Spatio-Temporal Markov Random Field for Video Denoising *

Jia Chen Chi-Keung Tang

Vision and Graphics Group
The Hong Kong University of Science and Technology
{jiachen,cktang}@cse.ust.hk

Abstract

This paper presents a novel spatio-temporal Markov random field (MRF) for video denoising. Two main issues are addressed in this paper, namely, the estimation of noise model and the proper use of motion estimation in the denoising process. Unlike previous algorithms which estimate the level of noise, our method learns the full noise distribution nonparametrically which serves as the likelihood model in the MRF. Instead of using deterministic motion estimation to align pixels, we set up a temporal likelihood by combining a probabilistic motion field with the learned noise model. The prior of this MRF is modeled by piece-wise smoothness. The main advantage of the proposed spatio-temporal MRF is that it integrates spatial and temporal information adaptively into a statistical inference framework, where the posteriori is optimized using graph cuts with alpha expansion. We demonstrate the performance of the proposed approach on benchmark data sets and real videos to show the advantages of our algorithm compared with previous single frame and multi-frame algorithms.

1. Introduction

Many dynamic vision algorithms are sensitive to noise, such as corner detection and feature tracking. Video denoising can substantially improve the results of these algorithms. It is also desirable to reduce noise caused during capturing and transmission of video in order to provide better video quality for viewing or manipulation. A number of previous video denoising methods are directly extended from image denoising, such as filtering [2, 7], wavelet shrinkage [21, 24] and PDE [15] based methods. However, treating video as an isotropic 3D volume may introduce motion blur and artifact because the spatial dimension has different resolution and continuity properties with the temporal dimension. Several methods [8, 30] use a motion compensation stage before spatial filtering. An adaptive spatio-temporal bilateral filtering with motion compensation is used in [1]. A Bayesian approach is presented in [13] to restore old film using MCMC.

At a first glance, video denoising could be much easier than the static image case because the temporal dimension provides much richer information. If the trajectory of one pixel is accurately known, the average of n samples along this trajectory provides a maximum likelihood estimator which reduces noise level by a factor of $\sqrt{n}$. Unfortunately, motion estimation can be inaccurate from a noisy sequence. On the other hand, long range motion cannot be estimated for all pixels. Another uncertainty is the level of noise which is usually needed to be known such that parameters can be tuned accordingly [19, 20]. Noise estimation algorithms have been proposed to estimate the *level* of noise from a single image [16] or image sequence [31]. All these methods are still limited to estimate an upper bound of noise or estimate parametric (Gaussian) noise.

In this paper, we address the video denoising problem as visual inference where three uncertainties: motion, noise and smoothness are integrated together using an MRF approach. Previous MRF based methods have demonstrated decent results in static image restoration [9, 14, 22], but no attempt has been made to video by considering temporal coherence. The main difficulty for video is how to construct the MRF in the spatio-temporal domain [27] while avoiding the complexity to grow exponentially.

The main theoretical contribution of this paper is a spatio-temporal Markov random field along with the well designed algorithms for its construction. Two main issues are addressed in this paper: the estimation of noise model and the proper use of motion estimation.

Different from previous methods which mainly estimate the noise level, we present an effective method to learn the non-parametric noise distribution from video sequence, which does not rely on a parametric representation of noise

*This research is supported by the Research Grant Council of Hong Kong Special Administration Region, China: HKUST620006.

1-4244-1180-7/07/$25.00 ©2007 IEEE

and is capable of coping with time-varying noise distribution. The learned noise model will serve as the likelihood model in the proposed MRF. In video denoising, temporal coherence is important. Here in our algorithm, the key of modeling temporal likelihood in this MRF relies on the use of a probabilistic motion field [26], where not only a dense motion field is provided, but also a probabilistic distribution is build to describe the motion of each pixel. It explicitly accounts for inaccuracy and ambiguity in motion estimation and provides an adaptive temporal neighborhood.

In the remainder of this paper, we will first formulate the spatio-temporal MRF in section 2. We then show how the noise model is automatically learned in section 3. The temporal likelihood is described in section 4. We provide experimental results in section 5 and conclude the paper in section 6.

2. Bayesian Spatio-Temporal Restoration

We treat denoising as a Bayesian inference problem: infer clean frames $\{\mathbf{x}^t\}$, from noisy frames $\{\mathbf{y}^t\}$, where $t = 1, ..., n$. One may choose to write the full posterior as

$$\begin{aligned} &P(\mathbf{x}^1, ..., \mathbf{x}^n | \mathbf{y}^1, ..., \mathbf{y}^n\} \\ \propto &P(\mathbf{y}^1, ..., \mathbf{y}^n | \mathbf{x}^1, ..., \mathbf{x}^n) P(\mathbf{x}^1, ..., \mathbf{x}^n) \end{aligned} \tag{1}$$

using the Bayes' law, where $P(\mathbf{x}^1, ..., \mathbf{x}^n)$ models both spatial and temporal smoothness prior. However this direct approach will result in a very large MRF which is intractable. Instead of the full Bayesian method, we use a factorized Bayesian approach: the posterior of individual frame is written as $P(\mathbf{x}^t | \mathbf{y}^{t-m}, ..., \mathbf{y}^t, ..., \mathbf{y}^{t+m})$ by assuming m-th order dependence on the bidirectional temporal neighborhoods. In this way, the posterior defined on each frame t can be written as:

$$P(\mathbf{x}^t | \mathbf{y}^{t-m}, ..., \mathbf{y}^t, ..., \mathbf{y}^{t+m}) \propto P(\mathbf{x}^t) \prod_{k=-m}^{m} P(\mathbf{y}^{t+k} | \mathbf{x}^t) \tag{2}$$

We explain equation 2 in more detail as follows:

- *Likelihood* - $P(\mathbf{y}^t | \mathbf{x}^t)$ provides data evidence from the observed frame at time t. This likelihood is essentially defined by a noise model. A complete likelihood model requires the estimation of noise distribution rather than a level of noise. We will show how to learn a noise distribution from the input video. Our noise model can also been adaptively updated along time to reflect non-stationary noise distribution.

- *Temporal likelihood* - $P(\mathbf{y}^{t+k} | \mathbf{x}^t)|_{k \neq 0}$ provides data evidence from temporal neighborhoods for frame t. Modeling these temporal likelihoods is not a trivial task because the observation $\mathbf{y}^{t+k}|_{k \neq 0}$ is not spatially aligned with the hidden variable $\mathbf{x}^t$ so we cannot directly use the noise model as the temporal likelihood. Obviously temporal likelihood is determined by both the motion and the noise model so we will introduce a probabilistic motion field together with the noise model to formulate the temporal likelihood in our MRF.

- *Prior* - $P(\mathbf{x}^t)$ models intra frame smoothness which follows the widely used piecewise smooth prior for natural images.

Compared with the full posterior model, this factorized Bayesian formulation shifts the modeling of temporal smoothness from prior $P(\mathbf{x}^1, ..., \mathbf{x}^n)$ term to temporal likelihood $P(\mathbf{y}^{t+k} | \mathbf{x}^t)|_{k \neq 0}$ term. The original problem of inferring all frames at once is transformed to inferring individual frames and thus the problem size is greatly reduced.

In later sections, we will describe the detailed modeling of the noise likelihood, temporal likelihood and graphical representation of the spatio-temporal MRF.

3. Noise Model Estimation

We assume the noise is additive:

$$y = x + n. \tag{3}$$

Our goal is to learn the distribution of the noise $n \sim P(n)$ at time t. We do not assume an actual parametric form of the noise distribution such as Gaussian; instead, we use a non-parametric approach to learn and represent the noise distribution which can account for a large variety of distributions.

Noise estimation from a single image is difficult because the clean image $\mathbf{x}$ is unknown. Noise level can be estimated by analyzing the gradients of image corrupted by Gaussian noise [31], or by making use of the piece-wise smoothness assumption and noise prior [16]. [18] uses static regions from tens of film frames to estimate the clean frame and a parametric noise distribution.

We estimate noise distribution from multiple video frames by using optical flow as the first step. We use the 2D CLG method described in [6] which has already shown satisfactory accuracy and robustness. The energy minimization formulation of optical flow is:

$$E(\mathbf{w}) = \int (\mathbf{w}^T J_\rho(\nabla_3 f)\mathbf{w} + \alpha \left|\nabla \mathbf{w}\right|^2) dxdy, \tag{4}$$

where $\mathbf{w} = (u, v, 1)^T, \nabla_3 f = (f_x, f_y, f_t)^T$, and $J_\rho(\nabla_3 f) = K_\rho * (\nabla_3 f \nabla_3 f^T)$. The purpose of the structure tensor $J_\rho(\nabla_3 f)$ is to diffuse the spatial and temporal derivatives at a spatial scale of ρ which makes the optical flow more robust under noise.

We assume the noise is stationary, *i.e.*, the noise distribution does not vary, within a small period of time, and spatially ergodic, *i.e.*, the noise distribution at different locations is the same. The first assumption enables us to compute the value of noise from motion estimation; the second assumption enables us to collect sufficient number of noise samples. Noise model estimation is performed frame by frame. At frame t, a number of sites $\Omega = \{\mathbf{p}_i\}$ need to be selected and their temporal correspondences will be found in frame $t-m, \cdots, t, \cdots, t+m$ by optical flow. There are three criteria for choosing one good location $\mathbf{p}$:

1. Motion estimation at $\mathbf{p}$ should be as accurate as possible. We use the energy computed by $\mathbf{w} = (u, v, 1)^T, \nabla_3 f = (f_x, f_y, f_t)^T$ as confidence measurement of motion accuracy [6]. The higher the energy is, the lower the confidence is.

2. Estimated motion vector should be close to integer *i.e.*, one location with motion $(u, v) = (0.9, 0.9)$ where current pixel will almost align to its temporal correspondent is preferable to that another with $(u, v) = (0.6, 0.6)$ where current pixel will only partially align to another pixel.

3. Pixel $\mathbf{p}$ should locate at low gradient region thus inaccurate motion will not cause significant color difference between $\mathbf{p}$ and its temporal correspondences. A derivative of Gaussian filter ($\sigma = 1.5$) is applied to compute the spatial gradient. The confidence is low when the gradient is large.

Both conditions 1 and 2 apply to all $2m$ temporal neighborhoods of frame t when choosing pixel $\mathbf{p}$. Once we have collected a stack of correspondences $\{y^{t-m}, \ldots, y^t, \ldots, y^{t+m}\}$ from motion estimation for site $\mathbf{p}$, the mean value is computed by $\bar{y} = \frac{1}{2m+1} \sum_{k=t-m}^{t+m} y^k$. We can then calculate noise samples $n_i = y^{i+t-m} - \bar{y}$, $i = 0, \cdots, t-1, t+1, \cdots, 2m$ and add them into the noise set Θ.

We use the Parzen-window [10] approach to derive a nonparametric representation of the noise distribution from $\Theta = \{n_j\}$:

$$p(n) = \frac{1}{|\Theta|} \sum_{j=1}^{|\Theta|} W(n - n_j), \tag{5}$$

where the kernel function W is taken to be a Gaussian function $W(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$.

The effectiveness of this nonparametric approach is evaluated by applying synthetic noise onto a motion sequence. Gaussian noise $n_g \sim N(0, 400)$ and uniform noise $n_u \sim U(-40, 40)$ are applied to the *Suzie* sequence respectively. We use second order temporal neighborhood ($m = 2$) in finding temporal correspondences. In Fig. 1, the confidence map of motion and gradient are shown together with the noisy frames. Noise samples are drawn from the *good* regions defined by the criteria listed above. The learned distributions are compared to the ground-truth and are also shown.

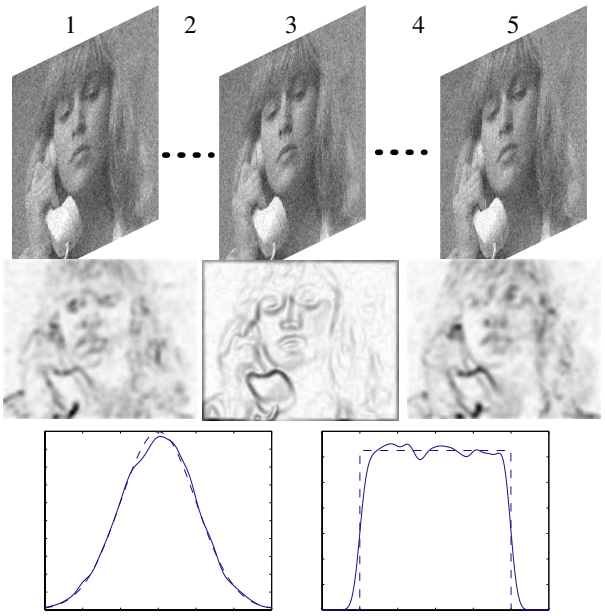


Figure 1. Noise estimation from video. Top row: three noisy frames from the Suzie sequence. Middle row: motion confidence from t to 1, gradient confidence of 3, motion confidence from 3 to 5. Bottom row: estimation of synthetic Gaussian noise and uniform noise. Solid line shows the estimated noise distribution and dashed line shows the ground-truth distribution.

4. Temporal Likelihood From Probabilistic Motion Field

A key issue in video denoising is how to exploit the temporal redundancy from motion estimation. In an ideal case, if long range [23], pixel-to-pixel trajectory is available for all pixels, per-trajectory average will produce a good estimation of clean pixel values. Unfortunately this is not true in practice because long range correspondence cannot be generated exactly at the pixel level - it can be either long range correspondence at a super-pixel level, or short range correspondence at a sub-pixel level. Besides, in continuous motion, one pixel is likely to match to part of another pixel in the next frame, or in other words, one pixel disassembles to several pixels. Fig. 2 compares the continuous physical motion, optical flow motion estimation and the probabilistic motion estimation which has a spatial uncertainty measurement. It has been well known that spatial and temporal information need to be adaptively integrated for effective noise removal in video [1], a good temporal neighborhood

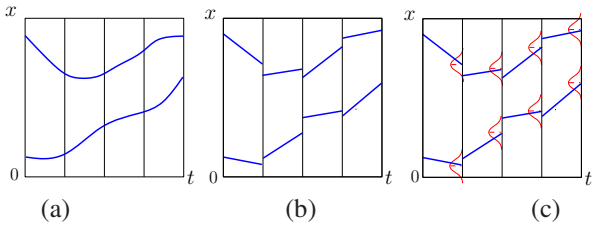


Figure 2. (a) Ideal long range trajectory. (b) Sub-pixel optical flow starts from pixel center and ends at a subpixel coordinate. (c) Motion with uncertainty starts from pixel center and ends at a subpixel coordinate with probability.

is essential in exploiting the temporal redundancy. In this section, we will show how to define and utilize the temporal neighborhood through a probabilistic motion field and how to use it to model the temporal likelihood.

4.1. Probabilistic Motion Field

Motion estimation is usually performed using deterministic methods, such as optical flow and correlation, which typically produce trajectories shown in Fig. 2 (b) where every pixel $\mathbf{p} = (x, y)$ is provided with a definite motion vector $\mathbf{v} = (u, v)$ [11, 17]. Recent development of optical flow techniques has achieved high accuracy even under certain of noise level [4, 5, 6].

We use optical flow as the base-line motion estimator, and propose to use a probabilistic motion field to model and utilize the motion uncertainty. By replacing the motion vector $\mathbf{v}$ of pixel $\mathbf{p}$ with a probabilistic density function $p(\mathbf{v})$, it means that one pixel $\mathbf{p}$ now moves to $\mathbf{p}+\mathbf{v}$ with probability $P(\mathbf{v})$.

Let us revisit the fundamental ambiguity of optical flow, the aperture phenomenon. Because there are two unknowns in the optical constancy equation: $I_x u + I_y v + I_t = 0$, there is one directional ambiguity perpendicular to image gradient. While optical flow algorithms have always been trying to resolve the ambiguity by applying local [17] or global [6, 11] constraints, we explicitly make use of this ambiguity to set up the temporal likelihood. Comparing to the deterministic method which commits to hard decisions on pixel motion, the probabilistic method is *soft* because it introduces a good spatial support, providing more robustness against noise and capability for a more general probabilistic framework for video denoising.

We repeat the Lucas-Kanade algorithm here by assuming a s by s neighborhood which have consistent motion, thus s^2 equations are obtained at pixel $\mathbf{p}$:

$$I_x^{\mathbf{q}} u + I_y^{\mathbf{q}} v + I_t^{\mathbf{q}} = 0, \mathbf{q} = 1, 2..., s^2 \tag{6}$$

where $\mathbf{q}$ is $\mathbf{p}$'s neighborhood. We add an uncertainty term ϵ in the optical flow constraint equation and we get a new set of equations:

$$I_x^{\mathbf{q}} u + I_y^{\mathbf{q}} v + I_t^{\mathbf{q}} = \epsilon_{\mathbf{q}}, \mathbf{q} = 1, 2..., s^2,$$

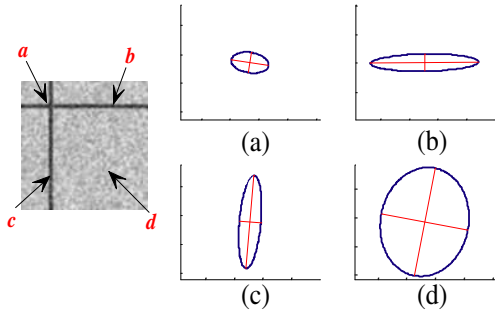


Figure 3. (a)–(d) represent the respective 2D Gaussian of the four locations a-d shown in the left image.

where $\epsilon_{\mathbf{q}}$ is assumed to be *i.i.d* Gaussian $\epsilon_{\mathbf{q}} \sim N(0, \sigma^2)$.

We use a matrix representation of these linear equations:

$$\mathbf{A} \begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{b}, \text{ where } \begin{cases} \mathbf{A} = \begin{bmatrix} I_x^1 & \cdots & I_x^s \\ I_y^1 & \cdots & I_y^s \end{bmatrix}^T \\ \mathbf{b} = \begin{bmatrix} \epsilon_1 - I_t^1, \cdots, \epsilon_s - I_t^s \end{bmatrix}^T \end{cases}, \tag{7}$$

and the probabilistic motion will follow a 2D Gaussian distribution:

$$\mathbf{v}^{\mathbf{p}} \sim N(A^+ E(\mathbf{b}), A^+ \Sigma A^{+T}) \tag{8}$$

where $\Sigma = diag\{\sigma^2, ..., \sigma^2\}$ and A^+, A^{+T} denote respectively the pseudo inverse of A and A^T. We can see that the distribution is Gaussian and the Lucas-Kanade approach produces a maximum likelihood estimation from this motion field.

An interesting property of this motion field is that the distribution is solely determined by the texture of the neighborhood around pixel $\mathbf{p}$. If $\mathbf{p}$ locates at an edge, the motion will have higher ambiguity along the edge but lower ambiguity perpendicular to the edge. For highly textured or corner pixels, the motion ambiguity is very low. This is intuitively reasonable and it is consistent with research on solvability issues in optical flow [25]. The motion ambiguity is well measured by the covariance matrix of the probabilistic motion field. If the motion is not very ambiguous, the spatial support is small, otherwise it is large. The shape of spatial support is anisotropic, which is determined by the shape of Gaussian distribution. Fig. 3 shows a synthetic image in which small planar motion, where different locations have different motion ambiguities. The observation is consistent with our analysis.

In our implementation, we take deterministic motion estimation approach [6] to estimate $\mathbf{v}^{\mathbf{p}}$, and associate it with the covariance matrix in Eqn. 8. The noise term ϵ is positively correlated to the noise level of the video. We set $\sigma = 0.3\sigma_n$ where σ_n is the standard deviation of the estimated noise in Section 3.

In order to reject outliers such as occlusion or regions with low motion confidence, we define an outlier rejection switch as in [28] based on the motion energy in Eqn. 4:

$$\rho(\mathbf{v}) = \begin{cases} \varepsilon, \text{ if } \mathbf{w}^T J_\rho(\nabla_3 f)\mathbf{w} > T \\ 1, \text{ otherwise} \end{cases}, \tag{9}$$

where T is a large threshold and ε is a small constant which are empirically chosen.

The motion distribution is finally given by:

$$p(\mathbf{v}) = \frac{1}{C} N(\mathbf{v}, A^{+T}\Sigma A^{+})\rho(\mathbf{v}) \tag{10}$$

where $C = \int N(\mathbf{v}, A^{+T}\Sigma A^{+})\rho(\mathbf{v})d\mathbf{v}$ is a normalization factor.

4.2. Temporal Likelihood

As we have mentioned earlier, noise model cannot be directly applied to the temporal likelihood without integration with motion estimation. We use $V^{t,t+k}$ to denote the deterministic motion field of the full image between t and $t+k$, followed by the distribution $p(V^{t,t+k})$. The temporal likelihood is a marginalization over the probabilistic motion field:

$$P(\mathbf{y}^{t+k}|\mathbf{x}^t) = \int P(\mathbf{y}^{t+k}|\mathbf{x}^t, V^{t,t+k})P(V^{t,t+k})dV^{t,t+k}. \tag{11}$$

The probabilistic motion field allows all pixels to contribute to the temporal likelihood statistically.

Because $V^{t,t+k}$ is defined over all pixel location pairs $\{(\mathbf{p},\mathbf{q})\}$, it is of very high dimension and Eqn. 11 needs to be simplified. Noticing that the motion field has local property, $i.e.$ the motion field of one pixel is a Gaussian of several pixel width for a majority of pixel pairs $(\mathbf{p},\mathbf{q})$, $P(V^{t,t+k}_{\mathbf{p}\to\mathbf{q}}) \approx 0$. We define the *temporal neighborhood* as:

$$E_t(\mathbf{p}) = \{\mathbf{q}|P(V^{t,t+k}_{\mathbf{p}\to\mathbf{q}}) > \epsilon\},$$

where ϵ is a small constant $(1e{-}2)$. The temporal likelihood of Eqn. 11 at pixel $\mathbf{p}$ is discretized as:

$$P(y^{t+k}|\mathbf{x}^t_{\mathbf{p}}) = \sum_{\mathbf{q}\in E_t(\mathbf{p})} P(y^{t+k}_{\mathbf{q}}|x^t_{\mathbf{p}}, V^{t,t+k}_{\mathbf{p}\to\mathbf{q}})P(V^{t,t+k}_{\mathbf{p}\to\mathbf{q}}) \tag{12}$$

where $P(y^{t+k}_{\mathbf{q}}|x^t_{\mathbf{p}}, V^{t,t+k}_{\mathbf{p}\to\mathbf{q}}) = P(y^{t+k}_{\mathbf{q}}|x^{t+k}_{\mathbf{q}})$ is exactly the noise-based likelihood estimated in Section 3.

The modeling of temporal likelihood by using the probabilistic motion field has a number of benefits:

1. The temporal neighborhood is adaptive both in size and shape represented by the probabilistic motion field. The adaptivity relies on local color and texture information, which is superior to using a fixed temporal neighborhood.

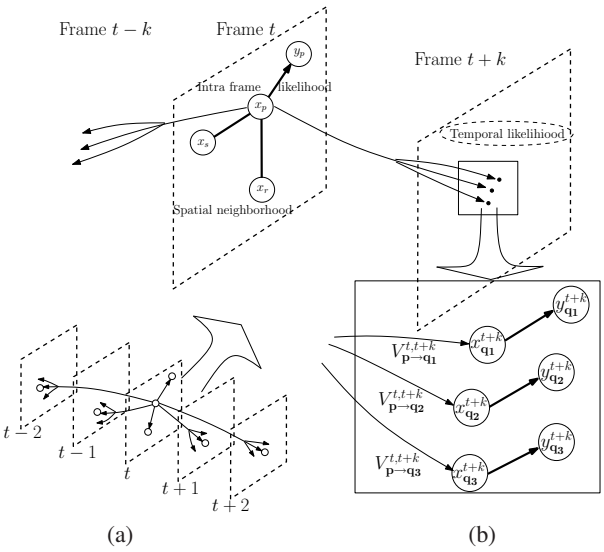


Figure 4. The graphical representation of the spatio-temporal MRF. (a) An overview of this MRF. (b) Closed view of temporal likelihood.

2. The temporal likelihood balances well with the intra-frame likelihood. If the motion estimation is accurate, the temporal likelihood will be higher thus more temporal information will be used in inference. Otherwise more intra frame smoothness will be used.

4.3. Graphical Model for the Spatio-Temporal MRF

The graphical model for the spatio-temporal MRF to denoise frame t is illustrated in Fig. 4. We usually use $m = 3$ in this MRF. Only parts of the spatial and temporal neighborhoods are shown in this figure. In this MRF, the data energy is defined to model the intra frame likelihood and temporal likelihood:

$$D(x^t_{\mathbf{p}}) = -\log\{\sum_{\substack{k=-m \\ k\neq 0}}^{m} \sum_{\mathbf{q}\in E_t(\mathbf{p})} P(y^{t+k}_{\mathbf{q}}|x^t_{\mathbf{p}}, V^{t,t+k}_{\mathbf{p}\to\mathbf{q}})P(V^{t,t+k}_{\mathbf{p}\to\mathbf{q}})\} - \log\{P(y^t_{\mathbf{p}}|x^t_{\mathbf{p}})\}. \tag{13}$$

A smoothness energy $S(x_{\mathbf{p}}, x_{\mathbf{q}})$ for spatial neighbors is used to model the piece-wise smooth prior with discontinuity preservation over 5 by 5 neighborhood $(\mathbf{p},\mathbf{q}) \in E_s$:

$$S(x_{\mathbf{p}}, x_{\mathbf{q}}) = \min\{|x_{\mathbf{p}} - x_{\mathbf{q}}|, \lambda\}. \tag{14}$$

The parameter λ is empirically set between 30 to 60 in our experiments which explicitly influences the level of contrast to be preserved. Finally, the full energy of this MRF is:

$$E = \sum_{\mathbf{p}} D(x^t_{\mathbf{p}}) + \alpha \sum_{(\mathbf{p},\mathbf{q})\in E_s} S(x_{\mathbf{p}}, x_{\mathbf{q}}). \tag{15}$$

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

Figure 5. (a) Original frame No. 83 from the Suzie sequence. (b) Noisy frame, 28.0dB. (c) Denoised result cropped from [8], 34.6dB. (d) Denoised by GSM algorithm [20], 35.04dB. (e) Denoised by the proposed approach, 36.9dB.

This spatio-temporal MRF is solved by graph cuts algorithm [3] with alpha expansion. The graph cuts is initialized to noisy observation. The running time for one 176 by 144 frame is approximately 50s on a P4 3GHz desktop PC.

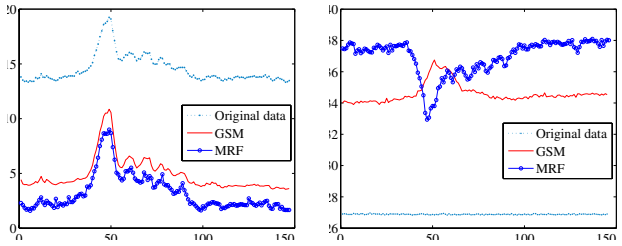
5. Experiments

We have conducted experiments on both synthetic and real noisy video to test the proposed algorithm. First we compare our method with other recent single frame and mult-frame methods on denoising sequences corrupted by Gaussian noise. The test sequences used are downloaded from [12]. Two levels of Gaussian noise are added: $\sigma = 10$ or PSNR of 28, and $\sigma = 16$ or PSNR of 24. The average PSNRs of denoised sequences are listed in Table 1. Our method consistently outperforms previous video denoising methods and single frame denoising methods. The visual quality can be evaluated in Fig. 5.

Data	σ/PSNR	(a)	(b)	(c)	(d)	(e)
Salesman	10/28	-	32.5	32.7	35.1	35.2
	16/24	-	-	30.0	32.6	33.7
Suzie	10/28	34.8	-	35.2	37.1	37.1
	16/24	32.0	-	33.0	35.1	35.4
Trevor	10/28	33.9	34.1	35.1	36.7	36.9
	16/24	31.3	-	32.8	34.8	35.1
Foreman	10/28	33.9	-	33.4	34.9	35.3
	16/24	31.1	-	30.7	32.9	33.4

Table 1. Peak signal-to-noise ratio (PSNR) in dB for benchmark sequences. (a) Joint Kalman and Wiener denoising with motion compensation [8]. (b) Adaptive K-NN space-time filter [29]. (c) Gaussian scale mixture method [20]. (d) Space-time patch based method [2]. (e) Proposed method.

Although in image processing, MRF based method is not always preferable to wavelet based methods, we can still see that in video denoising, the spatio-temporal MRF approach produces very competitive results by a providing a better interpretation of motion estimation and statistical integration of motion with noise model and smoothness. From the benchmark results, our proposed approach gives better PSNR than the state-of-the-art single image denoising algorithm, the Gaussian scale mixture (GSM) algorithm introduced by [20].



(a) (b)

Figure 6. Evaluation on Suzie data with uniform noise $U(-20, 20)$. (a) Mean absolute difference error between consecutive frames of the noisy and denoised sequence.(b) PSNR of noisy and denoised sequence.

We have also conducted experiments on uniform noise applied to the Suzie sequence. First, we evaluate the temporal coherence by computing the mean absolute difference between adjacent frames. Although motion also introduces intensity difference between frames, it is a common factor in the comparison. We can see from Fig. 6(a) that our algorithm have much better temporal consistence than the Gaussian scale mixture approach. We further compared the per-frame PSNR which shows both the advantage and disadvantage of our approach. As shown in Fig. 6(b), the proposed approach has much better overall PSNR than the than that of the GSM method. However, we can see that GSM method outperforms our algorithm between frame No. 40 to No. 60. The reason for this is when the motion is very large, there is much less temporal information that can be

(a) (b) (c)

Figure 7. (a) Top, noisy video frame. Bottom, denoised frame using the proposed approach. (b) Closed-up view of the noisy frame. (c) Closed-up view of proposed approach.

(a) (b) (c)

Figure 8. (a) Top, video frame output by the ASTA algorithm. Bottom, denoised frame using the proposed approach. (b) Closed-up view of ASTA output. (c) Closed-up view of proposed approach. (Please view this figure in color.)

used by our algorithm. On the contrary, because motion blur gives more smooth regions, GSM method works even better in this situation than it usually does.

We apply the proposed MRF to denoise a clip of an old film, where the noise distribution is unknown. One of the original noisy frame and corresponding denoised frame are shown in Fig. 7 which shows the effectiveness of our algorithm. We also compare our approach with the recent ASTA algorithm [1] although it is not a straight forward comparison. Because the ASTA aims to generate virtual exposure of video and process it by tone mapping operators which is unknown to us, we perform denoising on the output of the ASTA. The color frame is processed in RGB channels separately and the results are combined together. The comparison in Fig. 8 shows that our algorithm further reduces the amount of noise in this video.

6. Conclusions

To the best of our knowledge, we are the first to perform video denoising by exploiting Markov random field jointly in spatial and temporal domain. This MRF is available only when the probabilistic motion field is modeled and the noise model is learned. We have shown in this paper how to learn the noise model and probabilistic motion field to construct the spatio-temporal MRF.

The proposed probabilistic motion field has the advantage of automatically adapting spatial support in the temporal neighborhoods. Besides denoising, we predict our algorithm has more applications in video restoration problems.

The learning of the noise model is another contribution of our paper. Our method is capable of handling not only Gaussian noise but also other additive noise. The noise model is updated online in our algorithm, so our approach can also handle time-varying noise.

There are still several limitations of our approach: when the motion is large, the temporal likelihood becomes weak, we will investigate the use of stronger image prior model such as [22] to compensate this case; the noise is assumed to be additive in the proposed noise modeling algorithm, we will try to extend our algorithm to handle the non-additive case in future work.

References

[1] E. P. Bennett and L. McMillan. Video enhancement using per-pixel virtual exposures. In *ACM SIGGRAPH*, 2005.

[2] J. Boulanger, C. Kervrann, and P. Bouthemy. Space-time adaptation for patch-based image sequence restoration. Technical report, 2006.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[4] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. In *International Journal of Computer Vision*, volume 67, pages 141–158, April 2006.

[5] T. Brox, A. Bruhn, and J. Weickert. Variational motion segmentation with level sets. In *European Conference on Computer Vision (ECCV)*, 2006.

[6] A. Bruhn, J. Weickert, and C. Schnorr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 2005.

[7] A. Buades, B. Coll, and J.-M. Morel. Denoising image sequences does not require motion estimation. Technical report, Preprint CMLA 2005-18, 2005.

[8] R. Dugad and N. Ahuja. Video denoising by combining kalman and wiener estimates. In *IEEE International Conference on Image Processing*, 1999.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–261–I–268 Vol.1, 2004.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2001.

[11] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[12] http://media.xiph.org/video/derf/. Xiph.org test media, 2006.

[13] A. Kokaram and S. J. Godsill. Mcmc for joint noise reduction and missing data treatement in degraded video. *IEEE Transactions on Signal Processing, Special Issue on MCMC*, 50:189–205, 2002.

[14] X. Lan, S. Roth, D. Huttenlocher, and M. J. Black. Efficient belief propagation with learned higher-order markov random fields. In *European Conference on Computer Vision (ECCV)*, 2006.

[15] S. H. Lee and M. G. Kang. Spatio-temporal video filtering algorithm based on 3-d anisotropic diffusion equation. In *The IEEE International Conference on Image Processing (ICIP)*, 1998.

[16] C. Liu, W. T. Freeman, R. Szeliski, and S. B. Kang. Noise estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Workshop*, pages 121–130, 1981.

[18] T. M. Moldovan, S. Roth, and M. J. Black. Denoising archival films using a learned bayesian model. In *The IEEE International Conference on Image Processing (ICIP)*, 2006.

[19] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[20] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 2003.

[21] N. Rajpoot, Z. Yao, and R. Wilson. Adaptive wavelet restoration of noisy video sequences. In *The IEEE International Conference on Image Processing (ICIP)*, 2004.

[22] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 860–867, 2005.

[23] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[24] I. W. Selesnick and K. Y. Li. Video denoising using 2d and 3d dual-tree complex wavelet transforms. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[25] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994.

[26] E. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1991.

[27] O. Williams, M. Isard, and J. MacCormick. Estimating disparity and occlusions in stereo video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[28] J. Xiao, H. Cheng, H. Sawhney, C. Rao, and M. Isnardi. Bilateral filtering-based optical flow estimation with occlusion detection. In *European Conference on Computer Vision (ECCV)*, 2006.

[29] V. Zlokolica and W. Philips. Motion- and detail-adaptive denoising of video. In *Proceedings of SPIE - Image Processing: Algorithms and Systems IIIVladimir*, 2004.

[30] V. Zlokolica, A. Pizurica, and W. Philips. Wavelet-domain video denoising based on reliability measures. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(8):993– 1007, 2006.

[31] V. Zlokolica, A. Pizurica, E. Vansteenkiste, and W. Philips. Spatio-temporal approach for noise estimation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.