP) and coded independently. For the pre-filtering sub-sequences, the higher compression efficiency can be achieved from both intra and inter predictions. At the decoder side, the missing detail for $\{X_{n2}\}$, i.e. $\{\hat{X}_{n2}\}$ is predicted based on an inference from $\{Y_{n1}\}$.

A. Problem Formulation

We consider a generic stochastic discrete filtering problem in a dynamic system as:

$$\mathbf{Y}_n = \mathbf{F}_{n,n-1}(\mathbf{Y}_{n-1}) + \epsilon_n, \tag{1}$$

$$\mathbf{X}_n = \mathbf{G}_n(\mathbf{Y}_n) + \delta_n \tag{2}$$

where $\mathbf{Y}_n$ denotes the state vector at time n, which shall represent the video frames with complete information. $\mathbf{X}_n$ is an observation vector to represent the incomplete data generated by a function $\mathbf{G}$. ϵ and δ are random noise with probability $p(\epsilon) \backsim N(0, \mathbf{Q_1})$ and $p(\delta) \backsim N(0, \mathbf{Q_2})$, resepctively.

Eq. 1 characterizes the state transition probability prior $p(\mathbf{Y}_n|\mathbf{Y}_{n-1})$, which reflects the temporal correlation between frames. Eq. 2 describes the likelihood of the prediction comparing with the observation $p(\mathbf{X}_n|\mathbf{Y}_n)$, which measures the spatial accuracy of the restored content. The objective of the

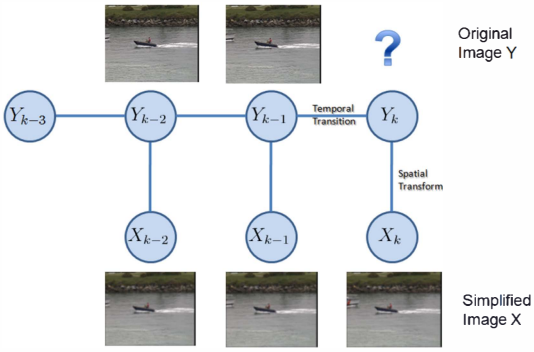


Fig. 2. Frame reconstruction through state estimation.

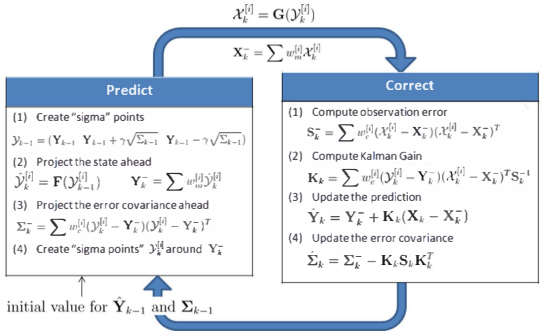


Fig. 3. A summary of the unscented Kalman Filter algorithm.

filtering is to estimate the optimal state value, i.e. $p(\hat{\mathbf{Y}}_n|\mathbb{X}_n)$, where $\mathbb{X}_n = \{\mathbf{X}_0, \mathbf{X}_1, \cdots \mathbf{X}_n\}$, as Fig. 2.

The state transition and measurement process $\mathbf{F}$ is considered as nonlinear while the observation model $\mathbf{G}$ is regularized as a prior. To solve such a non-linear problem, we apply UKF which uses a deterministic sampling (the unscented transform) to pick a set of sample points (sigma points). The sigma points will be propagated through the non-linear function for the estimation of mean and covariance. The overall process is summarized by Fig. 2. Given a proper model in the form of (1) and (2), we will get stable prediction after iterations.

B. UKF with Gaussian Process Prediction Model

The motion model $\mathbf{F}$ is a known non-linear function, and such a parametric model is usually unavailable. According to UKF from Fig. 2, we are interested in what the best forward projection $\mathbf{Y}_n = \mathbf{F}(\mathbf{Y}_{n-1})$ is, rather than a parametric estimation of $\mathbf{F}$, given a state $\mathbf{Y}_{n-1}$.

Here we introduce Gaussian Process (GP) for learning regression from sampled data [7]. It defines a distribution over functions, which is able to predict a projection from the noised training data as well as estimate the uncertainty.

Consider a function $b_i = f((\mathbf{a_i})) + \epsilon$ with training data $D = \{(\mathbf{a}_1, b_1), (\mathbf{a}_2, b_2), \cdots, (\mathbf{a}_n, b_n)\}$, a GP model is able to provide prediction on an unknown points $\mathbf{a}^*$ through defining a distribution of b^* with mean:

$$GP_\mu(\mathbf{a}_*, D =< A, \mathbf{b} >) = \mathbf{k}_*^T[K(A, A) + \sigma_n^2\mathbf{I}]^{-1}\mathbf{b} \quad (3)$$

and variance:

$$GP_\mu(\mathbf{a}_*, D) = k(a_*, a_*) - \mathbf{k}_*^T[K(A, A) + \sigma_n^2\mathbf{I}]^{-1}\mathbf{k}_* \quad (4)$$

where k is the kernel function:

$$k(\mathbf{a}, \mathbf{a}') = \sigma_f^2 \exp((-\frac{1}{2}(\mathbf{a} - \mathbf{a}')(\mathbf{a} - \mathbf{a}')^T)) \quad (5)$$

and $K(A, A)$ is the kernel matrix with elements defined as $K_{i,j} = K(\mathbf{a}_i, \mathbf{a}_j)$.

Our objective is to project the current state forward, i.e. to estimate the pixel value in the forward frame based on the backward motion field. The estimation error is supposed to be involved and corrected within a Gaussian system model. Since GP only outputs a scalar value, it cannot be directly applied to predict state vector with a high dimension in a video sequence. Here we consider to predict the motion vector instead of the state vector so that only two scalars will be estimated, i.e.

$$\mathbf{Y}_i = \mathbf{M}(\mathbf{Y}_{i-1}, V_{i-1}) \quad (6)$$

where $\mathbf{M}$ is a defined interpolation operator and V_{i-1} is the motion vector obtained through GP:

$$V_{i-1} = GP_\mu^F(\mathbf{Y}_{i-1}, D_F =< \mathbb{Y}_n, \mathbb{V}_n >). \quad (7)$$

The training set will be acquired using the preceding frames. In the block by block reconstruction, the motion between frames is consistent within a small region and the motion vector is correlated with the current state vector.

As for the observation model, it is correlated with the pre-filter at the encoder side. A Gaussian smooth filter is used to maintain an accurate observation model. Since it is linear, the observation can be expressed in the matrix form, i.e. $[\mathbf{x}]_{m^2\times 1} = [\mathbf{G}]_{m^2\times n^2}[\mathbf{y}]_{n^2\times 1}$, where $\mathbf{x}$ is the observation vector and $\mathbf{y}$ is the state vector that contains necessary information to generate x; $n = m + \text{filter\_size} - 1$.

Once we have all the parameters and models required, we shall make reconstruction in the following manner. Given a frame we are to restore $\mathbf{X}_R \in \{X_{n2}\}$ (the current observation, i.e. $R = n + 1$), we will select a set of frames as reference, which usually consist of its n neighboring frames $\mathbb{Y}_n = \{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n\} \subset \{Y_{n1}\}$ and its corresponding observations $\mathbb{X}_n = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n\} \subset \{X_{n1}\}$. For one size $m \times m$ observation block denoted as $\mathbf{X}_R^i$ (i denotes the coordinate of the interested block center), we apply motion estimation using $\mathbf{X}_R^i$ from $\mathcal{X}_n$ to generate a sequence of observations $\mathbb{X}_n^i = \{\mathbf{X}_k^{i_k}\}$, and the corresponding known $n \times n$ states $\mathbb{Y}_n^i = \{\mathbf{Y}_k^{i_k}\}(k = 1, 2, \cdots, n)$, together with the motion vector used for learning the motion model $D_{F,R}^i =< \{\mathbf{Y}_k^{i_k}\}, \{V_k\} >$.

We shall start UKF with initial state $\mathbf{Y}_0$ and variance $\Sigma_0 = E[(\mathbf{Y}_k^{i_k} - \mathbf{Y}_{k-1}^{i_{k-1}})(\mathbf{Y}_k^{i_k} - \mathbf{Y}_{k-1}^{i_{k-1}})^T](k = 1, 2, \cdots, n)$. During each iteration, a set of sigma points will be generated based on the variance. GP will be applied to predict the current state forward projection from summing the projection of sigma points in the form as shown by Fig. 3:

$$\mathcal{Y}_k^{[i]} = \mathbf{M}(\mathcal{Y}_{k-1}^{[i]}, GP_\mu^F(\mathcal{Y}_{k-1}^{[i]}, D_{F,R}^i)) \quad (8)$$

while the observations are generated from the filter matrix:

$$\mathcal{X}_k^{[i]} = [\mathbf{G}]_{m^2\times n^2}\mathcal{Y}_k^{[i]}. \quad (9)$$

After iterations that the system has become stable, we shall input the current observation $\mathbf{X}_{n+1}^i$, with which we shall get an optimal estimation of the desired state vector (the image block containing missing detail) from previous information, i.e. the maximum posteriori possibility $p(\hat{\mathbf{Y}}_{n+1}^i|\mathbb{X}_{n+1}^i, \mathbb{Y}_n^i)$.

For each $m \times m$ block, we will reconstruct a state vector with size $\mathfrak{N} \times \mathfrak{N}$. Since $\mathfrak{N} > m$, there will be overlap between the state vectors, which is equal to one half of the filter size. A min-cut algorithm is applied to the overlapped area to achieve an optimal seam and to avoid block artifact.

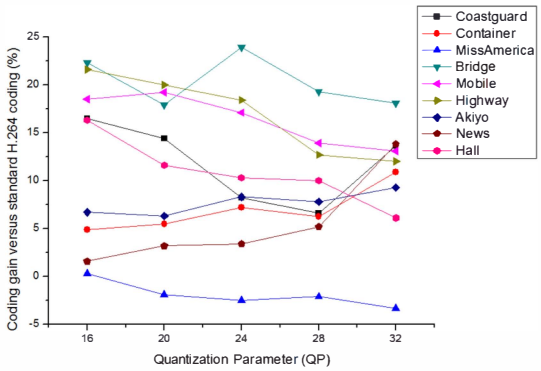


Fig. 4. The bit-rate saving percentages *versus* H.264/AVC.

III. Experimental Results

To illustrate the efficiency of the proposed video completion algorithm compared to existing vision-based methods, we perform the experiments on a variety of test video sequences with QCIF 176×144 resolution in YUV 4:2:0 format. Without loss of generality, the even frames are selected to be pre-filtered with a Gaussian smooth filter (with size 5 and $\sigma = 1$) and the odd frames as key frames. For comparison, we use the same coding configuration for the involved H.264/AVC coding: two B frames, CABAC entropy coding with a frame rate of 25 fps, and a GOP size of 9 frames. The decoded sequence with incomplete data is further reconstructed using the proposed algorithm. To restore one frame with low-quality data, seven frames are used as reference to create the training set. Fig. 4 summarizes the result on coding gain *versus* quantization parameter (QP). Within common QP interval [16, 32], up to

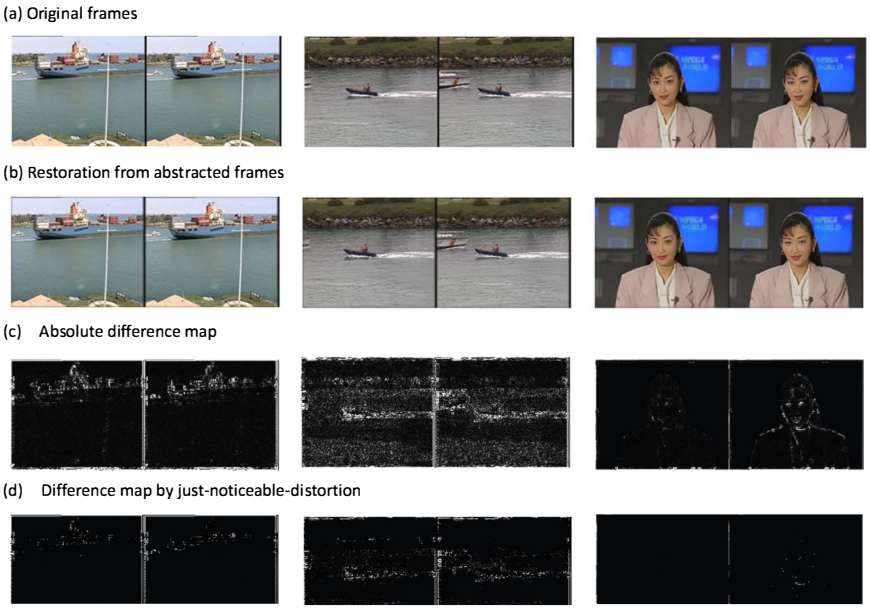
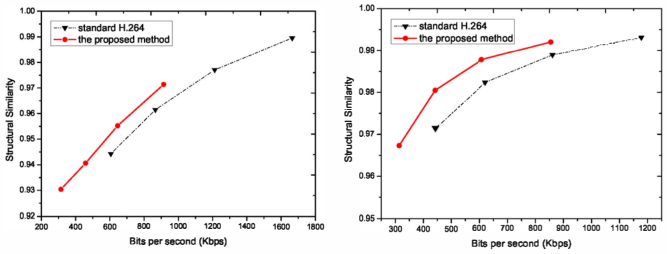


Fig. 5. The illustration of the super-resolution based completion quality as well as the difference map evaluated by JND.



(a) Container (b) Akiyo

Fig. 6. The SSIM-based R-D performance compared to H.264/AVC.

20% coding gain can be achieved beyond the H.264/AVC with similar visual quality. To show the similar visual quality, Fig. 5 shows the reconstructed quality and evaluates just-noticeable distortion (JND) quality. Also, the objective metrics including PSNR and Structural Similarity (SSIM) are provided in Table I (the left two columns). The SSIM index reveals the visual quality of the proposed reconstructed video is hardly distinguishable from the H.264/AVC coded video. Comparing with the space-time completion [8], the proposed completion shows distinguishable improvement without annoying artifacts. The SSIM-based R-D curves on test sequences "container" and "Akiyo" are shown in Fig. 6, which outperform the traditional H.264 scheme.

TABLE I

THE OBJECTIVE QUALITY FOR THE COMPLETED VIDEO.

	The proposed		Spatio-temporal completion	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Coastguard	32.56	0.9426	31.35	0.8483
Container	35.23	0.9902	35.55	0.9422
Akiyo	41.51	0.9949	37.47	0.9860
Foreman	31.94	0.9496	Not Applicable	

IV. CONCLUSION

To address the demand over resource-limited communication, this paper proposes a super-resolution video completion as a non-parametric sequential estimation of nonlinear system state (the restored detail) based on low-quality observation and its measurement (the prior). We consider a more general situation where the motion model is non-linear and expressed through Gaussian process regression in a non-parametric manner. Inheriting the unscented transform for linearization to the transition function and the observation function, the unscented Kalman filter uniformly solves the process and observation equations for the unknown state. Under ill-posed inverse context, Gaussian process regression could unite a sophisticated and consistent view with computational tractability given a finite number of observation pairs.

REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," *Proceeding of ACM SIGGRAPH*, New Orleans, USA, pp.259-263, Jul. 2000.

[2] D. Liu, X. Sun, F. Wu, et al., "Image compression with edge-based inpainting," *IEEE Trans. Circuits and Systems for Video Technology*, vol.17, no. 10, Oct. 2007.

[3] P. Ndjiki-Nya, T. Hinz, and T. Wiegand, "Texture synthesis method for generic video sequences," *Proc. of IEEE International Conference on Image Processing*, vol.3, pp. 397-400, Nov. 2007.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image Super-Resolution via Sparse Representation," *IEEE transactions on image processing*, May 2010.

[5] Y. Li, X.Y. Sun, H.K. Xiong and F. Wu, "Incorporating primal sketch based learning low bit-rate image compression," *Proc. of IEEE International Conference on Image Processing*, vol.3, pp. 173-176, Nov. 2007.

[6] W. Fan and D.Y. Yeung, "Image hallucination using neighbor embedding over visual primitive manifolds," *Proc. of IEEE Computer Vision and Pattern Recognition*, 2007.

[7] C.E. Rasmussen and C.K.I.Williams, *Gaussian Process for Machine Learning*, MIT press, 2005.

[8] Z. Yuan, H.K. Xiong, Y. F. Zheng, "A generic video coding framework based on anisotropic diffusion and spatio-temporal completion," *Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, Mar. 2010.