VIDEO MODELING BY SPATIO-TEMPORAL RESAMPLING AND BAYESIAN FUSION

Yunfei Zheng and Xin Li

West Virginia University
Lane Dept. of Computer Science and Electrical Engineering
Morgantown, WV 26505-6109

ABSTRACT

In this paper, we propose an empirical Bayesian approach toward video modeling and demonstrate its application in multiframe image restoration. Based on our previous work on spatio-temporall adaptive localized learning (STALL), we introduce a new concept of spatio-temporal resampling to facilitate the task of video modeling. Resampling produces a redundant representation of video signals with distributed spatio-temporal characteristics. When combined with STALL model, we show how to probabilistically combine the linear regression results of resampled video signals under a Bayesian framework. Such empirical Bayesian approach opens the door to develop a whole new class of video processing algorithms without explicit motion estimation or segmentation. The potential of our distributed video model is justified by considering its application into two multiframe image restoration tasks: repair damaged blocks and remove impulse noise.

Index Terms— Video signal processing, Bayes procedures, Statistics

1. INTRODUCTION

Motion plays a fundamental role in mathematical modeling of video signals. Despite that motion estimation has been extensively studied in the literature of signal processing and computer vision, uncertainty with motion representation and algorithms for extracting motion information from video signals remains poorly understood. As articulated in [1], deterministic representation of motion information by optical flow or motion vector field is the source of difficulty. In recent years, statistical modeling of video signals without explicit motion estimation have received increasingly more attention. Both nonparametric models (e.g., patch-based [2], [3]) and parametric models (e.g., STALL [4]) have been proposed and achieved promising results for low-level vision tasks.

To overcome the high dimensionality of video, locality assumption is often made - in patch-based models, 3D patches are localized in space and time; in our STALL model, training window used by Least-Square regression is also localized. Despite the convenience of such locality assumption, its validity remains questionable especially when video contains fast camera or object motion. Due to limited temporal sampling rate (typically less than 30Hz), any point along the motion trajectory could be easily located outside a 3D window of limited size. One possible solution to overcome the above difficulty is via spatio-temporal (ST) adaptation or layered representation [5] - i.e., adaptively choose patch shape or training window to match the local motion characteristics. However, such adaptation strategy inevitably involves motion segmentation, another notoriously challenging problem.

In this paper, we present an empirical Bayesian approach toward adaptive modeling of video without explicit motion segmentation. Motivated by the fundamental tradeoff between space and time, we introduce new ST resampling techniques in Section 2 to obtain a redundant representation of video signals. ST resampling is implemented by warping the video sequence in a reversible fashion [6]. Each resampled sequence can be viewed as a redundant version of the original but acquired by a different virtual camera. By the analogy between sample and population, our ST resampling aims at facilitating the modeling task by offering a computational alternative to acquiring more samples by exploiting the relativity of motion (e.g., a moving object would appear still to a camera moving at the same speed).

ST resampling gives rise to distributed or redundant representation of video signals. While modeling the array of virtual cameras by a discrete random variable, we can show how to probabilistically combine the statistical inference results from the distributed representation under a Bayesian framework in Section 3. Such Bayesian fusion offers a clean solution to ST adaptation and avoids the unknown impact of the uncertainty with layered representation. By combining localized STALL and Bayesian fusion, we show how to exploit ST dependency in multi-frame image restoration without explicit motion estimation or segmentation. Significant gain over STALL-based schemes has been achieved for the class of complex video sequences containing camera motion as we will show in Section 4.

2. EMPIRICAL BAYES MODELING VIA SPATIO-TEMPORAL RESAMPLING

Locality or Markovian assumption is often made when modeling high-dimensional signals such as video. Our previous work on STALL model [4] can be viewed as a localized version of previous ST-autoregressive (STAR) model [7]. For completeness, we will briefly review the STALL model to motivate the introduction of ST resampling. Like STAR, we consider a linear regression model

$$X(\vec{n}_0) = \sum_{k=1}^{N} a_k X(\vec{n}_k) + e(\vec{n}_0) \tag{1}$$

where $\mathcal{N} = \{\vec{n}_i\}_{i=1}^{N}$ denotes the ST neighbors of $\vec{n}_0$. However, unlike STAR which estimates model parameters globally, STALL updates AR coefficients $\vec{a}$ on a pixel-by-pixel basis by solving a local LS optimization problem

$$\vec{a} = \underset{\vec{a}}{argmin} ||\vec{y}_{M\times 1} - C_{M\times N}\vec{a}_{N\times 1}||^2 \tag{2}$$

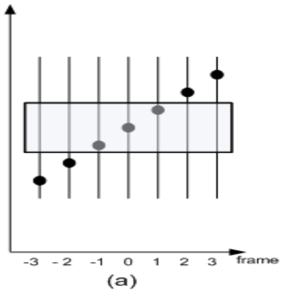

Fig. 1. An example of spatio-temporal resampling: vertical axis denotes y direction ($d_x = 0$).

where $\vec{y}, C$ denote local training window $\mathcal{M}$ and ST neighborhood $\mathcal{N}$ of each pixel in $\mathcal{M}$ respectively. The effectiveness of STALL largely depends on the choices of $\mathcal{N}$ and $\mathcal{M}$. When both $\mathcal{N}$ and $\mathcal{M}$ are fixed, STALL can only handle the class of video containing slow motion. Although adaptive selection of $\mathcal{N}$ and $\mathcal{M}$ by layered representation [5] is conceptually appealing, layer decomposition or motion segmentation remains difficult especially in the presence of complex motion.

An alternative solution to achieve ST adaptation is to realize the fundamental tradeoff between space and time - specifically motion is a $relative$ concept. Human perception of motion arises from the spatial displacement of the same physical point with respect to the camera. Therefore, a moving object might appear still if the camera is moving in parallel to the object at the same speed. Such observation motivates us to introduce a class of ST resampling techniques for video modeling. Based on the analogy between sample and population, we propose to obtain spatio-temporally resampled signal by “reversibly” transforming the original video into another perceptually meaningful one. For example, temporal reversing is a valid resampling operation (reversed signal is physically infeasible but perceptually convincing); while temporal shuffling usually destroys the motion continuity and is arguably not a valid resampling operation.

In this work, we consider the class of resampling via spatio-temporally warping as shown in Fig.1. Specifically, the n-th frame is shifted by $[(n-1)d_x, (n-1)d_y]$ (d_x, d_y are integers). Note that such warping is readily reversible since no interpolation is involved [6]. The impact of warping can be intuitively understood by referring to Fig. 1 - when the warping parameter matches the speed of a moving object, a slant trajectory could be transformed into a straight one (therefore better fit the STALL model with fixed $\mathcal{M}$ and $\mathcal{N}$). The warped video can also be viewed as the “new” sample acquired by a $virtual$ camera which records the same set of intensity values but in a different ordering. Note that the above resampling strategy does not affect the motion continuity and therefore the warped video is still perceptually meaningful (ignoring boundary artifacts).

Resampling produces a redundant or distributed representation for video signals. To manage the computational complexity, we make some assumption about the high-level knowledge about video such as camera motion type (e.g., panning vs. zoom), which is often available from video segmentation [8] or can be estimated from the phase-correlation function. Such high-level information is useful to the selection of resampling parameters. For instance, we can choose a 1D virtual camera array in the presence of horizonal camera panning and a 2D array in the case of camera zoom. Fig. 2 shows the temporal slices [8] of the original and resampled video for $garden$ - it can be clearly seen that the fast panning tree in the original sequence virtually moves slower and slower as the warping parameter d_y increases ($d_x = 0$).

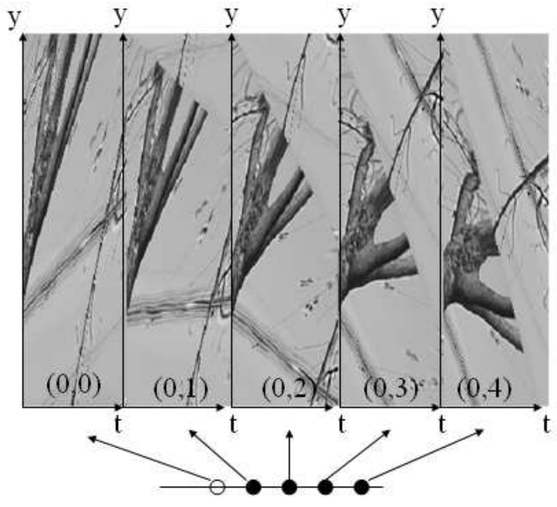

Fig. 2. Example of spatio-temporal resampling applied to $garden$ sequence containing horizontal camera panning.

Such observation with the relativity of motion is at the heart of our video modeling approach. It suggests an alternative approach of achieving adaptation by $soft$ fusion instead of $hard$ decision as in layered representation. Note that even when a cubic training window is used, ST adaptation can be achieved by resampling because the shape of training window varies from one resampled video to another. Apparently, we might choose the optimal training window (virtual camera) for a pixel, which assigns a deterministic label to each pixel (the layer index). Such strategy can be shown equivalent to a maximum-likelihood (ML) approach of making hard decisions for every pixel. Next, we will show how to softly combine the regression results from distributed virtual cameras under a Bayesian framework.

3. BAYESIAN FUSION OF LINEAR REGRESSION RESULTS

To simplify the notation, we drop 3D coordinated $\vec{n}$ from $X(\vec{n})$ and use $(X_1, X_2, ..., X_K)$ to denote the regression result by applying STALL to the K resampled sequences [9] respectively. Here virtual camera index $k = 1, ..., K$ is modeled by a discrete random variable reflecting our uncertainty about inferring $\hat{X}$ (clean and complete observation) from the k-th resampled sequence of X (noisy or incomplete observation). Using Bayesian modeling averaging technique [10], we can have

$$p(\hat{X}|X) = \sum_{k=1}^{K} p(\hat{X}|X, X_k)p(X_k|X), \tag{3}$$

Multiplying both sides by $\hat{X}$ and taking summations, we obtain the Bayesian LS estimation by

$$E[\hat{X}|X] = \sum_{k=1}^{K} \alpha_k X_k \tag{4}$$

where linear weight $\alpha_k = P(X_k|X)$ denotes the posterior model probability of inferring X from the k-th resampled sequence (note that $E[\hat{X}|X, X_k] = X_k$). According to Bayesian rule, the weighting coefficients can be calculated by

$$\alpha_k = P(X_k|X) = \frac{P(X|X_k)P(X_k)}{\sum_{i=1}^{K} P(X|X_i)P(X_i)} \tag{5}$$

where $P(X_k)$ is the prior probability and $P(X|X_k)$ is the likelihood function of observing X in the k-th resampled sequence. Based on STALL [4], we can model likelihood term $P(X|X_k)$ by a Gaussian probability function of regression error e_k

$$P(X|X_k) = P(e_k) \propto exp(\frac{-e_k^2}{2\sigma^2}) \tag{6}$$

where e_k is the regression error of the k-th camera as defined in Eq. (1) and σ^2 is a constant determined by heuristics (we use $\sigma^2 = 500$). If we use a uniform prior for convenience, Eq. (5) can be simplified into

$$\alpha_k = P(X_k|X) = \frac{exp(\frac{-e_k^2}{2\sigma^2})}{\sum_{i=1}^{K} exp(\frac{-e_i^2}{2\sigma^2})} \tag{7}$$

It is easy to see that a smaller regression error leads to a larger weighting coefficient in the Bayesian fusion model Eq.(5), which matches our intuition that α_k should reflect the confidence about the inference result from the k-th resampled sequence. Intuitively, as long as the array of virtual cameras is sufficiently large, any segment of a slant motion trajectory is likely to be warped to the straight position (aligned with the cubic training window) in some resampled sequence. Upon the alignment, localized regression by STALL will produce the smallest errors and therefore make the largest contribution during Bayesian fusion. When compared with hard-decision based layer representations, our distributed model systematically pools together the inference results from the virtual camera array and avoids the penalty of the uncertainty with any suboptimal labeling process.

4. EXPERIMENTAL RESULTS

In this section, We show that the modeling capability of STALL model can be dramatically improved by spatio-temporal resampling and Baysian fusion especially for complex video sequences containing camera motion. Four test sequences are used in our experiments: two containing fast camera panning (SIF-*garden* and CIF-*bus*) and two containing camera zoom (SIF-*tennis* and CIF-*mobile*). For camera panning sequences, we use a 1 × 9 horizontal virtual camera array ($d_x = 0$); for camera zoom sequences, we use a 3 × 3 virtual camera array. Due to space limitation, we will only report our experimental results in two applications related to multi-frame restoration here - error concealment and impulse noise removal. The MATLAB demo program can be accessed at `http://www.csee.wvu.edu/~xinl/demo/MALSTAR.html`.

•Video Error Concealment
Error concealment refers to the problem of repairing damaged blocks in block-based video communication systems. The block loss is assumed to occur at the same spatial location but for three consecutive frames (3rd-5th frames). Such consecutive block loss is particularly challenging for video containing camera motion because content contained in the damaged blocks varies from frame to frame. After specifying an appropriately chosen scanning order (3D extension of the rules suggested by [11]), we can sequentially recover the missing data by LS-based ST interpolation. The same model support and training window parameters of STALL model as [12] are used. For CIF-$mobile$ sequence, a 2D 3 × 3 camera array is used; for SIF-$garden$ sequence, the 9-point camera array is chosen to be $(0, -2), ..., (0, 6)$ (camera panning direction is towards the right in this sequence).

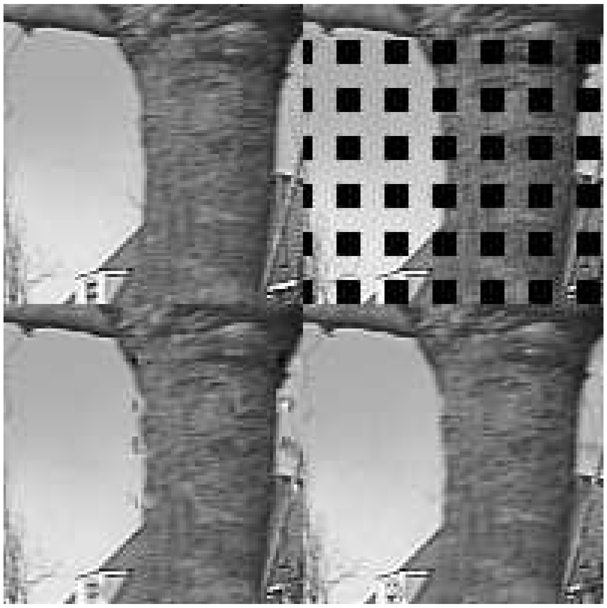

Fig. 3. Top-left: original; top-right: damaged; bottom-left: concealed result without fusion (PSNR= $27.32dB$); bottom-right: concealed result with fusion (PSNR= $30.51dB$)

Table 1 shows the comparison results of average PSNR across three frames without and with Bayesian fusion. Dramatic gain can be observed especially for the sequences with fast camera panning. Figs. 3 and 4 include the comparison among 100 × 100 portions of the original, damaged, concealed $3rd$ frames by STALL of $garden$ and $mobile$ sequences without and with Bayesian fusion. The improvements on visual quality are also convincing - e.g., most artifacts around occluded areas are suppressed after Bayesian fusion. This is because larger weights are assigned to virtual cameras with smaller regression errors. Again we note that such ST adaptation is achieved without any explicit segmentation.

	Error Concealment		Denoising (10% impulse)	
	w/o	w	w/o	w
Garden	25.93	31.87	32.14	35.54
Bus	28.34	34.20	33.19	36.55
Mobile	27.77	31.64	33.52	37.13
Tennis	32.54	34.45	37.90	39.93

Table 1. PSNR (dB) performance comparison between STALL-based error concealment and impulse noise removal algorithms (w/o - without fusion, w - with fusion).

•Video Impulse noise removal
We consider the impulse removal problem where video data is contaminated by random-valued impulses uniformly distributed between [0,255]. Similar to [3], we assume the known noisy pixel location

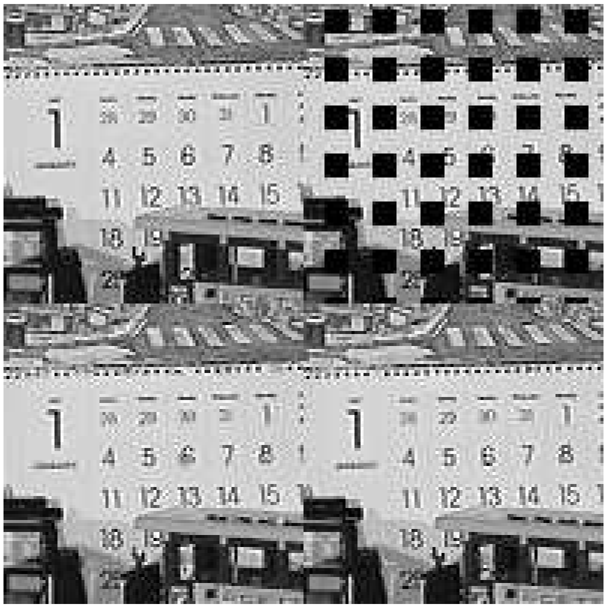

Fig. 4. Top-left: original; top-right: damaged; bottom-left: concealed result without fusion (PSNR= $29.07dB$); bottom-right: concealed result with fusion (PSNR= $34.10dB$)

and focus our comparison on different filtering strategies. Note that explicit ME is difficult for noisy video especially when the amount of impulse noise is high (to the best of our knowledge, most existing works on impulse noise removal deal with still images instead of video). Preliminary denoising results without Bayesian fusion have been reported in [4]. Table 1 also includes the PSNR performance comparison for impulse noise removal without and with Bayesian fusion. Again we have observed dramatic improvement ($> 2dB$) brought by the proposed fusion scheme. Fig. 5 contains the subjective quality comparison of denoising results for the *bus* sequence. We have also found our approach significantly outperforms epitome-based approach [3] on both subjective and objective qualities.

5. CONCLUSION

In this paper, we further improve the modeling capability of STALL model by an empirical Bayes approach based on ST resampling and probabilistic fusion. ST resampling is implemented by virtual cameras which warp a video sequence in a reversible fashion. Bayesian fusion probabilistically pool together the statistical inference results from resampled video signals (distributed virtual cameras). When combined with our previous STALL model, the empirical Bayes approach offers a new framework for distributed processing of video signals without explicit motion estimation or segmentation. Highly encouraging experimental results with multiframe image restoration have been reported to support the effectiveness of this new model. Our model also has potential applications into deinterlacing, temporal interpolation and super-resolution.

6. REFERENCES

[1] E. Simoncelli, *Distributed Representation and Analysis of Visual Motion*, Ph.D. thesis, MIT, Jan 1993.

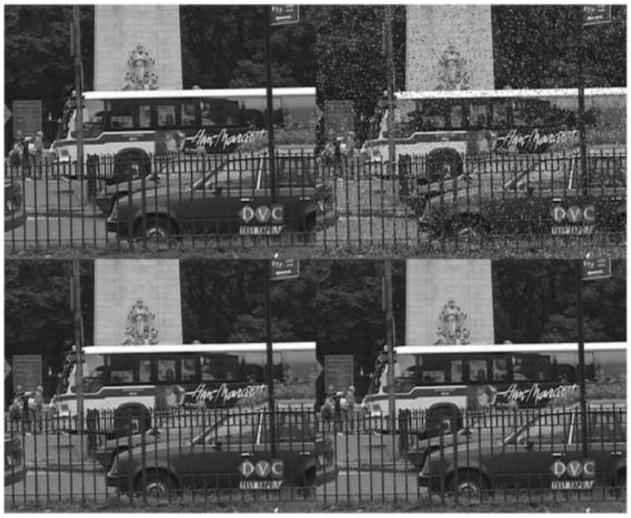

Fig. 5. Top-left: original frame; Top-right: noisy frame (10% impulse noise); Bottom-left: denoised result by STALL without fusion (PSNR= $31.54dB$); Bottom-right: denoised result by STALL with fusion (PSNR= $35.01dB$)

[2] Y. Wexler, E. Shechtman, and M. Irani, “Space-Time Video Completion,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[3] V. Cheung, B. J. Frey, and N. Jojic, “Video epitomes,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.

[4] Yunfei Zheng and Xin Li, “Video modeling via spatio-temporal adaptive localized learning (stall),” in *40th Asilomar Conference on Signals, Systems, and Computers*, 2006.

[5] J.Y.A. Wang and E.H. Adelson, “Representing moving images with layers,” *IEEE Transactions on Image Proc.*, vol. 3, no. 5, pp. 625–638, September 1994.

[6] D. Taubman and A. Zakhor, “Multi-rate 3d subband coding of video,” *IEEE Transactions on Image Processing*, vol. 3, pp. 572–588, 1994.

[7] M. Szummer and R.W. Picard, “Temporal texture modeling,” in *Proc. of Int. Conf. on Image Proc.*, 1996, pp. 65–70.

[8] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang, “Motion analysis and segmentation through spatio-temporal slices processing,” *IEEE Transactions on Image Processing*, vol. 12, pp. 341 – 355, 2003.

[9] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, 1994.

[10] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky, “Bayesian Model Averaging: A Tutorial,” *Statistical Science*, vol. 14, pp. 382–401, 1999.

[11] A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, September 2004.

[12] Y. Zheng, X. Li, and C. Dai, “Video error concealment based on implicit motion models,” in *SPIE Conf. on Multimedia Systems and Applications VIII*, 2005.