Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Reliable tracking algorithm for multiple reference frame motion estimation

Tsz-Kwan Lee
Yui-Lam Chan
Chang-Hong Fu
Wan-Chi Siu

Reliable tracking algorithm for multiple reference frame motion estimation

Tsz-Kwan Lee
Yui-Lam Chan
Chang-Hong Fu
Wan-Chi Siu
The Hong Kong Polytechnic University
Centre for Signal Processing
Department of Electronic and Information Engineering
Hung Hom, Kowloon, Hong Kong
E-mail: enylchan@polyu.edu.hk

Abstract. *Multiple reference frame motion estimation (MRF-ME) is one of the most crucial tools in H.264/AVC to improve coding efficiency. However, it disciplines an encoder by giving extra computational complexity. The required computation proportionally expands when the number of reference frames used for motion estimation increases. Aiming to reduce the computational complexity of the encoder, various motion vector (MV) composition algorithms for MRF-ME have been proposed. However, these algorithms only perform well in a limited range of reference frames. The performance deteriorates when motion vector composition is processed from the current frame to a distant reference frame. In this paper, a reliable tracking mechanism for MV composition is proposed by utilizing only the relevant areas in the target macroblock and taking different paths through a novel selection process from a set of candidate motion vectors. The proposed algorithm is especially suited for temporally remote reference frames in MRF-ME. Experimental results show that compared with the existing MV composition algorithms, the proposed one can deliver a remarkable improvement on the rate-distortion performance with similar computational complexity.* © *2011 SPIE and IS&T.*
[DOI: 10.1117/1.3605574]

1 Introduction

H.264/AVC is an international video coding standard jointly developed by ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group.[1,2] The H.264/AVC standard has dominated in the video coding standardization community for the past several years. It has achieved a significant improvement in rate-distortion efficiency relative to all previous video coding standards.[3,4] The coding gain mainly comes from many of new sophisticated techniques such as the integer transform, deblocking filtering, quarter-sample accuracy for motion compensation, multiple reference frame motion estimation (MRF-ME), etc.[2,5,6]

Motion estimation (ME) is a process to find a prediction of pixels in the current frame from a reference frame, and it is a key step of frame rate up-conversion[7–10] and video coding.[11,12] For frame rate up-conversion, the best prediction is the one that results in the highest accuracy of motion trajectories. Unlike frame rate up-conversion, ME in video coding needs to find the best prediction with minimum residual energy instead of the true motion trajectory. The magnitude of prediction errors, rather than the accuracy of motion trajectories, is of the greatest importance for video coding. Among various ME algorithms, the block matching algorithm is considered the most mature and practically useful one since its implementation is simple.

The consideration of MRF-ME within the H.264 codec plays a major role in delivering better coding gain.[13] MRF-ME allows the codec to predict a picture using more than one reference picture for ME and compensation. It achieves more accurate prediction and higher coding efficiency, especially in the cases of uncovered backgrounds, repetitive motions, highly textured areas, lighting changes, etc.[3] A scenario of MRF-ME using N reference frames is illustrated in Fig. 1. For each block of the encoded frame, the motion vector (MV) is obtained by searching all possible locations within the search window in each reference frame. The optimal location in the reference frame for the current block being encoded is located by minimizing the Lagrangian cost function J_{motion}

$$J_{motion}(MV, \lambda_{motion}) = SAD(s, r) + \lambda_{motion} \cdot R_{motion}(MV - PMV), \quad (1)$$

where PMV is the motion vector used for prediction, λ_{motion} is the Lagrangian multiplier for ME, $R_{motion}(MV - PMV)$ is the estimated number of bits for coding MV, and SAD is the sum of absolute differences between the original block s and its reference block r. Adopting the full-search scheme in a frame-by-frame manner of MRF-ME incurs a considerable computational complexity in the encoder. The increased complexity is in proportion to the number of searched frames.[14,15] The more number of reference frames the encoder uses, the more demanding complexity it needs.

Various fast algorithms[4,6,16–27] have been proposed in the literature to reduce the computational complexity caused by MRF-ME. These algorithms can be classified into two categories. The first category is to employ early termination in MRF-ME by the consideration of the specific condition at a certain reference frame.[6,14,15,17–21] The condition always depends on spatial and temporal correlation

Paper 10209R received Dec. 2, 2010; revised manuscript received Apr. 13, 2011; accepted for publication Jun. 9, 2011; published online Jul. 14, 2011.

1017-9909/2011/20(3)/033003/14/$25.00 © 2011 SPIE and IS&T

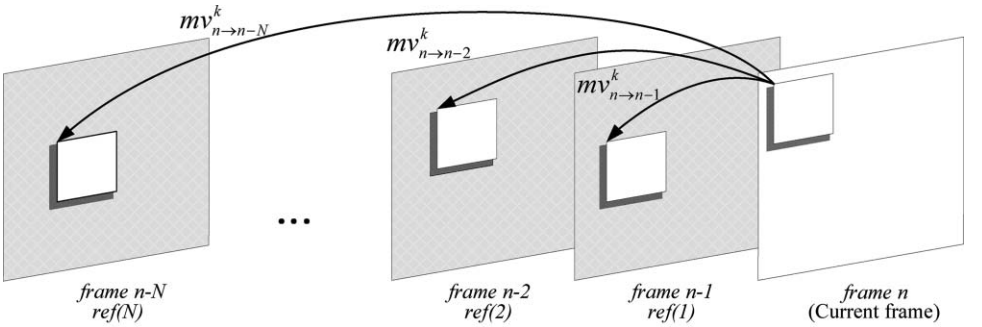

Fig. 1 ME with multiple reference frames.

in video sequences.[16,27] However, the early termination approach might select an undesirable reference frame, and the threshold value for early termination usually relies on the characteristics of video sequences. In other words, more reference frames are needed for sequences with fast motion activities, resulting in increased computational complexity. The performance of the early termination approach is, therefore, dependent on the characteristics of the video sequences.

The second category is to carry out ME with several candidate search points through MV composition for each reference frame.[4,22–27] This type of approach becomes one of the popular solutions for complexity reduction since it can be adaptively used in various video sequences and maintains coding efficiency. Different MV composition algorithms have been introduced to reduce the computational complexity in MRF-ME. In Refs. 23–25, MV composition has adopted forward dominant vector selection (FDVS) for vector selection criterion between frames, which is considered as one of the best methods in the MV composition algorithms of MRF-ME. It reuses the stored MVs between successive frames to synthesize the MVs of the second to fifth reference frames, as shown in Fig. 1. Without performing the ME and computing J_{motion}, its computational complexity can be greatly reduced. The algorithms in Refs. 4 and 26 have further used a weighted average of the neighboring MVs after MV composition by FDVS. In Ref. 27, the median MV in neighboring blocks has also been suggested for vector selection criterion in MV composition. They can achieve the desirable coding efficiency in most cases. In the circumstances of the long distance between the reference and current frames, these MV composition techniques do not work efficiently. The reason is that the new composed MVs might no longer represent the moving contents of the current macroblock (MB). As a result, prediction errors could not diminish as usual. In this case, the quality of the encoded videos deteriorates.

In this paper, a more faithful algorithm to compose new MVs has been proposed. The proposed algorithm is suitable for the case when the temporal distance between the reference and current frames is large. The success of the proposed method is based on examining the relevant area of the current MB in MV composition. It also tracks several possible candidates related to the current MB and selects the best candidate. The organization of this paper is as follows. In Sec. 2, we discuss the impacts on the performance of existing MV composition algorithms when the reference frame is temporally far away from the current frame. Section 3 describes our proposed algorithm for MRF-ME. Simulation results are evaluated in Sec. 4. Finally, some concluding remarks are provided in Sec. 5.

2 Impact on the Accuracy of Motion Vectors Composition in Multiple Reference Frame Motion Estimation

For MV composition, the MVs between successive frames are estimated by full-search motion estimation and are saved in a buffer for composing MVs in other reference frames of MRF-ME by means of various MV composition algorithms. For the sake of discussion, Fig. 2 defines a number of terms and notations for the rest of this paper. In Fig. 2, ref(i) is the *i*th reference frame from the current frame, frame *n*, and the number of searched reference frames, *N*, is equal to 5 in this example. Only four neighboring MBs within a frame are illustrated, and MB^k_n represents the *k*th MB in frame *n*. The MV of MB^k_n referencing to ref(i) is denoted by $mv^k_{n\to n-i}$. Figure 3 then shows an example of using FDVS in MRF-ME. Assume that *N* is equal to 3 in this example. To conduct MV composition between frame *n* and the target reference frame, ref(3), it is required to find the new MV of MB^1_n to ref(3), i.e., $mv^1_{n\to n-3}$ in dotted arrow shown in Fig. 3(a). For every MB, FDVS selects one dominant MV carried by a dominant MB that has the largest overlapping segment with the motion-compensated MB of MB^1_n in the previous reference frame. Considering the motion-compensated MB of MB^1_n overlaps with four MBs, MB^1_{n-1}, MB^2_{n-1}, MB^3_{n-1}, and MB^4_{n-1}, in frame $n-1$ of Fig. 3(a), MB^3_{n-1} is chosen as the dominant MB while its MV, $mv^3_{n-1\to n-2}$, is selected as the dominant MV. This dominant vector selection process is repeated until the desired reference frame is reached, i.e.. Ref. 3 in this

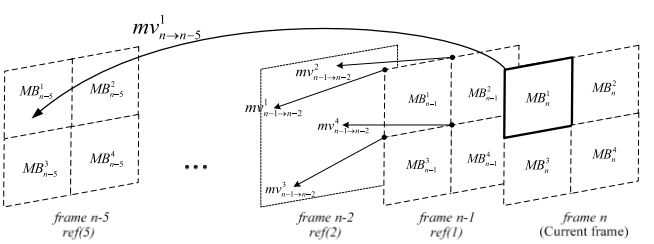

Fig. 2 MV composition for ME.

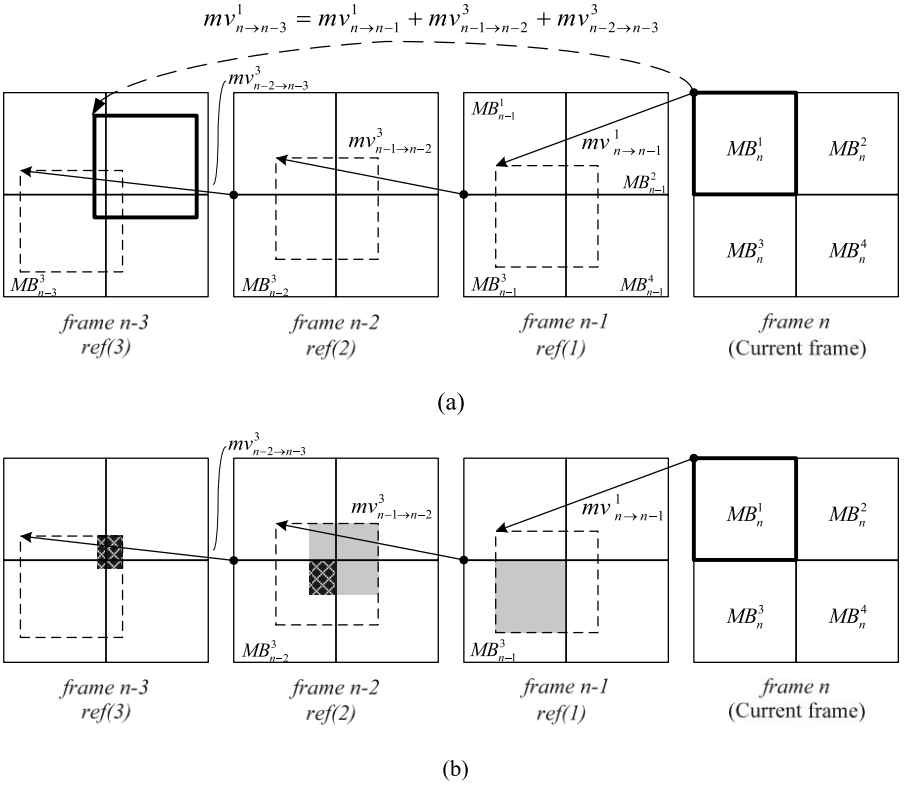

Fig. 3 FDVS in MRF-ME.

example. Here, $mv^1_{n\to n-3}$ is therefore composed by summing up the selected dominant MVs across the in-between frames and can be written as

$$mv^1_{n\to n-3} = mv^1_{n\to n-1} + mv^3_{n-1\to n-2} + mv^3_{n-2\to n-3}. \quad (2)$$

FDVS can provide promising results for MV composition for MRF-ME.[3,4] However, in fast-motion video sequences, the temporal distant reference frame is always used for MRF-ME due to existence of the fast moving objects. FDVS does not work well for this scenario. This phenomenon can be explained as portrayed in Fig. 3(b), which is redrawn from Fig. 3(a). In frame $n-1$, MB^3_{n-1} is selected to be the dominant MB and the corresponding $mv^3_{n-1\to n-2}$ is used to determine the dominant MB in frame $n-2$. It is observed that only the shaded area of MB^3_{n-1} is actually relevant to target MB, MB^1_n. Nevertheless, FDVS also utilizes the irrelevant nonshaded area in MB^3_{n-1} to compute the dominant MB in frame $n-2$. At the same time, the relevant area of MB^1_n further diminishes when far away reference frames are used. The cross-hatch shaded area only occupies a very minor portion of the dominant MB, MB^3_{n-2} as depicted in Fig. 3(b). It seriously affects the accuracy of the composed MVs since a large irrelevant area to the target MB is used to decide the dominant MB in frame $n-3$. Figure 4 then illustrates an example of performing FDVS on the 288th MB extracted from the 201st frame of the "Mobile" sequence. FDVS composes the new MV of this MB to ref(5), as depicted in Fig. 4(b). From Fig. 4(b), it is clear that the relevant areas to the 288th MB (indicated by the cross-hatch shaded areas) quickly diminish throughout the whole FDVS process. It incurs the inaccuracy of the resultant composed MV, which is demonstrated in Fig. 4(c). In Fig. 4(c), it is found that the resultant composed MV, $(-20,8)$, from FDVS is significantly deviated from the MV obtained by full-search ME. It is mainly due to the use of large irrelevant area for dominant vector selection in FDVS.

Tables 1–6 then show the percentage of MBs using different reference frames of FDVS and the full-search (FS) algorithm in the case of $N = 5$. From Tables 1–6, the use of the first and second reference frames dominates in FDVS due to the low accuracy of the composed MVs in temporal remote reference frames. On the other hand, FS exhibits a different tendency. It is likely to use temporal remote references as shown in Tables 1–6. The discrepancy between FS and FDVS is more obvious for sequences with complex motion activities such as Mobile and "Tempete." It is because the relevant area of the target MB further lessens in far away reference frames for these complex motion sequences. Tables 1–6 also show the results of using the median (MED) algorithm[27] for vector composition, and this algorithm encounters the same situation under the circumstance of using temporal distant reference frames in MV composition for MRF-ME. In other words, both FDVS and the median algorithms could not achieve satisfactory performance in sequences with complex motion since it could not fully track the relevant information for vector selections. Consequently, the accuracy of composed results may deteriorate since some irrelevant information is being contemplated for vector selections.

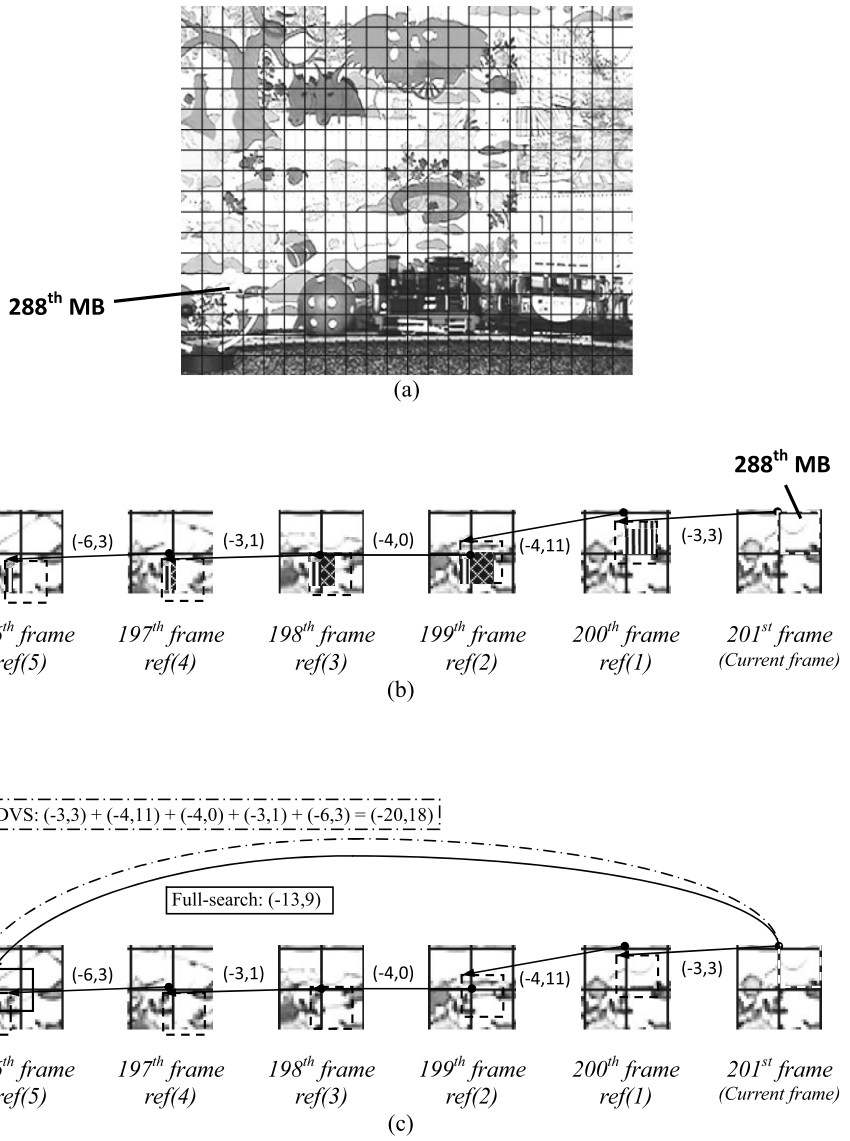

Fig. 4 Example from Mobile of MV composition by FDVS: (a) the 201st frame of Mobile, (b) diminishing relevant areas to 288th MB during FDVS, and (c) the resultant composed MV of FDVS as compared to full-search.

3 Proposed Vector Selection Algorithm

To overcome the aforementioned drawback of the existing vector composition algorithms, we propose an efficient vector composition algorithm for MRF-ME in which only the relevant area to the target MB is contributed for dominant vector selection, as illustrated in the example shown in Fig. 5. Similar to the example of FDVS shown in Fig. 3(a), MB_{n-1}^{3} is chosen as the dominant MB in the first MV

Table 1 Distributions of the final selected reference frame for various vector composition algorithms at QP20 for Mobile.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	53.70%	71.33%	79.30%	64.68%
ref(2)	13.81%	12.26%	10.18%	12.47%
ref(3)	13.71%	7.46%	4.51%	10.03%
ref(4)	9.38%	5.24%	4.04%	7.06%
ref(5)	9.40%	3.71%	1.97%	5.76%

Table 2 Distributions of the final selected reference frame for various vector composition algorithms at QP20 for Tempete.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	53.18%	75.75%	84.18%	68.39%
ref(2)	15.90%	10.96%	7.90%	11.62%
ref(3)	16.00%	7.70%	4.51%	10.79%
ref(4)	8.20%	3.19%	2.12%	5.16%
ref(5)	6.72%	2.40%	1.30%	4.04%

composition step. In Fig. 5, it is found that only the shaded area in MB^3_{n-1} is the relevant region to MB^1_n. In the second step of the proposed algorithm, only this shaded area is used to select the next dominant MB in frame $n-2$. By only considering the relevant region to the target MB^1_n, the proposed vector selection algorithm is different from FDVS. For instance, MB^2_{n-2} is selected as the dominant MB in frame $n-2$, which shows a different selection result in comparison with the original FDVS, where MB^3_{n-2} is picked. In the last step, only the cross-hatch shaded area in frame $n-2$ is used to determine the next dominant MB in frame $n-3$. Consequently, the resultant MV, $mv^1_{n-1\to n-3}$, is different from the result obtained by using FDVS in Eq. (2), and can be formed as

$$mv^1_{n\to n-3} = mv^1_{n\to n-1} + mv^3_{n-1\to n-2} + mv^2_{n-2\to n-3}. \quad (3)$$

The selection process of the proposed algorithm ensures that only the relevant area of MB^1_n is employed in MV composition. Since only the relevant area is used, the area for determining the dominant MB becomes smaller. The situation is more serious after MV composition in a temporally remote reference frame. As a consequence, it reduces the reliability of the resultant MVs. To further improve their reliability, another contribution of the proposed algorithm is to maintain the relevant area to the target MB as large as possible during MV composition. To do so, other nondominant areas in the reference frames, but relevant to MB^1_n, are also taken into consideration to enhance the use of the relevant area in MB^1_n. Figure 6 demonstrates the proposed way of enlarging the relevant area in MV composition by considering homogeneity of MVs within a moving object in a video sequence. In the example shown in Fig. 6(a), assume that $mv^1_{n-1\to n-2}$ is equal to $mv^3_{n-1\to n-2}$ in frame $n-1$. In this case, the shaded areas overlapped with MB^1_{n-1} and MB^3_{n-1} are combined, and the new combined area is for determining the next dominant MB in frame $n-2$ as depicted in Fig. 6(b). By this merging process, the proposed algorithm could keep the area relevant to the target MB as large as possible in the MV composition process since homogeneity of MVs is further considered. Then, the selected MB in frame $n-2$ is MB^2_{n-2}, where the area relevant to MB^1_n is larger and more reliable to decide the dominant MB in frame $n-3$. This merging mechanism is suitable for areas with homogeneous motion such as MBs in the background and inside the moving objects.

Nevertheless, the merging process cannot benefit the object boundary of a video object since their MVs of MBs are diverse. In the proposed algorithm, more than one candidate MB can be adopted in the MV composition process. The use of multiple candidates is to augment the area relevant to the target MB in MV composition, as demonstrated in the example of Fig. 7. In Fig. 7, C^i_{n-k} denotes the ith candidate in frame $n-k$ sorted according to the area of the overlapping segment. For instance, the overlapping area of C^1_{n-k} is the largest, while C^4_{n-k} is the smallest in each individual MV composition step, where $k=1$, 2, and 3. For the sake of simplicity, the example in Fig. 7 only picks up two-candidate MBs for composing the MV in each step. In frame $n-1$, C^1_{n-1} and C^2_{n-1} are the largest and second largest overlapping segments with the motion-compensated MB of MB^1_n. With the use of multiple candidates, both of MB^3_{n-1} and MB^4_{n-1} are used to determine the next dominant MBs in frame $n-2$, respectively. It is because both of the shaded areas in MB^3_{n-1} and MB^4_{n-1} are relevant to MB^1_n. This forms two different paths for MV composition as illustrated in Figs. 7(a) and 7(b), respectively.

Table 3 Distributions of the final selected reference frame for various vector composition algorithms at QP20 for Foreman.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	63.29%	75.78%	84.37%	71.10%
ref(2)	14.59%	10.62%	8.10%	11.26%
ref(3)	10.24%	6.57%	4.09%	8.07%
ref(4)	6.02%	3.66%	1.94%	4.94%
ref(5)	5.87%	3.38%	1.50%	4.64%

Table 4 Distributions of the final selected reference frame for various vector composition algorithms at QP20 for Salesman.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	30.15%	37.25%	49.29%	34.73%
ref(2)	66.80%	56.48%	42.61%	57.63%
ref(3)	0.76%	0.95%	1.95%	1.19%
ref(4)	1.85%	4.90%	5.27%	5.86%
ref(5)	0.44%	0.42%	0.88%	0.60%

Figure 7(a) shows the path due to the use of C_{n-1}^{1} for further MV composition. In this case, four possible candidates including C_{n-2}^{2}, C_{n-2}^{3}, C_{n-2}^{4}, and C_{n-2}^{5} in frame $n-2$ are considered in the next step. On the other hand, Fig. 7(b) depicts another path due to the contribution from C_{n-1}^{2}. This path only generates one possible candidate, C_{n-2}^{1}, in frame $n-2$ for the consideration of MV composition in the next step. Among these five candidates, C_{n-2}^{1} and C_{n-2}^{2} are the largest and second largest overlapping segments with the corresponding MBs, MB_{n-2}^{3} and MB_{n-2}^{2}, respectively. This process continues until reaching the desired reference frame. In addition, the path using the candidate MB with the second largest overlapping segment, C_{n-1}^{2}, in frame $n-1$ of Fig. 7(b) provides an alternative path to compose the new MV. In Fig. 7(b), it is found that the cross-hatch shaded area in frame $n-2$ for determining the dominant MB in frame $n-3$ is even larger than that of Fig. 7(a). In spite of the largest overlapping segment of C_{n-1}^{1} in frame $n-1$, there is no guarantee that it is still the largest overlapping segment in frame $n-2$, as shown in Fig. 7(b). With the help of multiple-candidate MBs for each reference frame, the possibility of keeping the MBs with the large relevant area to the target MB is higher during MV composition. It is because of there being only three reference frames in this working example, that two candidates for each step are sufficient. The proposed algorithm can also adopt a flexible number of possible candidates if more reference frames are needed in ME.

In conclusion, the proposed algorithm only uses the area relevant to the target MB for dominant vector selection and keeps the area relevant to the target MB as large as possible during MV composition. In contrast, FDVS only considers the largest overlapping segment with motion-compensated MB of the target MB. Both relevant and nonrelevant areas are taken into account for dominant vector selection, which may reduce the reliability of the resultant MVs.

Figure 8 then shows a flowchart and a working example for the proposed algorithm. For simplicity, two candidate MBs are selected for each MV composition step and the target reference frame is assumed to be ref(3). For each MB in the current frame, say MB_{n}^{1}, its motion compensated MB in frame $n-1$ is divided into four segments that overlap four MBs (MB_{n-1}^{1}, MB_{n-1}^{2}, MB_{n-1}^{3}, and MB_{n-1}^{4}), as denoted by the four shaded segments in Fig. 8(a). The proposed algorithm mainly consists of three steps: 1. merging, 2. multiple-candidate selection, and 3. MV composition, which are then summarized as follows:

1. Merging: Check the homogeneity of neighboring MVs. If yes, merge the segments with the same MV. Otherwise, skip the merging process.

Example: In Fig. 8(b), the shaded segments of MB_{n-1}^{1} and MB_{n-1}^{3} are merged in frame $n-1$.

2. Multiple-candidate selection: Label all merged and nonmerged segments as C_{n-1}^{i}, where $i=\{1,2,\ldots,k\}$ and k represent the total number of segments. Calculate all areas of C_{n-1}^{i}, and then rank C_{n-1}^{i} according to the area. Select the largest and second largest shaded segments.

Example: In Fig. 8(b), the three shaded segments after merging are C_{n-1}^{1}, C_{n-1}^{2}, and C_{n-1}^{3} in frame $n-1$. Then, C_{n-1}^{1} and C_{n-1}^{2} are picked for possible candidates in the next step.

Table 5 Distributions of the final selected reference frame for various vector composition algorithms at QP20 for Container.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	88.35%	93.61%	95.02%	92.68%
ref(2)	5.69%	3.53%	3.13%	3.72%
ref(3)	3.10%	1.59%	0.91%	1.90%
ref(4)	1.42%	0.94%	0.75%	1.08%
ref(5)	1.45%	0.34%	0.20%	0.63%

Table 6 Distributions of the final elected reference frame for various vector composition algorithms at QP20 for Carphone.

Reference frames	FS_ref5	FDVS_ref5 (Refs. 23–25)	MED_ref5 (Ref. 27)	PROPOSED_ref5
ref(1)	66.37%	79.89%	91.11%	76.21%
ref(2)	13.19%	8.90%	5.00%	9.81%
ref(3)	9.84%	5.61%	2.11%	6.76%
ref(4)	5.47%	3.07%	0.99%	3.85%
ref(5)	5.13%	2.52%	0.78%	3.38%

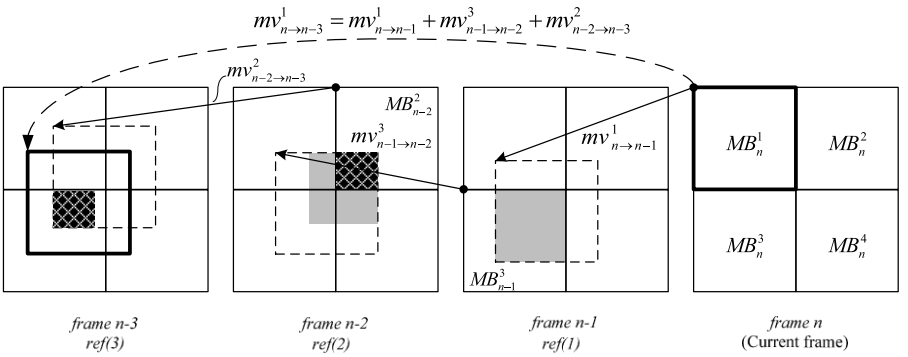

Fig. 5 Only the relevant area to the target MB is adopted in motion vector selection.

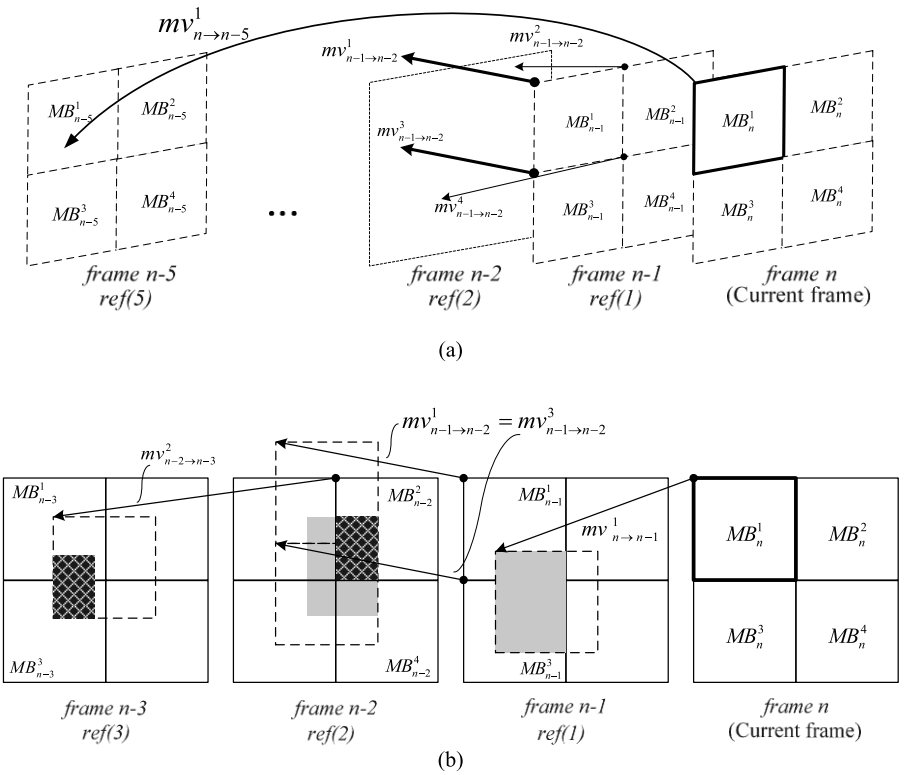

Fig. 6 (a) Scenario in homogeneous partitions where the neighboring MBs contain the same motion vector and (b) merging process with neighboring MBs of same motion vectors.

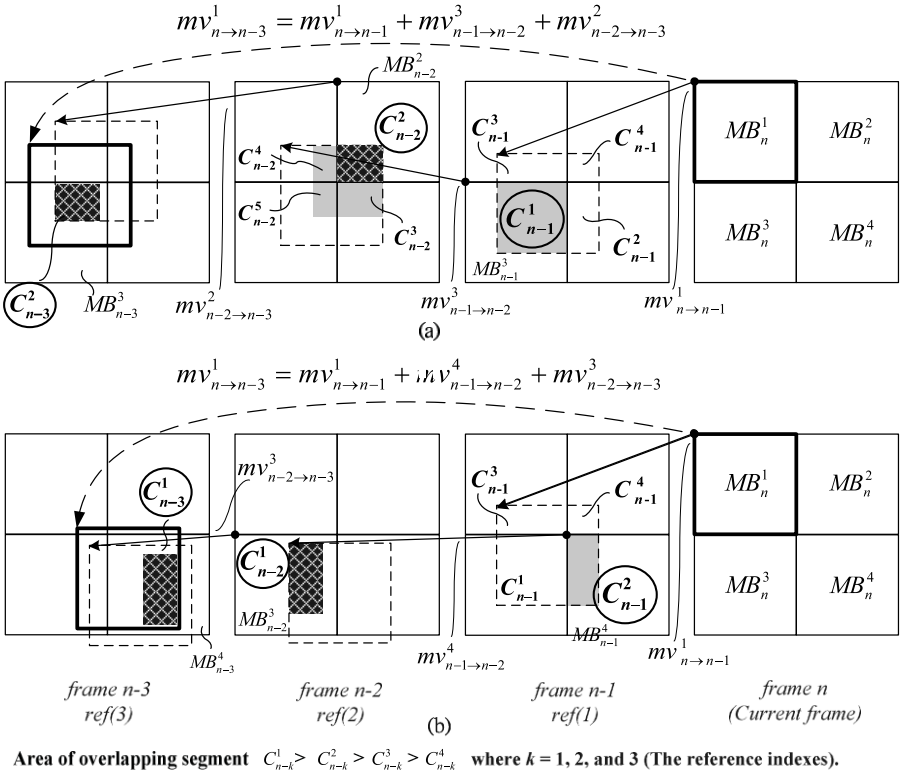

Fig. 7 Multiple-candidate MB selection: path starting from (a) C^1_{n-1} and (b) C^2_{n-1}.

3. MV composition: Compose MVs according to the candidates obtained from the previous step.

Example: In Fig. 8(b), C^1_{n-1} is chosen as a possible candidate and its corresponding MV sum up with $mv^1_{n\to n-1}$ to compose a new MV between frame n and frame $n-2$, $mv^1_{n\to n-2}(C^1_{n-1})$. MV composition also applies to the second largest shaded segment, C^2_{n-1}, to compose a new MV between frames n and $n-2$, $mv^1_{n\to n-2}(C^2_{n-1})$. In the example shown in Fig. 8(b), the possible candidate vectors from C^1_{n-1} and C^2_{n-1} are $mv^3_{n-1\to n-2}$ and $mv^4_{n-1\to n-2}$, respectively. Then, $mv^1_{n\to n-2}(C^1_{n-1})$ and $mv^1_{n\to n-2}(C^2_{n-1})$ are written as

$$mv^1_{n\to n-2}\left(C^1_{n-1}\right) = mv^1_{n\to n-1} + mv^3_{n-1\to n-2} \quad (4)$$

and

$$mv^1_{n\to n-2}\left(C^2_{n-1}\right) = mv^1_{n\to n-1} + mv^4_{n-1\to n-2}. \quad (5)$$

To compose the MV in the next reference frame (frame $n-3$) as shown in Fig. 8(c), according to $mv^1_{n\to n-2}(C^1_{n-1})$ and $mv^1_{n\to n-2}(C^2_{n-1})$, a number of C^i_{n-2}'s are obtained in frame $n-2$ after the previous MV composition step. Among these C^i_{n-2}'s, again, the candidates belong to the largest and second largest segments are selected and their corresponding MVs are employed for composing the final MVs by using frame $n-3$ as the reference. In Fig. 8(c), assume that all MVs in MB^1_{n-2}, MB^2_{n-2}, MB^3_{n-2}, and MB^4_{n-2} are different where no merging is required in frame $n-2$, the possible final MVs, $mv^1_{n\to n-3}(C^1_{n-2})$ and $mv^1_{n\to n-3}(C^2_{n-2})$, in this example can be given by

$$mv^1_{n\to n-3}\left(C^1_{n-2}\right) = mv^1_{n\to n-2}\left(C^1_{n-1}\right) + mv^2_{n-2\to n-3} = mv^1_{n\to n-1} + mv^3_{n-1\to n-2} + mv^2_{n-2\to n-3} \quad (6)$$

and

$$mv^1_{n\to n-3}(C^2_{n-2}) = mv^1_{n\to n-2}(C^2_{n-1}) + mv^3_{n-2\to n-3} = mv^1_{n\to n-1} + mv^4_{n-1\to n-2} + mv^3_{n-2\to n-3}. \quad (7)$$

After the MV composition, J_{motion} in Eq. (1) between MB^1_n and the MBs pointed by the two final MV $[mv^1_{n\to n-3}(C^1_{n-2})$ and $mv^1_{n\to n-3}(C^2_{n-2})]$ inframe $n-3$ are computed. The one with smaller J_{motion} is considered to be the new composed MV of MB^1_n with frame $n-3$ as the reference.

4 Simulation Results

We evaluated the coding performance and the coding complexity of the proposed algorithm using six test sequences when MRF-ME is activated. These sequences include Mobile (CIF), Tempete (CIF), "Foreman" (CIF), "Salesman" (CIF),

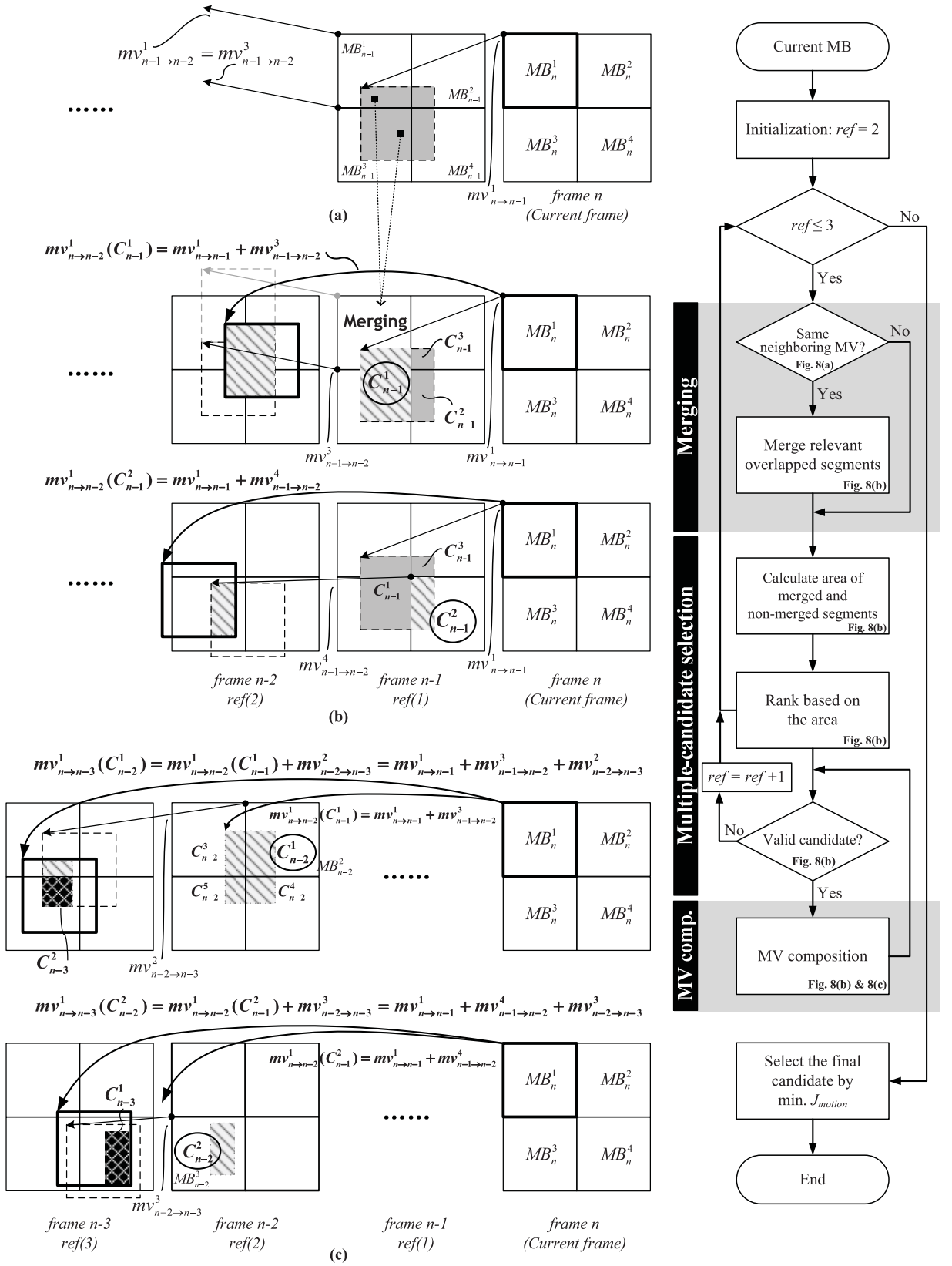

Fig. 8 Flowchart and working example of the proposed algorithm: (a) merging process, (b) selected two-candidate MBs in ref(1), and (c) selected two-candidate MBs in ref(2).

Table 7 Rate-distortion performances of various algorithms in different sequences.

		FS_ref5		FDVS_ref5 (Refs. 23–25)		MED_ref5 (Ref. 27)		PROPOSED_ref5	
	Sequences	PSNR (dB)	Bitrate (kbits/s)	PSNR (dB)	Bitrate (kbits/s)	PSNR (dB)	Bitrate (kbits/s)	PSNR (dB)	Bitrate (kbits/s)
QP20	Carphone	42.26	424.95	42.2	448.87	42.16	464.69	42.21	444.22
	Foreman	41.56	1693.80	41.51	1816.24	41.48	1858.78	41.52	1795.70
	Salesman	41.16	738.98	40.85	908.26	40.84	1136.48	40.99	841.89
	Tempete	41.09	3326.94	40.96	3726.78	40.95	3822.89	40.99	3663.78
	Mobile	40.53	3827.20	40.47	4110.65	40.46	4185.21	40.49	4023.82
	Container	41.50	977.60	41.46	1000.31	41.46	1003.93	41.47	992.97
	Average	41.35	1831.58	41.24	2001.85	41.23	2078.66	41.28	1960.40
QP24	Carphone	39.27	264.52	39.18	279.14	39.12	289.58	39.20	275.94
	Foreman	38.69	925.00	38.61	997.08	38.58	1023.43	38.63	985.60
	Salesman	38.24	390.12	38.13	406.02	38.09	437.13	38.14	401.99
	Tempete	37.72	2074.54	37.57	2443.49	37.56	2437.61	37.62	2320.85
	Mobile	36.99	2454.10	36.92	2656.30	36.89	2727.34	36.94	2589.07
	Container	38.41	492.10	38.34	512.75	38.35	514.99	38.35	509.04
	Average	38.22	1100.06	38.13	1215.80	38.10	1238.35	38.15	1180.42
QP28	Carphone	36.41	159.65	36.33	167.43	36.20	172.85	36.33	165.90
	Foreman	36.11	515.75	36.03	547.13	35.97	565.68	36.05	544.55
	Salesman	35.80	233.51	35.76	235.07	35.72	242.56	35.75	235.11
	Tempete	34.60	1195.93	34.44	1425.65	34.41	1454.55	34.48	1402.37
	Mobile	33.70	1453.06	33.63	1581.14	33.58	1640.19	33.65	1536.20
	Container	35.75	254.61	35.68	266.90	35.64	271.56	35.67	264.43
	Average	35.40	635.42	35.31	703.89	35.25	724.57	35.32	691.43
QP32	Carphone	33.49	93.09	33.39	97.11	33.22	99.30	33.42	96.03
	Foreman	33.55	295.90	33.40	311.84	33.34	316.84	33.47	313.67
	Salesman	33.19	152.28	33.15	152.80	33.14	153.36	33.16	152.75
	Tempete	31.41	595.97	31.20	744.64	31.17	740.77	31.23	742.38
	Mobile	30.30	732.25	30.22	795.49	30.11	830.39	30.26	776.37
	Container	33.15	151.56	33.06	159.84	33.01	162.80	33.10	155.76
	Average	35.52	336.84	32.40	376.95	32.33	383.91	32.44	372.83

"Container" (CIF), and "Carphone" (QCIF) with the frame rates of 30 frames/s. For the implementation, the proposed vector selection algorithm was built based on the H.264/AVC JM9.2 codec[28] for performance evaluation in MRF-ME. The results of the proposed algorithm were compared with the FS motion estimation, FDVS,[23–25] and median algorithms.[27] The following five test cases were then included for comparison:

1. FS_ref1: FS with one reference frame
2. FS_ref5: FS with five reference frames
3. FDVS_ref5: FDVS with five reference frames

4. MED_ref5: the median algorithm with five reference frames
5. PROPOSED_ref5: the proposed algorithm with five reference frames.

Basically, the bitstreams were encoded with IPPP. . . structure for 300 frames by different algorithms. It is noted that only 260 frames were encoded for Tempete due to its maximum length. Four different quantization parameters (QP = 20, 24, 28, and 32) were used. In all MV composition algorithms, MVs between consecutive frames by full-search motion estimation with a search range of -16 to $+16$ pixels were computed. These MVs were reused by various MV composition algorithms to compose the new MVs to different reference frames. For the proposed algorithm, the number of candidate MBs selected for each stage was 4. For all ME algorithms, J_{motion} was adopted as the cost function.

4.1 *Distribution of Final Selected Reference Frames*

Full-search motion estimation is always used as the benchmark for MRF-ME and it gives an optimal solution of choosing the best reference frame. Tables 1–6 show the distributions of the final selected reference frames of the tested algorithms for various video sequences. It is found that more MBs in FDVS_ref5 and MED_ref5 are predicted from ref(1) and ref(2) than that obtained from FS_ref5, which causes a quite different trend of the distribution on the final selected reference to FS_ref5. It is due to the fact that the composition of MVs in FDVS_ref5 and MED_ref5 to temporally remote reference frames may not represent the current MB anymore. It results in low reliability of the composed MVs and diminishes the benefit of MRF-ME. Tables 1–6 also list the distributions of the final selected reference frames of PROPOSED_ref5. It is clear that more MBs end up with being predicted using temporally remote reference frames such as ref(4) and ref(5), which is closer to the results of FS_ref5. From these statistics, we conclude that the proposed vector composition process is likely to utilize more benefit of MRF-ME by obtaining more accurate composed MVs.

4.2 *Results of Coding Efficiency*

To evaluate the coding efficiency, the rate-distortion (R-D) curves by using different algorithms for Mobile (CIF), Tempete (CIF), Foreman (CIF), Salesman (CIF), Container (CIF), and Carphone (QCIF) are shown in Figs. 9–14, respectively. From Figs. 9–14, PROPOSED_ref5 clearly outperforms FDVS_ref5 and MED_ref5, especially in the high bitrate cases. The gaps in both peak signal-to-noise ratio (PSNR) and generated bits between FS_ref5 and PROPOSED_ref5 become remarkably narrower compared to other algorithms. It is clear that the R-D performance of the proposed algorithm remarkably improves in Mobile, Tempete, Foreman, Salesman, and Carphone. It is expected, since FDVS does not work well for these sequences. Even though a slight improvement could also be seen for the sequence with a still background such as Container in Fig. 13, it is not so significant as compared with other sequences. It is due to the probability that temporally remote reference frames being used in Container is unlikely, as shown in Table 5. The room for improvement of the proposed algorithm is then limited. To further evaluate the

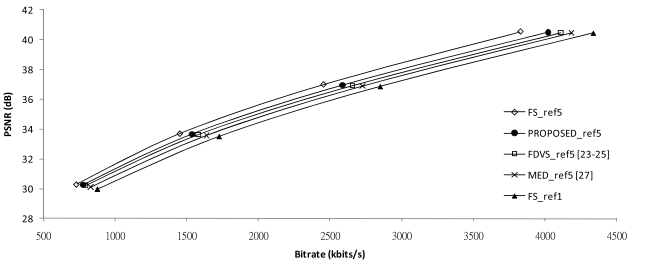

Fig. 9 R-D curves for the five test cases in Mobile.

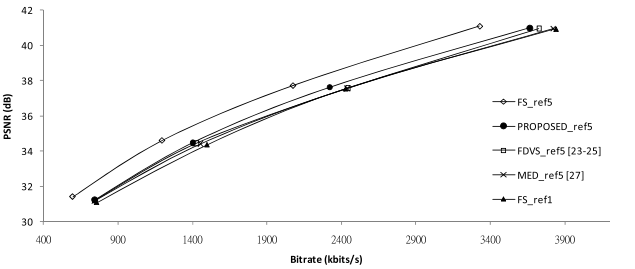

Fig. 10 R-D curves for the five test cases in Tempete.

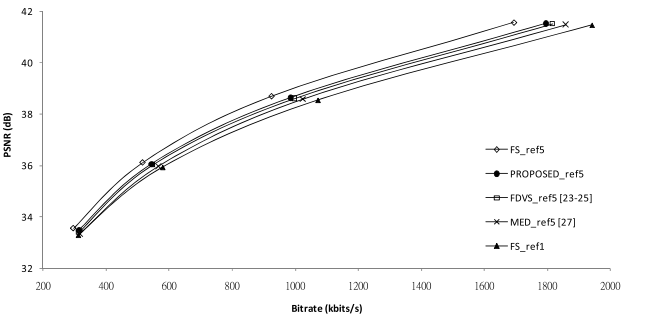

Fig. 11 R-D curves for the five test cases in Foreman.

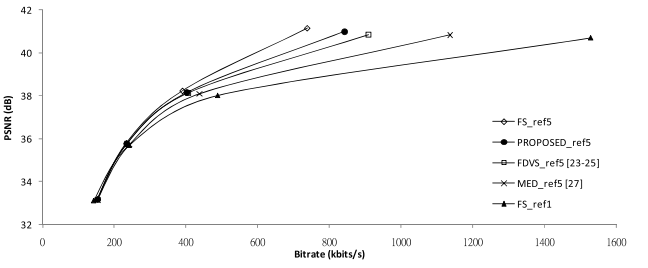

Fig. 12 R-D curves for the five test cases in Salesman.

Table 8 Computational comparison in terms of ME encoding time (unit: second).

Sequences	FS_ref5 Time, in seconds	FDVS_ref5 (Refs. 23–25) (Δ*Time*)	MED_ref5 (Ref. 27) (Δ*Time*)	PROPOSED_ref5 (Δ*Time*)
Carphone	233.79	49.22	47.55	51.12
		(−78.95%)	(−79.66%)	(−78.13%)
Foreman	951.29	186.44	199.10	215.64
		(−80.40%)	(−79.07%)	(−77.33%)
Salesman	963.33	180.45	198.55	205.81
		(−81.27%)	(−79.39%)	(−78.64%)
Tempete	846.82	159.40	159.27	173.55
		(−81.18%)	(−81.19%)	(−79.51%)
Mobile	935.97	187.02	186.48	195.33
		(−80.02%)	(−80.08%)	(−79.13%)
Container	913.31	182.79	192.20	182.27
		(−79.99%)	(−78.96%)	(−80.04%)
Average	807.42	−80.30%	−79.72%	−78.80%

results, Table 7 shows the PSNR and required bitrate of various sequences at four different QPs. Table 7 also shows that PROPOSED_ref5 has a consistent gain in coding efficiency for all video sequences. It is because PROPOSED_ref5 considers only the area related to the target MB and tries to keep the relevant partition as large as possible in every MV composition step. It ensures that the resultant MV is highly correlated to the contents of the target MB in the current frame, which cannot be achieved by FDVS_ref5 and MED_ref5. The proposed algorithm can then provide outstanding performance in the view point of rate-distortion in comparison with the other MV composition algorithms.

4.3 *Results of Computational Complexity*

To compare the computational complexity of the proposed algorithm, FS_ref5 is used as a reference method, and all simulations were carried out on an Intel(R) Xeon(R) CPU X5550 at 2.66 GHz PC with 12 GB memory. It is noted that the only major overhead of PROPOSED_ref5 compared to FDVS_ref5 and MED_ref5 is the calculation of J_{motion} required for the final selection of MVs from different candidates in the last step of the MV composition process. The average ME time per video sequence were then measured and tabulated in Table 8. The $\Delta Time$ in Table 8 is calculated as follows:

$$\Delta Time(\%) = \frac{Time_{Test} - Time_{FS\_ref5}}{Time_{FS\_ref5}} \times 100, \tag{8}$$

where $Time_{FS\_ref}$ and $Time_{Test}$ denote the ME coding time used by FS_ref5 and the tested algorithms. It can be easily seen that all algorithms can substantially reduce the computational complexity of FS_ref5 by nearly 80%. Although the ME time of the proposed algorithm is slightly increased with the number of multiple candidates, the increase is only around 1.5% in average, and the ME time of the proposed algorithm is still quite similar with that of FDVS_ref5 and MED_ref5.

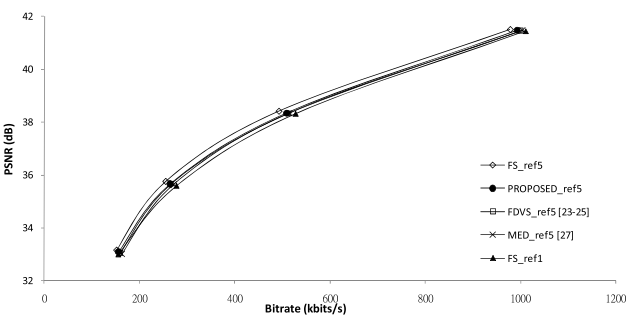

Fig. 13 R-D curves for the five test cases in Container.

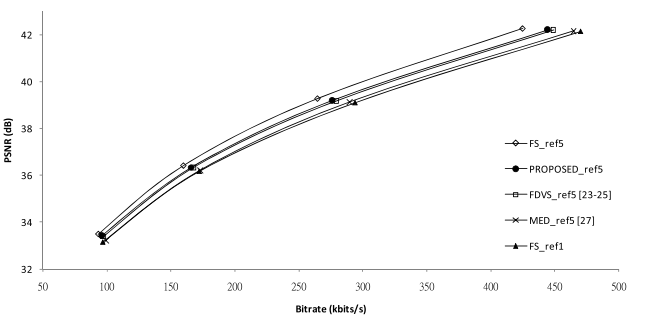

Fig. 14 R-D curves for the five test cases in Carphone.

5 Conclusion

In this paper, we have proposed a novel MV composition algorithm for MRF-ME. The proposed algorithm is beneficial to perform ME to a reference frame with a large temporal distance. It can entirely make use of the relevant area to the target MB by two vector selection criteria. First, only the actually relevant area of the target MB is contributed to dominant MB selections. Second, the relevant area to the target MB is kept as large as possible in MV composition by adopting the concept of MV merging and multiple candidates in the vector selection process. These techniques can increase the reliability of the final composed MVs.

The performance of the proposed algorithm, experimentally verified in terms of both quality and bitrate, is remarkably better than that of the conventional approach, such as FDVS and the median algorithms. The distribution of the final selected reference frame obtained from the proposed algorithm is very similar to full-search. It indicates that the proposed algorithm is highly probable to get the benefit of using MRF-ME. Besides, the proposed algorithm is adaptive in nature, and the number of candidate MBs can be adjusted according to the number of reference frames.

In addition, the proposed algorithm is not restricted to MRF-ME, it can also be beneficial to MV composition in frame-skipping transcoding, which is a process of skipping some frames in order to change the frame rate of a video sequence. Our proposed algorithm is specially suited for the scenario of skipping a large number of frames in transcoding. As a concluding remark, it is believed that the results of the present work will certainly be useful for the future development of digital video coding and transcoding.

Acknowledgments

The work described in this paper is partially supported by the Centre for Signal Processing, Department of Electronic and Information Engineering, Hong Kong Polytechnic University. and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant No. PolyU 5125/10E). Tsz-Kwan Lee acknowledges the research studentships provided by the University.

References

1. A. Luthra, G. J. Sullivan, and T. Wiegand, "Introduction to the special issue on the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.* **13**(7), 557–559 (2003).
2. S. K. Kwon, A. Tamhankar, and K. R. Rao, "Overview of H.264/MPEG-4 Part 10," *J. Visual Commun. Image Represent.* **17**, 186–216 (2006).
3. H. Shim and C. M. Kyung, "Selective search area reuse algorithm for low external memory access motion estimation," *IEEE Trans. Circuits Syst. Video Technol.* **19**(7), 1044–1050 (2009).
4. Y. Su and M. T. Sun, "Fast multiple reference frame motion estimation for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.* **16**(3), 447–452 (2006).
5. H. M. Wong, O. C. Au, A. Chang, S. K. Yip, and C. W. Ho, "Fast mode decision and motion estimation for H.264 (FMDME)," in *Proc. IEEE Int. Symp. Circuits Syst., ISCAS'06*, pp. 473–476, Island of Kos, Greece (2006).
6. K. Lee, G. Jeon, R. Falcon, C. Ha, and J. Jeong, "An adaptive fast multiple reference frame selection algorithm for H.264/AVC using reference region data," in *Proc. IEEE Workshop Signal Process. Syst., SiPS'09*, pp. 93–96, Tampere, Finland (2009).
7. S. J. Kang, S. Yoo, and Y. H. Kim, "Dual motion estimation for frame rate up-conversion," *IEEE Trans. Circuits Syst. Video Technol.* **20**(12), 1909–1914 (2010).
8. C. Wang, L. Zhang, Y. He, and Y. P. Tan, "Frame rate up-conversion using trilateral filtering," *IEEE Trans. Circuits Syst. Video Technol.* **20**(6), 886–893 (2010).
9. D. Wang, L. Zhang, and A. Vincent, "Motion-compensated frame rate up-conversion - part I: fast multi-frame motion estimation," *IEEE Trans. Broadcast.* **56**(2), 133–141 (2010).
10. D. Wang, A. Vincent, P. Blanchfield, and R. Klepko, "Motion-compensated frame rate up-conversion–part II: new algorithms for frame interpolation," *IEEE Trans. Broadcast.* **56**(2), 142–149 (2010).
11. G. Sullivan and T. Wiegand, "Video compression – from concepts to the H.264/AVC video coding standard," *Proc. IEEE* **93**(1), 18–31 (2005).
12. D. Marpe, T. Wiegand, and G. J. Sullivan, "The H.264/MPEG-4 advanced video coding standard and its applications," *IEEE Commun. Mag.* **44**(8), 134–143 (2006).
13. Y. W. Huang, B. Y. Hsieh, S. Y. Chien, S. Y. Ma, and L. G. Chen, "Analysis and complexity reduction of multiple reference frames motion estimation in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.* **16**(4), 507–522 (2006).
14. Y. W. Huang, B. Y. Hsieh, T. C. Wang, S. Y. Chen, S. Y. Ma, C. F. Shen, and L. G. Chen, "Analysis and reduction of reference frames for motion estimation in MPEG-4 AVC/JVT/H.264," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* **3**, pp. 145–148, Hong Kong, China (2003).
15. K. Lee, G. Jeon, and J. Jeong, "Fast reference frame selection algorithm for H.264/AVC," *IEEE Trans. Consum. Electron.* **55**(2), 773–779 (2009).
16. G. N. Rao and P. Gupta, "Temporal motion prediction for fast motion estimation in multiple reference frames," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, pp. 817–820, Vancouver, Canada (2006).
17. N. Ozbek and A. M. Tekalp, "Fast H.264/AVC video encoding with multiple frame references," in *Proc. IEEE Signal Process. Comm. Appl. Conf.*, pp. 573–576, Kayseri, Turkey (2005).
18. Z. Wang, J. Yang, Q. Peng, and C. Zhu, "An efficient algorithm for motion estimation with multiple reference frames in H.264/AVC," in *Proc. Int. Conf. Image Graph., ICIG'07*, pp. 259–262, Washington, DC, USA (2007).
19. T. Y. Kuo and H. J. Lu, "Efficient reference frame selector for H.264," *IEEE Trans. Circuits Syst. Video Technol.* **18**(3), 400–405 (2008).
20. L. Shen, Z. Liu, Z. Zhang, and G. Wang, "An adaptive and fast multiframe selection algorithm for H.264 video coding," *IEEE Signal Process. Lett.* **14**(11), 836–839 (2007).
21. D. S. Jun and H. W. Park, "An efficient priority-based reference frame selection method for fast motion estimation in H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.* **20**(8), 1156–1161 (2010).
22. M. J. Chen, G. L. Li, Y. Y. Chiang, and C. T. Hsu, "Fast multiframe motion estimation algorithms by motion vector composition for the MPEG-4/AVC/H.264 standard," *IEEE Trans. Multimedia* **8**(3), 478–487 (2006).
23. J. Youn and M. T. Sun, "A fast motion vector composition method for temporal transcoding," in *Proc. IEEE Int. Symp. Circuits Syst.* **4**, pp. 243–246, Orlando, FL, USA (1999).
24. M. J. Chen, Y. Y. Chiang, H. J. Li, and M. C. Chi, "Efficient multi-frame motion estimation algorithms for MPEG-4 AVC/JVT/H.264," in *Proc. IEEE Int. Symp. Circuits Syst., ISCAS'04*, pp. 737–740, Vancouver, Canada (2004).
25. S. Zhang, Y. Wang, J. Kang, and H. Li, "A new approach to fast multiple reference frame motion estimation for H.264," in *Int. Symp. Comput. Sci. Computational Technol.* **2**, pp. 254–258, Shanghai, China (2008).
26. Y. Su and M. T. Sun, "Fast multiple reference frame motion estimation for H.264," in *Proc. IEEE Int. Conf. Multimedia Expo.* **1**, pp. 695–698, Taipei, Taiwan (2004).
27. S. E. Kim, J. K. Han, and J. G. Kim, "An efficient scheme for motion estimation using multireference frames in H.264/AVC," *IEEE Trans. Multimedia* **8**(3), 457–466 (2006).
28. Reference Software JM9.2 from http://iphome.hhi.de/suehring/tml/download/.

Tsz-Kwan Lee received her BSc (Honors) degree from Department of Electronic and Information Engineering of The Hong Kong Polytechnic University in 2008. She is currently pursuing an MPhil degree at the Center for Signal Processing under the supervision of Dr. Y. L. Chan and Professor W. C. Siu in the same Department and University. Her research interests include multimedia technologies, signal processing, image, and video compression and transcoding.

Yui-Lam Chan received his BEng with a First Class Honors degree and his PhD degree from the Hong Kong Polytechnic University in 1993 and 1997, respectively. During his studies, he was the recipient of more than ten famous prizes, scholarships, and fellowships for his outstanding academic achievement, such as being the champion in Varsity Competition in Electronic Design, the Sir Edward Youde Memorial Fellowship, and the Croucher Foundation Scholarships. Chan joined the Hong Kong Polytechnic University in 1997, and is now an associate professor in the Department of Electronic and Information Engineering. Chan is also actively involved in professional activities. In particular, he serves as a reviewer and session chairman for many international journals/conferences. He was the secretary of the 2010 IEEE International Conference on Image Processing (ICIP'2010). He has published over 60 research papers in various international journals and conferences. His research and technical interests include multimedia technologies, signal processing, image and video compression, video streaming, video transcoding, video conferencing, digital TV/HDTV, 3DTV, multi-view video coding, future video coding standards, error-resilient coding, and digital VCR.

Chang-Hong Fu entered South East University in 1998 and transferred to the Hong Kong Polytechnic University in 1999, by the support of Hong Kong Jockey Club Scholarship for outstanding mainland students. He received his BEng (with first class honors) and PhD degrees from the Hong Kong Polytechnic University in 2002 and 2008, respectively. During his studies, he was the recipient of several scholarships for his outstanding academic achievements. He joined Nanjing University of Science and Technology in 2011 and is now an associate professor in School of Electronic and Optical Engineering. He has published over 20 research papers in various international journals and conferences. His research and technical interests include the area of multimedia technologies, signal processing, image and video compression, video transcoding, video streaming, bitstream switching, digital video cassette recoding, multi-view/3D video coding, and future video coding standards.

Wan-Chi Siu received his MPhil and PhD degrees from CUHK and Imperial College, UK, in 1977 and 1984, respectively. He joined the Hong Kong Polytechnic University in 1980 and was head of Electronic and Information Engineering Department and subsequently dean of Engineering Faculty between 1994 and 2002. He has been chair professor since 1992 and is now director of the Centre for Signal Processing. He is an expert in DSP, specializing in fast algorithms, video coding, transcoding, 3D-videos and pattern recognition, and has published 380 research papers, over 160 of which appeared in international journals, such as *IEEE Transactions on Image Processing*. He has been guest/associate editor of many journals, and keynote speaker (e.g., PCM'2002) and chief organizer (ISCAS'1997, ICASSP'2003, and ICIP'2010) of many important international conferences. His work on fast computational algorithms (such as DCT) and motion estimation algorithms have been well received by academic peers, with good citation records, and a number of which are now being used in hi-tech industrial applications, such as modern video surveillance and video codec design for HDTV systems.