Front. Comput. Sci.
DOI 10.1007/s11704-015-4184-0

RESEARCH ARTICLE

A pixel-based outlier-free motion estimation algorithm for scalable video quality enhancement

Xuan DONG (✉), Jiangtao WEN

1 Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract Scalable video quality enhancement refers to the process of enhancing low quality frames using high quality ones in scalable video bitstreams with time-varying qualities. A key problem in the enhancement is how to search for correspondence between high quality and low quality frames. Previous algorithms usually use block-based motion estimation to search for correspondences. Such an approach can hardly estimate scale and rotation transforms and always introduces outliers to the motion estimation results. In this paper, we propose a pixel-based outlier-free motion estimation algorithm to solve this problem. In our algorithm, the motion vector for each pixel is calculated with respect to estimate translation, scale, and rotation transforms. The motion relationships between neighboring pixels are considered via the Markov random field model to improve the motion estimation accuracy. Outliers are detected and avoided by taking both blocking effects and matching percentage in scale-invariant feature transform field into consideration. Experiments are conducted in two scenarios that exhibit spatial scalability and quality scalability, respectively. Experimental results demonstrate that, in comparison with previous algorithms, the proposed algorithm achieves better correspondence and avoids the simultaneous introduction of outliers, especially for videos with scale and rotation transforms.

Keywords motion estimation, scalable video coding, video super resolution.

Received April 9, 2014; accepted September 21, 2014

E-mail: dongx10@mail.tsinghua.edu.cn

1 Introduction

With the increasing use of MPEG dynamic adaptive streaming (DASH) [1] and scalable video coding (SVC) standards [2], streaming servers on the sender side often adjust the video rate to prevent the playback on the receiver side from stalling. As a result, on the receiver side, video bitstream quality varies with time. This kind of video is called scalable video, where some pre-buffered frames are of high quality while other frames are of low quality, in which the fine details of the enhancement layers are missing due to lower resolution or quality. Scalable video quality enhancement refers to the process of enhancing low quality (LQ) frames using the corresponding details of the high quality (HQ) frames, thus improving the quality of the live streamed or pre-encoded video bitstreams with time-varying quality.

The key problems in scalable video quality enhancement are to search for the correspondence between HQ and LQ frames and select the high frequency (HF) components of HQ frame to recover those of LQ frame. It remains a difficult problem for the following reasons: 1) motions between HQ and LQ frames are often complicated in practical videos, including translation, scale, and rotation. 2) In textureless regions, the intensity based metric of motion estimation (ME) is useless. 3) Many occlusions exist when the intervals between HQ and LQ frames are large. 4) Lighting conditions and compression ratios may have large variation between frames due to a change of scene lighting or camera parameters, change of resolution, or quantization parameter (QP) values, etc. The goal of an ME algorithm for scalable video quality enhance-

ment is to exploit the correspondence between HQ and LQ frames as much as possible while avoiding introducing outliers to the motion estimation results. In this paper, outliers are points where the wrong correspondence is calculated by the ME algorithm.

Previous algorithms often utilize block-based motion estimation (BBME) with a sum of absolute differences (SAD) strategy to search for correspondence. For example, in [3], the motion-compensated super resolution (MSR) method utilizes hierarchical ME and bi-directional ME with overlapped block motion compensation (OBMC). In the learning stage, the learn-based super resolution (LSR) method trains blocks using K-means clustering to obtain K clusters of blocks. In the inference stage, for each block, the method performs a full search of the K clusters to get the best match. In [4], variable-block-size ME with OBMC are utilized. In [5], block-based Scale-Invariant Feature Transform (SIFT) feature [6] matching is utilized, followed by warping each block. However, BBME is not good at estimating complicated motions because the pixels within each block may have different motions and the best-match block may have different scale or rotation. BBME also does not perform robustly in textureless regions because it does not consider the spatial relationships of pixels. The SAD metric is affected by changes in lighting or compression ratio and is not sensitive enough to outliers because many outliers have similar information in low frequency (LF) components but different information in HF components, as introduced in [7]. As a result, the available areas in a HQ frame that can help enhance an LQ frame are not fully exploited and the BBME results have many outliers, leading to introducing artifacts into the enhanced video.

In order to exploit more available areas in HQ frames when there exist scale and rotation transforms, and textureless regions while at the same time avoid introducing outliers when lighting conditions and compression ratio change, we present a pixel-based outlier-free motion estimation algorithm. First, our pixel-based model is able to describe complicated motions including scale and rotation transforms because it generates a motion vector for each pixel instead of a block of pixels. Second, our model considers the motion relationship between neighboring pixels so that information from highly textured regions can be propagated into textureless regions, and some incorrect pixel correspondence can be repaired using neighbor pixels results. Third, based on the observation that outliers are always distinguishably different from correct pixels in HF components, our outlier detection model uses blocking effects and the matching percentage in the SIFT field, for the reason that these two factors are sensitive to differences in HF components while insensitive to differences in LF components, such as change of lighting conditions. Thus the outlier detection model is not only sensitive to outliers but also robust to change of lighting and compression ratio.

We conduct experiments on two scenarios: spatial scalability and quality scalability. In the first scenario, the LQ frames have lower resolution than the HQ frames. The problem in this scenario is the same as that in video super resolution discussed in [3–5, 8–10]. In the second scenario, the LQ frames have higher QP values than the HQ frames. The problem in this scenario can be regarded as a cross segment enhancement problem as discussed in [11, 12]. The general diagram for enhancing a spatial scalability sequence is shown in Fig. 1, and a general diagram for enhancing a quality scalability sequence is shown in Fig. 2. The difference between the two is that the received LQ frames need to be upsampled at first in the spatial scenario. Then, in both scenarios, we search for correspondence and HQ frames are warped according to ME results, outliers are detected and HF components of LQ frames are repaired.

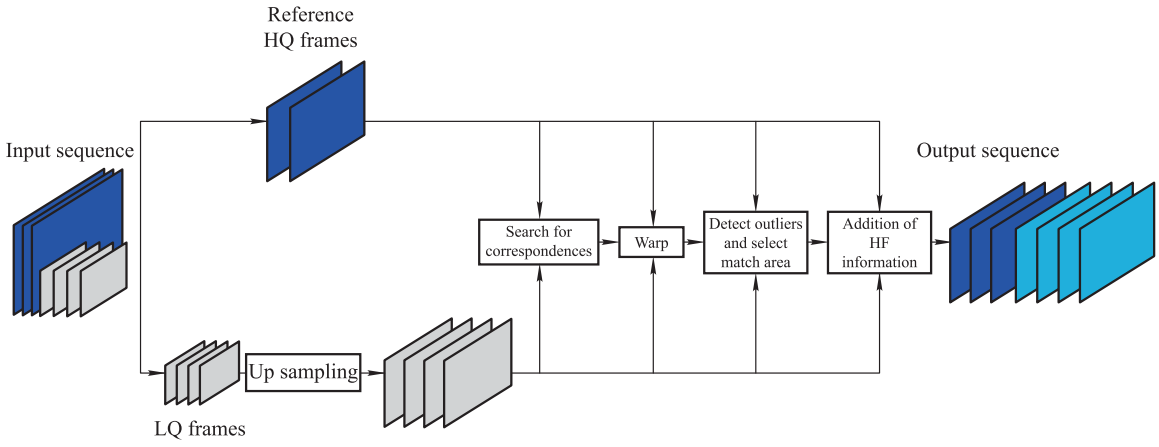

Fig. 1 General diagram of decoded video quality enhancement for spatial scalability sequence

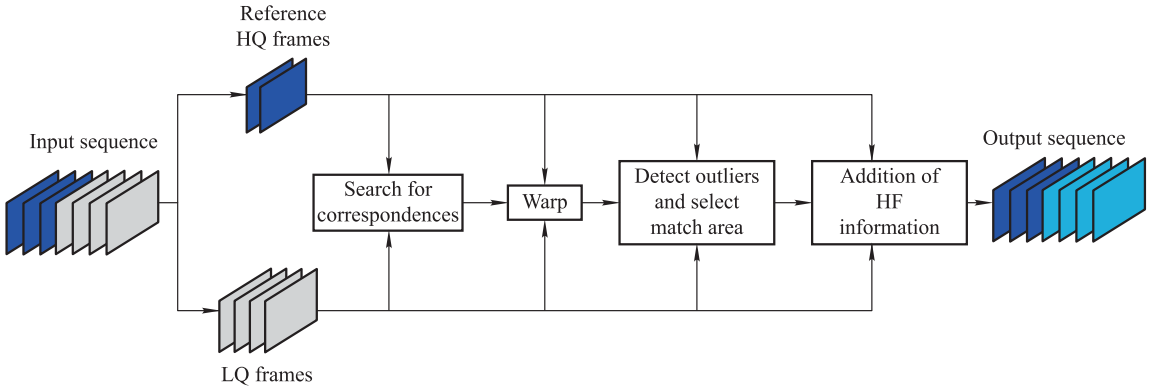

Fig. 2 General diagram of decoded video quality enhancement for spatial scalability sequence

Our experimental results demonstrate that, in comparison with previous algorithms, our algorithm can search for correspondence in many more areas and avoid introducing outliers at the same time, especially for videos with scale and rotation transforms.

Section 2 introduces related work, and Section 3 formulates the problem of pixel-based outlier-free motion estimation. Section 4 introduces the proposed algorithm. Experimental results are demonstrated and discussed in Section 5. Finally, conclusions are presented in Section 6.

2 Related work

In this section, we review related example-based image super resolution and video super resolution algorithms.

Image super resolution refers to the process in which a higher-resolution enhanced image is synthesized from one or more low-resolution images. Example-based image super resolution is a learning-based approach. In the learning stage, the algorithm learns the scene details that correspond to different image regions observed in the input. In the inference stage, those learned relationships are utilized to predict missing details in the target high-resolution image. The algorithm in [7] considers both intensity similarity and spatial smoothness of patches, and models the problem as a Markov network. The work in [13] extends the Markov network to handle the estimation of point spread function (PSF) parameters. The methods in [14] consider the symmetry of a cropped human face in the Markov network. The algorithm [15] utilizes vector quantization to organize example patches. We refer the reader to a more detailed taxonomy of example-based image super resolution algorithms in [16].

Video super resolution is more related to this paper's topic. Different from example-based image super resolution, in the problem of video super resolution, because some frames are of higher quality due to SVC, the algorithms learn from the HQ frames of the sequence to enhance the LQ frames. Related work can be found in [3–5, 8–10]. In [3], motion-compensated super resolution (MSR) and learn-based super resolution (LSR) are proposed. MSR utilizes hierarchical motion estimation (ME) and bi-directional ME with overlapped block motion compensation (OBMC) to search for correlation and replaces the whole block with the best matching block of the HQ frame. LSR builds an on-the-fly training dictionary using K-means clustering in the learning stage. In the inference stage, for each block, the algorithm performs a full search of the K clusters of blocks to find the best match. Then, a least mean squares filter is utilized to enhance details. In [4, 5], the authors extend their work in [8–10]. The algorithm in [4] utilizes variable-block-size motion estimation and OBMC in low-pass filtered blocks to discover K similar blocks, and replace the HF components of the block using the average of the HF components of the K blocks in the HQ frames. The work in [5] considers the projective transform of scenes. It utilizes block constrained SIFT feature matching for each block, warps each block respectively, and selects the best matching block with minimum SAD in the gradient field.

The absence of an efficient stochastic computing method, especially a method that can search for correspondence accurately and avoid introducing outliers, makes scalable video quality enhancement less attractive. In this paper, we propose a pixel-based outlier-free motion estimation algorithm so as to achieve better enhancement results.

3 Problem formulation of pixel-based outlier-free motion estimation

In this section, we formulate the problem of pixel-based outlier-free motion estimation for scalable video quality en-

hancement. This can help us understand the fundamental problems that we are facing, how to describe them mathematically, and expose the key factors to solve these challenges.

3.1 Pixel-based motion estimation

In this subsection we formulate the problem of pixel-based motion estimation between a LQ frame and a HQ frame. The goal is to solve three problems: complicated transforms, textureless regions, and noise.

In videos, because multiple objects move in different directions with different speeds, and the camera itself also moving, motion transforms can be very complicated. Depending on the motion types, the transform can be classified into three levels: euclidean translation, affine transform, and projective transform. Since motion in practical videos always belongs to projective transform, a robust algorithm should be able to find correspondence where scene motions belong to a projective transform that includes scale and rotation transforms. In addition, an intensity-based metric like SAD is useless in textureless regions. Thus, information from highly textured regions needs to be propagated into textureless regions so as to find the correct correspondence in textureless regions. Noise always exist in HQ and LQ frames because of unavoidable light variations, compression ratio variations, image blurring, and so forth. So the algorithm should also be robust and insensitive to noise.

Unlike traditional BBME, the proposed ME estimates motion vectors for each pixel and considers the motion relationship between neighboring pixels. This is a Bayesian problem and it can be modeled as a conditional probability

$$P(u|L, H; \Omega), \tag{1}$$

where u is the estimated motion vector for each pixel, L is the LQ frame, H is the HQ frame, and Ω is the model parameters. According to Bayesian rule, we have

$$P(u|L, H; \Omega) \propto P(u|\Omega_S)P(L, H|u; \Omega_D). \tag{2}$$

The model parameter Ω_S describes a spatial term: how the motion vector is expected to vary across the picture, i.e. the relationship between neighboring pixel motion vectors. The model parameter Ω_D describes a data term: how the pixel of an LQ frame is different from the pixel of an HQ frame with motion vector u. And the best estimation can be obtained by maximizing the *a posteriori* probability (MAP), i.e.,

$$\arg\max_{u \in S} P(u|\Omega_S)P(L, H|u; \Omega_D), \tag{3}$$

where S is the set of all the possible ME values. Then the problem of searching for correspondence can be transferred to computing the MAP of Bayesian Labeling.

The advantage of pixel-based motion estimation is that it can describe more complicated motions than BBME such as scale and rotation transforms. By utilizing the Markov Random Field (MRF) model, the accuracy can be improved by considering the neighboring motion relationships because information from highly textured regions can be propagated into textureless regions and some wrong results in small areas can be corrected with the help of neighboring pixels.

3.2 Outlier detection

Pixel-based ME can usually give the best estimation for each pixel. However, due to scene changes, occlusions, background clutter and incorrectly calculated correspondence, the best estimation result always contains outliers. Without detecting the outliers, the enhanced results will produce artifacts in corresponding regions.

The problem of outlier detection can be modeled as a conditional probability

$$P(c|L, R; \Omega), \tag{4}$$

where L is the LQ frame, R is the warped HQ frames, and the warping is performed according to the pixel-based ME results. c is each chosen pixel from the LQ frame or the warped HQ frames. And, Ω is the outlier detection model parameters. When the warped HQ frames are detected as outliers according to Ω, the pixels will choose the LQ frame. Otherwise, the pixels will choose the HQ frame enhancement details.

The best choice can be obtained by maximizing the conditional probability, i.e.,

$$\max_{c \in \{L,R\}} P(c|L, R; \Omega). \tag{5}$$

4 Algorithm description of pixel-based outlier-free motion estimation

In this section, the pixel-based outlier-free motion estimation algorithm will be illustrated in detail.

4.1 Pixel-based motion estimation

The problem of pixel-based motion estimation between LQ and HQ frames is equivalent to computing the MAP of Bayesian labeling (see Eq. 3). This is similar to the optical flow problem. The difference is that, different degrees of

degradation introduced by encoding process and long frame intervals between LQ frame and reference HQ frame mean that the commonly used brightness constancy and spatial smoothness assumptions in optical flow are not valid for correspondence between the LQ frame and the reference HQ frame. The key to solving this problem is how to model the Bayesian labeling and how to define the model parameters including data term and spatial term. MRF provides a tool for analyzing spatial or contextual dependencies of physical phenomena. According to the Hammersley Clifford theorem [17], the model is an MRF if and only if the probability distribution of the configuration is a Gibbs distribution. So we adopt MRF to model the problem. According to the Gibbs distribution, we have

$$P(u|\Omega_S)P(L,H|u;\Omega_D)_i = Z^{-1}e^{-\frac{E}{T}}, \tag{6}$$

where Z and T are constants, and the objective function is $E(u) = V_D + V_S$. V_D is the data term's objective function and V_S is the spatial term's objective function. Thus, the MAP solution is found by

$$\min_{u\in S} E(u), \tag{7}$$

and a proper definition of data term V_D and spatial term V_S is key to determining accurate correspondence between LQ and HQ frames. As mentioned in previous work [18], we define the spatial term V_S as

$$V_S = \sum_{s\in S_4} \|u - u_s\|, \tag{8}$$

where S_4 is the set of 4-neighbor pixels, and u_s is the motion vector of pixel s's 4-neighbor pixels. This term can help search for correspondence of pixels in textureless regions. Inspired by the idea of SIFT flow as described in [19], we define the data term V_D as

$$V_D = \sum_x \|s(L(x)) - s(R(x))\|, \tag{9}$$

where $s(x)$ is the SIFT descriptor of pixel x, $R(x)$ is the warped HQ frame's pixel, and L is the LQ frame. Because SIFT has an accuracy advantage and is stability scale and radiationally invariant, it can help measure intensity similarity between two pixels, in spite of the effects introduced by encoding process, projective transform due to motion, and noises. The SIFT descriptor is initially designed to be utilized to find correspondence among a large database of images. To distinguish the feature among the database of images, only a small part of the pixels is selected as feature points. However, in this paper case, the HQ and LQ frames are temporally correlated so we can estimate dense correspondence between them. In addition, we perform outlier detection when selecting matching pixels for enhancement. Thus, the SIFT descriptor is utilized at every pixel.

To solve the MRF, we adopt standard coarse-to-fine loopy belief propagation (LBP) from [19] that significantly improves performance. It roughly estimates the flow at a coarse level of image grid, then gradually propagates and refines the flow from coarse to fine. Details can be found in [19].

Even though optical flow methods always give a best estimate for each pixel, we cannot use them directly in our work. Due to the existence of occlusions and wrong estimations, not all of the ME results are correct. So we need to detect these outliers and avoid introducing them to the enhanced results. To improve the performance of the pixel-based ME and make the ME process outlier-free, we propose an outlier detection method in the following subsection.

4.2 Outlier detection

The outlier detection problem is equivalent to the problem of maximizing the conditional probability of Eq. (5). Because we utilize fast fourier transform (FFT) to divide the HF and LF components to later replace the HF components of each LQ frame with the HF components of the warped reference HQ frame, the minimum computation unit in the conditional probability is defined as a block instead of pixel. To simplify the computation, we assume that the probability that each block's estimated correspondence contains outliers is independent and identically distributed. So the problem can be solved by calculating the maximum likelihood estimation (MLE) of the conditional probability:

$$\max_{c\in\{L,R\}} \prod_c P(c|L,R;\Omega), \tag{10}$$

i.e.,

$$\prod_c \arg\max_{c\in\{L,R\}} P(c|L,R;\Omega), \tag{11}$$

where c is the choice from LQ and warped HQ frames of each block according to outlier detection model. We can define the likelihood $P(c|L,R;\Omega)$ as:

$$P(c|L,R;\Omega) \propto \exp(-E_{OD}(c|L,R;\Omega)), \tag{12}$$

where $E_{OD}(c|L,R;\Omega)$ is the objective function to describe the similarity between block c of LQ frame and the corresponding block of the candidate warped reference HQ frame. Thus, the calculation of MLE of Eq. (11) is equivalent to calculating the minimum energy of the objective function, i.e.,

$$\prod_c \arg\min_{c\in\{L,R\}} E_{OD}(c). \tag{13}$$

The objective function considers two factors: 1) blocking effects value if replacing the HF components of an LQ frame in block c using HF components of the candidate warped reference HQ frame's block, and 2) the percentage of false matching pixels in the SIFT field of each block. These two terms are based on the observation that 1) if the ME result is an outlier, replacing the HF components will produce significant blocking effects, and that 2) the fewer matching pixels in the SIFT field, the lower the probability that two blocks are in the same scene. An example of the blocking effects caused by outliers is shown in Fig. 4. In Fig. 4, we directly replace the HF components of the HQ frame with HF components of the LQ frame using FFT. Because we do not search for correspondence and warp the HQ frame, the motion areas, i.e., the face areas of the man, are outliers. As a result, as shown in the enhancement result, the face areas produce many blocking effects.

Thus, we propose to define E_{OD} as

$$E_{OD}(b) = Q_{BLK}(b) + s_{BLK}(b), \tag{14}$$

where $Q_{BLK}(b)$ is block b's blocking effect value and $s_{BLK}(b)$ is the percentage of false matching pixels in the SIFT field of block b.

To evaluate the blocking effects, after block c's HF components are replaced by those of each candidate block, we propose term Q_{BLK} to measure the blocking effects of the reconstructed block, inspired by [20],

$$Q_{BLK}(b) = \begin{cases} \max\limits_{e} \frac{N_e(b)}{D_e(b)} : \max\limits_{e} \frac{N_e(b)}{D_e(b)} < T_q, y \notin L; \\ T_q + 1 \quad : \max\limits_{e} \frac{N_e(b)}{D_e(b)} \geqslant T_q, y \notin L; \\ T_q \qquad\qquad\qquad\quad : y \in L, \end{cases} \tag{15}$$

where e is one of the four edges of block b, y is the selection from LQ or warped HQ frames, and T_q is the blocking effects value threshold. At each edge of block b, as denoted in Fig. 3, we define $N(b)$ and $D(b)$ between blocks A and B as

Fig. 3 Example of blocking effects introduced by outliers. (a) LQ frame; (b) HQ frame; (c) enhancement result by replacing HF components of HQ frame to HF components of LQ frame directly

$$N(b) = \sum_{i=1}^{L_b} |a_{i1} - b_{iL_b}|, \tag{16}$$

and

$$D(b) = N(b) + \sum_{i=1}^{L_b} \Big(\sum_{j=\frac{L_b}{2}+1}^{L_{b-1}} |b_{i(j+1)} - b_{ij}| + \sum_{j=1}^{\frac{L_b}{2}-1} |a_{i(j+1)} - a_{ij}| \Big), \tag{17}$$

where a and b are edge pixels, as shown in Fig. 3, L_b is the length of the block edge.

To measure the percentage of false matching pixels in the SIFT field, we propose another term s_{BLK}. Comparing the distance of the closest neighbor with that of the second-closest neighbor performs well and has been tested by much work in the computer vision fields [21]. Inspired by this strategy, for each pixel x in block b, except for the pixel itself, we search for the pixel y with the most similar SIFT value in the whole image $W(H)$. If

$$\frac{s(x)}{s(y)} < T_v, \tag{18}$$

pixel x is a correct match. Otherwise, it is a false match. $s(x)$ is pixel x's SIFT value. T_v is the matching threshold in the SIFT field. The percentage of false matching pixels is G_p. We define

$$s_{BLK} = \begin{cases} G_p \quad\quad : G_p < T_p, y \notin L; \\ T_p + 1 : G_p \geqslant T_p, y \notin L; \\ T_p \qquad\quad : y \in L, \end{cases} \tag{19}$$

where T_p is the threshold of false matching pixels. According to the definition of s_{BLK} and Q_{BLK}, G_p and $\max\limits_{e} \frac{N_e(b)}{D_e(b)}$ lie on the range [0, 1]. According to Eqs. (15) and (19), if either of Q_{BLK} or s_{BLK} is larger than T_p or T_q, this block will not be selected as the best match because the total energy will be larger than the energy of choosing the block b itself, i.e., $T_p + T_q$. Only if both of these two terms are lower than the threshold, can the block be chosen as the best matching block. If all candidate blocks of the warped reference HQ frame are detected as outliers, the minimum E_{OD} will be that of the block c itself and the reconstruction of this block will not be performed at all. Otherwise, if there exist candidate blocks that are not outliers, the proposed algorithm will choose the block with the minimum E_{OD} and enhance the block of the LQ frame with the chosen one.

5 Experimental results

In order to evaluate performance, we used popular H.264 and high efficiency video coding (HEVC) test video sequences:

four 352 × 288 sequences (*Container*, *Hall*, *Mobile*, *News*), one 640×480 sequence (*ParkScene*), one 720×480 sequence (*Chromakey*), and two 832×480 sequences (*BasketballDrill*, *Flowervase*). Of these videos, *BasketballDrill*, *News*, *Hall* are of simple motions. *Flowervase*, *ParkScene*, *Chromakey*, *Container* belong to projective motion, under scale or rotation transform. *Chromakey*, *ParkScene* contain a lot of noise and many outliers. *Mobile* contains large textureless regions. The experiments are conducted on a Windows PC (Intel Core 2 Duo T6500 at 2.0 GHz with 3 GB of RAM). The implementation language is C++. The average processing time is 2.82 seconds per frame for 352 × 288 sequences, 7.06 seconds per frame for 640 × 480 sequences, 8.79 seconds per frame for 720 × 480 sequences, and 9.39 seconds per frame for 832 × 480 sequences.

In the spatial scalability experiments, this paper sets the scaling ratio to 2. To obtain the LQ frames, the original frames are down-sampled by 2 at height and width and then up-sampled to the original resolution with Bi-cubic interpolation. In the quality scalability experiments, this paper sets the QP difference between HQ and LQ frames to 5. The HQ frames QP value is set to 37 while the LQ frames QP value is set to 42. The videos are encoded and decoded by HEVC HM6.0. The coding mode is *lowdelay P*. For each video, there are two groups of pictures (GOP). the first GOP consists of the first 30 frames encoded as HQ frames while the second GOP consists of all of the following frames encoded as LQ frames. In our experiments, all the algorithms utilize the HQ frames to enhance the LQ frames.

We compared the proposed algorithm with a baseline, motion-compensated super resolution (MSR) [3], learning-based super resolution (LSR) [3], Hung et al.'s algorithm [4], and Horn and Schunck's (HS) pixel-based motion estimation algorithm [22]. The baseline algorithm is Bi-cubic interpolation in spatial scalability and standard decoding in quality scalability. Subjective quality as well as PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index) are compared. To compare with HS, we first use HS to estimate the motions between pixels of different frames. Then HQ frame is warped according to the estimated motions. Then, FFT is performed to divide the LQ frame and warped HQ frame into HF and LF components to replace the HF components of LQ frame with the HF components of warped HQ frame. In Hung et al ME algorithm, a single matching block size of 16 × 16 that provides the best performance was evaluated. In implementing MSR, the motion search range for MSR was vertically and horizontally set to 64 pixels. $L \times L$ and $M \times M$ were set to 4 × 4 and 16 × 16, respectively. In implementing LSR, K was set to 512. The parameters of our proposed algorithm are listed below. Block size is 16 × 16, blocking effects value threshold T_q in (15) is 0.5, matching threshold in SIFT field T_v in Eq. (18) is 0.8, and the threshold of false matching pixels T_p in Eq. (19) is 0.3.

Tables 1 and 2 show the PSNR and SSIM results of applying the proposed algorithm and the other algorithms to all test videos in spatial scalability. Tables 3 and 4 show the PSNR and SSIM results of applying the proposed algorithm and the other algorithms to all test videos in quality scalability. The first 30 frames of the sequences are encoded as a training set of HQ frames, and the following frames are encoded as LQ frames that we propose to enhance. As shown in Tables 1, 2, 3, and 4, our proposed algorithm gets better PSNR and SSIM than all the other algorithms including the baseline and the algorithms in ref. [3, 4, 22] for almost all the test videos. The LSR and Hung et al algorithm could enhance better than the baseline for *Container*, *News*, *Hall*, and *BasketballDrill*, while perform worse than the baseline for *Mobile*, *ParkScene*, *Flowervase*, and *Chromakey*. MSR performs relatively worse than LSR. HS performs well for *News*, *Hall*, and *BasketballDrill*. But its performance for the other sequences is not good.

Table 1 Average PSNR [dB] in spatial scalability, including the baseline, MSR, LSR, Hung et al. algorithm, HS, and the proposed algorithm

Sequence	Resolution	Baseline	MSR	LSR	Hung et al.'s algorithm	HS	Proposed algorithm
Container	352x288	31.37	30.81	32.22	32.48	30.61	**32.67**
Hall	352x288	32.32	35.08	37.14	37.28	37.32	**37.42**
Mobile	352x288	27.06	26.81	26.42	27.02	26.21	**27.41**
News	352x288	33.62	34.29	35.78	36.29	36.33	**36.11**
Parkscene	640x480	35.83	31.27	33.91	34.56	30.85	**37.27**
Flowervase	832x480	33.51	32.24	32.67	33.40	32.07	**34.71**
Chromakey	720x480	31.56	29.75	30.87	31.41	30.37	**31.55**
BasketballDrill	832x480	36.06	35.75	38.60	38.62	38.65	**39.67**

Table 2 Average SSIM in spatial scalability, including the baseline, MSR, LSR, Hung et al. algorithm, HS, and the proposed algorithm

Sequence	Resolution	Baseline	MSR	LSR	Hung et al.'s algorithm	HS	Proposed algorithm
Container	352x288	0.86	0.79	0.88	0.88	0.78	**0.90**
Hall	352x288	0.91	0.88	0.96	0.96	0.96	**0.98**
Mobile	352x288	0.74	0.66	0.69	0.73	0.65	**0.76**
News	352x288	0.92	0.88	0.93	0.94	0.94	**0.96**
Parkscene	640x480	0.96	0.79	0.89	0.90	0.78	**0.97**
Flowervase	832x480	0.92	0.84	0.88	0.90	0.84	**0.93**
Chromakey	720x480	0.86	0.73	0.79	0.83	0.74	**0.87**
BasketballDrill	832x480	0.96	0.90	0.96	0.96	0.97	**0.99**

Table 3 Average PSNR [dB] in quality scalability, including the baseline, MSR, LSR, Hung et al. algorithm, HS, and the proposed algorithm

Sequence	Resolution	Baseline	MSR	LSR	Hung et al.'s algorithm	HS	Proposed algorithm
Container	352x288	32.10	30.85	31.56	31.88	30.53	**32.26**
Hall	352x288	33.26	33.82	33.72	33.65	33.84	**33.85**
Mobile	352x288	27.37	27.15	26.72	27.25	26.61	**27.38**
News	352x288	32.75	32.08	32.67	32.72	32.78	**32.91**
Parkscene	640x480	31.37	30.03	31.20	31.25	29.89	**31.35**
Flowervase	832x480	34.23	32.71	32.26	33.61	32.01	**34.17**
Chromakey	720x480	29.80	29.15	29.13	29.74	29.31	**29.80**
BasketballDrill	832x480	32.42	32.02	32.15	32.62	32.62	**32.64**

Table 4 Average SSIM in quality scalability, including the baseline, MSR, LSR, Hung et al. algorithm, HS, and the proposed algorithm

Sequence	Resolution	Baseline	MSR	LSR	Hung et al.'s algorithm	HS	Proposed algorithm
Container	352x288	0.83	0.80	0.81	0.81	0.80	**0.84**
Hall	352x288	0.86	0.85	0.87	0.87	0.87	**0.88**
Mobile	352x288	0.80	0.77	0.78	0.79	0.76	**0.81**
News	352x288	0.86	0.80	0.86	0.84	0.86	**0.88**
Parkscene	640x480	0.81	0.78	0.81	0.81	0.78	**0.83**
Flowervase	832x480	0.87	0.82	0.84	0.85	0.81	**0.88**
Chromakey	720x480	0.79	0.76	0.78	0.78	0.78	**0.80**
BasketballDrill	832x480	0.85	0.82	0.84	0.85	0.86	**0.87**

(a) (b) (c)

Fig. 4 PSNR performance with the change of down-sampling ratio of LQ frames for sequences (a) *Hall* and (b) *News*

For simple sequences such as *BasketballDrill*, *News*, *Hall*, our algorithm and the algorithms of LSR, Hung et al. and HS can enhance better than the baseline, and our algorithm gets the best performance. This is because in simple sequences where the motions between different frames are small and simple, there are many pixels in HQ frames that can be used to enhance LQ frames. So most algorithms can get better PSNR and SSIM than the baseline. Since the proposed algorithm can exploit more correct motions and avoid more wrong motions, our algorithm achieves the best performance. For complicated sequences such as *Chromakey*, *ParkScene*, *Flowervase*, *Mobile*, the proposed algorithm can also get better or equal PSNR and SSIM compared with the baseline, while MSR, LSR, Hung et al.'s algorithms and HS are not always as good as the baseline. This is because in complicated sequences where motions between different frames are large and complicated, there are not many pixels in HQ frames that can be used to enhance LQ frames. So the improvement of our algorithm becomes smaller. In the quality scalability case, since the quality differences between HQ and LQ frames are small in our experiments, for each matching pixel, the improvement is not significant. As a result, for complicated sequences in quality scalability, the improvements will be limited. Due to the outlier-detection module, performance of the proposed algorithm is almost equal to the baseline while the other algorithms are sometimes worse than the baseline. In short, we notice the robustness and effectiveness of our algorithm. When there are few pixels that can be enhanced, the robustness of our proposed algorithm can get equal performance to that of the baseline while MSR, LSR, Hung et al.'s algorithm and HS perform worse than the baseline due to outliers in the ME results. When there are many pixels that can be enhanced, our proposed algorithm achieves much higher performance than the baseline, MSR, LSR, Hung et. al algorithm, and HS.

In Figs. 5 and 6, we also show the PSNR and SSIM performance of different algorithms with the change of down-sampling ratio in spatial scalability. The video sequences are *Hall* and *News*. As shown in Figs. 5 and 6, when the down-sampling ratio becomes smaller, MSR, LSR, Hung et. al algorithm, HS and our algorithm achieve higher quality improvement than the baseline because the quality of HQ frames becomes better than the quality of LQ frames. When the down-sampling ratio becomes larger, MSR, LSR, Hung et.al algorithm and HS will get less quality improvement or even worse results than the baseline because the differences between HQ and LQ frames become smaller and there might be some outliers in ME results. However, the proposed algorithm can always perform better than the other algorithms and even the LQ frames have the same quality with HQ frames, the proposed algorithm will not get worse results than the baseline. This is because our outlier detection model can avoid introducing outliers in the ME results.

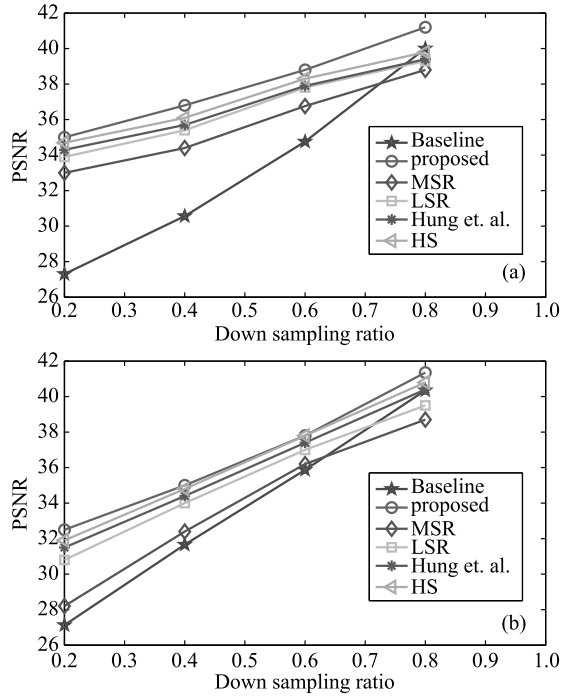

Fig. 5 SSIM performance with the change of down-sampling ratio of LQ frames for sequences (a) *Hall* and (b) *News*

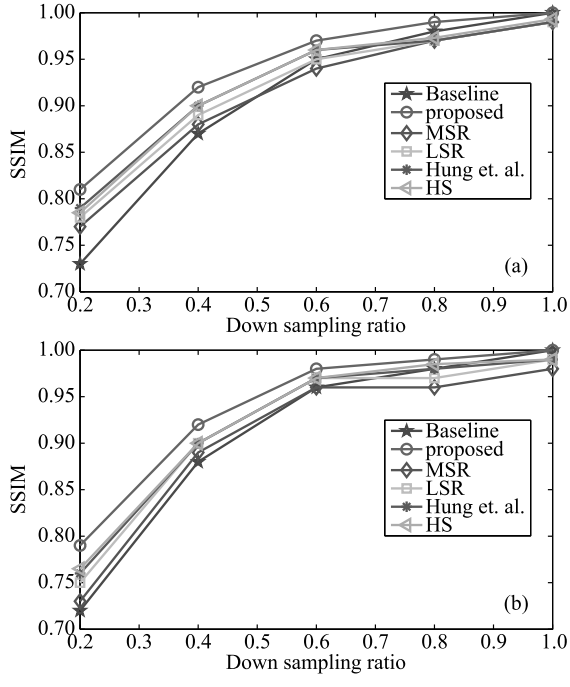

Fig. 6 SSIM performance with the change of down-sampling ratio of LQ frames for sequences (a) *Hall* and (b) *News*

Figures. 7, 8, and 9 show the subjective results of our proposed algorithm and the other algorithms. In Fig. 7, results of MSR, LSR, Hung et al algorithm and HS produce artifacts in textless regions, while the results for our algorithm are accurate. This is because for textureless regions, our algorithm propagates information from highly textured regions into neighboring regions and achieves global optimization by solving spatial term of Eq. (8). In comparison, the other algorithms do not take neighbor relationship into consideration, leading to artifacts in textureless regions. This not only decreases the PSNR, but has negative subjective effects. In Fig. 8, results of MSR, LSR and Hung et al. algorithm fail to enhance the face due to rotation. HS can enhance some areas of the face but also introduce some artifacts at the edge region of the face and the background region. The result of our algorithm contains more detail and sharpness in the face region without introducing artifacts. This is because pixel-based ME can deal with rotation and scale transforms and our algorithm is robust to outliers. Although the PSNR is not much higher than the baseline due to the limited area for enhancement, we can see that our algorithm outperforms the baseline and other algorithms in subjective quality. In Fig. 9, the results of the other algorithms produce many artifacts in the region of greens, while our algorithm avoids introducing artifacts. This is because the outlier detection module helps detect and avoid outliers in the ME results. In the regions of background greens, recognizing outliers is difficult because the details are too complicated. A simple method such as SAD is not accurate enough to detect them. Our algorithm solves this well and few artifacts are produced because the correct match in the SIFT field and blocking effects detection can efficiently recognize the outliers. However, the results of the other algorithms exhibit many artifacts in the regions of green. This not only causes a large PSNR decrease, but also greatly affects the subjective quality. To sum up, because the problems of projective transforms, outliers, textureless regions, and noise are well solved, our algorithm can get more area enhanced and avoid artifacts, resulting in the higher objective and subjective quality.

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

Fig. 7 Decoding result of one frame of sequence Mobile. (a) Resutt of the baseline; (b) MSR; (c) LSR; (d) Hung et al. algorithm; (e) HS; (f) Our algorithm

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

Fig. 8 Decoding result of one frame of sequence Chromakey. (a) Resutt of the baseline; (b) MSR; (c) LSR; (d) Hung et al. algorithm; (e) HS; (f) Our algorithm

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

Fig. 9 Decoding result of one frame of sequence ParkScene. (a) Resutt of the baseline; (b) MSR; (c) LSR; (d) Hung et al. algorithm; (e) HS; (f) Our algorithm

Figures. 10 and 11 show the PSNR with frame number. This directly shows how the algorithms perform as frame number increases. For motion sequences such as *Container*, the PSNR gains decrease when the frame number increases. This is because the region available to be enhanced decreases when the frame number increases. Because our algorithm is more accurate for a variety of motions, the PSNR gains are higher than the other algorithms. In addition, for complicated sequences such as *ParkScene*, because our method takes the factors of outliers into consideration, it can enhance LQ frames for longer. Even if there is no area available to be enhanced, our algorithm will not reduce the decoded video quality. And, because the other algorithms are not robust to outliers, their performance is even lower than the performance of the baseline. For noisy sequences such as *BasketballDrill* where the background light changes over time, our algorithm, LSR, Hung et al's algorithm and HS perform better than MSR and the baseline. This is because they enhance the details of a block by replacing the HF components of the block instead of replacing the whole block. This helps avoid introducing in noise in LF components. To sum up, when the movements between HQ and LQ frames are small, the enhancement areas will be large and the enhancement is always noticeable, but when the movements are large, for example, when new objects or scenes appear, the enhancement areas will decrease because there is no correspondence between HQ frames and the new objects of LQ frames.

30 human observers were asked to rate enhanced videos of the different algorithms and the mean opinion score is calculated to show the benefits of our algorithm. The scores range from 1 (worst) to 5 (best). Both scenarios of spatial scalability and quality scalability are tested with the algorithms of the baseline, Hung et al., MSR, LSR, HS and our algorithm. The results are shown in Table III. As shown in the table, our algorithm's results have higher subjective quality.

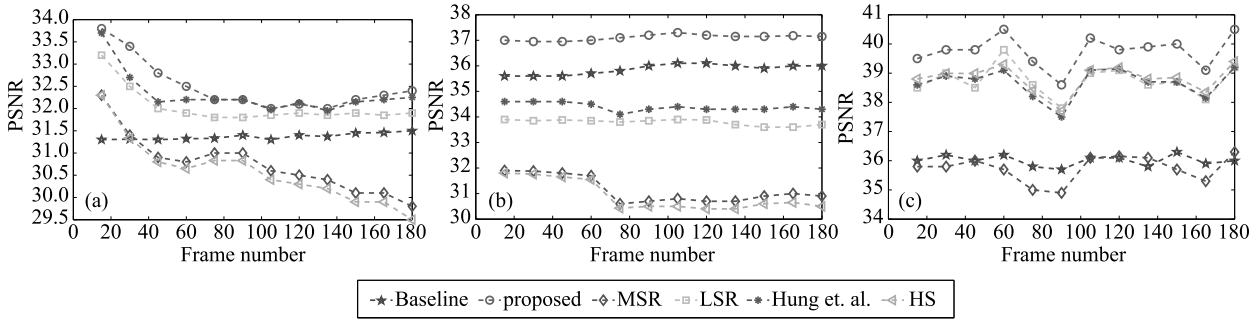

Fig. 10 PSNR/Frame number performance of our algorithm, Hung et al. algorithm, MSR, LSR, and HS in spatial scalability for (a) Container, (b) Parkscene, (c) BasketballDrill

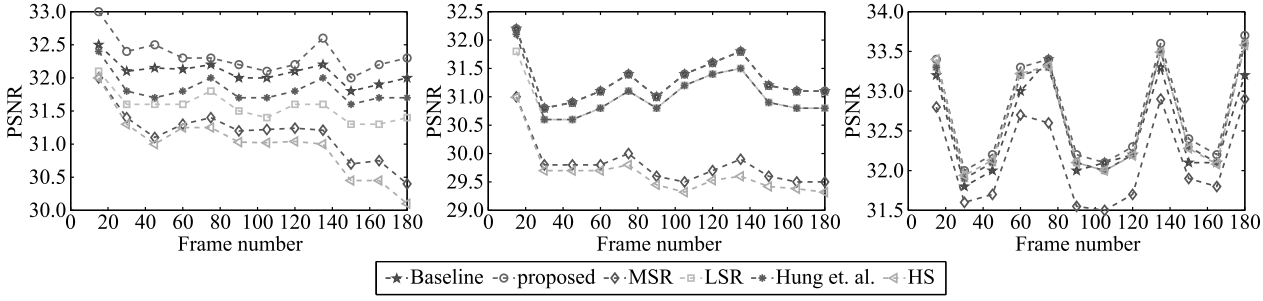

Fig. 11 PSNR/Frame number performance of our algorithm, Hung et al. algorithm, MSR, LSR and HS in quality scalability for (a) Container, (b) Parkscene, (c) BasketballDrill

Table 5 Mean opinion score given by human observers for different algorithms in scenarios of spatial scalability and quality scalability

	Spatial scalability	Quality scalability
Baseline	3.1	2.5
Hung et al.	3.7	2.9
MSR	2.6	2.1
LSR	3.3	2.7
HS	3.2	2.5
Our algorithm	4.1	3.3

6 Conclusions and future work

We have proposed a pixel-based outlier-free motion estimation algorithm for enhancing scalable video quality. Based on the observation that neighboring pixels in space should often have similar motions, our proposed method searches for correspondence between HQ and LQ frames for each pixel by computing the MAP of a Bayesian Labeling. This method is capable of searching for correspondences with scale and rotation transforms and improving the ME accuracy in textureless regions. In addition, our algorithm will detect motion outliers by taking blocking effects and matching percentage in SIFT field into consideration so as to get outlier-free ME results. Our experimental results demonstrate that our algorithm provides significantly better subjective visual quality as well as higher objective quality than previous algorithms with the improvement of rideo quality, many high-level visionproblems can also be better solved, such as Ref [23–27].

Our future work will include the acceleration of our algorithm. Although processing is not real-time at present, our algorithm can be accelerated by parallel processing in the future since the enhancement of different frames are independent. A possible solution is to use the GPU for acceleration.

Acknowledgements This work was supported by the National Science Fund for Distinguished Young Scholans of China (61125102), and the State Key Program of National Natural Science Foundation of China (Grant No. 61133008).

References

1. Sodagar I. The MPEG-DASH standard for multimedia Streaming Over the Internet. IEEE Multimedia, 2011, 18(4): 62-67
2. Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(9): 1103-1120
3. Song B C, Jeong S C, Choi Y. Video super-resolution algorithm using bi-directional overlapped block motion compensation and onthefly dictionary training. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(3): 274-285
4. Hung E M, de Queiroz R L, Brandi F, de Oliveira K F, Mukherjee D. Video super-resolution using codebooks derived from keyframes. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(9): 1321-1331
5. Ferreira R U, Hung E M, de Queiroz R L. Video super resolution based on local invariant features matching. In: Proceedings of 19th IEEE International Conference on Image Processing. 2012, 877-880
6. Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 17th IEEE International Conference on Computer Vision. 1999, 1150-1157
7. Freeman W T, Jones T R, Pasztor E C. Example-based superresolution. IEEE Computer Graphics and Applications, 2002, 22(2): 56-65
8. Brandi F, de Queiroz R, Mukherjee D. Super resolution of video using key-frames. In: Proceedings of the IEEE International Symposium on Circuits Systems. 2008, 1608-1611
9. Brandi F, de Queiroz R L, Mukherjee D. Super-resolution of video using key-frames and motion estimation. In: Proceedings of the 15th IEEE International Conference on Image Processing. 2008, 321-324
10. Oliveira K F, Brandi F, Hung E M, de Queiroz R L, Mukherjee D. Bipredictive video super-resolution using key-frames. In: Proceedings of SPIE Symposium on Electronic Image, Visual Information Processing and Communication. 2010, 1-5
11. Hung E M, de Queiroz R L, Mukherjee D. Inter-frame postprocessing for intra-coded video. Journal of Communication and Information Systems, 2013, 28(1): 1-7
12. Wen J, Li S, Lu Y, Fang M, Dong X, Chang H, Tao P. Cross segment decoding for improved quality of experience for video applications. In: Proceedings of the 2013 IEEE Data Compression Conference. 2013, 231-240
13. Wang Q, Tang X, Shum H. Patch based blind image super resolution. In: Proceedings of the 10^{th} IEEE International Conference on Computer Vision. 2005, 709-716
14. Stephenson T A, Chen T. Adaptive Markov random fields for example-based super-resolution of faces. Journal on Applied Signal Processing, 2006, 2006: 1-11
15. Qiu G. Interresolution look-up table for improved spatial magnification of image. Journal of Visual Communication and Image Representation. 2000, 11: 360-373
16. Elad M, Datsenko D. Example-based regularization deployed to superresolution reconstruction of single image. The Computer Journal Advance Access, 2007, 20: 15-30
17. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B, 1974, 36: 192-293
18. Sun D, Roth S, Lewis J, Black M J. Learning optical flow. Lecture Notes in Computer Science, 2008, 5304: 83-97
19. Liu C, Yuen J, Torralba A, Sivic J, Freeman W T. SIFT flow: dense correspondence across different scenes. Lecture Notes in Computer Science, 2008, 5304: 28-42
20. Pan F, Lin X, Rahardja S, Lin W, Ong E, Yao S, Lu Z, Yang X. A locally adaptive algorithm for measuring blocking artifacts in images and videos. Signal Processing: Image Communication, 2004, 19(6): 499-506
21. Brown M, Lowe D G. Automatic panoramic image stitching using invariant features. International Journal of Computer Vision, 2007,

74(1): 59-73

22. Horn B, Schunck B. Determining optical flow. Artificial Intelligence, 1981, 16: 185-203
23. Wang S, Uchida S, Liwicki M, Feng Y K. Part-based methods for handwritten digit recognition. Frontiers of Computer Science, 2013, 7(4): 514–525
24. Mehrotra H, Majhi B. Local feature based retrieval approach for iris biometrics. Frontiers of Computer Science, 2013, 7(5): 767–781
25. PRIYA R, Shanmugama T H. Comprehensive review of significant researches on content based indexing and retrieval of visual information. Frontiers of Computer Science, 2013, 7(5): 782–799
26. Wang Y W, Zhou Y C, Liu Y, Luo Z, Guo D H, Shao J, Tan F, Wu L, Li J H, Yan B P. A grid-based clustering algorithm for wild bird distribution. Frontiers of Computer Science, 2013, 7(4): 475–485
27. Kang L, Wu L D, Yang Y H. A Novel Unsupervised Approach for Multilevel Image Clustering from Unordered Image Collection. Frontiers of Computer Science, 2013, 7(1): 69-82

Xuan Dong received his BS in computer science and technology from Beihang University, China, in 2010. He is a PhD candidate in the Department of Computer Science and Technology, Tsinghua University. His current research interests include computational photography, video processing, video coding, and image segmentation.

Jiangtao Wen received his BS, MS, and PhD deqrees all in electrical engineering from Tsinghua University, China, in 1992, 1994, and 1996, respectively. From 1996 to 1998, he was a staff research fellow at the University of California, Los Angeles (UCLA). After UCLA, he served as the principal scientist at PacketVideo Corp., chief technical officer at Morphbius Technology Inc., director of Video Codec Technologies at Mobilygen Corp., and as a technology advisor at Ortiva Wireless and Stretch, Inc. Since 2009, he has been a professor in the Department of Computer Science and Technology, Tsinghua University. His research focuses on multimedia communication over challenging networks and computational photography.