Novel Integration of Frame Rate Up Conversion and HEVC Coding Based on Rate-Distortion Optimization

Guo Lu, *Student Member, IEEE*, Xiaoyun Zhang, *Member, IEEE*, Li Chen, *Member, IEEE*, and Zhiyong Gao

***Abstract*— Frame rate up conversion (FRUC) can improve the visual quality by interpolating new intermediate frames. However, high frame rate videos by FRUC are confronted with more bitrate consumption or annoying artifacts of interpolated frames. In this paper, a novel integration framework of FRUC and high efficiency video coding (HEVC) is proposed based on rate-distortion optimization, and the interpolated frames can be reconstructed at encoder side with low bitrate cost and high visual quality. First, joint motion estimation (JME) algorithm is proposed to obtain robust motion vectors, which are shared between FRUC and video coding. What's more, JME is embedded into the coding loop and employs the original motion search strategy in HEVC coding. Then, the frame interpolation is formulated as a rate-distortion optimization problem, where both the coding bitrate consumption and visual quality are taken into account. Due to the absence of original frames, the distortion model for interpolated frames is established according to the motion vector reliability and coding quantization error. Experimental results demonstrate that the proposed framework can achieve 21% ∼ 42% reduction in BDBR, when compared with the traditional methods of FRUC cascaded with coding.**

***Index Terms*— Frame rate up conversion, HEVC, motion estimation, rate-distortion optimization, high frame rate.**

I. Introduction

TEMPORAL redundancy is one of the most important properties of video signals, many video processing algorithms and video coding systems exploit this characteristic. In frame rate up conversion (FRUC), temporal redundancy is utilized to improve visual quality, while HEVC coding [1] employs it to increase coding efficiency.

Frame rate up conversion interpolates new frames between two consecutive original frames, thus it can be utilized to reduce the motion blur of hold-type displays, e.g., liquid crystal display. In addition, recent growth of multimedia devices and display technology has also led to the demand for FRUC, where it generates high frame rate (HFR) video with better visual quality [2].

Most FRUC methods utilize motion information to perform interpolation along motion trajectory [3]–[14], these methods are also called motion-compensated FRUC (MC-FRUC). A typical MC-FRUC consists of two procedures: motion estimation (ME) and motion compensation interpolation (MCI). ME aims at estimating motion vectors (MVs) to represent the motion trajectory between consecutive frames, and MCI uses the estimated MVs to generate interpolated frames. Among these ME methods, block matching algorithm is most widely used [3]–[8] due to its simplicity and hardware-friendly implementation. Kang *et al.* [4] proposed a dual ME algorithm to enhance the accuracy of motion vectors. Liu *et al.* [5] utilized a multiple hypotheses FRUC scheme for estimating the intermediate frame with maximum *a posteriori* probability. Dikbas *et al.* [7] tracked the true motion trajectory by imposing implicit and explicit smoothness constrains on block matching algorithm. Recently, optical flow estimation is employed to provide more accurate motion vector field (MVF). In [9], multiple frames are generated based on the optical flow and fused to get the final interpolated frame.

In addition to motion estimation, a lot of MCI techniques have also been carried out for developing effective interpolation [3], [11], [12]. Choi *et al.* [3] proposed an adaptive overlapped block motion compensation (AOBMC), which determines weighting coefficients according to the reliability levels of neighboring MVs. Wang *et al.* [11] utilized the trilateral filter to correct the unreliable pixels and reducing ghost artifacts. Besides, a motion-aligned autoregressive model (MAAR) was proposed in [12], where each pixel is approximated by a linear combination of the pixels in reference frames.

However, in order to obtain accurate MVs or faithful interpolated frames, the traditional FRUC algorithms have to employ complicated ME methods and elaborated interpolation schemes, which lead to high computational complexity. Therefore some research [15], [16] exploited the MVs in compressed bitstreams to alleviate the computation burden of FRUC. Since received MVs are generated at encoder side by minimizing the prediction error instead of finding the true

Manuscript received December 24, 2016; revised August 13, 2017 and October 11, 2017; accepted October 18, 2017. Date of publication October 30, 2017; date of current version November 14, 2017. This work was supported in part by National Natural Science Foundation of China under Grant 61771306, Grant 61521062, and Grant 61527804, in part by the Chinese National Key S&T Special Program under Grant 2013ZX01033001-002-002, in part by the STCSM under Grant 17DZ1205602, in part by the 111 Project under Grant B07022, and in part the Shanghai Key Laboratory of Digital Media Processing and Transmissions under Grant STCSM15DZ2270400. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Catarina Brites. *(Corresponding author: Xiaoyun Zhang; Li Chen.)*

The authors are with the Institute of Image Communication and Network Engineering, Department of Electronic Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China (e-mail: luguo2014@sjtu.edu.cn; xiaoyun.zhang@sjtu.edu.cn; hilichen@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.2017.2767782

motion, FRUC method that directly employs the received MVs may suffer from annoying artifacts. Hence a lot of methods have been proposed to refine the received MVs at decoder side. In [15], Huang *et al.* proposed a hierarchical MV processing method that exploits both the residual information and bidirectional prediction difference. In [16], a correlation-based motion vector processing method was proposed to detect and correct those unreliable motion vectors. However, the reconstructed video quality depends on the decoded frames. If videos are transmitted in limited bandwidth and the quality of decoded frames has already been deteriorated, it is difficult to reconstruct good quality intermediate frames even a sophisticated post-processing method is adopted. In [17], Krishnamurthy *et al.* proposed sending side information to inform the decoder which interpolation scheme should be used. A similar concept of this encoder assisted framework was employed in [18], where the encoder decides how to interpolate the frame and sends that side information to the decoder. These methods [17], [18] can achieve better performance than the FRUC only using decoder information. Nonetheless, this may not be a feasible approach in practice. First, it is difficult for an encoder to estimate MVs accurately by solely utilizing block-based motion estimation. More importantly, these approaches [17], [18] are not standard compatible and need a non-standard decoder, therefore their further applications are impeded.

In addition, either the decoder based FRUC [15], [16] or encoder assisted FRUC [17], [18] methods, their hardware implementations remain a challenging issue. For example, in the terminal devices such as TVs, high quality FRUC has been traditionally designed and implemented using Application Specific Integrated Circuit (ASIC), while the video is decoded by specific hardware module integrated within a system on chip (SoC). Such hardware architecture makes it difficult to share the MVs between the decoder and FRUC. Therefore, although the computational complexity is reduced significantly, these methods [15]–[18] are not widely used in practical video applications.

Recently, temporal frame interpolation technique is employed by some scalable video coding methods [19], [20]. In [19], HoangVan *et al.* utilized a motion compensated temporal interpolation scheme to generate the reference frames. It should be mentioned that their methods aimed at enhancing the existing B-slices coding performance rather than generating non-exist intermediate frames. In [20], Rufenacht *et al.* proposed a novel motion vectors anchoring scheme, where FRUC function can be performed at decoder side with arbitrary framerates and high quality.

A worthwhile alternative is to make an effort to conduct FRUC at encoder side, which means the converted high frame rate video can be displayed directly for consumers. Most display devices, such as the mobile device, do not provide hardware implemented FRUC function due to the strict limit on power consumption. In this approach, the computational complexity is transferred from decoder side to encoder side, which is very friendly for low power display devices. To generate high frame rate video at encoder side, a straightforward way is to carry out traditional FRUC algorithms [3]–[13] before encoding in a cascaded way. However, the data volume at encoder side will increase significantly and more transmission bandwidth are demanded, which poses a challenge to the transmission network.

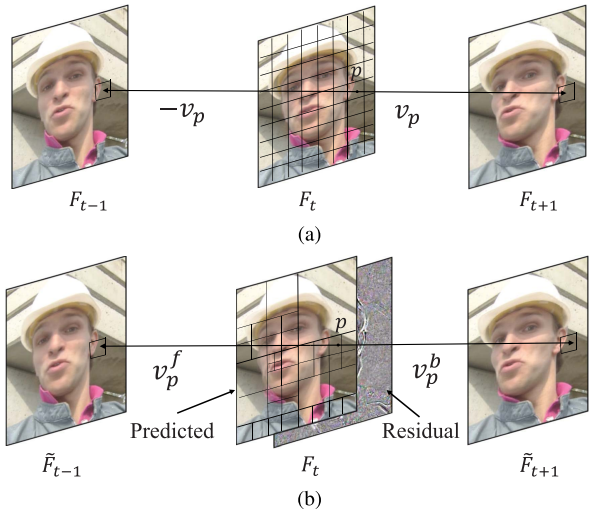


Fig. 1. Comparison between bi-prediction in FRUC and HEVC coding. (a) Bilateral ME for interpolated frame F_t in FRUC. (b) Bi-prediction for F_t in HEVC coding.

Therefore it is necessary to combine FRUC and coding in a more tightly coupled way, which aims at providing high quality FRUC and reducing the bitrate cost at encoder side simultaneously. To the best of our knowledge, there is no such a framework that explicitly integrates FRUC into the encoder loop while maintains standard compatible. In this paper, we present the insight into how video coding and frame rate up conversion can be tackled within an integration framework. In the proposed method, MVs are shared between coding and FRUC for the interpolated frame, and a joint motion estimation method that embedded into the encoder loop is utilized to track true motion trajectory. In addition to traditional block based ME, feature matching and motion segmentation are adopted in joint ME to improve MV accuracy for complicated motion regions and textureless regions. More importantly, due to the absence of original frames, a novel distortion model for interpolated frames is established according to MV reliability and coding quantization error. Then interpolated frames can be optimized by a rate-distortion optimization (RDO) procedure with high coding efficiency. Experimental results show that the jointly optimizing of HEVC and FRUC in this way can improve the video quality and reduce the bitrate cost for interpolated frames.

This paper is organized as follows. Section II presents the integration framework of FRUC and HEVC. Section III introduces the joint motion estimation. The novel distortion model for interpolated frames is described in Section IV. Section V presents the experimental results. Finally, Section VI concludes this paper.

II. Proposed Integration Framework

A. Motion-Compensated Frame Rate Up Conversion

In Fig. 1(a), F_{t-1} and F_{t+1} are the forward and backward reference frames for F_t. In general, motion-compensated frame

rate up conversion method assumes that the motion trajectory is continuous between consecutive frames. Therefore, motion from F_t to F_{t-1} and motion from F_t to F_{t+1} are antisymmetric. To obtain accurate MVs, block based motion estimation (BME) is most widely used. There are two kinds of BME methods: unilateral ME [5], [7], [8] and bilateral ME [3], [4]. Bilateral ME estimates the MVs for interpolated frame directly, while unilateral ME estimates MVs for forward or backward reference frames.In this paper, we take bilateral ME as example to show the basic FRUC model.

Let $\boldsymbol{v_p}$ denote the MV at location $\boldsymbol{p}$ for interpolated frame F_t. To obtain $\boldsymbol{v_p}$, F_t is divided into non-overlapping blocks and then, for each block, $\boldsymbol{v_p}$ is estimated as follows,

$$\boldsymbol{v_p} = \arg\min_{\boldsymbol{v}\in R} \mathcal{C}(\boldsymbol{v}, \mathcal{B}_{\boldsymbol{p}}) \tag{1}$$

where $\mathcal{C}(\boldsymbol{v}, \mathcal{B}_{\boldsymbol{p}})$ is the cost function to measure the similarity of blocks that indicating by $\boldsymbol{v}$. $\mathcal{B}_{\boldsymbol{p}}$ is the block that containing $\boldsymbol{p}$. R is the motion candidate set. Various cost functions [3], [4], [6], [7] have been proposed to improve the accuracy of ME. However, without loss of generality, $\mathcal{C}$ can be basically defined as follows,

$$\mathcal{C}(\boldsymbol{v}, \mathcal{B}_{\boldsymbol{p}}) = \sum_{\boldsymbol{p}\in\mathcal{B}_{\boldsymbol{p}}} |F_{t-1}(\boldsymbol{p}-\boldsymbol{v}) - F_{t+1}(\boldsymbol{p}+\boldsymbol{v})| \tag{2}$$

Generally, the pixel values along motion trajectory change little in a very short time, therefore interpolated frame $\hat{F}_t(\boldsymbol{p})$ can be obtained by averaging the pixel values in its reference frames, which is given by,

$$\hat{F}_t(\boldsymbol{p}) = \frac{1}{2}\cdot[F_{t-1}(\boldsymbol{p}-\boldsymbol{v_p}) + F_{t+1}(\boldsymbol{p}+\boldsymbol{v_p})] \tag{3}$$

B. Bi-Prediction in HEVC

Similar to prior video coding standards, such as H.264 [21], the HEVC design follows the classic block-based hybrid video coding architecture [1], where inter prediction is employed to exploit temporal statistical dependencies. In order to improve the compression efficiency, a lot of novel coding tools have been adopted in HEVC. In particular, the coding structure of HEVC has been extended from a traditional macroblock concept to a hierarchical block partitioning scheme [22] that supports block size up to 64×64 pixels. As depicted in Fig. 1(b), motion estimation for Bi-prediction performs in various block sizes. $\tilde{F}_{t-1}$ and $\tilde{F}_{t+1}$ are the reconstructed reference frames which are used to predict F_t. $\boldsymbol{v_p^f}$ and $\boldsymbol{v_p^b}$ are the forward and backward MVs that estimated for position $\boldsymbol{p}$ by HEVC coding.

In HEVC, MVs can be estimated through *Merge mode* or *Non-merge mode*. *Merge mode* derives the motion information from spatial or temporal neighboring candidates. Therefore Prediction Units (PUs) can share the same motion information, which improves the coding efficiency. In *Non-merge mode*, encoder has to estimate new MVs by minimizing the prediction error of current PU. Since the full search algorithm brings high computational complexity, a lot of methods have been proposed [23], [24] to accelerate the ME procedure for *Non-merge mode*. After the motion estimation in coding, reconstructed block can be acquired by averaging the matching blocks that indicted by estimated MVs as follows,

$$\tilde{F}_t(\boldsymbol{p}) = \frac{1}{2}\cdot[\tilde{F}_{t-1}(\boldsymbol{p}+\boldsymbol{v_p^f}) + \tilde{F}_{t+1}(\boldsymbol{p}+\boldsymbol{v_p^b})] + Resi(\boldsymbol{p}) \tag{4}$$

where $Resi(\boldsymbol{p})$ is the residual value in position $\boldsymbol{p}$ and is compensated to reduce the prediction error in coding loop.

C. Proposed Integration of FRUC and HEVC

Comparing (3) and (4), we can discover that FRUC and Bi-prediction in HEVC both adopt the similar motion compensated technique in interpolating the intermediate frame or reconstructing the encoded frame. If $Resi(\boldsymbol{p})$ in (4) is not compensated, the encoded frame is actually interpolated by the adjacent reference frames. In this case, the procedure of encoding a frame is equivalent to the interpolation of a frame in FRUC. This motivates us to integrate FRUC within HEVC coding by a more compact manner.

Our framework aims at performing coding and FRUC simultaneously. Specifically, a video sequence at low frame rate (LFR) is input into the system and then it is up-converted and encoded in the proposed framework. Finally, a high frame rate (HFR) video sequence is output as a compressed bitstream which can be decoded by a standard HEVC decoder. The complete framework is shown in Fig. 2.

There are three contributions of the proposed framework. First, comparing to traditional cascaded methods, FRUC is integrated within HEVC encoder in a tightly coupled way, which provides the potential to generate interpolated frames with low bitrate cost and high visual quality. Second, as depicted in Fig. 2, a joint ME algorithm is embedded into the encoder loop and targeted at obtaining accurate and robust motion vector field. Specifically, the MVs derived from feature matching are employed to improve the robustness for complicated motions and large displacement motions. Meanwhile, motion segmentation map is extracted from encoder information and utilized to preserve the spatial piecewise smoothness. Third, we propose a novel distortion model for interpolated frames by considering both MV reliability and quantization error. Based on this distortion model, a new RDO criterion for interpolated frame is incorporated into the encoder gracefully. Then the frame interpolation can be solved by the rate-distortion optimization within HEVC coding loop.

III. Joint Motion Estimation

Motion estimation plays an important role in FRUC since the estimated MVs greatly influence the quality of interpolated frames. Recently, abundant research of motion estimation algorithms, especially the optical flow estimation, have been proposed [25]–[28]. Although these methods improve the accuracy of MVs and solve some challenging problems, they also lead to high computational complexity.

It is noticeable that the encoder itself already has a block based ME as mentioned in II-B, therefore it will be more efficient to exploit the original ME rather than performing an additional complicated ME algorithm. However, ME in coding requires original frames which are not available for FRUC.

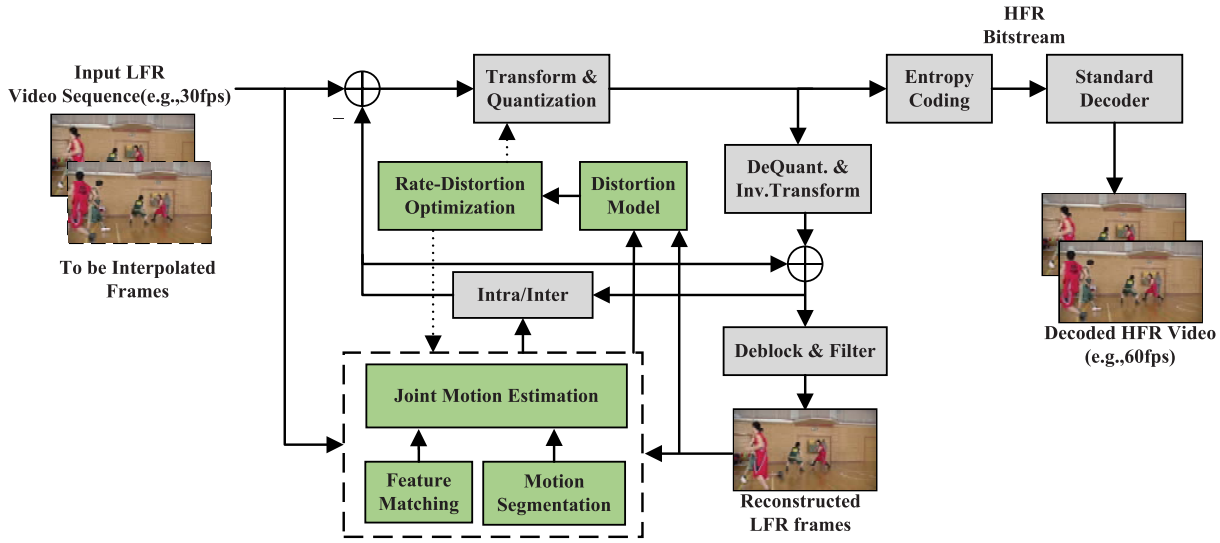


Fig. 2. The proposed novel integration framework of FRUC and HEVC. The key contributions are highlighted in green boxes. *(Best viewed in color)*

What's more, the criterion of ME in coding aims at minimizing the prediction error rather than tracking the true motion trajectory.

To address these problems, we only utilize the motion search strategy in coding process and combine multiple information to estimate accurate and robust MVs. Namely,

$$E(\boldsymbol{v}) = E_D(\boldsymbol{v}) + \alpha E_M(\boldsymbol{v}) + \beta E_S(\boldsymbol{v}) \tag{5}$$

where $\boldsymbol{v}$ is the generated MV through *Merge mode* or *Non-merge mode* in HEVC coding loop. $E(\boldsymbol{v})$ represents the energy of $\boldsymbol{v}$ and is used to measure the MV reliability. $E_D(\boldsymbol{v})$, $E_M(\boldsymbol{v})$ and $E_S(\boldsymbol{v})$ are the data term, matching term and smoothness term, respectively. α and β denote the weighting parameters. The MVs estimated through the proposed joint ME can be encoded by HEVC and utilized to reconstruct the interpolated frames at the same time, i.e., both the coding procedure in HEVC and prediction (reconstruction) procedure for interpolated frames share the same MVF. It should be mentioned that the novelty of our joint ME algorithm lies in the efficient and graceful integration of ME procedure in FRUC and coding. The advantages of the proposed energy cost come from that these three terms can complement each other and the disadvantage of one term can be overcome by the other two terms.

A. Data Term

As mentioned in Section II-A, FRUC generally assumes that the motion is continuous between consecutive frames and the intensity of pixel along motion trajectory remains unchanged. Therefore the data term can be defined as follows:

$$E_D(\boldsymbol{v}) = \frac{1}{N} \sum_{\boldsymbol{p} \in \mathcal{B}_p} |\tilde{F}_{t-1}(\boldsymbol{p} - \boldsymbol{v}) - \tilde{F}_{t+1}(\boldsymbol{p} + \boldsymbol{v})| \tag{6}$$

Here, $\tilde{F}_{t-1}$ and $\tilde{F}_{t+1}$ are the forward and backward reconstructed reference frames, respectively. $|\cdot|$ denotes the L_1 norm. $\boldsymbol{p}$ is the pixel location in the current Block $\mathcal{B}_p$. Since the HEVC adopts the hierarchical partition scheme [1], [22], thus the block size of $\mathcal{B}_p$ can vary from 8×8 to 64×64 pixels.

In block based ME, block size affects the performance significantly [5]. In general, a small block size is useful to estimate MVs in complex scene and describe detail motion accurately. Nevertheless it is also likely to introduce motion ambiguity. On the contrary, large block size can reduce this artifact by containing more structure information, however it may lead to inaccuracy at motion boundary. Therefore, recent block matching methods [5], [29] employ multiple block sizes to obtain a more robust MVF. Considering that the joint motion estimation is embedded into the HEVC coding loop, it is efficient and straightforward to perform variable block size motion estimation by utilizing the hierarchical partition scheme in coding and improve the accuracy of MVF.

$E_D(\boldsymbol{v})$ in (6) measures how well the MV matches by comparing the intensity consistency between the reference frames. It performs well for regions where have modest structural information and simple linear rigid motion [30]. However, the data term only utilizes the blockwise intensity information and cannot provide accurate MVs in all cases, especially for the complicated scenes. Another drawback of this bidirectional estimation is that the date term in (6) may fall into the minimum point when an object with complex texture and a background with homogeneous texture are simultaneously present. In this situation, MVs of homogeneous texture blocks are more likely to be chosen since they have a relative small date term cost [31]. Although a lot of methods have been proposed to tackle these problems [3], [8], [10], [15], the performance of ME still calls for improvement. In the next two subsections, we will introduce how to overcome the inherent limitations of date term and improve the proposed joint ME algorithm by descriptor matching and motion region segmentation.

B. Matching Term

Descriptor matching approaches, such as SIFT [32] or HOG [26], [33] matching have been extensively studied and

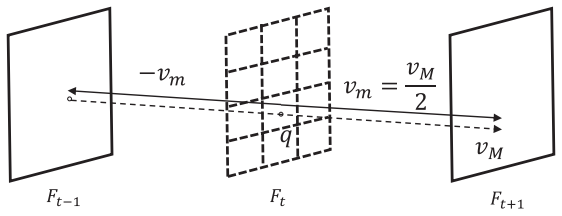


Fig. 3. Project the MV v_M from F_{t-1} to F_{t+1} into two halved MVs $-v_M/2$ and $v_M/2$ for Bi-prediction of the interpolated frame F_t.

applied to various domains [34]. The main reason is that these descriptors are highly distinctive. For example, a 128 dimensional vector is used to represent SIFT feature in each pixel. Therefore, matching by descriptors can provide more robust MVs when compared with methods utilizing raw pixel values.

In this paper, a matching term is incorporated into the proposed joint motion estimation method and the estimated MV $\boldsymbol{v}$ for interpolated frame is encouraged to be similar with the MV derived from descriptors matching. In our implementation, Deepmatching [35] is chosen as the matching algorithm.

Since the descriptor matching can only be performed between available frames, we firstly project MVs from Deepmatching to the interpolated frame. As depicted in Fig. 3, F_{t-1} and F_{t+1} are the reference frames, $\boldsymbol{v}_M$ is the MV obtained through Deepmatching. $\boldsymbol{q}$ is the pixel location that passed through by motion trajectory between reference frames. Therefore, MV $v_{\boldsymbol{m}}$ of Bi-prediction for position $\boldsymbol{q}$ can be obtained by halving $\boldsymbol{v_M}$ as shown in Fig. 3. Then the matching term is formulated as follows:

$$E_M(\boldsymbol{v}) = \mu\psi(\boldsymbol{v_m})\Psi(|\boldsymbol{v} - \boldsymbol{v_m}|) \tag{7}$$

where $\boldsymbol{v}$ and $\boldsymbol{v_m}$ are the estimated MV in HEVC coding and projected MV from Deepmatching. $\Psi(\cdot)$ is a robust penalizer and $\Psi(\boldsymbol{v}) = \sqrt{\boldsymbol{v^2} + \boldsymbol{\epsilon^2}}$ with $\boldsymbol{\epsilon} = 0.1$. Since the projected MVF for the interpolated frame is sparse, $\boldsymbol{v_m}$ may not be available for every block. Therefore μ is used to indicate the existence of $\boldsymbol{v_m}$. If no MV from Deepmatching is projected into $\mathcal{B}_p$, μ is equal to 0, otherwise it is equal to 1.

Additionally, in order to prevent imposing penalty due to an inaccurate MV from descriptor matching, the corresponding matching term is modified according to the accuracy of $\boldsymbol{v_m}$. Therefore each matching term is multiplied by a weight $\psi(\boldsymbol{v_m})$, which is defined as follows:

$$\psi(\boldsymbol{v_m}) = max(0, 1 - exp(\frac{E_D(\boldsymbol{v_m}) - Th}{2\sigma_1^2})) \tag{8}$$

where $E_D(\cdot)$ is defined in (6) and used to measure the accuracy of MV. σ_1 is a control parameter. If $E_D(\boldsymbol{v_m})$ is large than the threshold Th, then it is not necessary to add this penalty. However, if $\boldsymbol{v_m}$ is accurate enough, $\psi(\boldsymbol{v_m})$ will increase proportionally and the estimated MV $\boldsymbol{v}$ will approximate $\boldsymbol{v_m}$. In addition, if more than one MVs are projected into the current block $\mathcal{B}_p$, only the optimal $\boldsymbol{v_m}$ with the minimum $E_D(\cdot)$ value is selected for the matching term in (7).

The matching term contributes in the following ways. First, Deepmatching employs more descriptive features than data term, and can obtain robust motion vectors, especially for the complicated motion scenes. In addition, HEVC coding employs the TZSearch to perform ME for *Non-merge* mode. One major disadvantage of this predictive search method is that the estimated MV may not converge quickly or fall into local minimum point for large displacement. However, MVs from descriptor matching show the ability to capture large displacement [26], [27], [36], [37] and can facilitate ME in coding by providing accurate convergence direction.

C. Smoothness Term

Both data term (6) and matching term (7) only utilize local features, therefore the estimated $\boldsymbol{v}$ may be inaccurate when few local features can be employed, especially for the homogeneous region. However, it is known that the true MVF for natural video is spatial piecewise smoothness, which is consistent in most video sequences and uncorrelated with video content. Therefore, we take advantage of this global prior knowledge and employ the smoothness term to improve MV accuracy. To prevent over-smoothing at motion boundary, a straightforward way is to utilize motion segmentation information. Although recent algorithms show desirable performance [38], [39], they are too complex. Considering the proposed framework is integrated within HEVC encoder, it is feasible and more efficient to employ the bitstream information (e.g., encoded MVF) to perform motion segmentation and improve the reliability of estimated MVF. The proposed motion segmentation method consists of two procedures: matching MVs based coarse segmentation and encoded MVs based fine segmentation. In Section III-B, after Deepmatching, the matching MVs between the forward and backward reference frames are obtained. First, we extract several dominant MVs $\mathcal{D} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_n}\}$ by utilizing RANSAC algorithm [40] and the corresponding labels are denoted as $\mathcal{L} = \{\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, \ldots, \boldsymbol{\omega_n}\}$. Then the reference frame is divided into non-overlapped blocks and the MV for each block is assigned to the MV extracted by Deepmatching. After the division, a median filter is utilized to fill the blocks that have no MV assignment. Let $\boldsymbol{v_M}$ denote the MV in current block. Then the label $\boldsymbol{\omega}$ for current block is determined as follows:

$$\boldsymbol{\omega} = \boldsymbol{\omega}_j, \quad where\ j = \arg\min_{i \in \mathcal{H}} \|\boldsymbol{v_M} - \boldsymbol{v_i}\|^2 \tag{9}$$

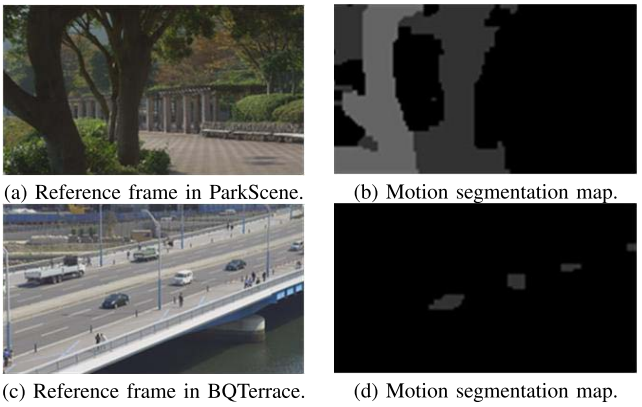
Here, $\mathcal{H} = \{1, 2, \ldots n\}$. After this procedure, a coarse segmentation map is constructed. Then it is refined by utilizing the encoded MVF in reference frame. The pipeline is similar to Chen *et al.* [41]. Let $\boldsymbol{v_c}$ denotes the MV that obtained through HEVC coding in current location. The goal is to find a label value with the maximum probability for each block, based on the encoded MV $\boldsymbol{v_c}$. Namely,

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{w} \in \mathcal{L}} P(\boldsymbol{\omega}|\boldsymbol{v_c}) \tag{10}$$

Based on the Bayes' theorem, the maximum *a posterior* estimate of $\boldsymbol{\omega}$ can be reformulated as

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{w} \in \mathcal{L}} P(\boldsymbol{v_c}|\boldsymbol{\omega}) P(\boldsymbol{\omega}) \tag{11}$$

The first term $P(\boldsymbol{v_c}|\boldsymbol{\omega})$ is the likelihood function that represents the deviation between $\boldsymbol{v_c}$ and the dominant MV of the



(a) Reference frame in ParkScene. (b) Motion segmentation map.
(c) Reference frame in BQTerrace. (d) Motion segmentation map.

Fig. 4. The motion segmentation results. (a), (c) represent the original reference frames from ParkScene and BQTerrace sequences. (b), (d) are their corresponding motion segmentation results, where different motion regions are represented by different gray values.

corresponding motion label. The second term $P(\boldsymbol{\omega})$ is the prior motion region label model describing the correlation of spatial label values inside a frame. To solve (11), $P(\boldsymbol{v_c}|\boldsymbol{\omega})$ can be modeled as the Gaussian distribution [41],

$$P(\boldsymbol{v_c}|\boldsymbol{\omega}) = \frac{1}{\sqrt{2\pi}\sigma_2} exp\{-\frac{\|\boldsymbol{v_c} - \boldsymbol{v_d}\|^2}{2\sigma_2^2}\} \quad \boldsymbol{v_d} \in \mathcal{D},\ \omega \in \mathcal{L} \tag{12}$$

where $\boldsymbol{v_d}$ is the dominant MV of corresponding label $\boldsymbol{\omega}$, σ_2 is a control parameter. In addition, $P(\boldsymbol{\omega})$ is assumed to be a Markov random field and formulated as follows,

$$P(\boldsymbol{\omega}) = \frac{1}{Z}\prod_C e^{-V(C)} \tag{13}$$

where Z is the normalizing constant, C is a clique (i.e., a group of connected label values) and $V(C)$ is the clique potential. Let $\boldsymbol{\omega_N}$ be the neighboring label value of $\boldsymbol{\omega}$. $V(C)$ is given by,

$$V(C) = \begin{cases} -T, & if\ \boldsymbol{\omega} = \boldsymbol{\omega_N} \\ T, & otherwise \end{cases} \tag{14}$$

T is a threshold controlling the degree of homogeneity.

Then the coarse segmentation map is set as the initial value and (11) can be optimized by employing the iterative conditional modes [42] algorithm. Finally, the refined segmentation map $\mathcal{S}$ can be obtained after several iterations. Fig. 4 shows the reference frames and their corresponding motion segmentation maps.

Based on the piecewise smoothness characteristic, the estimated $\boldsymbol{v}$ should be similar with its neighboring MVs when they indicate the same motion region in the reference frame. Therefore the smoothness term is defined as follows,

$$E_S(\boldsymbol{v}) = \sum_{\boldsymbol{v_n} \in \mathcal{N}} \rho(\boldsymbol{n}) \|\boldsymbol{v} - \boldsymbol{v_n}\|^2 \tag{15}$$

where $\mathcal{N}$ is a set of neighboring MVs of $\boldsymbol{v}$. In the proposed framework, MVs from above and left neighboring blocks are included in $\mathcal{N}$. Here, $\rho(\boldsymbol{n})$ denotes a weight parameter controlling the interaction between neighboring MVs and current MV, and is formulated as follows,

$$\rho(\boldsymbol{n}) = exp\{-\frac{\sum_{\boldsymbol{p} \in \mathcal{B}_p} \mathcal{I}(\mathcal{S}(\boldsymbol{p} - \boldsymbol{v}), \mathcal{S}(\boldsymbol{p} - \boldsymbol{v_n}))}{N \cdot \sigma_3}\} \tag{16}$$

where σ_3 is a control parameter and N is the number of pixels in $\mathcal{B}_p$. $\mathcal{I}(x, y)$ is an indicator function and $\mathcal{I}(x, y) = 0$ when x is equal to y , otherwise $\mathcal{I}(x, y) = 1$.

In (16) , $\rho(\boldsymbol{n})$ will increase when $\boldsymbol{v}$ and $\boldsymbol{v_n}$ point to the same motion region in reference frame, and then the smoothness term in (15) will prevent the estimated $\boldsymbol{v}$ deviating a lot from neighboring MVs. At the same time, if $\boldsymbol{v}$ and $\boldsymbol{v_n}$ indicate different motion regions, it implies that $\boldsymbol{v}$ may locate at motion boundary. In this case, smoothness term will decrease, which avoids an over-smoothed MVF.

Fig. 5 confirms the effectiveness of proposed joint motion estimation method. Compared with method which only utilizes data term, the joint motion estimation algorithm significantly reduces the outliers, especially for homogeneous regions (river in Fig. 5(e) or blue sky in Fig. 5(g)). Meanwhile, our method also preserves the discontinuity at motion boundary (moving car in Fig. 5(e) or walking woman in Fig. 5(g)) and improves the robustness of MVF.

IV. Rate-Distortion Optimization For Interpolated Frames

It is well known that HEVC coding [1] and previous coding standards such as H.264 [21] all adopt the Rate-Distortion Optimization technique to improve the compression efficiency. Generally, RDO criterion is given by,

$$\arg\min_{\{Para\}} J, \quad J = D + \lambda R \tag{17}$$

where $\{Para\}$ is the coding parameter set, including mode, coding size, motion, etc. J is the total RD cost. D is the distortion between an original frame and its reconstructed frame, and Sum Square Error (SSE) is the most commonly used. R denotes the generated bits. λ is the Lagrange multiplier to balance the distortion and bitrate.

In the proposed framework, due to the absence of original frames, the distortion cannot be directly calculated for interpolated frames. However, RDO is one of the fundamental considerations in the compression system and its performance significantly influences the coding efficiency. Therefore we analyze the distortion for interpolated frames and put forward a novel distortion model based on the MV reliability and the corresponding quantization error. Then the frame interpolation, i.e, encoding procedure for interpolated frames can be solved in a rate-distortion optimization manner with high coding efficiency.

A. Distortion Model for Interpolated Frame

Let us consider the frame interpolation procedure in proposed framework. The interpolated frame $\hat{F}_t$ at location $\boldsymbol{p}$ can be estimated by,

$$\hat{F}_t(\boldsymbol{p}) = \frac{1}{2}(\tilde{F}_{t-1}(\boldsymbol{p} - \boldsymbol{v}) + \tilde{F}_{t+1}(\boldsymbol{p} + \boldsymbol{v})) \tag{18}$$

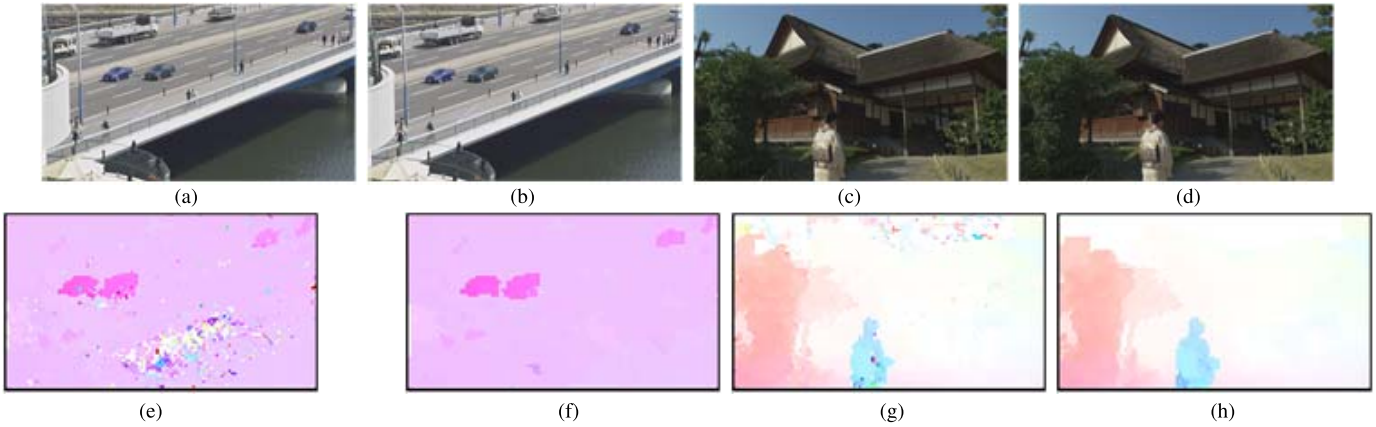


Fig. 5. Comparison of the obtained MVF results using different terms. (a), (b) The 258^{th} and 260^{th} frames of the BQTerrace sequences. (c), (d) The 186^{th} and 188^{th} frames of the Kimono sequences. (e)-(h) Color maps for the estimated MVFs, where the hue and saturation denote the direction and magnitude of the MV, respectively. (e), (g) Data term only. (f), (h) The proposed joint ME algorithms.

$\boldsymbol{v}$ is the estimated MV. $\tilde{F}_{t-1}$ and $\tilde{F}_{t+1}$ are the reconstructed frames for F_{t-1} and F_{t+1} in coding loop. Namely,

$$\begin{aligned} \tilde{F}_{t-1} &= F_{t-1} + \eta_{c,t-1} \\ \tilde{F}_{t+1} &= F_{t+1} + \eta_{c,t+1} \end{aligned} \tag{19}$$

Here, $\eta_{c,t-1}$ and $\eta_{c,t+1}$ are the random disturbance noises that are introduced by video compression and assumed to be independent of each other.

Generally, the assumption for FRUC implies that pixel value in the interpolated frame can be represented by its adjacent pixels along motion trajectory. However, the estimated MV $\boldsymbol{v}$ in the proposed framework cannot be always accurate enough even if an efficient ME algorithm is employed. Besides, the evolvement of natural scene such as the lighting change, object deformations and camera noise may bring in disturbance along motion trajectory [43]. Therefore the motion compensated interpolation is noisy,

$$F_t(\boldsymbol{p}) = \frac{F_{t-1}(\boldsymbol{p}-\boldsymbol{v}) + F_{t+1}(\boldsymbol{p}+\boldsymbol{v})}{2} + \eta_{v,t}(\boldsymbol{p}) \tag{20}$$

where $\eta_{v,t}$ denotes the noise that caused by imperfect motion compensation. Combining (18), (19) and (20), the prediction error $e(\boldsymbol{p})$ of the interpolated frame is formulated as

$$e(\boldsymbol{p}) = F_t(\boldsymbol{p}) - \hat{F}_t(\boldsymbol{p}) = \eta_{v,t}(\boldsymbol{p}) + \eta_{c,t}(\boldsymbol{p}) \tag{21}$$

where $\eta_{c,t}(\boldsymbol{p}) = -(\eta_{c,t-1}(\boldsymbol{p}) + \tilde{a}\text{Ł}\text{Ł}\eta_{c,t+1}(\boldsymbol{p}))/2$.

Considering current block $\mathcal{B}_p$, the distortion D of whole block can be given by

$$D = \sum_{\boldsymbol{p}\in\mathcal{B}_p} e^2(\boldsymbol{p}) \tag{22}$$

The distribution of prediction error $e(\boldsymbol{p})$ is usually modeled as a Gaussian distribution [5] and it has been pointed out that the prediction error has a weak spatial correlation [44]. A natural assumption is thus made that $e(\boldsymbol{p})$ in $\mathcal{B}_p$ can be modeled as the independent identically Gaussian distribution,

$$P(e) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{e^2}{2\sigma^2}) \tag{23}$$

σ^2 is the variance of prediction error. Combining (22) and (23), the distortion D in $\mathcal{B}_p$ complies with the Chi-squared distribution in the following way,

$$\frac{D}{\sigma^2} \sim \mathcal{X}^2(N) \tag{24}$$

N is the number of pixels in $\mathcal{B}_p$. $\mathcal{X}^2(N)$ is Chi-squared distribution with N degrees of freedom. Since the expectation value for $\mathcal{X}^2(N)$ is only determined by the degrees of freedom, therefore the expectation $E[D]$ for distortion can be derived by,

$$E[D] = N \cdot \sigma^2 \tag{25}$$

In the proposed framework, $E[D]$ is used to estimate the distortion for interpolated frames.

B. Variance Estimation of Prediction Error

In order to calculate the distortion in (25), we need to adopt an appropriate model to describe the prediction error variance σ^2. In (21), σ^2 is closely related with compression error $\eta_{c,t}$ and motion compensation error $\eta_{v,t}$. For location $\boldsymbol{p}$, the variance of compression error $\eta_{c,t}$ can be approximated by the corresponding mean quantization error in reference blocks as follows [45],

$$\begin{aligned} Q(\boldsymbol{p}) = \frac{1}{4N} \sum_{p\in\mathcal{B}_p} [(\tilde{F}_{t-1}(\boldsymbol{p}-\boldsymbol{v}) - F_{t-1}(\boldsymbol{p}-\boldsymbol{v}))^2 \\ + (\tilde{F}_{t+1}(\boldsymbol{p}+\boldsymbol{v}) - F_{t+1}(\boldsymbol{p}+\boldsymbol{v}))^2] \end{aligned} \tag{26}$$

Here, $Q(\boldsymbol{p})$ denotes the mean quantization error.

In (20), $\eta_{v,t}(\boldsymbol{p})$ indicates the reliability of mean pixel value along motion trajectory as a prediction of the interpolated frame $F_t(\boldsymbol{p})$. Intuitively, $\eta_{v,t}(\boldsymbol{p})$ mainly depends on how reliable it is that the estimated MV $\boldsymbol{v}$ form a true motion trajectory. Therefore, $\eta_{v,t}(\boldsymbol{p})$ is expected to be highly correlated with MV energy $E(\boldsymbol{v})$ defined in (5).

For prediction error in location $\boldsymbol{p}$ with estimated MV $\boldsymbol{v}$, to examine the relationship between prediction error variance $\sigma^2(\boldsymbol{p},\boldsymbol{v})$ and $Q(\boldsymbol{p})$, $E(\boldsymbol{v})$, we conduct an experiment

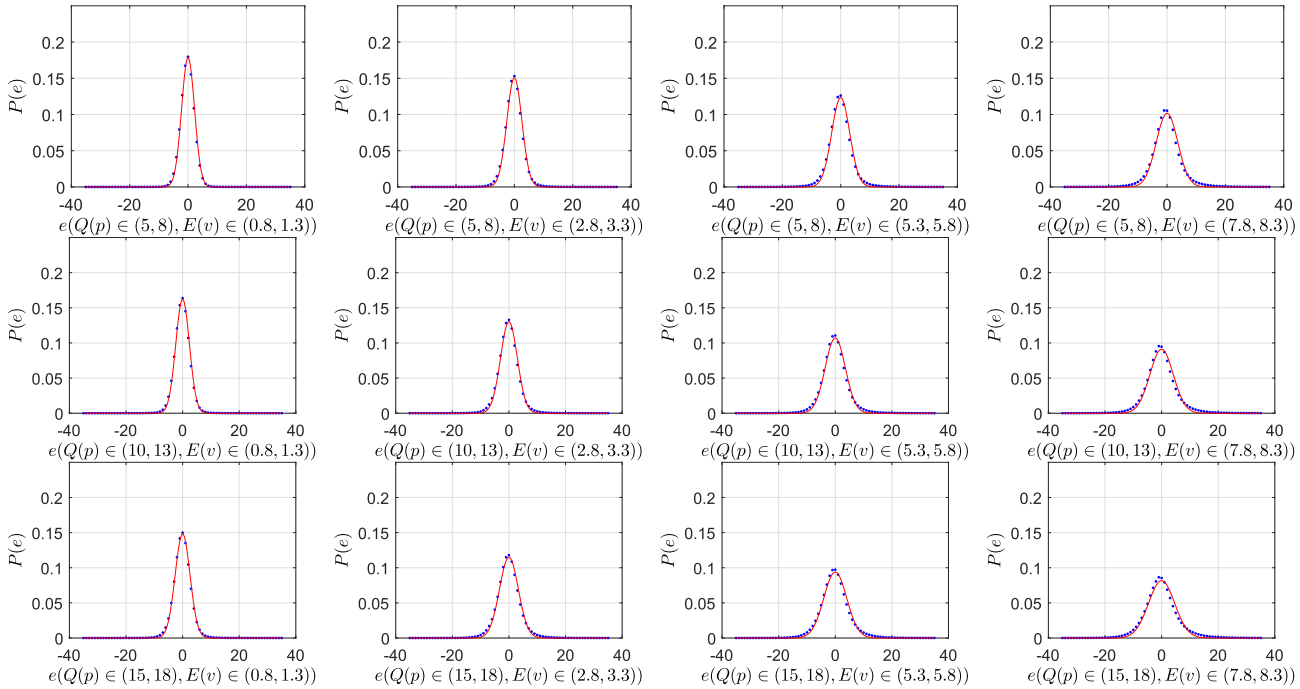


Fig. 6. Histograms and their corresponding fitted Gaussian curves according to the range of MV reliability and quantization error.

on five video sequences, including Basketball, BQTerrace, Cactus, Kimono and ParkScene. After interpolating the intermediate frame, histograms of the prediction error between the interpolated frame and original frame are constructed according to the range of $Q(\boldsymbol{p})$ and $E(\boldsymbol{v})$, respectively. As shown in Fig. 6, columns correspond to different MV reliabilities and rows correspond to different quantization errors. From the figure it is noted that the prediction error with high MV reliability (i.e., small value for $E(\boldsymbol{v})$) tends to be concentrated on zero in the same row. At the same time, the prediction error with less quantization error tends to be concentrated on zero in the same column. Therefore the statistical characteristic of the prediction error variance $\sigma^2(\boldsymbol{p}, \boldsymbol{v})$ can be described in terms of MV reliability $E(\boldsymbol{v})$ and quantization error $Q(\boldsymbol{p})$.

In order to quantify this relationship, offline training is utilized. Specifically, considering $e(\boldsymbol{p})$ is assumed to be Gaussian, $\sigma^2(\boldsymbol{p}, \boldsymbol{v})$ in different ranges of MV reliability and quantization error can be obtained based on the maximum likelihood estimation. Then each $\sigma^2(\boldsymbol{p}, \boldsymbol{v})$ and its corresponding $E(\boldsymbol{v})$ and $Q(\boldsymbol{p})$ construct a training sample. In the proposed integration framework, $\sigma^2(\boldsymbol{p}, \boldsymbol{v})$ is modeled as a polynomial function of $Q(\boldsymbol{p})$ and $E(\boldsymbol{v})$, i.e.,

$$\sigma^2(\boldsymbol{p}, \boldsymbol{v}) = \sum_{i=0}^{n} \sum_{j=0}^{m} g_{i,j} E^i(\boldsymbol{v}) Q^j(\boldsymbol{p}) \tag{27}$$

where m and n represent the orders for this polynomial. $g_{i,j}$ is the weight coefficient. It is found that second-order polynomial is accurate enough to describe this inherent relationship, and the corresponding coefficients are estimated based on the least square regression. The simulation result is illustrated in Fig. 7. Thereby, prediction error variance σ^2 in the proposed framework can be estimated by MVs reliability and quantization error.

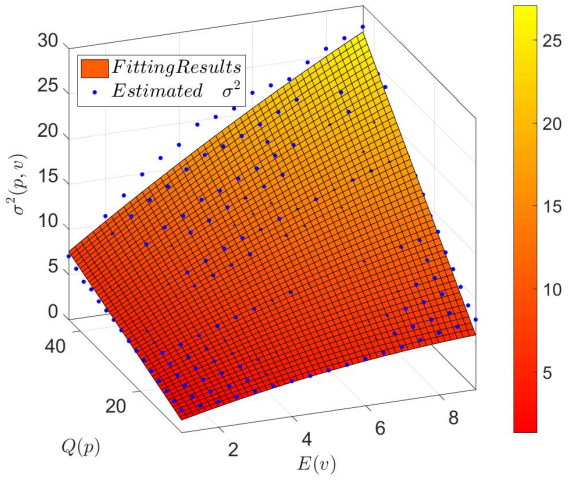


Fig. 7. Illustration of the relationship between $\sigma^2(\boldsymbol{p}, \boldsymbol{v})$ and $E(\boldsymbol{v})$, $Q(\boldsymbol{p})$.

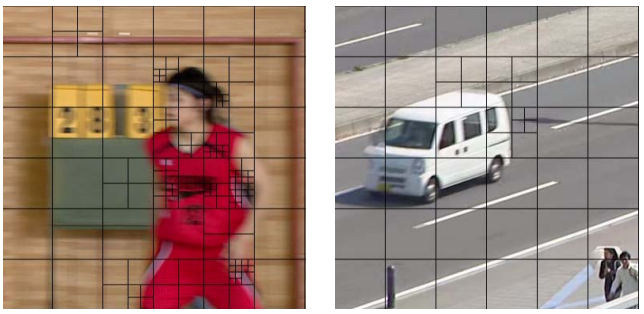
In the integration framework, $E[D]$ in (25) is used to estimate the distortion for interpolated frames. Comparing with the original RDO criterion in (17), we only need to replace the calculation method for distortion. Therefore no modification is necessary for λ or R, and the new RDO criterion can be integrated within the encoder gracefully. In fact, this new RDO bridges the gap between HEVC coding and FRUC by constructing a criterion that accounts for both the coding parameters and visual quality of FRUC. More importantly, the frame interpolation in this integration framework is solved based on rate-distortion optimization, which guarantees that the output high frame rate video can achieve high quality with few bits.

Fig. 8 presents the interpolation results based on the proposed distortion model. In motion boundary or complex

TABLE I

OBJECTIVE EVALUATION OF THE PROPOSED FRAMEWORK COMPARED WITH THREE TRADITIONAL FRUC ALGORITHMS [3], [7], [13] CASCADED WITH HEVC CODING

Sequence			Choi [3]		Dikbas [7]		Yoo [13]	
			BDBR	BD-PSNR	BDBR	BD-PSNR	BDBR	BD-PSNR
BasketballDrive	1080p	50fps	-48.93	1.02	-25.13	0.44	-25.34	0.45
BQTerrace	1080p	60fps	-58.64	1.11	-30.44	0.46	-43.89	0.75
Cactus	1080p	50fps	-41.01	0.80	-19.03	0.33	-23.83	0.42
Kimono1	1080p	24fps	-22.79	0.54	-6.48	0.14	-14.90	0.35
ParkScene	1080p	24fps	-28.59	0.77	-10.05	0.24	-16.06	0.41
BlueSky	1080p	25fps	-72.31	3.59	-42.81	1.64	-42.49	1.61
Sunflower	1080p	25fps	-36.77	0.81	-11.15	0.20	-0.76	0.00
Stockholm	720p	60fps	-46.51	1.18	-34.34	0.81	-7.16	0.14
Shields	720p	50fps	-49.58	0.89	-24.02	0.38	-23.43	0.30
Parkrun	720p	60fps	-18.96	0.60	-27.96	0.90	-12.73	0.39
Average			-42.51	1.13	-23.14	0.55	-21.06	0.48



(a) (b)

Fig. 8. The different coding block sizes for interpolated frames by the novel RDO criterion in different sequences. (a) Basketball. (b) BQTerrace.

motion region, such as the head of player in Fig. 8(a) or the edge of car in Fig. 8(b), small coding block size, e.g., 16×16, is chosen. On the other hand, large coding block size is selected for background region, such as the wall in Fig. 8(a) and the road in Fig. 8(b). It confirms that the new RDO criterion can effectively optimize the coding parameters (e.g., block size) for interpolated frames. More quantitative experimental results are detailed in Section V-C.

V. EXPERIMENTAL RESULTS

To examine the performance of the proposed integration framework of FRUC and HEVC, experiments are conducted on various videos, including seven 1080p (1920x1080) sequences (BasketballDrive@50fps, BQTerrace@60fps, Cactus@50fps, Kimono1@24fps, ParkScene@24fps, BlueSky@25fps and Sunflower@25fps) and three 720p (1280x720) sequences (ParkRun@50fps, Shields@50fps and Stockholm@60fps). These sequences are widely used in the video processing literature.

Our integration framework is implemented in the open source HEVC encoder x265,[1] which is an efficient practical encoder. The GOP structure used in experiments is IBBBP, where the first and third B frames are the to be interpolated frames. Odd frames are used as input for the proposed framework, even frames are skipped and saved as the ground truth to evaluate the performance of output interpolated frames.

[1] https://bitbucket.org/multicoreware/x265/wiki/Home/

It should be mentioned that although our proposed method is designed to achieve an unsampling factor of two in this paper, it is not difficult to change the GOP structure and obtain other frame rate videos based on the frame rate of target devices.

The proposed algorithm uses several parameters, including σ_1, σ_2, σ_3, α and β. All the parameters are empirically set to fixed values. α and β in (5) are employed to balance the matching term and smoothness term, and are set to 0.4 and 0.2, respectively. σ_1 in (8) is used for measuring the matching MV reliability and is set to 2. σ_2 in (12) is the variance of Gaussian distribution and is set to 3. σ_3 in (16) is used for calculating the similarity between different motion regions and is set to 0.3. In order to analyze the parameter sensitivity, we conduct several experiments in different parameter settings. As shown in Table II, the average PSNR of the interpolated frames remains stable in a wide range of parameter settings ($\alpha \in [0.2, 0.8]$, $\beta \in [0.1, 0.4]$). In other words, we can obtain similar experimental results in this parameter scope, which proves the robustness of our method.

To compare with the cascaded approaches, three traditional FRUC methods [3], [7], [13] are tested. These three algorithms employ block based ME scheme and achieve desirable performance. For Choi's algorithm [3], the weighting coefficient μ is set to 0.6. For Dikbas's algorithm [7], the weighting coefficient λ for smoothness term is set to 0.05. Finally, the block size is set to 16×16 and the search range is fixed to $[-32, 32]\times[-32, 32]$ for all the algorithms.

A. Objective Assessment

In this subsection, we compare the objective quality of the proposed integration framework with three cascaded methods. For the traditional cascaded methods, odd frames in original video sequences are upconverted by FRUC algorithms [3, 7, 13] then the high frame rate videos are encoded with x265 in different quantization parameters (QP = 22, 27, 32, 37). These results are used as the benchmarks. In the proposed method, the same coding configuration is employed but the input are low frame rate videos. Table I shows the comparison results, where BDBR and BD-PSNR [46] are adopted to assess the overall performance. Compared with the cascaded approaches, the proposed method can provide better coding performance. For example, the integration framework achieves

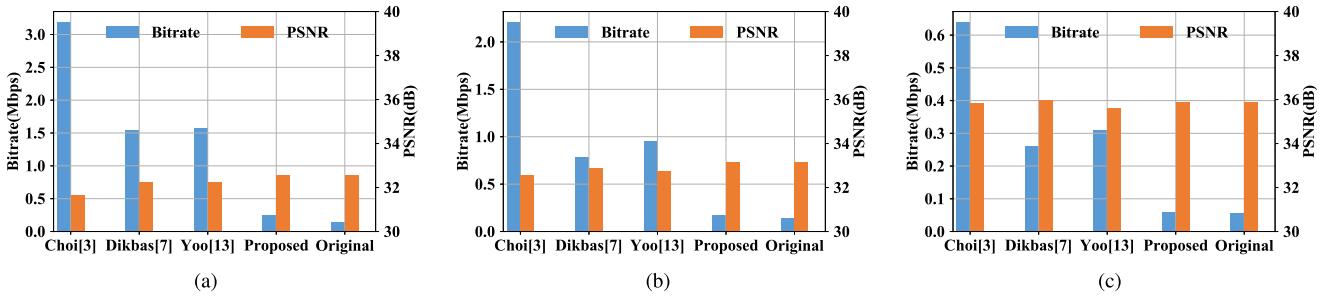


Fig. 9. The average bitrate and PSNR for interpolated frames in different methods. QP for methods [3], [7], [13] and the proposed framework is fixed to 32. The original high frame rate videos at the similar bitrate is also provided for comparison. (a) Basketball. (b) Cactus. (c) Kimono.

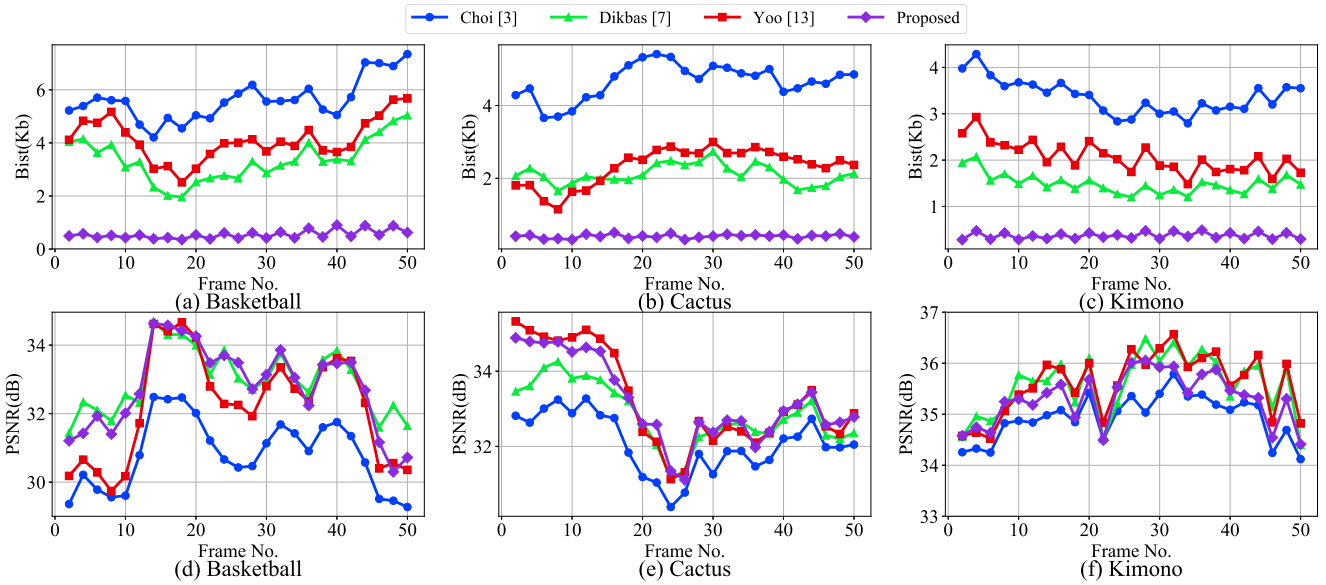


Fig. 10. The generated bits and PSNR for the first 50 interpolated frames in different methods when QP is set to 32. Bits costs are shown in (a)-(c), PSNR results are shown in (d)-(f).

TABLE II

AVERAGE PSNR VALUES OF THE PROPOSED METHODS IN DIFFERENT WEIGHT PARAMETERS α AND β

β \ α	0.0	0.1	0.2	0.4	0.6	0.8	1.0	3.0
0.0	30.85	30.95	31.14	31.14	31.15	31.15	31.15	31.04
0.1	30.95	31.35	31.44	31.44	31.45	31.45	31.38	31.30
0.2	30.97	31.33	31.44	31.45	31.45	31.45	31.36	31.30
0.3	30.98	31.33	31.43	31.43	31.43	31.43	31.36	31.28
0.4	30.98	31.30	31.42	31.43	31.43	31.44	31.35	31.26

an average 42% reduction in BDBR or 1.1 dB increase in BD-PSNR when compared with the Choi's method [3]. Even if a more complex algorithm such as [7] is used for comparison, the proposed method can achieve an average 23% reduction in BDBR or 0.55 dB increase in BD-PSNR.

The advantage of proposed framework lies in its ability for reconstructing the interpolated frame with low bitrate and high visual quality. In Fig. 9, the average bits and PSNR for the interpolated frames in different methods are depicted. The PSNR of interpolated frames in the proposed method is similar or higher than the cascaded approaches. At the same time, the bits used for interpolated frame decrease significantly. Taking Fig. 9(c) as an example, the average bitrate for interpolated frames in proposed method is 0.06 Mbps, however three cascaded methods consume 4∼10 times more bits. In addition, the average bitrate consumption of original high frame rate video sequences at the similar PSNR is also provided in Fig. 9 for comparison. It can be concluded that these two methods are comparable. Considering the interpolated frames are not available in our method, it is expectable and reasonable that our method may consume more bitrates at a similar PSNR level.

Fig. 10 shows the bits and PSNR for interpolated frames on Basketball, Cactus and Kimono sequences when QP is set to 32. These curves also prove that the proposed method can achieve similar or better interpolation results in most cases with few bits.

B. Subjective Assessment

Subjective assessment of the up-converted video sequences is presented in Fig. 11, where rows correspond to different video sequences and columns correspond to (in order): original frames, interpolated frames of Choi *et al.* [3], Dikbas *et al.* [7], Yoo *et al.* [13] and the proposed method. Since the joint motion estimation employs multiple information to enhance

Fig. 11. Comparison of the subjective interpolation results for 202^{th} frame in ParkScene (a)-(e), 12^{th} frame in Cactus (f)-(g), 470^{th} frame in BQTerrace (k)-(o). From left to right, the frames in each column represent the original frames, interpolated frames by [3], [7], [13] and the proposed method, respectively. *(Best viewed in color)*

the MV reliability, therefore the generated interpolation results show better visual quality.

For the 202^{th} frame in ParkScene, the wheel of bicycle is a complicated motion scene. Traditional FRUC methods such as [3], [7], [13] only utilize the information in pixel domain to estimate the MVs, therefore the shape of wheel are distorted as depicted in (b)-(d). However, the proposed joint motion estimation in (e) can handle this situation by taking advantage of the feature matching and alleviate these artifacts significantly.

For the 12^{th} frame in Cactus, the toy has a fast movement which brings a challenge for accurate motion estimation. Choi *et al.* [41] and Yoo *et al.* [13] adopt the full search strategy, where the estimated MVs may fall into the local minimum and cannot represent the true motion trajectory. Dikbas *et al.*'s algorithm employs the predictive motion estimation method, however the estimated MVs may not converge for the fast motion. In contrast, the MVs obtained through Deepmatching are robust to large displacement and employed to accelerate the MVs convergence in the joint motion estimation. Therefore, the proposed method in (j) outperforms all the traditional FRUC methods [3], [7], [13] in (g)-(i) and are free from noticeable artifacts.

For the 470^{th} frame in BQTerrace, the bridge railings can be considered as the repetitive texture region and tracking the true motion of this region is not an easy task. As shown in (l)-(n), the interpolated frames in the yellow rectangle are considerably blurred because the estimated MVs for this region are not accurate enough. An effective approach to address this issue is to impose smoothness constrains in ME process, especially for the textureless region or repetitive texture region. Choi *et al.*'s algorithm do not consider smoothness constrain, and the smooth schemes in [7], [13] do not reduce this blur significantly. However, our method utilize the motion segmentation information and provides a clear bridge railings image by preserving the piecewise smoothness characteristic of the MVF.

C. Effectiveness of the Proposed RDO Criterion

In Section IV, a new RDO criterion is proposed based on the novel distortion model for interpolated frames. To examine the effectiveness of this RDO criterion, we compare it with other straightforward method, which only use MV reliability to represent the distortion for interpolated frames. In others words, the RD cost used for comparison is defined as $J = E(\boldsymbol{v}) + \lambda R$. $E(\boldsymbol{v})$ is given by (5). In addition, since the interpolated frame has no original frame used for calculating the distortion, we also compare the RDO criterion where no distortion model is included, which is defined as $J = \lambda R$.

In Fig. 12, three video sequences including Basketball, Kimono and ParkScene are tested for comparison. It is

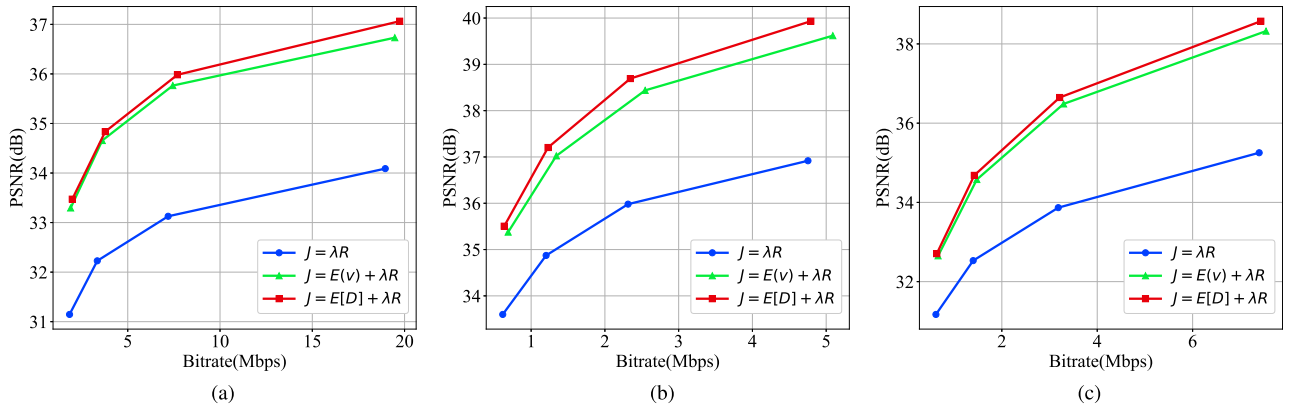


Fig. 12. Different RDO criterions for interpolated frames are used for comparison in various video sequences. (a) Basketball. (b) Kimono. (c) ParkScene.

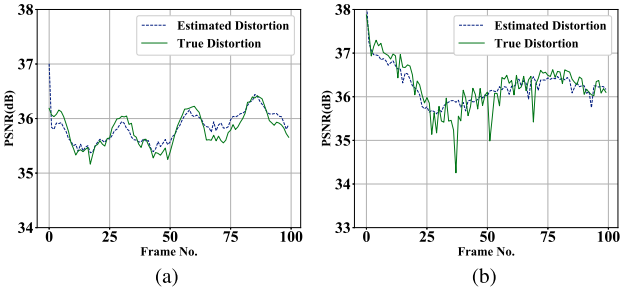


Fig. 13. The comparison between estimated distortion and true distortion in different video sequences. The dash line represents the estimated distortion. (a) Cactus. (b) ParkScene.

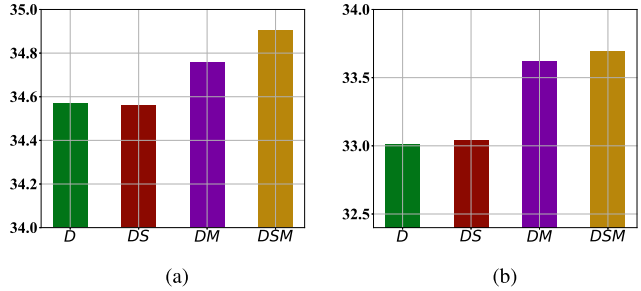


Fig. 14. The average PSNR(dB) of interpolated frames through different combinations of data term, matching term and smoothness term. D, DS, DM, DSM represent the E_D, $E_D + \beta E_S$, $E_D + \alpha E_M$ and $E_D + \alpha E_M + \beta E_S$, respectively. (a) PSNR for Kimono1. (b) PSNR for ParkScene.

clear that the new RDO provides better RD performance than other straightforward approaches. Taking Fig. 12(b) as an example, the proposed method can achieve more than 19% reduction in BDBR when compared with $J = E(\boldsymbol{v}) + \lambda R$. Even for the video sequences with a global smooth motion such as ParkScene in Fig. 12(c), the proposed RDO criterion still achieves 9% reduction in BDBR when compared with $J = E(\boldsymbol{v}) + \lambda R$. Obviously, the proposed method is significantly better than the RDO criterion $J = \lambda R$ which cannot measure the distortion for the interpolation frames. This experiment proves that the novel distortion based RDO manner can effectively optimize the coding procedure for interpolated frames and improve coding efficiency.

One of the major contributions in our method is the estimation of the true distortion for interpolated frames. To examine the reliability of the estimated distortion, we compare the estimated values with the true distortion. The average distortion in a frame is calculated. As shown in Fig. 13, the dash line represents the estimated distortion. According to the figure, although the gaps between estimated distortion and true distortion fluctuate in different frames (e.g., Fig. 13(c)), the estimated distortions present strong consistency with the true values. This proves that the estimation method which replies on both the motion vector reliability and quantization error can produce satisfactory results.

D. Effectiveness of the Proposed JME

To evaluate the effectiveness of proposed terms in the joint ME algorithm quantitatively, several experiments with different settings are used for comparison. As shown in Fig. 14, the smoothness term or matching term can improve the quality of interpolated frames. More importantly, the combination of these three terms performs better than other settings. This implies that these three terms can be complementary as we mentioned in Section III.

TABLE III

COMPARISON OF THE COMPUTATIONAL COMPLEXITY FOR 1080P VIDEO IN DIFFERENT METHODS

Methods	Run Time
Choi[3]	1000s
Dikbas[7]	6s
Yoo[13]	75s
DeepMatching(1/4 downscaling)	40s
DeepMatching(1/8 downscaling)	2s

E. Analysis of Computational Complexity

In this subsection, we compare the computational complexity between the proposed method and the other related work. In traditional cascaded methods, the computational complexity comes from two aspects: FRUC procedure and the coding procedure. While the motion estimation and motion compensation procedure are shared between coding and FRUC in our method, the extra computation complexity is mainly introduced by the Deep Matching algorithm [35]. Therefore we compare the computational complexity between FRUC methods [3], [7], [13] and deepmatching [35] in Table III.

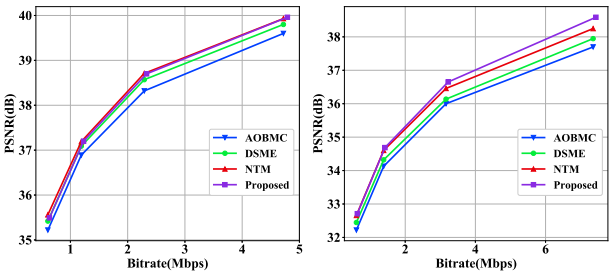


Fig. 15. Comparison of RDO performance between post-processing pipeline and the proposed method in different sequences. (a) Kimono1. (b) ParkScene.

These FRUC methods are implemented in MATLAB and DeepMatching algorithm uses the online code provided by the authors. Deepmatching is an efficient feature matching algorithm with high accuracy and low complexity. Considering HEVC utilizes block based motion estimation, Deepmatching algorithm can downscale the reference frames for further computational complexity reduction with negligible accuracy decrease. It only costs 2 seconds to calculate the matching MVF for the 240x135 (1/8 downscaling for 1080p) reference frame in our method. Experimental results demonstrate that the computational complexity of the proposed method are competitive with traditional methods.

F. Comparison Between the Post-Processing Methods and the Proposed Method

In this experiment, low frame rate videos are encoded and transmitted to the decoder side, then up-converted to high frame rates video based on the traditional methods (AOBMC, DSME, NTM). The RDO performances of high frame rate video generated through post-processing are compared with the proposed method. As shown in Fig. 15, the proposed method is competitive and outperforms these post-processing methods. One major reason is that FRUC in decoder side can only utilize the reconstructed or distorted reference frames while the proposed method in encoder side can leverage the original frames (e.g., matching term) and provide high quality interpolated frames.

VI. Conclusion

In this paper, we have proposed a novel integration framework of frame rate up conversion and HEVC coding based on rate-distortion optimization. The integration framework can generate interpolated frames at encoder side with low bitrate cost and high visual quality. In order to improve the reliability for shared MVF between FRUC and coding, the proposed joint motion estimation utilizes multiple information, including data term, matching term and smoothness term. More importantly, a novel distortion model for interpolated frames has been established, where both the MV reliability and coding quantization error are taken into account. Then frame interpolation is solved by rate-distortion optimization with high coding efficiency. Extensive experimental results have demonstrated that the proposed framework provides better subjective and objective performance than conventional cascaded methods. Our framework shows the ability in delivering high frame rate video at limited bandwidth and can be applied to various applications.

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] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 441–446, Mar. 2007.

[3] 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.

[4] S.-J. Kang, S. Yoo, and Y. H. Kim, "Dual motion estimation for frame rate up-conversion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1909–1914, Dec. 2010.

[5] H. Liu, R. Xiong, D. Zhao, S. Ma, and W. Gao, "Multiple hypotheses Bayesian frame rate up-conversion by adaptive fusion of motion-compensated interpolations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1188–1198, Aug. 2012.

[6] D. Kim, H. Lim, and H. Park, "Iterative true motion estimation for motion-compensated frame interpolation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 3, pp. 445–454, Mar. 2013.

[7] S. Dikbas and Y. Altunbasak, "Novel true-motion estimation algorithm and its application to motion-compensated temporal frame interpolation," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 2931–2945, Aug. 2013.

[8] 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.

[9] 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.

[10] H. Lim and H. W. Park, "A region-based motion-compensated frame interpolation method using a variance-distortion curve," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 518–524, Mar. 2015.

[11] 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.

[12] 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.

[13] D.-G. Yoo, S.-J. Kang, and Y. H. Kim, "Direction-select motion estimation for motion-compensated frame rate up-conversion," *J. Display Technol.*, vol. 9, no. 10, pp. 840–850, Oct. 2013.

[14] Y. Guo, L. Chen, Z. Gao, and X. Zhang, "Frame rate up-conversion method for video processing applications," *IEEE Trans. Broadcast.*, vol. 60, no. 4, pp. 659–669, Dec. 2014.

[15] A.-M. Huang and T. Q. Nguyen, "A multistage motion vector processing method for motion-compensated frame interpolation," *IEEE Trans. Image Process.*, vol. 17, no. 5, pp. 694–708, May 2008.

[16] A.-M. Huang and T. Q. Nguyen, "Correlation-based motion vector processing with adaptive interpolation scheme for motion-compensated frame interpolation," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 740–752, Apr. 2009.

[17] R. Krishnamurthy, J. W. Woods, and P. Moulin, "Frame interpolation and bidirectional prediction of video using compactly encoded optical-flow fields and label fields," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 5, pp. 713–726, Aug. 1999.

[18] G. Dane, K. El-Maleh, and Y.-C. Lee, "Encoder-assisted adaptive video frame interpolation," in *Proc. ICASSP*, 2005, pp. 349–352.

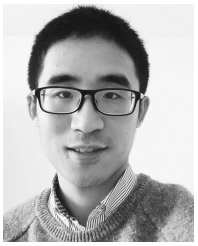
[19] X. HoangVan, J. Ascenso, and F. Pereira, "Improving scalable video coding performance with decoder side information," in *Proc. IEEE PCS*, Dec. 2013, pp. 229–232.

[20] D. Rüfenacht, R. Mathew, and D. Taubman, "A novel motion field anchoring paradigm for highly scalable wavelet-based video coding," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 39–52, Jan. 2016.

[21] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[22] I.-K. Kim, J. Min, T. Lee, W.-J. Han, and J. Park, "Block partitioning structure in the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1697–1706, Dec. 2012.

[23] J. Kim *et al.*, "An SAD-based selective bi-prediction method for fast motion estimation in high efficiency video coding," *ETRI J.*, vol. 34, no. 5, pp. 753–758, 2012.
[24] W. Dai, O. C. Au, C. Pang, L. Sun, R. Zou, and S. Li, "A novel fast two step sub-pixel motion estimation algorithm in HEVC," in *Proc. IEEE ICASSP*, Mar. 2012, pp. 1197–1200.
[25] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. ECCV*, 2004, pp. 25–36.
[26] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.
[27] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. CVPR*, Jun. 2015, pp. 1164–1172.
[28] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," in *Proc. ICCV*, 2015, pp. 4015–4023.
[29] S.-G. Jeong, C. Lee, and C.-S. Kim, "Motion-compensated frame interpolation based on multihypothesis motion estimation and texture optimization," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4497–4509, Nov. 2013.
[30] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, "Multi-level video frame interpolation: Exploiting the interaction among different levels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 7, pp. 1235–1248, Jul. 2013.
[31] J.-H. Kim, Y.-H. Ko, H.-S. Kang, S.-W. Lee, and J.-W. Kwon, "Frame rate up-conversion method based on texture adaptive bilateral motion estimation," *IEEE Trans. Consum. Electron.*, vol. 60, no. 3, pp. 445–452, Aug. 2014.
[32] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, vol. 1. Jun. 2005, pp. 886–893.
[34] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards Internet-scale multi-view stereo," in *Proc. IEEE CVPR*, Jun. 2010, pp. 1434–1441.
[35] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "DeepMatching: Hierarchical deformable dense matching," in *Proc. IJCV*, 2015, pp. 1–24.
[36] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. ICCV*, 2013, pp. 1385–1392.
[37] 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.
[38] J. Lezama, K. Alahari, J. Sivic, and I. Laptev, "Track to the future: Spatio-temporal video segmentation with long-range motion cues," in *Proc. IEEE CVPR*, Jun. 2011, pp. 3369–3376.
[39] M. Narayana, A. Hanson, and E. Learned-Miller, "Coherent motion segmentation in moving camera videos using optical flow orientations," in *Proc. ICCV*, 2013, pp. 1577–1584.
[40] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[41] Y.-M. Chen and I. V. Bajic, "A joint approach to global motion estimation and motion segmentation from a coarsely sampled motion vector field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1316–1328, Sep. 2011.
[42] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 48, no. 3, pp. 259–302, 1986.
[43] J. Konrad and E. Dubois, "Bayesian estimation of motion vector fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 9, pp. 910–927, Sep. 1992.
[44] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.
[45] Y. Dar and A. M. Bruckstein, "Motion-compensated coding and frame rate up-conversion: Models and analysis," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2051–2066, Jul. 2015.
[46] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33 ITU-T Q6/16, Austin, TX, USA, Apr. 2001.



Guo Lu received the B.S. degree in electrical engineering from the Ocean University of China, Shandong, China, in 2014. He is currently pursuing the Ph.D. degree with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include video coding and processing.

Xiaoyun Zhang received the B.S. and M.S. degrees in applied mathematics from Xian 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 has been nominated as National 100 Best Ph.D. Theses of China. Her research interests include computer vision and pattern recognition, image and video processing, 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, Xian, 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.

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 took several senior technical positions in England, including a Principal Engineer with Snell and Wilcox, Petersfield, U.K., from 1995 to 2000, a Video Architect with 3DLabs, Egham, U.K., from 2000 to 2001, a Consultant Engineer with Sony European Semiconductor Design Center, Basingstoke, U.K., from 2001 to 2004, and a Digital Video Architect with Imagination Technologies, Kings Langley, U.K., 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.