High-Order Model and Dynamic Filtering for Frame Rate Up-Conversion

Wenbo Bao, *Student Member, IEEE*, Xiaoyun Zhang, *Member, IEEE*, Li Chen, *Member, IEEE*, Lianghui Ding, and Zhiyong Gao

***Abstract*— This paper proposes a novel frame rate up-conversion method through high-order model and dynamic filtering (HOMDF) for video pixels. Unlike the constant brightness and linear motion assumptions in traditional methods, the intensity and position of the video pixels are both modeled with high-order polynomials in terms of time. Then, the key problem of our method is to estimate the polynomial coefficients that represent the pixel's intensity variation, velocity, and acceleration. We propose to solve it with two energy objectives: one minimizes the auto-regressive prediction error of intensity variation by its past samples, and the other minimizes video frame's reconstruction error along the motion trajectory. To efficiently address the optimization problem for these coefficients, we propose the dynamic filtering solution inspired by video's temporal coherence. The optimal estimation of these coefficients is reformulated into a dynamic fusion of the prior estimate from pixel's temporal predecessor and the maximum likelihood estimate from current new observation. Finally, frame rate up-conversion is implemented using motion-compensated interpolation by pixel-wise intensity variation and motion trajectory. Benefited from the advanced model and dynamic filtering, the interpolated frame has much better visual quality. Extensive experiments on the natural and synthesized videos demonstrate the superiority of HOMDF over the state-of-the-art methods in both subjective and objective comparisons.**

***Index Terms*— Frame rate up conversion, high order model, dynamic filtering, energy minimization, maximum *a posteriori*.**

I. Introduction

HIGH-Frame-Rate (HFR) videos have been pervasively demanded by numerous applications in the past decades. Traditionally, because of the high bandwidth requirement and limited communication resources, video streams are compressed by encoders [1] and provided at a relatively low frame rate (24/30Hz). However, low-frame-rate (LFR) videos may lead to some visual artifacts including motion blurriness, frame flickering, *etc*, especially for moving scenes. To provide high-quality videos to TV users, the high-end TV chips may be integrated with frame interpolation modules. Frame interpolation or frame rate up conversion (FRUC) is a technique to temporally super-resolve LFR video into HFR one, such that the motion smoothness is strengthened. Recently, the demand for HFR videos has also arisen in other applications [2], [3]. Typical scenarios are online videos, live sports, video gaming [2], *etc*. These new applications are much more sensitive to motion smoothness between frames since users are encouraged to have more frequent interactions with the subjects in videos. Customers may even be immersed in videos when equipped with head-mounted displays, such as AR and 3D movies where artifact-free HFR videos are highly desired. The rising demands in these emerging applications have stimulated the study for better FRUC methods [4]–[24].

In FRUC, except for the frame repetition or frame averaging scheme, mainstream algorithms adopt the idea of Motion Compensated Frame Interpolation (MCFI). It usually contains the Motion Estimation (ME), Motion Refinement (MR) and Motion Compensation (MC) procedures. The ME and MR are responsible for estimating the true Motion Vector (MV) of pixels between original reference frames of LFR videos, and MC utilizes these MVs to interpolate the pixel intensities of intermediate frame and produce HFR videos.

Plenty of algorithms have been developed in literature for each of the three procedures. For ME, except for some Feature based [4]–[8] and Phase Plane Correlated algorithms [9]–[11], commonly used methods can be categorized into Block Matching based Algorithms (BMA) [12]–[18] and Optical Flow based Algorithms (OFA) [19]–[24]. The BMA methods usually divide reference frames into small pixel blocks and exploit certain search strategies [10], [13], [15]–[17] with a selection criteria [13], [18] to obtain their MVs. Representative search strategies are spatial/temporal search [13], hierarchical search [14], *etc*. While sum of absolute block difference is widely adopted as the selection criteria. In contrast, the OFA methods aim to generate dense pixel-wise MV Field (MVF), and mainstream methods are based on a variational framework [19]. The variational model usually imposes an energy constraint on the pixel difference between reference frames. For MR [25]–[29], it is often assumed that the MVF is spatially piece-wise smooth such that neighboring MVs are correlated. Then, the smoothness prior is exploited to refine the initial MVF by ME. As for the MC procedure, the most

Manuscript received May 10, 2017; revised December 20, 2017; accepted April 2, 2018. Date of publication April 9, 2018; date of current version April 26, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61771306, Grant 61301116, Grant 61521062, and Grant 61420106008, in part by STCSM under Grant 17DZ1205602, in part by the Chinese National Key S&T Special Program under Grant 2013ZX01033001-002-002, and in part by the Shanghai Key Laboratory of Digital Media Processing and Transmissions under Grant STCSM 18DZ2270700. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Oleg V. Michailovich. *(Corresponding author: Xiaoyun Zhang.)*

The authors are with the Department of Electronic Engineering, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: baowenbo@sjtu.edu.cn; xiaoyun.zhang@sjtu.edu.cn; hilichen@sjtu.edu.cn; lhding@sjtu.edu.cn; zhiyong.gao@sjtu.edu.cn).

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

Digital Object Identifier 10.1109/TIP.2018.2825100

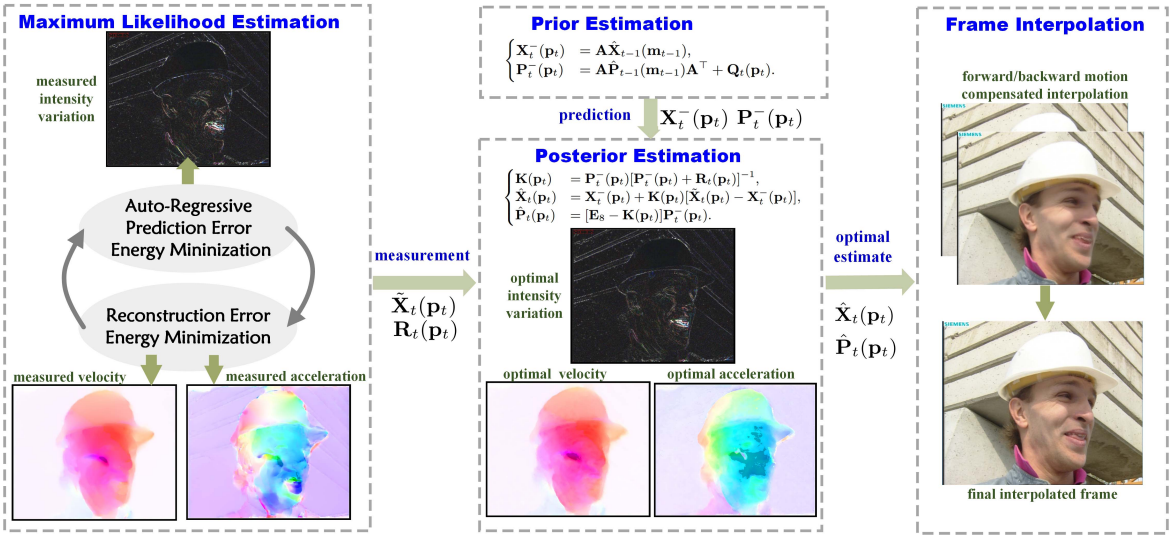


Fig. 1. Flowchart of the proposed HOMDF method for FRUC.

widely adopted approach is to take the blocks or pixels of the reference frame directly to compensate for the intermediate frame by their MVs. In consideration of the occlusions, blurriness or blockiness, some other methods improve the compensation by overlapped block MC [30], [31] or patch-based MC [24].

In the above methods, most algorithms are derived based on two fundamental assumptions. One is that pixel motion in-between two consecutive reference frames is linear. The motion pattern of a pixel which may vary greatly in practical scenarios is simply specified by a motion vector. Consequently, the linear motion assumption leads to that the displacement between an intermediate frame and a reference frame is proportional to their time interval, which is not always true. Another assumption lies in that the intensities of pixels connected by their motion vector are constant. It is reflected in both ME and MC. In ME, the criteria for optimal MV is usually dominated by absolute block difference in BMA or absolute pixel difference in OFA. Some works also exploit the intensity gradients [18] or color channel constancy [18], [32] to improve the capability for estimating correct MV in ME. However, in the stage of MC, they still project the intensity value of reference frame directly to the corresponding position in the intermediate frame without considering the gradients or color channel information. To overcome the limitations of constant intensity and linear motion assumption, we propose to use better models for pixels. With the proposed model, more accurate frame interpolation results will be obtained, especially for complicated motion and brightness cases.

In this paper, we firstly propose the high order model for pixel's intensity and position to enhance its capability to provide accurate descriptions for videos. The intensity is approximated by 1st-order polynomial composed of the constant part and a new inter-frame variation part. And the position is approximated by 2nd-order motion trajectory which comprises an explicit acceleration term besides linear motion. After that, we propose two energy functions for these polynomial coefficients, which are used for estimating intensity variation and motion trajectory, respectively. One function minimizes the auto-regressive prediction error of intensity variation by its past samples. And the other minimizes frame's reconstruction error along the accelerated motion trajectory. By optimizing the energy functions, the optimal estimation for these coefficients are generated and can be used to produce more visually pleasant interpolated frames.

However, this multi-variate optimization problem is complicated due to the pixel-wise dense field requirement, local minimum problem, sensitivity to variable initialization and so on. Inspired by video's temporal coherence that the pixel's states such as motion, brightness are temporally correlated, we reformulate the problem with maximum *a posteriori* (MAP) estimation of these coefficients. The MAP estimate can be factorized into the prior and maximum likelihood (ML) parts. The prior part exploits the information of pixel's temporal predecessor to make a prediction for the current state, while the ML part maximizes the likelihood of new frame pixel observation and acquire a fresh measurement through a reduced energy minimization problem. Then the optimal estimate is accomplished by fusing the prediction and measurement adaptively. Eventually, with the optimal estimation of intensity variation and motion trajectory, the forward and backward motion compensated frame interpolation are conducted to generate the final interpolated frames. The flowchart of the proposed method based on high order model and dynamic filtering (HOMDF) is illustrated in Fig.1.

Compared with the existing FRUC algorithms in the literature, our method in this paper makes the following contributions:

1) Both the intensity and position of pixels in a video are modeled with high-order polynomials, which is superior to conventional methods with constant motion and brightness assumptions.
2) The high order model is solved with proposed dynamic filtering, which exploits video's temporal coherence and

achieves the optimal estimation in an efficient and robust way.

3) During the dynamic filtering, the measurement of intensity variation and motion is decoupled into two subproblems, which are iteratively optimized using energy minimization method.

The rest of this paper is organized as follows. Section II discusses some of the related works. Section III explains the framework of the proposed method for FRUC. Section IV gives details of the optimal estimation to the high order model. Experiments are conducted in Section V. Finally, Section VI makes a conclusion for this paper.

II. Related Work

Frame rate up conversion has been an active research topic in video processing. There are a large number of publications focusing on this topic. In this section, we introduce some of the most relevant ones from the perspective of the motion's high order model, pixel's intensity model, dynamic filtering method, and optical flow estimation adopted in FRUC.

A. Motion's High Order Model

Regarding the modeling of high order motion, the research by Tsai and Lin [33] proposed to estimate the acceleration and calibrated existing motion trajectories so as to improve the precision of motion. In contrast, we take the acceleration estimation as a key element at the foremost stage of modeling instead of regarding it a post-processing refinement technique. Moreover, their calibration highly depends on the accuracy of initial block based MVF and no guarantee is made for the quality of acceleration vector. As a contrast, our energy minimization and dynamic filtering solution provide more truthful pixel-wise MVF, which is assured by optimization and estimation theories.

B. Pixel's Intensity Model

Zhang *et al.* have proposed two models, the spatial-temporal auto-regressive model (STAR) [34] and motion-aligned auto-regressive model (MAAR) [35]. Basically, these two models assumed that pixel intensity of the interpolated or the original frames could be approximated by a weighted sum of pixels within the spatial and/or temporal neighborhoods. Besides, Zhang *et al.* [36] also proposed to model the brightness along the motion trajectory with a polynomial approximation method. However, they solved the polynomial derivatives with a trivial weighted average of adjacent pixel differences. In contrast, we model the intensity variation instead of intensity itself with a locally stationary auto-regressive model and use the energy minimization method to accomplish a more robust estimate.

C. Dynamic Filtering

The adopted Kalman filter in dynamic filtering is an optimal filtering method [37] and has been applied to image/video processing including object tracking [38], [39], depth estimation [40], motion estimation [41], video error concealment [42], *etc*. Although, a linear quadratic motion estimation (LQME) algorithm [27] for FRUC has been proposed based on Kalman filter for quadratic estimation of MV, it simply uses the co-located position's motion vector of previous frame to make a prior prediction for current pixel block, which is not appropriate for non-stationary objects. In contrast, our method performs filtering along the motion trajectory and is more reasonable.

D. Optical Flow Estimation

BMA methods are hardware-friendly and widely applied in integrated circuit chips, but they suffer from many defects such as incorrect motion vectors, slow convergence in searching for the best MV, *etc*. Recently, several OFA methods [21]–[24] introduced some of the excellent optical flow algorithms [19], [20] into the application of FRUC and significant improvements have been obtained due to the dense MVF. As most of them directly use the existing optical flow estimation [19] in FRUC, Lee *et al.* [23] refined pixel's optical flow by choosing the flows of its neighboring pixels, which can be regarded as a spatial filter. However, since the spatial smoothness information has already been exploited as a regularization term in the energy function of the variational model, their refinement may have very limited effectiveness. In our method, we also obtain pixel-wise MVF, but it is composed of two subfields corresponding to velocity and acceleration. And the proposed dynamic filter can be viewed as a temporal filter to improve the accuracy of MVF, which is more reasonable than Lee's spatial filter scheme [23].

III. Framework of Frame Rate Up Conversion

A. High Order Model

Formally, let $\mathbf{p}_t := (x, y)^\top$ indexes a pixel's two-dimensional spatial coordinate in the video frame at time step t. The pixel's spatial location $\mathbf{p}_t$ is restricted in the image plane $\Omega \subset \mathbb{R}^2$. By our high order model for pixels, it is assumed that pixels have their velocities and accelerations, which comprise the motion vector fields (MVF) $\mathbf{p}'_t := (u, v)^\top$ and $\mathbf{p}''_t := (w, z)^\top$, respectively. According to the law of accelerated motion, the new position for the pixel in a small time interval Δt is then expressed as $\mathbf{p}_t + \mathbf{p}'_t \Delta t + \frac{1}{2}\mathbf{p}''_t \Delta t^2$. Therefore, the position $\mathbf{p}_{t+\Delta t}$ along trajectory at time $t + \Delta t$ is represented as

$$\mathbf{p}_{t+\Delta t} = \left(\mathbf{p}_t, \mathbf{p}'_t, \mathbf{p}''_t\right) \times \left(1, \Delta t, \frac{\Delta t^2}{2}\right)^\top. \tag{1}$$

The time interval Δt can be any real number such as 0, ± 1 or $\pm 1/2$. For instance, $\mathbf{p}_{t+1}$ refers to the pixel's position in the next neighboring frame while $\mathbf{p}_{t+1/2}$ refers to the pixel's position in the intermediate frame.

Let $\mathrm{I}_t(\mathbf{p}_t) : \mathbb{R}^2 \to \mathbb{R}$ represent the pixels' intensities. In contrast to traditional methods which assume that the pixel in consecutive frames keeps the same brightness, we assume that there exists an intensity variation from frame to frame. Therefore, the intensity of a pixel in a time interval Δt is modeled as

$$\mathrm{I}_{t+\Delta t}\left(\mathbf{p}_{t+\Delta t}\right) = \mathrm{I}_t(\mathbf{p}_t) + \mathrm{I}'_t(\mathbf{p}_t)\Delta t. \tag{2}$$

The high order model of this paper is reflected by the two equations (1) and (2). The equation (1) expresses that, from time t to $t+\Delta t$, the pixel follows a second-order motion trajectory and equation (2) means that the pixel intensity along the trajectory can also vary with time.

Overall, in this high order model for pixel position and intensity, there are three coefficient terms to be determined. They are $\mathbf{p}'_t$, $\mathbf{p}''_t$ and I'_t, representing pixel's velocity, acceleration and the intensity variation, respectively. By contrast, conventional FRUC algorithms usually assume that objects are moving linearly and hold a brightness constancy assumption. In these algorithms, only the MVF $\mathbf{p}'_t$ is to be determined, lacking the flexibility on dealing with nonlinear motions. In this paper, however, we take a step further by higher-order approximation for more accurate modeling of pixels to increase our algorithm's capability for handling more complicated videos.

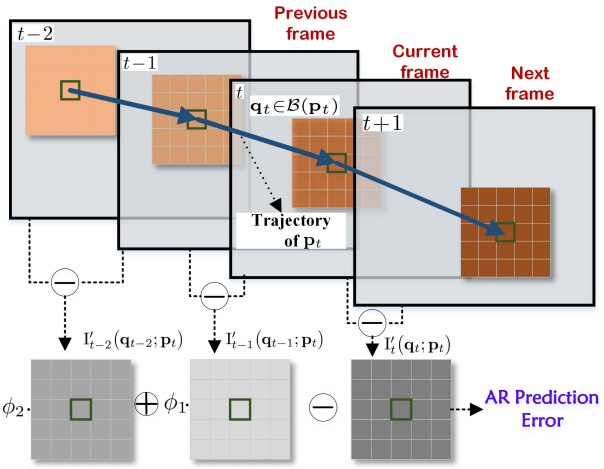


Fig. 2. Illustration of prediction error of AR model.

B. Intensity Variation and Motion Trajectory Estimation

To obtain these coefficients of the high order model, energy objective functions are defined based on two models, namely, auto-regressive (AR) model for intensity variation and variational model for motion trajectory.

1) Auto-Regressive Model for Intensity Variation: In natural videos, it is observed that scene illumination usually changes smoothly or consistently [36]. For example, the light source may gradually go dark or bright, leading to a decreasing or increasing intensity value of pixels. Therefore, intensity variations can be seen as predictable. In this paper, the intensity variation of a pixel is considered as a stationary random process. The following auto-regressive model describes this process,

$$\mathrm{I}'_t(\mathbf{p}_t) = \sum_{l=1}^{k} \phi_l(\mathbf{p}_t)\mathrm{I}'_{t-l}\big(\mathbf{p}_{t-l}\big) + \mathrm{n}(\mathbf{p}_t), \quad \forall \mathbf{p}_t \in \Omega. \tag{3}$$

The k-order AR model takes the weighting average of pixel $\mathbf{p}$'s past samples $\mathrm{I}'_{t-l}\big(\mathbf{p}_{t-l}\big)$ to approximate current intensity variation $\mathrm{I}'_t(\mathbf{p}_t)$. And the weights for each order are noted as $\phi_1(\mathbf{p}_l)$ correspondingly. Besides, there is an prediction error $\mathrm{n}(\mathbf{p})$ in AR process. In our model we empirically set the AR order to be $k = 2$. By this AR model, the estimation of intensity variation is turned into the problem of estimating the AR parameters ϕ_1 to ϕ_k.

To determine these parameters, the samples of each target pixel $\mathbf{p}_t$ along its motion trajectory are used to train the model by minimizing the prediction error. However, if we only use the single set of samples of the target pixel itself along the trajectory, the AR process will not have a unique solution since the samples are insufficient. Considering that the neighboring pixels usually belong to the same object, their intensity variations share the same regressive process with the target pixel and thus can help determine the AR parameters. The pixel block $\mathcal{B}(\mathbf{p}_t)$ centered at target pixel $\mathbf{p}_t$ are simultaneously predicted along the trajectory of $\mathbf{p}_t$. Then, the pixel-wise AR prediction error for all the target pixels in image space Ω makes up of the following energy cost:

$$E_{AR,D}(\phi_1,\cdots,\phi_k) = \int_{\Omega} \sum_{\mathbf{q}_t \in \mathcal{B}(\mathbf{p}_t)} \Psi\Big(\mathrm{n}^2\big(\mathbf{q}_t;\mathbf{p}_t\big)\Big)\mathrm{d}\mathbf{p}_t. \tag{4}$$

This energy is regarded as the data term of our AR model. Fig.2 illustrates how AR prediction error is constructed in details. Firstly, guided by the motion trajectory of $\mathbf{p}_t$, the sample set of pixels $\mathbf{q}_t$ of block $\mathcal{B}(\mathbf{p}_t)$ are extracted with the direct subtraction of pixel intensity between two frames. Thus, the samples $\mathrm{I}'_t\big(\mathbf{q}_t;\mathbf{p}_t\big)$, $\mathrm{I}'_{t-1}\big(\mathbf{q}_{t-1};\mathbf{p}_t\big)$ and $\mathrm{I}'_{t-2}\big(\mathbf{q}_{t-2};\mathbf{p}_t\big)$ of frames at t, $t-1$ and $t-2$ can be obtained. Finally, the AR prediction error $\mathrm{n}(\mathbf{q}_t;\mathbf{p}_t)$ is calculated as the difference between $\hat{\mathrm{I}}'_t(\mathbf{q}_t;\mathbf{p}_t)$ and $\mathrm{I}'_t(\mathbf{q}_t;\mathbf{p}_t)$. Namely,

$$\begin{aligned}\mathrm{n}(\mathbf{q}_t;\mathbf{p}_t) &= \mathrm{I}'_t(\mathbf{q}_t;\mathbf{p}_t) - \hat{\mathrm{I}}'_t(\mathbf{q}_t;\mathbf{p}_t)\\ &= \mathrm{I}'_t(\mathbf{q}_t;\mathbf{p}_t) - \sum_{l=1}^{k}\phi_l(\mathbf{p}_t)\mathrm{I}'_{t-l}\big(\mathbf{q}_{t-l};\mathbf{p}_t\big).\end{aligned} \tag{5}$$

All the squared prediction error of block pixels $\mathcal{B}(\mathbf{p}_t)$ over all pixel position of the entire frame $\mathbf{p}_t \in \Omega$ is summed up, resulting in the total data energy cost $E_{AR,D}$ of equation (4). The convex function $\Psi(s^2) = \sqrt{s^2+\varepsilon}$ with ε being a small constant of 0.0001 is used to avoid singularity problem.

Since the data energy of AR model in equation (4) is minimized pixel-by-pixel, the obtained AR parameters are spatially independent. Therefore, the data energy minimization can produce spatial deviation of AR parameters in the current frame. For example, an abrupt changing of training samples of neighboring target pixels would lead to tremendously different AR parameter, which may not be in conformity with the intuition that neighboring AR processes are spatially similar. Therefore, the AR parameter planes $\phi_l(l = 1,\cdots,k)$ are viewed as Markov Random Fields. In addition to the data term, the smoothness term that imposes constraint on the parameter planes is utilized. Namely,

$$E_{AR,S}(\phi_1,\cdots,\phi_k) = \textstyle\sum_{l=1}^{k}\int_{\Omega}\Psi\Big(||\nabla\phi_l(\mathbf{p}_t)||_2^2\Big)\mathrm{d}\mathbf{p}. \tag{6}$$

It can prevent parameter planes from being affected by inconsistent brightness change or undesired abrupt deviation. $||\nabla\phi_l||_2$ denotes the L_2-norm of spatial gradient of ϕ_l.

From the two energy terms in equation (4) and (6) of AR model, it is expected to obtain the parameter plane ϕ_l. And the pixel-wise intensity variation $\hat{\mathrm{I}}'_t(\mathbf{p}_t)$ will be obtained by equation (3) neglecting the noise term. However, in this AR model, the motion trajectory is required because the training samples are obtained along pixels' trajectories. Therefore, the total energy of AR model for the high order coefficients are as follows:

$$E_{AR}(\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t) = E_{AR,D} + \lambda E_{AR,S}. \tag{7}$$

The regularization factor λ is to make a balance between the data and smoothness terms.

2) Variational Model for Motion Trajectory: To estimate the motion, we employ the idea of variational model which regards the to-be-estimated MVF as a function over the two-dimensional grid space and defines an energy objective for the function. This model has been widely studied in optical flow estimation [19], [20] where only one MVF is required. In contrast, by our high order model, since we adopt a second-order approximated motion trajectory for pixels, two MVFs of velocity and acceleration are to be determined. Moreover, as pixel's intensity varies among neighboring frames, the data term of variational model for flow field implicitly imposes constraints on intensity variation. Therefore, the energy function of variational model is established on the three coefficients as follows:

$$E_{VA}(\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t) = E_{VA,D} + \alpha E_{VA,S}, \tag{8}$$

where the data term $E_{VA,D}$ penalizes the error of observed and reconstructed frame. In order to solve the two MVFs of velocity and acceleration, the data term consists of both forward and backward frame reconstruction errors at time $t-1$ and $t+1$. Namely,

$$\begin{aligned} E_{VA,D} = &\int_\Omega \Psi\left(\left|\mathrm{I}_{t+1}(\mathbf{p}_{t+1}) - \left(\mathrm{I}_t(\mathbf{p}_t) + \mathrm{I}'_t(\mathbf{p}_t)\right)\right|^2\right)\mathrm{d}\mathbf{p}_t \\ &+ \int_\Omega \Psi\left(\left|\mathrm{I}_{t-1}(\mathbf{p}_{t-1}) - \left(\mathrm{I}_t(\mathbf{p}_t) - \mathrm{I}'_t(\mathbf{p}_t)\right)\right|^2\right)\mathrm{d}\mathbf{p}_t. \end{aligned} \tag{9}$$

Moreover, in equation (8), the smoothness term is defined as

$$E_{VA,S} = \int_\Omega \Psi\left(||\nabla u||_2^2 + ||\nabla v||_2^2 + ||\nabla w||_2^2 + ||\nabla z||_2^2\right)\mathrm{d}\mathbf{p}_t, \tag{10}$$

which penalizes the spatial gradients of four components u, v, w, z of $\mathbf{p}'_t$ and $\mathbf{p}''_t$. The parameter α in (8) is used to balance the data and smoothness terms.

3) Joint Energy Minimization: By combining the two energy functions in (7) and (8) together, the high order coefficients can be optimally determined by solving the joint energy minimization problem,

$$\langle \hat{\mathrm{I}}'_t, \hat{\mathbf{p}}'_t, \hat{\mathbf{p}}''_t \rangle = \mathrm{argmin}\{E_{AR} + E_{VA}\}. \tag{11}$$

Obviously, it is non-trivial to solve this multi-variate optimization problem, especially for that each of the high order coefficients is a pixel-wise field. Also, it is hard to get a robust estimation by numerical minimization due to the problems such as local minimum, sensitivity to variable initialization and so on. Moreover, since the calculation of intensity variation and motion are interdependent with each other, the minimization of E_{AR} and E_{VA} will be alternatively performed in the optimization process. It will be found that each time after the minimization of E_{VA}, training samples required by the E_{AR} have to be re-extracted, which is quite computationally inefficient. Hence, to overcome these difficulties and achieve an optimal estimation, we put up forward a dynamic filtering approach based on video's temporal coherence. The details of the approach are presented in Section IV.

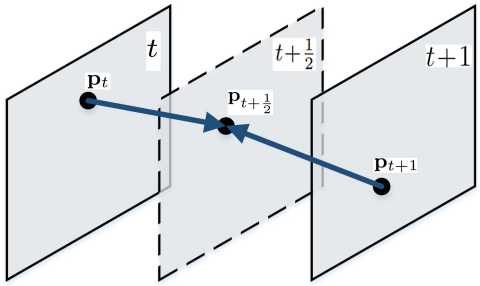


Fig. 3. Illustration of the frame interpolation. Two corresponding pixels $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ from forward and backward reference frames are used to compensate for $\mathbf{p}_{t+\frac{1}{2}}$ of the intermediate frame. Namely, $\mathbf{p}_{t+\frac{1}{2}} = \mathbf{p}_t + \frac{1}{2}\hat{\mathbf{p}}'_t + \frac{1}{8}\hat{\mathbf{p}}''_t$ or $\mathbf{p}_{t+\frac{1}{2}} = \mathbf{p}_{t+1} - \frac{1}{2}\hat{\mathbf{p}}'_{t+1} + \frac{1}{8}\hat{\mathbf{p}}''_{t+1}$.

C. Frame Interpolation

After the estimation of motion trajectory and intensity variation of current time t, according to the equation (1) of high order model, the intermediate frame at time step $t+\frac{1}{2}$ can then be obtained as follows:

$$\mathrm{I}^f_{t+\frac{1}{2}}(\mathbf{p}_{t+\frac{1}{2}}) = \mathrm{I}_t(\mathbf{p}_t) + \frac{1}{2}\hat{\mathrm{I}}'_t(\mathbf{p}_t), \tag{12}$$

where $\mathbf{p}_{t+\frac{1}{2}} = \mathbf{p}_t + \frac{1}{2}\hat{\mathbf{p}}'_t + \frac{1}{8}\hat{\mathbf{p}}''_t$. This is a forward unilateral motion compensated interpolation (MCI) by current estimation of the high order model. Generally, due to the relative movement, some pixels shown in one reference frame may be covered in another frame, which is known as the occlusion problem. It will result in hole pixels of the intermediate frame if we only use a unilateral interpolation. Therefore, it's better to use both the forward and backward unilateral MCI, which will make up most of the missing pixels for each other. The backward interpolation $\mathrm{I}^b_{t+\frac{1}{2}}(\mathbf{p}_{t+\frac{1}{2}})$ is projected along the inverted motion of pixels at $t+1$. Then, as illustrated in Fig.3, both $\mathrm{I}^f_{t+\frac{1}{2}}$ and $\mathrm{I}^b_{t+\frac{1}{2}}$ are fused together to produce the final frame. It is formulated as follows (the coordinate index $\mathbf{p}_{t+\frac{1}{2}}$ of intermediate frame is omitted for simplicity):

$$\mathrm{I}_{t+\frac{1}{2}} = \begin{cases} \left(\mathrm{I}^f_{t+\frac{1}{2}} + \mathrm{I}^b_{t+\frac{1}{2}}\right)/2, & \textit{both } \mathrm{I}^f \textit{ and } \mathrm{I}^b \textit{ exist}, \\ \mathrm{I}^f_{t+\frac{1}{2}}, & \textit{only } \mathrm{I}^f_{t+\frac{1}{2}} \textit{ exists}, \\ \mathrm{I}^b_{t+\frac{1}{2}}, & \textit{only } \mathrm{I}^b_{t+\frac{1}{2}} \textit{ exists}, \\ \textit{hole filling}, & \textit{neither } \mathrm{I}^f_{t+\frac{1}{2}} \textit{ nor } \mathrm{I}^b_{t+\frac{1}{2}} \textit{ exists}. \end{cases} \tag{13}$$

Actually, compared with traditional FRUC methods which only compensate the constant intensity along linear motion

trajectory, our method is superior in modeling pixels with high order polynomials for intensity variation (I'_t) and accelerated motion ($\mathbf{p}'_t$ and $\mathbf{p}''_t$). Moreover, the high order model is solved with a dynamic filtering method. It exploits video's temporal coherence and achieves the optimal estimation ($\hat{\mathrm{I}}'_t$, $\hat{\mathbf{p}}'_t$ and $\hat{\mathbf{p}}''_t$) both efficiently and robustly. Besides, by using a sliding window of four frames for estimating the high order coefficients, the forward and backward interpolated frames $\mathrm{I}^f_{t+\frac{1}{2}}$ and $\mathrm{I}^b_{t+\frac{1}{2}}$ are more consistent with each other. It leads to a more sharp and clear intermediate frame instead of a severe blurry one caused by frame averaging. For the hole pixels that neither $\mathrm{I}^f_{t+\frac{1}{2}}$ nor $\mathrm{I}^b_{t+\frac{1}{2}}$ provides an interpolation, we use trilateral filtering [18] to fill them up for the method's efficiency on providing clear textures and sharp edges.

IV. Optimal Estimation for High Order Model

Let's reconsider the problem in equation (11), which aims at using four video frames $\{\mathrm{I}\}_{t-2\to t+1}$ to determine the three coefficients $\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t$ of high order model. The straightforward numerical approximation solution to (11) is obviously inefficient. To address this problem, we firstly reformulate the problem from the perspective of probability theory. In fact, the energy minimization can be expressed as maximum *a posteriori* (MAP) estimation of the conditional probability as follows:

$$\langle \hat{\mathrm{I}}'_t, \hat{\mathbf{p}}'_t, \hat{\mathbf{p}}''_t \rangle = \operatorname{argmax}\ p\left(\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t \middle| \{\mathrm{I}\}_{t-2\to t+1}\right). \quad (14)$$

As is known to all, the moving objects of videos usually exhibits high temporal coherence that the pixel's states such as motion, brightness are temporally correlated. Inspired by this knowledge, we factorize the MAP estimation into the prior and maximum likelihood (ML) parts according to Bayes' Theorem. Namely,

$$\langle \hat{\mathrm{I}}'_t, \hat{\mathbf{p}}'_t, \hat{\mathbf{p}}''_t \rangle = \operatorname{argmax}\ p\left(\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t \middle| \{\mathrm{I}\}_{t-2\to t}\right) \cdot p\left(\mathrm{I}_{t+1} \middle| \mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t, \{\mathrm{I}\}_{t-2\to t}\right). \quad (15)$$

The prior part $p\left(\mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t \middle| \{\mathrm{I}\}_{t-2\to t}\right)$ represents that the probability of random variables I'_t, $\mathbf{p}'_t$ and $\mathbf{p}''_t$ should be maximized according to previous information $\{\mathrm{I}\}_{t-2\to t}$ without presence of frame I_{t+1}. While the ML part $p\left(\mathrm{I}_{t+1} \middle| \mathrm{I}'_t, \mathbf{p}'_t, \mathbf{p}''_t, \{\mathrm{I}\}_{t-2\to t}\right)$ refers to that with given values of these random variables, the probability of newly observed frame I_{t+1} should be maximized.

In other words, the prior part exploits the information of pixel's temporal predecessor to make a prediction for current state and the ML part maximizes the likelihood of new frame observation to obtain a fresh measurement. Each of the two parts is to obtain an estimate, but both are sub-optimal. To achieve the optimality, a dynamic filtering procedure adaptively fuses the two estimations according to corresponding noise levels. In this paper, the Kalman filter [37] is employed since its temporal iterative characteristic coincides well with our temporal filtering demands.

A. Dynamic Filtering

For better readability and ease of presentation, we use a compact state vector to represent pixel's multiple properties including position, velocity, acceleration, intensity and intensity variation as follows:

$$\mathbf{X}_t(\mathbf{p}_t) := \left(\mathbf{p}_t, \mathbf{p}'_t, \mathbf{p}''_t, \mathrm{I}_t(\mathbf{p}_t), \mathrm{I}'_t(\mathbf{p}_t)\right)^\top.$$

It is an 8-dimensional vector for every pixel location $\mathbf{p}_t$. In the following subsections, the prior estimate, maximum likelihood estimate and posterior estimate for this state vector, noted by $\mathbf{X}^-_t(\mathbf{p}_t)$, $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$, $\hat{\mathbf{X}}_t(\mathbf{p}_t)$, respectively, will be determined.

1) Prior Estimation: The prior part of the MAP expresses that one can obtain a prior estimation of state vector without given the observation of frame I_{t+1} but only the previous information. During the processing of videos over time, the previous frames as well as intermediate results such as $\hat{\mathbf{X}}_{t-1}$ are already know when proceeding to time step t. Supposing that the pixel $\mathbf{m}_{t-1}$ of time step $t-1$ will move to $\mathbf{p}_t$, *i.e.*, $\mathbf{p}_t = \mathbf{m}_{t-1} + \hat{\mathbf{m}}'_{t-1} + \frac{1}{2}\hat{\mathbf{m}}''_{t-1}$, we use this temporal predecessor pixel $\mathbf{m}_{t-1}$ to make a prediction for current pixel $\mathbf{p}_t$ through the following transition rules: (i) current velocity is the sum of previous estimated velocity and acceleration, (ii) current pixel intensity is the sum of previous intensity and its estimated intensity variation, and (iii) current acceleration and intensity variation are the same as before. The transition is a noisy process since these rules are not strictly obeyed in practice and the noise is modeled as *system noise*. With the compact form of state vectors, the above transition process as well as corresponding noise covariances can be formulated as:

$$\begin{cases} \mathbf{X}^-_t(\mathbf{p}_t) = \mathbf{A}\hat{\mathbf{X}}_{t-1}(\mathbf{m}_{t-1}), \\ \mathbf{P}^-_t(\mathbf{p}_t) = \mathbf{A}\hat{\mathbf{P}}_{t-1}(\mathbf{m}_{t-1})\mathbf{A}^\top + \mathbf{Q}_t(\mathbf{p}_t). \end{cases} \quad (16)$$

The transition matrix $\mathbf{A}$ can be easily derived out as

$$\mathbf{A} = \begin{bmatrix} \mathbf{E}_2 & \mathbf{E}_2 & \frac{1}{2}\mathbf{E}_2 & & \\ & \mathbf{E}_2 & \mathbf{E}_2 & & \\ & & \mathbf{E}_2 & & \\ & & & 1 & 1 \\ & & & & 1 \end{bmatrix}, \quad (17)$$

where $\mathbf{E}_2$ is a 2×2 identity matrix (similarly hereinafter). The prior estimate $\mathbf{X}^-_t(\mathbf{p}_t)$ and noise covariance $\mathbf{P}^-_t(\mathbf{p}_t)$ are transited from predecessor's posterior estimation $\hat{\mathbf{X}}_{t-1}(\mathbf{m}_{t-1})$ and corresponding noise covariance $\hat{\mathbf{P}}_{t-1}(\mathbf{m}_{t-1})$. The system noise $\mathbf{Q}_t(\mathbf{p}_t) := \sigma^2_{\mathbf{Q}}(\mathbf{p}_t)\mathbf{E}_8$ is defined as an identity matrix parameterized by $\sigma^2_{\mathbf{Q}}(\mathbf{p}_t)$ which will be detailed in section IV-B.

2) Maximum Likelihood Estimation: The purpose of ML estimation is to maximize the likelihood of new observation of frame I_{t+1}, which can be equally defined as minimizing the reconstruction error of this observation. With this definition, the following data term of variational model is established,

$$E^{ml}_{VA,D} = \int_\Omega \Psi\left(\left|\mathrm{I}_{t+1}(\mathbf{p}_{t+1}) - \left(\mathrm{I}_t(\mathbf{p}_t) + \mathrm{I}'_t(\mathbf{p}_t)\right)\right|^2\right) \mathrm{d}\mathbf{p}_t. \quad (18)$$

In this equation, not only the MVFs of velocity and acceleration, but also the variation intensity are to be updated, which means that the newly observed frame contributes to

new measurement of the state vector. To obtain a complete updating of the state vector, the optimization of AR parameters is also engaged. However, there are some differences between the AR data term with its previous definition in equation (4), because the motions before current time step t are taken as deterministic signal in ML estimation. Therefore, the training samples can be obtained along the predecessor of $\mathbf{p}_t$, namely, $\mathbf{m}_{t-1}$'s trajectory, instead of $\mathbf{p}_t$'s trajectory. The new AR prediction error for ML estimation is thus written as

$$\mathrm{n}^{ml}(\mathbf{q}_t;\mathbf{p}_t)=\mathrm{I}'(\mathbf{q}_t;\mathbf{p}_t)-\sum_{l=1}^{k}\phi_l(\mathbf{p}_t)\mathrm{I}'_{t-l}\big(\mathbf{q}_{t-l};\mathbf{m}_{t-1}\big). \quad (19)$$

Then, the new AR data term $E^{ml}_{AR,D}$ is given by replacing $\mathrm{n}(\mathbf{q}_t;\mathbf{p}_t)$ of $E_{AR,D}$ with $\mathrm{n}^{ml}(\mathbf{q}_t;\mathbf{p}_t)$ in equation (4). The ML estimation is turned into an energy minimization problem,

$$\tilde{\mathbf{X}}_t(\mathbf{p}_t)=\mathrm{argmin}\{E^{ml}_{AR}+E^{ml}_{VA}\}, \quad (20)$$

where

$$\begin{aligned}E^{ml}_{AR}&=E^{ml}_{AR,D}+\lambda E_{AR,S},\\ E^{ml}_{VA}&=E^{ml}_{VA,D}+\alpha E_{VA,S}+\beta E_{VA,C},\end{aligned} \quad (21)$$

The smoothness terms $E_{AR,S}$ and $E_{VA,S}$ have been described in equation (6) and (10). Additionally, a new regularization term $E_{VA,C}$ with a factor of β is incorporated. This new term imposes a consistency constraint on the MVF of velocity as follows:

$$E_{VA,C}=\int_{\Omega}\Psi\big(||\mathbf{p}'_t-(\hat{\mathbf{m}}'_{t-1}+\hat{\mathbf{m}}''_{t-1})||_2^2\big)\mathrm{d}\mathbf{p}_t. \quad (22)$$

It minimizes the difference between the measured velocity and the predicted velocity $(\hat{\mathbf{m}}'_{t-1}+\hat{\mathbf{m}}''_{t-1})$.

Apparently, the new energy in (20) has a similar form with the energy in equation (11). However, they are different from each other because ML estimation takes the previous states of pixels as deterministic but MAP estimation does not. This significantly simplifies the energy optimization problem from several aspects. Firstly, the backward reconstruction error used in $E_{VA,D}$ is abandoned by $E^{ml}_{VA,D}$, reducing the computation complexity of the iterative optimization of energy functions. Secondly, the new consistency term $E_{VA,C}$ implicitly encourages new motion field of velocity to be drawn near to previous estimations, which makes the optimization robust to variable initialization. Thirdly, the training samples in $E^{ml}_{AR,D}$ is fixed, leaving the optimization free of re-extracting training samples in each iteration.

For $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$ which can be viewed as a new measurement with the newly observed frame at $t+1$, the measurement noise variance is noted as a parameterized identity matrix $\mathbf{R}_t(\mathbf{p}_t):=\sigma^2_{\mathbf{R}}(\mathbf{p}_t)\mathbf{E}_8$ analogous to the system noise $\mathbf{Q}_t(\mathbf{p}_t)$.

3) Kalman Filter: As both prior estimate $\mathbf{X}^-_t(\mathbf{p}_t)$ and ML estimate $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$ are noisy, we take use of the Kalman filtering tool and the following updating functions to obtain an optimal estimation. We refer the readers concerned about the deducing of updating equations of Kalman filter to the literature [37].

$$\begin{cases}\mathbf{K}(\mathbf{p}_t)=\mathbf{P}^-_t(\mathbf{p}_t)[\mathbf{P}^-_t(\mathbf{p}_t)+\mathbf{R}_t(\mathbf{p}_t)]^{-1},\\ \hat{\mathbf{X}}_t(\mathbf{p}_t)=\mathbf{X}^-_t(\mathbf{p}_t)+\mathbf{K}(\mathbf{p}_t)[\tilde{\mathbf{X}}_t(\mathbf{p}_t)-\mathbf{X}^-_t(\mathbf{p}_t)],\\ \hat{\mathbf{P}}_t(\mathbf{p}_t)=[\mathbf{E}_8-\mathbf{K}(\mathbf{p}_t)]\mathbf{P}^-_t(\mathbf{p}_t).\end{cases} \quad (23)$$

In equation (23), the Kalman gain $\mathbf{K}(\mathbf{p}_t)$ is firstly calculated as a matrix division between $\mathbf{P}^-_t(\mathbf{p}_t)$ and $\mathbf{P}^-_t(\mathbf{p}_t)+\mathbf{R}_t(\mathbf{p}_t)$. Then, the ML estimate $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$ and prior estimate $\mathbf{X}^-_t(\mathbf{p}_t)$ are fused adaptively to reach a final posterior estimate $\hat{\mathbf{X}}_t(\mathbf{p}_t)$ in an analytical way that is considerably efficient. Besides, the posterior noise covariance $\hat{\mathbf{P}}_t(\mathbf{p}_t)$ for $\hat{\mathbf{X}}_t(\mathbf{p}_t)$ is also updated. In this dynamic fusion of prior and ML estimates, the filtered result is dependent on the noise of them. If any one of them has a relatively high noise level, the other will be much more preferred, so that the optimal estimation for $\hat{\mathrm{I}}'_t$, $\hat{\mathbf{p}}'_t$, $\hat{\mathbf{p}}''_t$ can be extracted from state vector $\hat{\mathbf{X}}_t(\mathbf{p}_t)$. In a temporally iterative way, similar to what $\hat{\mathbf{X}}_{t-1}(\mathbf{m}_{t-1})$ has been used for making a prior prediction $\mathbf{X}^-_t(\mathbf{p}_t)$, $\hat{\mathbf{X}}_t(\mathbf{p}_t)$ will also be used for the successor at $t+1$ of pixel $\mathbf{p}_t$.

During the procedure of dynamic filtering, the previous information provided by past frames is exploited by the prior part, while the future information in the new frame is utilized by the ML part. The adopted temporal coherence information makes great contributions to temporally fluent motion and intensity with a highly efficient utilization of all video frames. Moreover, our method mitigates the difficulty of estimating motions in occlusion areas, because the motions of such areas can be predicted from pixels' past states, although these occluded pixels have no correspondences in the next frame. More details are presented in the experimental section V-B.

B. Implementation Details

From the descriptions above for dynamic filtering, the quality of filtering is related mainly to two issues, one is the evaluation metric of noise level and the other is the energy minimization in ML estimation.

1) Noise Covariance: In dynamic filtering, the system noise $\mathbf{Q}_t$ and measurement noise $\mathbf{R}_t$ related to prior and ML estimates are used to determine a Kalman gain, which should be well-balanced to acquire the final posterior estimation. The noise metric is optional, but a proper selection of it is profitable for robust filtering and fast convergence of the iterative Kalman filter. Empirically, if the reconstruction error by an estimated state vector is high or the estimated motion field is not locally smooth, its reliability is very likely to be low [23]. We suggest that the noise variances $\mathbf{Q}_t$ and $\mathbf{R}_t$ are related to the bidirectional reconstruction error and local motion smoothness terms. Instead of using the two terms over a single pixel, the block reconstruction error (BRE) and block motion smoothness (BMS) are utilized to improve the robustness of noise evaluation. Taking $\sigma^2_{\mathbf{R}}(\mathbf{p}_t)$ of $\mathbf{R}_t$ as an example, the calculation of the noise parameter is defined as follows:

$$\sigma^2_{\mathbf{R}}(\mathbf{p}_t)=1-\exp\big(-\eta BRE(\mathbf{p}_t)-\kappa BMS(\mathbf{p}_t)\big), \quad (24)$$

where

$$\begin{aligned}&BRE(\mathbf{p}_t)\\ &=\sum_{\mathbf{q}_t\in\mathcal{B}(\mathbf{p}_t)}\Big(\big|\mathrm{I}_{t+1}(\mathbf{q}_{t+1};\mathbf{p}_t)-\big(\mathrm{I}_t(\mathbf{q}_t;\mathbf{p}_t)+\tilde{\mathrm{I}}'_t(\mathbf{q}_t;\mathbf{p}_t)\big)\big|\\ &\quad+\big|\mathrm{I}_{t-1}(\mathbf{q}_{t-1};\mathbf{p}_t)-\big(\mathrm{I}_t(\mathbf{q}_t;\mathbf{p}_t)-\tilde{\mathrm{I}}'_t(\mathbf{q}_t;\mathbf{p}_t)\big)\big|\Big),\end{aligned} \quad (25)$$

and

$$BMS(\mathbf{p}_t) = \sum_{\mathbf{q}_t \in \mathcal{B}(\mathbf{p}_t)} ||\nabla \tilde{u}||_2 + ||\nabla \tilde{v}||_2 + ||\nabla \tilde{w}||_2 + ||\nabla \tilde{z}||_2. \tag{26}$$

In equation (24), η and κ are used to balance the two terms. The tilde marks $\sim$ in $\tilde{\mathrm{I}}'_t$, $\tilde{u}$ and so on represent that they are the components of measurement $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$.

2) Energy Minimization Method: For the problem of equation (20), we propose to alternatively optimize two sub-problems aimed at minimizing E^{ml}_{AR} and E^{ml}_{VA} respectively. Each sub-problem is solved with gradient descent algorithms. Therefore, we derive out the gradients of each objective function with respect to their variables. For the function of E^{ml}_{AR}, with fixed motion trajectory, the derivatives with respect to the parameters of each order at each pixel position are given as

$$\frac{\partial E^{ml}_{AR}}{\partial \phi_l(\mathbf{p}_t)} = \sum_{\mathbf{q}_t \in \mathcal{B}(\mathbf{p}_t)} \Psi'\Big(\mathrm{n}^2\big(\mathbf{q}_t; \mathbf{p}_t\big)\Big)\mathrm{I}'_{t-l}\big(\mathbf{p}_{t-l}; \mathbf{m}_{t-1}\big) - \lambda \cdot div\bigg(\Psi'\Big(||\nabla \phi_l(\mathbf{p}_t)||_2^2\Big)\nabla \phi_l(\mathbf{p}_t)\bigg). \tag{27}$$

where the operator $div(\cdot)$ is the *divergence* of a vector field. These derivatives of the variables in (27) are used by BFGS gradient descent algorithm [43] in numerical optimization.

For the second sub-problem, with fixed intensity variation $\mathrm{I}'_t(\mathbf{p}_t)$, the derivatives of E^{ml}_{VA} with respect to the variables $\mathbf{p}'_t$ and $\mathbf{p}''_t$ are also required. Without loss of generality, we take the horizontal component u of $\mathbf{p}'_t$ as an illustration as follows:

$$\begin{aligned}
&\frac{\partial E^{ml}_{VA}}{\partial u} \\
&= \Psi'\Big(\Big|\mathrm{I}_{t+1}(\mathbf{p}_{t+1}) - \big(\mathrm{I}_t(\mathbf{p}_t) + \mathrm{I}'_t(\mathbf{p}_t)\big)\Big|^2\Big) \\
&\quad \cdot \Big|\mathrm{I}_{t+1}(\mathbf{p}_{t+1}) - \big(\mathrm{I}_t(\mathbf{p}_t) + \mathrm{I}'_t(\mathbf{p}_t)\big)\Big| \cdot \frac{\partial \mathrm{I}_{t+1}(\mathbf{p}_{t+1})}{\partial u} \\
&\quad - \alpha \cdot div\bigg(\Psi'\Big(||\nabla u||_2^2 + ||\nabla v||_2^2 + ||\nabla w||_2^2 + ||\nabla z||_2^2\Big)\nabla u\bigg) \\
&\quad + \beta \cdot \Psi'\Big(||\mathbf{p}'_t - (\hat{\mathbf{m}}'_{t-1} + \hat{\mathbf{m}}''_{t-1})||_2^2\Big) \\
&\quad \times \frac{\partial ||\mathbf{p}'_t - (\hat{\mathbf{m}}'_{t-1} + \hat{\mathbf{m}}''_{t-1})||_2^2}{\partial u}.
\end{aligned} \tag{28}$$

In a similar way, we can also get the derivatives $\frac{\partial E^{ml}_{VA}}{\partial v}$, $\frac{\partial E^{ml}_{VA}}{\partial w}$, and $\frac{\partial E^{ml}_{VA}}{\partial z}$ correspondingly. We use the successive over relaxation algorithm which is also adopted by [19] to solve the variational optimization problem. By alternating the minimization of E^{ml}_{AR} and E^{ml}_{VA}, the ML estimation $\tilde{\mathbf{p}}'_t$, $\tilde{\mathbf{p}}''_t$ and $\tilde{\mathrm{I}}'_t(\mathbf{p}_t)$ of $\tilde{\mathbf{X}}_t(\mathbf{p}_t)$ in equation (20) will be obtained.

V. Experimental Results

We conduct several experiments to demonstrate the effectiveness of the proposed HOMDF method. In these experiments, videos of the CIF (352x288) and 1080p (1920x1080) formats are used as the test sources. Specifically, the CIF videos contain *Akiyo*, *Bus*, *City*, *Coastguard*, *Container*, *Flower*, *Football*, *Foreman*, *Hall*, *Ice*, *Mobile*, *News*, *Paris*, *Silent*, *Soccer*, *Stefan* and *Waterfall*. The 1080p videos are *Bluesky*, *Kimono*, *Sunflower*, *Parkscene*. In addition to these naturally captured videos, we further synthesized two video sequences named as *AccMot* and *VarIllu*. They are constructed based on the natural video *Flower* and *Akiyo*. *AccMot* is an sequence with accelerated motion. It is composed of a stationary background from *Flower* sequence and an accelerative foreground cropped from *Akiyo*. The foreground moves from the bottom-right to the top-left in the video and then turns back. The motion has an obvious deceleration pattern with an acceleration of -2 $pel/frame$ and initial velocity of 40 $pel/frame$. *VarIllu* is the variant illumination sequence originated from the *Flower* sequence. However, the frames are processed with gamma transformation at different gamma values (see more details in Section V-C). By this transformation, *VarIllu* simulates a prominent variant illumination scenario.

The proposed method uses some parameters. In the extraction of training samples for AR model or the noise parameter evaluation of (24), the pixel block size of $\mathcal{B}(\cdot)$ is set to 5×5, the order of AR model is $k = 2$, and the regularization parameter λ is empirically set to be 10.0. In the energy function (21) of E^{ml}_{VA}, α is 30 and β is 3. In the dynamic filtering, the noise covariance evaluation (24) for $\mathbf{Q}_t$ and $\mathbf{R}_t$ is dependent on η and κ, which are 0.03 and 0.3, respectively.

Several state-of-the-art FRUC algorithms [13], [23], [24], [27], [28] are used for comparison. 3DRS [13] and Kim and Sunwoo's [28] methods belong to the BMA category. Lee *et al.* [23] and Kaviani and Shirani's [24] methods use optical flow estimation. Guo *et al.*'s [27] method uses Kalman filter under a bidirectional BMA framework.

To evaluate the performance of FRUC, the even frames of test videos are skipped and then interpolated by HOMDF as well as the referenced methods. Thus, the PSNR (Peak-Signal-to-Noise Ratio) and SSIM (Structural SIMilarity) [44] of the interpolated frame are calculated with respect to the original frame.

A. Comparison With Existing Algorithms

For the CIF format videos, we make comparisons with four algorithms, including Lee *et al.*'s [23], Kaviani and Shirani's [24], Guo *et al.*'s [27], and Kim and Sunwoo's [28],methods. While for the 1080p format videos, due to the unavailable results of Lee *et al.*'s [23], Kaviani and Shirani's [24], and Guo *et al.*'s [27] methods, we implement Kaviani and Shirani's [24] and additionally incorporate the 3DRS [13] algorithm for more comprehensive comparison. TABLE I and II compare the average PSNR and SSIM values over CIF sequences (with total frames indicated in brackets after sequence name) between the proposed HOMDF and the other FRUC methods. In the two tables, the best performance for each sequence is highlighted in **bold**. It can be seen that our method outperforms the state-of-the-art algorithms on most sequences. Although Guo *et al.*'s [27] and our method both take use of the Kalman filtering tool, we achieve a remarkably improved performance up to 2.1dB gain (for *Stefan*), because of the dense motion vector field as well as the unilateral motion compensation used by HOMDF. Moreover, HOMDF consistently outperforms both

TABLE I
COMPARISON OF VARIOUS FRUC ALGORITHMS IN TERMS OF AVERAGE PSNR (DB) OVER CIF FORMAT VIDEOS

Sequence	*Acc -Mot*	*Var -Illu*	*Akiyo*	*Bus*	*City*	*Coast -guard*	*Conta -iner*	*Flower*	*Foot -ball*	*Fore -man*	*Hall*	*Ice*	*MobileNews*	*Paris*	*Silent*	*Soccer*	*Stefan*	*Water -fall*
(total frames)	(40)	(100)	(300)	(150)	(300)	(300)	(300)	(300)	(260)	(300)	(300)	(300)	(300)	(300)	(300)	(300)	(300)	(300)
Kim [28]	29.95	39.82	44.30	24.38	31.91	31.80	43.04	26.90	21.85	30.47	36.39	29.84	25.23	34.34	32.41	34.39	25.97	25.19	34.14
Kaviani [24]	34.48	55.79	47.03	26.59	33.79	31.29	39.70	30.01	23.73	32.91	34.55	32.57	28.80	35.73	33.91	35.65	27.81	26.81	36.76
Lee [23]	35.76	56.14	46.73	27.53	35.43	34.14	43.26	30.45	23.55	34.06	37.22	33.34	29.11	35.50	34.03	36.36	28.11	27.18	39.99
Guo [27]	35.41	55.42	**47.32**	26.68	34.29	34.72	**43.91**	31.72	23.33	33.78	37.63	33.37	29.22	35.75	35.40	36.62	26.15	25.76	**40.59**
HOMDF	**37.54**	**56.50**	47.28	**28.03**	**35.99**	**35.45**	43.68	**32.03**	**24.36**	**34.72**	**37.75**	**34.43**	**30.22**	**36.28**	**35.65**	**37.03**	**28.86**	**27.85**	40.26

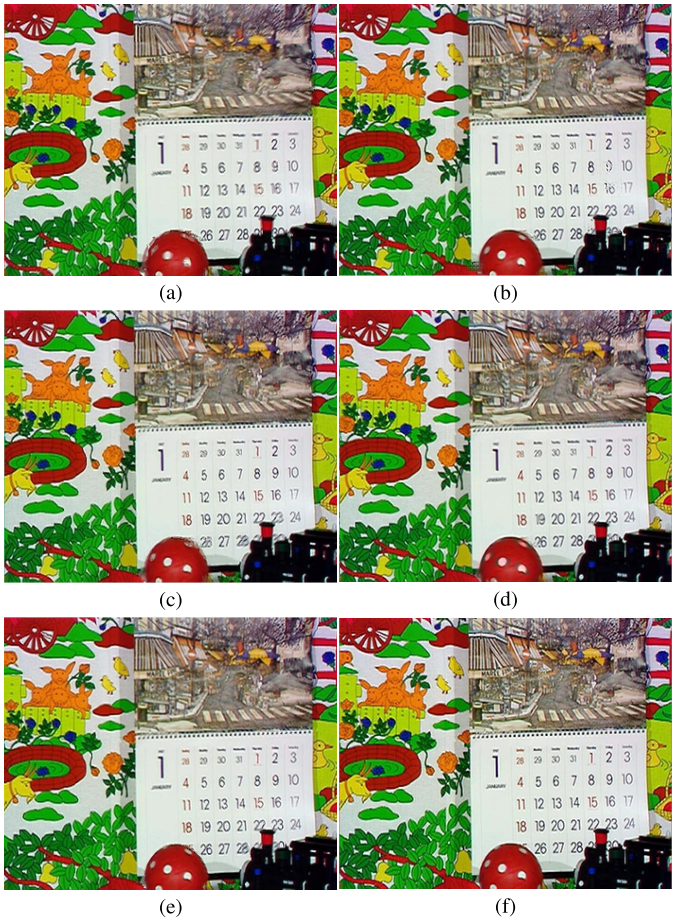


Fig. 4. Sujective comparison of *Mobile*'s 12-th frame. (a) Kim and Sunwoo [28], 24.88dB. (b) Kaviani and Shirani [24], 28.31dB. (c) Lee *et al.* [23], 27.87dB. (d) Guo *et al.* [27], 27.22dB. (e) HOMDF, 28.74dB. (f) Groundtruth.

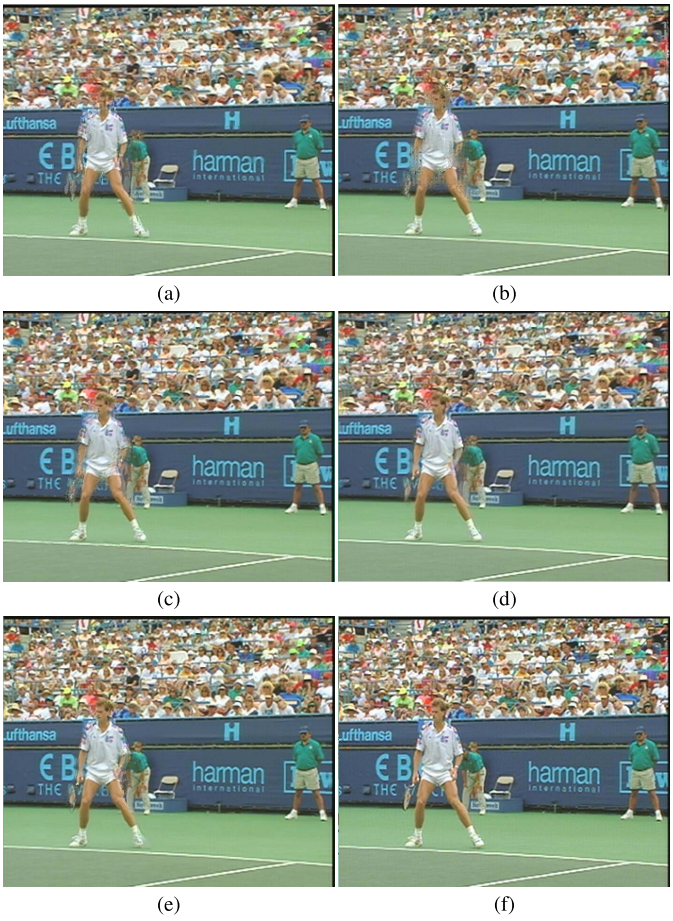


Fig. 5. Sujective comparison of *Stefan*'s 48-th frame. (a) Kim and Sunwoo [28], 27.97dB. (b) Kaviani and Shirani [24], 29.93dB. (c) Lee *et al.* [23], 29.81dB. (d) Guo *et al.* [27], 30.70dB. (e) HOMDF, 30.80dB. (f) Groundtruth.

Lee *et al.* [23] and Kaviani and Shirani [24] on all testing sequences, and also achieves up to 2.2dB (for *City*) and 1.5 dB (for *Flower* and *Paris*) PSNR gain respectively. Since the pixel-wise motion field is used in HOMDF as well as Lee *et al.*'s [23] and Kaviani and Shirani's [24] methods, these advantages should be attributed to the desirable property of a higher model for intensity and position in improving the precision for pixel. For the synthesized videos, our method obtains superior results with the PSNR of 37.54dB and 56.50dB for *AccMot* and *VarIllu* respectively, while Lee *et al.* [23] achieves 35.76dB and 56.14dB. It shows that our method has much more advantages in dealing with complicated motion scenarios.

In terms of SSIM, as presented in TABLE II, HOMDF obtains better performance among the competing methods for most of the sequences except the *Waterfall*. In the *Waterfall*

TABLE II
COMPARISON OF VARIOUS FRUC ALGORITHMS IN TERMS OF AVERAGE SSIM OVER CIF FORMAT VIDEOS

Sequence	*Acc -Mot*	*Var -Illu*	*Akiyo*	*Bus*	*City*	*Coast -guard*	*Conta -iner*	*Flower*	*Foot -ball*	*Fore -man*	*Hall*	*Ice*	*MobileNews*	*Paris*	*Silent*	*Soccer*	*Stefan*	*Water -fall*
(total frames)	(40)	(100)	(300)	(150)	(300)	(300)	(300)	(300)	(260)	(300)	(300)	(300)	(300)	(300)	(300)	(300)	(300)	(300)
Kim [28]	0.941	0.988	0.994	0.864	0.910	0.932	**0.988**	0.918	0.644	0.894	0.952	0.944	0.912	0.973	0.969	0.955	0.838	0.864	0.940
Kaviani [24]	0.978	0.993	**0.996**	0.896	0.936	0.866	0.956	0.958	0.704	0.929	0.945	0.967	0.950	0.977	0.977	0.962	0.862	0.875	0.961
Lee [23]	0.984	**0.999**	**0.996**	0.930	0.957	0.951	**0.988**	0.969	0.712	0.945	0.961	0.972	0.962	0.979	0.978	0.969	0.899	0.892	0.982
Guo [27]	0.983	0.991	**0.996**	0.903	0.940	0.952	0.987	0.965	0.708	0.925	0.955	0.973	0.962	0.980	0.982	0.970	0.840	0.807	**0.984**
HOMDF	**0.988**	**0.999**	**0.996**	**0.938**	**0.960**	**0.957**	**0.988**	**0.978**	**0.755**	**0.948**	**0.962**	**0.978**	**0.969**	**0.981**	**0.984**	**0.973**	**0.907**	**0.901**	0.983

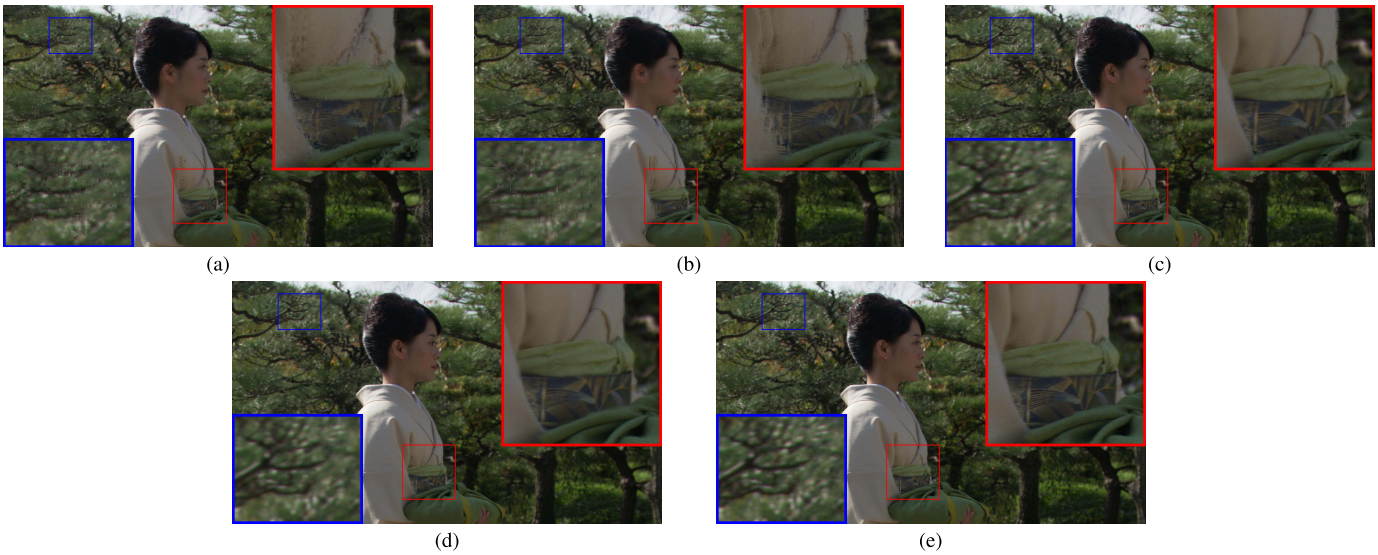


Fig. 6. Sujective comparison of *Kimono*'s 24-th frame. (Zoom in for better inspection). (a) 3DRS [13], 31.07dB. (b) Kim and Sunwoo [28], 30.96dB. (c) Kaviani and Shirani [24], 32.71dB. (d) HOMDF, 34.96dB. (e) Groundtruth.

TABLE III
AVERAGE PSNR (DB) AND SSIM RESULTS OF DIFFERENT FRUC ALGORITHMS OVER 1080P FORMAT VIDEOS (100 FRAMES)

Sequence	3DRS [13]		Kim [28]		Kaviani [24]		HOMDF	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Kimono	31.45	0.858	31.50	0.873	32.77	0.896	**34.96**	**0.929**
ParkScene	31.73	0.897	31.65	0.905	31.92	0.905	**35.61**	**0.934**
Sunflower	31.58	0.910	32.17	0.927	33.16	0.949	**35.46**	**0.959**
Bluesky	35.93	0.958	33.98	0.957	36.28	0.967	**39.09**	**0.974**

sequence, the irregular motions cannot be estimated correctly. However, the adaptive overlapped block MC method [31] used by Guo *et al.* [27] obtains smoothed interpolation and thereby achieves better SSIM result. Overall, since the SSIM index is known to approximate the structural similarity perceived by the human visual system, it shows that our method is able to preserve better structures.

The results of 1080p video sequences are illustrated in TABLE III. Our method has consistently obtained the best performances over all of the four sequences. Comparing our method with Kaviani and Shirani [24] which has the second best performance, an average 2.0dB gain over 1080p sequences is observed. The precise description for motion in our high order model plays a more important role in interpolating high resolution videos than interpolating low resolution ones. The accelerated motion can be captured and determined better in high resolution, and finally, contribute to improved performances.

Besides these objective results, the subjective comparisons are depicted in Figs.4∼6 on various scenes. Fig.4(a)∼(f) illustrate the interpolated results of *Mobile*'s 12-th frame, where multiple types of motion exist such as translation, rotation, acceleration as well as global motion. Kim and Sunwoo's [28] method produces obvious blocking artifacts around both the numbers and the rolling ball. Kaviani and Shirani's [24] result illustrated in (b) contains cracked numbers due to that their mismatch mechanism is sensitive to small structures. As for Lee *et al.*'s [23] result in (c), the occlusion area around the rolling ball is interpolated with erroneous MVs because the used outward MV by their method is not symmetrical with

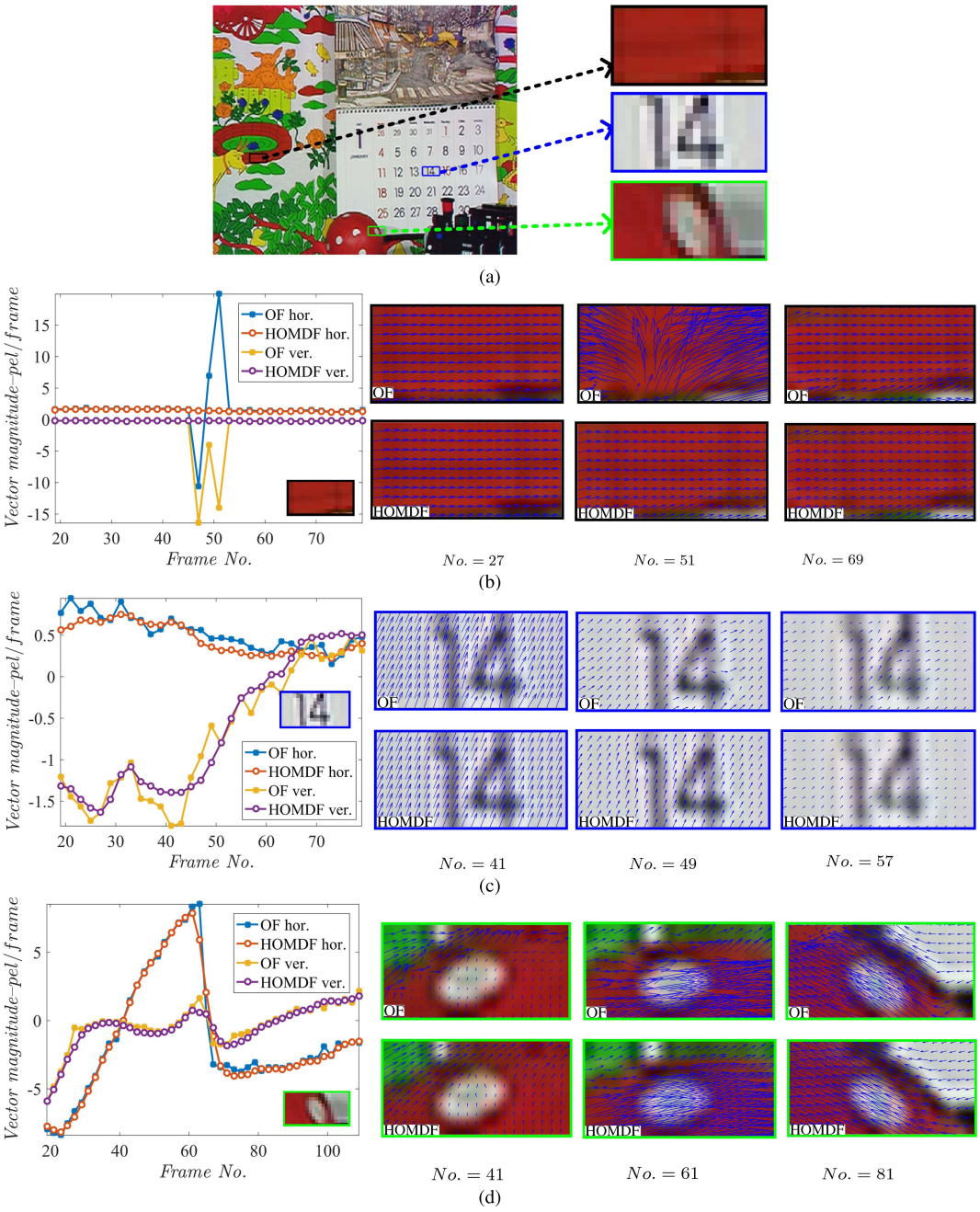


Fig. 7. Illustration of the estimated MVs of HOMDF and OF over multiple frames. (a) shows the three selected objects highlighted in black, blue, green boxes. The estimation details for objects are presented in (b), (c) and (d), respectively. (a) The three boxes containing various motions (global motion and zooming for black, accelerated motion for blue and rotation for green). (b) Wall in black box. (c) Number "14" in blue box. (d) Ball in green box.

inward MV under non-rigid cases. Guo *et al.*'s [27] result in (d) also shows cracked numbers because of the adopted bidirectional framework. In (e), HOMDF can alleviate these artifacts above, because the motion vector field is more dense and the dynamic filtering copes well with non-rigid motion by its adaptive noise variance evaluation.

Fig.5 presents the subjective comparison of the *Stefan* sequence, where a running player is with diverse motions among different parts (head, hand, foot, and body). Kaviani and Shirani's [24] and Kim and Sunwoo's [28] methods suffer from blockiness, and the tennis player's faces cannot be recognized very well. Lee *et al.* [23], Guo *et al.* [27] and HOMDF produce relatively clear faces. However, Lee *et al.* [23] does not preserve the minor structure of the player's hand well and Guo *et al.*'s [27] shows blurriness on the ball kid behind the player. In contrast, our method maintains both the textures for face and hands, and the clear boundaries of different objects are well kept.

Fig.6 shows the subjective comparison on 1080p videos. By inspecting the details of the walking woman closely, it can be found that there are blocking artifacts in Fig.6(a) and (b) caused by incorrect MVs of 3DRS [13] and

Kim and Sunwoo [28] around the woman. In contrast, Kaviani and Shirani's [24] results are more pleasant than theirs and HOMDF even produces better outcome in terms of PSNR. This is attributed to that the temporal coherence information leveraged by HOMDF is helpful to eliminate the occlusions occurred near motion boundaries. Moreover, for the homogeneous motion areas such as the background trees, there is also severe blockiness by block based method, but they are avoided by HOMDF since the delicate motion details are captured by our pixel-wise motion estimation.

Overall, either in objective or subjective comparison, over CIF or 1080p videos, our algorithm can produce favorable intermediate frames comparing to the state-of-the-art methods. Since our method can be analyzed from the motion and intensity variation parts, more experiments are conducted to show the details of HOMDF and show the effectiveness of motion estimation and intensity variation estimation.

B. Effectiveness of the Motion Estimation

1) Dynamic Filtering: As is known that the *Mobile* contains various motion types such as translational movement, rotation, zooming, acceleration, *etc*, this sequence is qualified enough to show the complexity of motion estimation and reveal how Kalman filter works. Fig.7 gives the results of tracked true motion trajectories of three objects over multiple frames. The motion vectors generated by Brox's Optical Flow (OF) [19] and our HOMDF are compared.

The black box contains a portion of the wall with continuous translational motion and zooming caused by camera movement. The line chart in the left of Fig.7(b) plots the horizontal/vertical vector magnitude of OF/HOMDF of the center pixel in the box at different *Frame No.* (frame number). While the six images in the right of Fig.7(b) depict the MVFs of OF/HOMDF at selected *Frame No.*s that we are particularly concerned about. As depicted by the line chart in Fig.7(b), the OF method produces an outlier at *No.* = 51. But HOMDF is robust and produces stable and accurate MVFs from *No.* = 27 to *No.* = 69 thanks to the Kalman filter's temporally iterative characteristic. The second box in blue contains number "14" on the calendar. The motion of this number by HOMDF is coherent in temporal and smooth in spatial comparing to the jittering trajectory and inhomogeneous MVF by OF as in Fig.7(c). The last box in green tracks a rolling ball as in Fig.7(d). The MVFs at *No.*= 41, 61 and 81 by OF is disordered near the rotating ball. However, by our HOMDF, the estimated motion shows great robustness. It can be found that our algorithm is able to handle various kinds of motions and produces both spatially and temporally robust MVFs.

2) True Motion of Occluded Regions: The proposed HOMDF is capable of estimating better MVF for the occluded areas. We compare HOMDF with Brox *et al.* [19] and Lee *et al.* [23]. The optical flow by Brox *et al.* [19] ignores the occlusions in two neighboring frames and usually produces over-smoothed flow fields. As illustrated in Fig.8(a) and (b), the vector flow fields by Brox *et al.* [19] are represented by colors with the visualization method suggested by Liu *et al.* [45]. The motion vectors in the occluded regions such as the boundary of the tree are noisy and don't align well with object

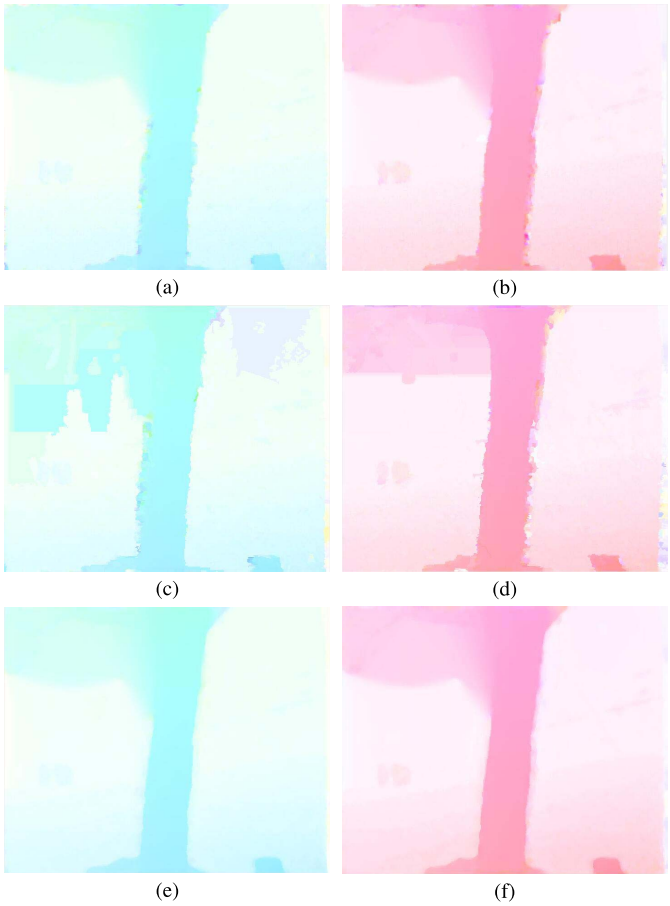


Fig. 8. Illustration of different optical flow estimation algorithms. Relative movement between the tree and the background causes occluded regions near the boundary of the tree. Image (a), (c), (e) in the left column are the forward flow while image (b), (d), (f) in the right column are the backward flow.

edges. Lee *et al.*'s [23] method attempt to refine Brox's result by selecting the best MV from spatial neighborhoods. It is effective for refining the motion boundaries at the cover side, but fails at the uncovered side as shown in Fig.8(c) and (d). More importantly, their method can not guarantee a smooth vector field at coherent motion areas. In contrast, in (e) and (f) by our method, the flow field is well estimated for both the cover and uncover side and the flow boundary coincides well with the object boundary. This achievement is attributed to the temporal information used by dynamic filtering. In the occluded regions, although the pixels have no correspondences in the next frame, they have truthful motion in the previous frames. Therefore, the prior estimation of dynamic filtering can utilize this temporal information to make a robust estimate.

C. Effectiveness of Intensity Variation Estimation

To better illustrate the effectiveness and accuracy of estimating intensity variation, a synthesized video sequence named as *VarIllum* is used to mimic a variant illumination scenario. We take a frame of *Flower* and perform gamma transformation $\left(\mathrm{I}' = 255(\frac{\mathrm{I}}{255})^{\gamma}\right)$ on it with gamma coefficient γ ranging from 0.1 to 10.0 at the step size of 0.1. These generated 100 frames are then assembled to constitute *VarIllum* sequence. In Fig.9, we show the estimated AR model parameters during the interpolation of 12-th frame. As shown by Fig.9(a), three

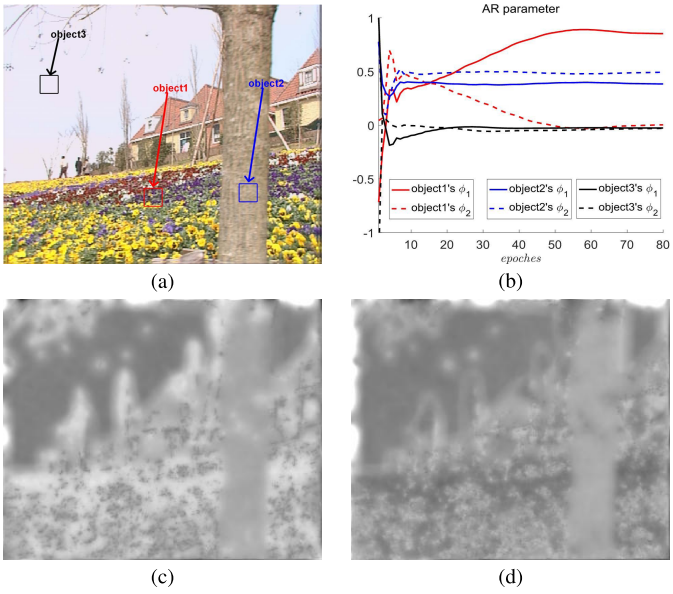


Fig. 9. AR model parameters of intensity variation by HOMDF. (a) Ground truth intermediate frame. (b) Convergence over epochs. (c) The first AR parameter plane ϕ_1. (d) The second AR parameter plane ϕ_2.

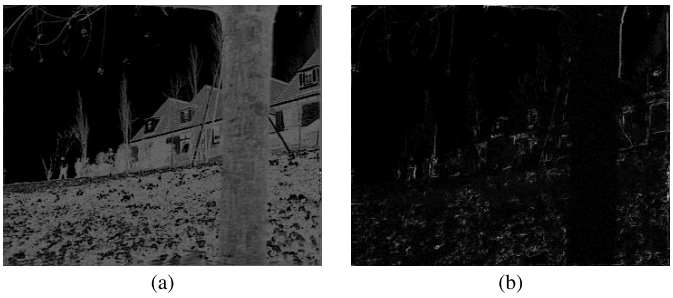


Fig. 10. Illustration of the reconstruction error of the intermediate frame. (a) Kaviani and Shirani [24]. (b) HOMDF.

objects, namely object1 in red, object2 in blue, and object3 in black are selected as examples to illustrate the intermediate results in iterative minimization by gradient descent algorithm. The pixel-wise parameter planes for ϕ_1 and ϕ_2 reach to convergence and are illustrated in Fig.9(c) and (d). And with these AR parameters, the 12-th frame can be reconstructed. We compare the reconstruction error of Kaviani and Shirani's [24] and ours methods in Fig.10. For black object3 in the sky with small intensity variation, the estimated ϕ_1 and ϕ_2 are zeros, and both methods work fine with imperceptible reconstruction errors. While for red object1 in the garden and blue object2 on the tree, with a good estimation of intensity variation, our method's reconstruction error is reduced compared to Kaviani and Shirani's [24]. It shows that by estimating the intensity variation, our algorithm can produce better results for frame interpolation.

VI. Conclusion

This paper has proposed a novel method for frame rate up conversion. A high order model for video pixel's intensity and position has been established. And to solve the multivariate energy objective for the coefficients in this high order model, we have proposed a dynamic filtering method based on the temporal coherence of videos. With the Bayes' Theorem, the optimal estimation for these coefficients is divided into prior and maximum likelihood estimation parts. It not only simplifies the estimation but also accomplishes an effective utilization of temporal information. It has been shown that our method is capable of handling the frame interpolation problems, even at some complicated scenarios such as motion acceleration and brightness variation. The estimated motion has been more fluent in temporal, and the interpolated frames have much better visual quality. Experimental results and analysis have validated the advantages of our method over the state-of-the-art ones.

References

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[2] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, and C. W. Chen, "Enabling adaptive high-frame-rate video streaming in mobile cloud gaming applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1988–2001, Dec. 2015.

[3] J. Wu, C. Yuen, N.-M. Cheung, J. Chen, and C. W. Chen, "Modeling and optimization of high frame rate video transmission over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2713–2726, Apr. 2016.

[4] Y.-H. Cho, H.-Y. Lee, and D.-S. Park, "Temporal frame interpolation based on multiframe feature trajectory," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2105–2115, Dec. 2013.

[5] D. Guo, L. Shao, and J. Han, "Feature-based motion compensated interpolation for frame rate up-conversion," *Neurocomputing*, vol. 123, pp. 390–397, Jan. 2014.

[6] S. Guo, C. Qiu, and X. Ye, "A kind of global motion estimation algorithm based on feature matching," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, Aug. 2009, pp. 107–111.

[7] C. Tang, R. Wang, and W. Wang, "Adaptive motion estimation order for frame rate up-conversion," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 2992–2995.

[8] K. Kim, M. Kim, D. Kim, and W. W. Ro, "True motion compensation with feature detection for frame rate up-conversion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2260–2264.

[9] R. Castagno, P. Haavisto, and G. Ramponi, "A method for motion adaptive frame rate up-conversion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 5, pp. 436–446, Oct. 1996.

[10] S. J. Kang, D. G. Yu, and Y. H. Kim, "Phase correlation-based motion estimation using variable block sizes for frame rate up-conversion," in *Proc. Int. Tech. Conf. Circuits/Syst., Comput., Commun.*, vol. 3. 2007, pp. 1399–1400.

[11] M. Biswas and T. Nguyen, "A novel motion estimation algorithm using phase plane correlation for frame rate conversion," in *Proc. Conf. Rec. 36th Asilomar Conf. Signals, Syst. Comput.*, vol. 1. Nov. 2002, pp. 492–496.

[12] M.-J. Chen, L.-G. Chen, and T.-D. Chiueh, "One-dimensional full search motion estimation algorithm for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 5, pp. 504–509, Oct. 1994.

[13] G. de Haan, P. W. A. C. Biezen, H. Huijgen, and O. A. Ojo, "True-motion estimation with 3-D recursive search block matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, no. 5, pp. 368–379, Oct. 1993.

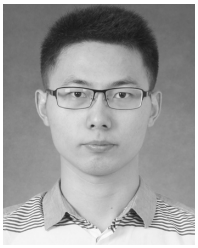
[14] K. M. Nam, J.-S. Kim, R.-H. Park, and Y. S. Shim, "A fast hierarchical motion vector estimation algorithm using mean pyramid," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 4, pp. 344–351, Aug. 1995.

[15] S. Zhu and K.-K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.

[16] X. Q. Gao, C. J. Duanmu, and C. R. Zou, "A multilevel successive elimination algorithm for block matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 501–504, Mar. 2000.

[17] S. Dikbas, T. Arici, and Y. Altunbasak, "Fast motion estimation with interpolation-free sub-sample accuracy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 7, pp. 1047–1051, Jul. 2010.

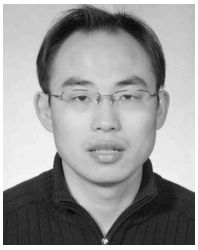
[18] C. Wang, L. Zhang, Y. He, and Y.-P. Tan, "Frame rate up-conversion using trilateral filtering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 886–893, Jun. 2010.
[19] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2004, pp. 25–36.
[20] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1744–1757, Sep. 2012.
[21] S. H. Keller, F. Lauze, and M. Nielsen, "Temporal super resolution using variational methods," in *High-Quality Visual Experience*. New York, NY, USA: Springer, 2010, pp. 275–296.
[22] C. Tang, R. Wang, W. Wang, and W. Gao, "A new frame interpolation method with pixel-level motion vector field," in *Proc. IEEE Vis. Commun. Image Process. Conf. (VCIP)*, Dec. 2014, pp. 350–353.
[23] W. H. Lee, K. Choi, and J. B. Ra, "Frame rate up conversion based on variational image fusion," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 399–412, Jan. 2014.
[24] H. R. Kaviani and S. Shirani, "Frame rate upconversion using optical flow and patch-based reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1581–1594, Sep. 2016.
[25] J. Astola, P. Haavisto, and Y. Neuvo, "Vector median filters," *Proc. IEEE*, vol. 78, no. 4, pp. 678–689, Apr. 1990.
[26] L. Alparone, M. Barni, F. Bartolini, and V. Cappellini, "Adaptively weighted vector-median filters for motion-fields smoothing," in *Proc. IEEE ICASSP*, vol. 4. May 1996, pp. 2267–2270.
[27] Y. Guo, L. Chen, Z. Gao, and X. Zhang, "Frame rate up-conversion using linear quadratic motion estimation and trilateral filtering motion smoothing," *J. Display Technol.*, vol. 12, no. 1, pp. 89–98, Jan. 2016.
[28] U. S. Kim and M. H. Sunwoo, "New frame rate up-conversion algorithms with low computational complexity," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 384–393, Mar. 2014.
[29] D. Choi, W. Song, H. Choi, and T. Kim, "MAP-based motion refinement algorithm for block-based motion-compensated frame interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 10, pp. 1789–1804, Oct. 2016.
[30] M. T. Orchard and G. J. Sullivan, "Overlapped block motion compensation: An estimation-theoretic approach," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 693–699, Sep. 1994.
[31] B.-D. Choi, J.-W. Han, C.-S. Kim, and S.-J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 407–416, Apr. 2007.
[32] D. Wang, L. Zhang, and A. Vincent, "Motion-compensated frame rate up-conversion—Part I: Fast multi-frame motion estimation," *IEEE Trans. Broadcast.*, vol. 56, no. 2, pp. 133–141, Jun. 2010.
[33] T. H. Tsai and H. Y. Lin, "High visual quality particle based frame rate up conversion with acceleration assisted motion trajectory calibration," *J. Display Technol.*, vol. 8, no. 6, pp. 341–351, Jun. 2012.
[34] Y. Zhang, D. Zhao, X. Ji, R. Wang, and X. Chen, "A spatio-temporal autoregressive frame rate up conversion scheme," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep./Oct. 2007, pp. 441–444.
[35] Y. Zhang, D. Zhao, S. Ma, R. Wang, and W. Gao, "A motion-aligned auto-regressive model for frame rate up conversion," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1248–1258, May 2010.
[36] Y. Zhang, L. Xu, X. Ji, and Q. Dai, "A polynomial approximation motion estimation model for motion-compensated frame interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 8, pp. 1421–1432, Aug. 2016.
[37] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, D, J. Basic Eng.*, vol. 82, pp. 35–45, 1960.
[38] S.-K. Weng, C.-M. Kuo, and S.-K. Tu, "Video object tracking using adaptive Kalman filter," *J. Vis. Commun. Image Represent.*, vol. 17, no. 6, pp. 1190–1208, Dec. 2006.
[39] Y. Motai, S. K. Jha, and D. Kruse, "Human tracking from a mobile agent: Optical flow and Kalman filter arbitration," *Signal Process., Image Commun.*, vol. 27, no. 1, pp. 83–95, 2012.
[40] C. Paramanand and A. N. Rajagopalan, "Depth from motion and optical blur with an unscented Kalman filter," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2798–2811, May 2011.
[41] C.-M. Kuo, S.-C. Chung, and P.-Y. Shih, "Kalman filtering based rate-constrained motion estimation for very low bit rate video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 3–18, Jan. 2006.
[42] W. N. Lie and Z. W. Gao, "Video error concealment by integrating greedy suboptimization and Kalman filtering techniques," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 8, pp. 982–992, Aug. 2006.
[43] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1, pp. 503–528, 1989.
[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[45] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.



Wenbo Bao recieved the B.S. degree in electronic information engineering from the Huazhong University of Science and Technology, Hubei, China, in 2014. He is currently pursuing the Ph.D. degree in electrical engineering with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision, machine learning, and video processing.

Xiaoyun Zhang received the B.S. and M.S. degrees in applied mathematics from Xi'an Jiaotong University in 1998 and 2001, respectively, and the Ph.D. degree in pattern recognition from Shanghai Jiao Tong University, China, in 2004. Her Ph.D. thesis was nominated as the National 100 Best Ph.D. Theses of China. Her research interests include computer vision and pattern recognition, image and video processing, and digital TV system. Her current research focuses on image processing and video compression.

Li Chen received the B.S. and M.S. degrees from Northwestern Polytechnical University, Xi'an, China, in 1998 and 2000, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, China, in 2006, all in electrical engineering. His current research interests include image and video processing, DSP and VLSI for image, and video processing.



Lianghui Ding received the Ph.D. degree from Shanghai Jiao Tong University, China, in 2009. From 2009 to 2010, he was a Researcher in signals and systems with Uppsala University, Sweden. He is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University. His research interests include wireless communication, wireless power transfer, and so on.

Zhiyong Gao received the B.S. and M.S. degrees in electrical engineering from the Changsha Institute of Technology, Changsha, China, in 1981 and 1984, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 1989. From 1994 to 2010, he held several senior technical positions in U.K., including a Principal Engineer with Snell and Wilcox, Petersfield, from 1995 to 2000, a Video Architect with 3DLabs, Egham, from 2000 to 2001, a Consultant Engineer with Sony European Semiconductor Design Center, Basingstoke, from 2001 to 2004, and a Digital Video Architect with Imagination Technologies, Kings Langley, from 2004 to 2010. Since 2010, he has been a Professor with Shanghai Jiao Tong University. His research interests include video processing and its implementation, video coding, digital TV, and broadcasting.