Said et al. EURASIP Journal on Image and Video Processing (2017) 2017:15
DOI 10.1186/s13640-017-0164-7

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Prediction architecture based on block matching statistics for mixed spatial-resolution multi-view video coding

Hany Said[1*], Mansour Moniri[2] and Claude C. Chibelushi[3]

Abstract

The use of mixed spatial resolutions in multi-view video coding is a promising approach for coding videos efficiently at low bitrates. It can achieve a perceived quality, which is close to the view with the highest quality, according to the suppression theory of binocular vision. The aim of the work reported in this paper is to develop a new multi-view video coding technique suitable for low bitrate applications in terms of coding efficiency, computational and memory complexity, when coding videos, which contain either a single or multiple scenes. The paper proposes a new prediction architecture that addresses deficiencies of prediction architectures for multi-view video coding based on H.264/AVC. The prediction architectures which are used in mixed spatial-resolution multi-view video coding (MSR-MVC) are afflicted with significant computational complexity and require significant memory size, with regards to coding time and to the minimum number of reference frames. The architecture proposed herein is based on a set of investigations, which explore the effect of different inter-view prediction directions on the coding efficiency of multi-view video coding, conduct a comparative study of different decimation and interpolation methods, in addition to analyzing block matching statistics. The proposed prediction architecture has been integrated with an adaptive reference frame ordering algorithm, to provide an efficient coding solution for multi-view videos with hard scene changes. The paper includes a comparative performance assessment of the proposed architecture against an extended architecture based on the 3D digital multimedia broadcast (3D-DMB) and the Hierarchical B-Picture (HBP) architecture, which are two most widely used architectures for MSR-MVC. The assessment experiments show that the proposed architecture needs less bitrate by on average 13.1 Kbps, less coding time by 14% and less memory consumption by 31.6%, compared to a corresponding codec, which deploys the extended 3D-DMB architecture when coding single-scene videos. Furthermore, the codec, which deploys the proposed architecture, accelerates coding by on average 57% and requires 52% less memory, compared to a corresponding codec, which uses the HBP architecture. On the other hand, multi-view video coding which uses the proposed architecture needs more bitrate by on average 24.9 Kbps compared to a corresponding codec that uses the HBP architecture. For coding a multi-view video which has hard scene changes, the proposed architecture yields less bitrate (by on average 28.7 to 35.4 Kbps), and accelerates coding time (by on average 64 and 33%), compared to the HBP and extended 3D-DMB architectures, respectively. The proposed architecture will thus be most beneficial in low bitrate applications, which require multi-view video coding for video content depicting hard scene changes.

Keywords: H.264/AVC, Mixed spatial-resolution, Multi-view video coding, Prediction architecture

* Correspondence: hany.said.1980@ieee.org
[1]College of Engineering, Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt
Full list of author information is available at the end of the article

1 Introduction

1.1 Context and related work

The mixed spatial-resolution coding approach provides a better solution for multi-view video than the symmetric coding approach, at low bitrates. It has been reported that mixed spatial-resolution stereoscopic video coding has less coding complexity and provides better rate-distortion than symmetric coding [1–3]. These advantages are desirable attributes towards meeting the requirements of low bitrate applications, as in handheld devices and telemedicine [4, 5]. According to the suppression theory of binocular vision, the total perceived quality for mixed spatial-resolution stereoscopic video is close to the view with the highest quality (the view with full spatial-resolution frames) [2, 6]. This is due to the high frequency components (which exist in the full spatial-resolution frames) which compensate the corresponding components in the lower spatial-resolution frames [7]. Asymmetric temporal-resolution and asymmetric quality are other alternatives for asymmetric coding. The former causes flickering artifacts especially when coding sequences, which contain fast object motion, while the latter produces inevitable blocking artifacts when coding videos at low bitrates [7, 8]. Still, the mixed spatial-resolution approach provides better perceived quality than other coding approaches when coding multi-view videos at low bitrates [2, 9].

The prediction architecture is a central part of multi-view coding, which exploits the temporal and cross-view correlations among neighbouring frames. Prediction architecture is described by the reference frame selection and reference frame ordering. Reference frame selection identifies a set of reference frames, where they are stored in decoded picture buffer. Reference frame ordering defines how the indices of these frames are placed inside the list buffer, where Exponential Golomb is used to code indices of reference frames [10]. Selecting reference frames, which have a most significant role for inter-picture prediction, alongside providing a suitable reference frame ordering, would improve coding efficiency. This is due to the block matching process, which targets the optimization of the actual bitrate and distortion through a Lagrangian method, which estimates $J\ (ref\ |\lambda_{\mathrm{Motion}})$ [11]. The latter is defined by the equation:

$$J(\mathrm{ref}|\lambda_{\mathrm{Motion}}) = \mathrm{SAD}(s,r) + \lambda_{\mathrm{Motion}} * R\ (\mathrm{MVD},\ \mathrm{REF})$$

where the sum of absolute difference (SAD) is the prediction absolute error between the current block (s) and the corresponding reference block (r), $\lambda_{\mathrm{Motion}}$ is a Lagrange multiplier and R is the number of bits required to code both the motion vector difference (MVD) and the reference frame (REF). The latter is the decoded frame, which is available at both the encoder and decoder sides [11].

Several prediction architectures have been proposed in the literature, for use in the context of MSR-MVC. The first prediction architecture is 3D digital multimedia broadcast (3D-DMB) which is based on the IPPP coding structure, as shown in Fig. 1a. The objective behind this architecture is to fit the ITU-T recommendations for DMB where the coded video streams should comply with the baseline profile (IPPP coding structure) and the number of reference frames is up to three [12]. A multi-view video codec, which is based on this prediction architecture, was used in several studies [3, 13, 14]. Part of these studies include assessing the coding efficiency for the mixed spatial-resolution coding approach and symmetric coding and to investigate the decoding and up-sampling optimization of low spatial-resolution frames [3, 13]. This architecture was also used to propose two sampling directions (horizontal and vertical sampling) for frames, which belong to the dependent view [14].

The hierarchical B-picture (HBP) is another prediction architecture. It is based on the IBBP coding structure, which is inspired from the typical prediction architecture of the multi-view coding standard as depicted in Fig. 1b. This well-known prediction architecture provides efficient coding since it allows inter-picture prediction from all directions for frames, which belong to the odd views. This architecture was used in the context of MSR-MVC to propose a low complexity motion compensation algorithm [15]. Other studies have used this prediction architecture to study the effect of using different inter-view prediction directions (by using full spatial-resolution and low spatial-resolution frames in the base view) upon the coding efficiency of multi-view coding, and to propose different decimation methods for full spatial-resolution frames and to explore the down-sample threshold where suppression theory is valid [16–18].

HBP and 3D-DMB are the most widely used prediction architectures for mixed spatial-resolution multi-view videos. The HBP prediction architecture relies on B frames for the majority of frames (92% are B frames, for typical prediction architecture of multi-view coding) [19]. Consequently, it achieves higher coding efficiency compared to architectures based on the IPPP coding structure, at the expense of demanding significant coding complexity and memory size. The former is due to allowing forward, backward and bi-prediction for temporal and spatial frames during inter-picture prediction [19]. The large memory size is due to the need to store these reference frames in the decoded picture buffer (34 frames are stored when coding 8 views for 8 groups-of-pictures) [19]. On the other hand, the 3D-DMB prediction architecture relies mainly on P-frames, which support unidirectional prediction. Therefore, this prediction architecture needs less coding time and memory size compared to the HBP architecture. The literature offers no justification for the reference frame

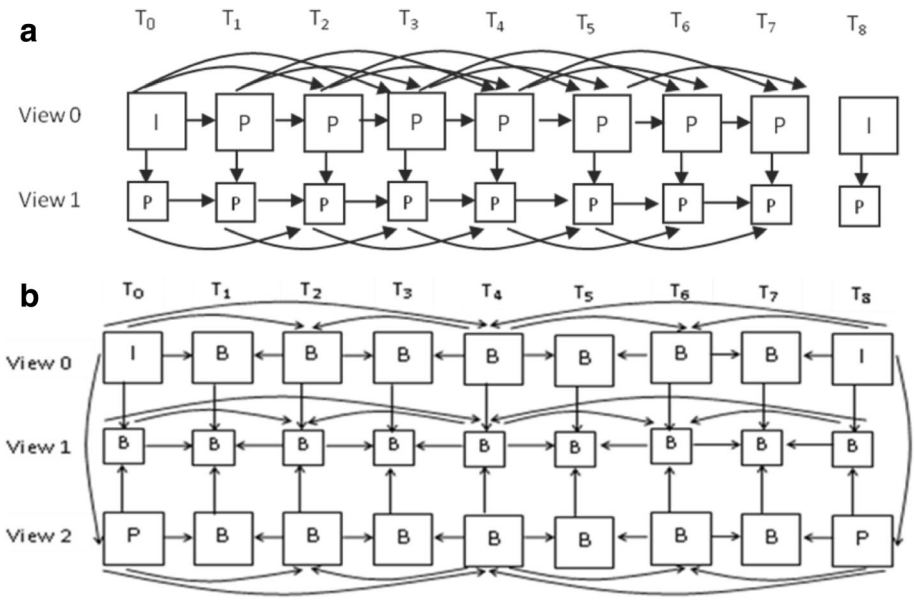

Fig. 1 Prediction architectures. **a** 3D-DMB. **b** Hierarchical B-picture

selection used for this prediction architecture in addition to how it can handle coding videos efficiently with different scene characteristics, such as object motion and scene complexity. This increases the coding challenges when relying on a fixed reference frame selection. In the context of coding videos which have multiple scenes, both prediction architectures would not provide an efficient coding solution since these architectures apply a non-adaptive reference frame ordering which sorts the reference frame indices in a particular way. This leads to the inability to adapt the reference frame ordering when coding videos, which have hard scene changes.

1.2 Contributions of the paper

The challenges highlighted above open an opportunity to investigate prediction architectures for MSR-MVC at low bitrates. This paper presents such an investigation and proposes suitable prediction architecture. The target is to achieve comparable coding efficiency, while reducing both computational and memory complexity, compared to 3D-DMB and HBP prediction architectures for multi-view videos, whether they contain single or multiple scenes.

Several points are addressed in this paper during the investigations of the prediction architecture for multi-view videos, which contain frames with mixed spatial-resolution. The first point is finding whether each group of frames should use a similar reference frame selection and reference frame ordering or not. Enabling inter-view prediction among these reference frames is mandatory, to exploit cross-view correlation. Therefore, it is essential to define suitable methods for decimating full spatial-resolution frames and interpolating low spatial-resolution reference frames, where suitability is defined in terms of computational complexity and coding efficiency. The second point concerns how to derive reference frame selection and reference frame ordering, for the prediction architecture to be able to code efficiently videos, which depict a variety of scene characteristics. The last point is how to provide prediction architecture with the ability to compress efficiently videos with hard scene changes.

The first point was answered through studying the effect of inter-view prediction direction on the coding efficiency of mixed spatial-resolution stereoscopic video coding. A comparative study was then conducted to assess different decimation and interpolation methods. The second point was tackled by performing a statistical analysis of block matching for MSR-MVC. Statistical analysis is a reliable technique to derive a prediction architecture, as it has been used for symmetric multi-view coding, where reference frame selection and reference frame ordering are derived by analyzing the amount of inter-picture prediction across reference frames [20–24]. This statistical analysis technique has not been applied for the mixed spatial-resolution coding approach. Therefore, this technique was used in the work reported in this paper, to propose a prediction architecture. Finally, to code efficiently multi-view videos with hard scene changes, the proposed prediction architecture needs to be integrated with an algorithm which can adapt the reference frame ordering. The adaptive reference frame ordering algorithm (which was developed in earlier work [24]) was integrated with the proposed prediction architecture as it proved its efficiency in coding

symmetric multi-view video, which contains videos from several scenes.

The remainder of this paper is organised as follows: Section 2 presents the experimental setup and performance parameters, while Section 3 discusses the empirical foundation of the proposed prediction architecture. It covers the effect of inter-view prediction direction on the coding efficiency of multi-view coding in Section 3.1. Different decimation and interpolation methods are evaluated in Section 3.2. A new prediction architecture is then proposed in Section 4, and it is integrated with the adaptive reference frame ordering algorithm. Results and Discussions of the performance evaluation of the prediction architecture are reported in Section 5, and Conclusions are summarised in Section 6.

2 Experimental setup and performance parameters

This section outlines the data preparation, coding configuration and the performance parameters used in the investigations reported in this paper. Six multi-view videos have been used in the paper; they are Break-dancers, Akko & Kayo, Ballroom, Exit, Race1 and Rena. These videos are recommended as the multi-view coding common test conditions [25]. Table 1 provides a brief description for each video. They cover a wide range of scene characteristics and object motion. The Akko & Kayo and Rena multi-view videos have less disparity compared to the remaining videos since both have less inter-camera distance and scene complexity [20]. The motion of objects in Exit videos is slow while it is fast in Race1 videos. Since this paper focuses on low bitrate applications, the original spatial-resolution of the luminance components was decimated using the MPEG-4 filter by a factor of two in the horizontal and vertical directions. The resulting videos are then treated as views which contain full spatial-resolution frames. The spatial-resolution for frames which belong to one of the views is further decimated in order to generate low spatial-resolution frames. In order to generate a single stream among multi-view videos, frames with different spatial-resolutions are multiplexed in a time-first coding order [19]. The coded low spatial-resolution frames are interpolated using an AVC interpolation filter. Table 2 shows the filter coefficients for the MPEG and AVC filters; these filters are recommended in asymmetric video coding [2, 16]. Three-view videos have been considered during the testing of the proposed prediction architecture in the context of a single scene scenario. To generate multi-view videos with hard scene changes, frames that belong to Akko & Kayo, Ballroom, Exit, Race1 and Rena videos were multiplexed. The video starts with the first nine frames from Akko & Kayo, followed by six frames from each of the other videos. Frames which belong to the middle view were decimated while frames that belong to the surrounding views were full spatial-resolution frames.

The experiments were carried out on a computer with an Intel i7-880 processor (8 M cache, 3.06 GHz) and 16 GB of memory. The H.264/AVC reference software JM 18.0 software was used to conduct the experiments, where all coding modes are enabled [26]. A sequential view prediction structure was used for the experiments presented in the next section. This architecture allows two reference frames (the nearest temporal and spatial frames) to be used for inter-picture prediction. The quantization settings which represent coding videos at lowest acceptable quality were adjusted according to the predefined values that are reported in the common test conditions [25]. Table 3 lists the settings of the quantization, where a symmetric quality was applied among neighbouring views.

Three performance parameters were used to measure coding efficiency, computational complexity and memory complexity. The average bitrate reduction and the average video quality improvement were used to measure coding efficiency. Both were exploited from rate-distortion curves using the average differences for bitrate and PSNR (for the luminance component) when applying two different prediction architectures. The total coding time was used to reflect the computational complexity of a particular prediction architecture, since most of the coding time is consumed during the prediction stage. Average coding time reduction was calculated by measuring the running time when applying a prediction architecture (A) compared to corresponding time from another prediction architecture (B). Therefore, coding time reduction is the result of dividing the difference between coding times for these architectures by the coding time consumed when

Table 1 Description of multi-view videos used in the investigations reported in this paper

Multi-view video	Number of cameras/setup	Camera spacing (cm)	Frame rate (fps)	Provider
Break-dancers	8/arc	20	15	Microsoft
Ballroom	8/1D linear	20	25	MERL
Exit	8/1D linear	20	25	MERL
Race1	8/1D linear	20	30	KDDI
Akko & Kayo	100/2D array	5×20	30	Tanimoto Lab
Rena	8/1D linear	5	30	Tanimoto Lab

Table 2 Low pass filter coefficients, which are used in decimating and interpolating the video frames

Filter	Coefficients
MPEG filter	{2, 0, −4, −3, 5, 19, 26, 19, 5, −3, −4, 0, 2}/64
AVC filter	{1, −5, 20, 20, −5, 1}/32

deploying prediction architecture (B). Similarly, memory complexity was calculated; it is defined by the minimum number of reference frames stored in the decoded picture buffer (taking into account full spatial-resolution frames which would be decimated prior to predicting frames with lower spatial-resolution and vice versa).

3 Empirical foundation of the proposed prediction architecture

3.1 Effect of inter-view prediction direction on the coding efficiency of multi-view video coding

This section seeks to answer the question whether or not frames with different spatial-resolution should use a similar reference frame selection and reference frame ordering. To answer this question, the coding efficiency for mixed spatial-resolution stereoscopic video coding is examined when it uses different inter-view prediction directions. Figure 2 shows two inter-view prediction directions, where the first inter-view prediction direction uses full spatial-resolution frames in the base view. Each frame is low pass filtered (LPF) and sub-sampled prior to predicting low spatial-resolution frame. The second direction relies on low spatial-resolution frames in the base view, where each frame is up-sampled and filtered when predicting a full spatial-resolution frame.

The coding efficiency of H.264/AVC-based multi-view coding is evaluated using these inter-view prediction directions, where a sequential-view prediction structure is used as shown in Fig. 3. The rate-distortion curves for six stereoscopic videos are presented in Fig. 4. From this figure, it is clear that the coding efficiency for the codec which uses full spatial-resolution frames in the base view is superior to a corresponding codec which uses low spatial-resolution frames, at low bitrates. Mixed spatial-resolution stereoscopic video coding saves bitrate by on average 6.2% while the video quality is improved (on average 0.63 dB) when it uses full spatial-resolution frames rather than low spatial-resolution frames in the base view. This improvement would be explained through the degree of consistency among reference frames. When low spatial-resolution frames are used in the base view, the interpolated reference frames have a certain degree of blurriness which has a negative effect for inter-view prediction. On the contrary, using full spatial-resolution frames in the base view, to predict frames with lower spatial-resolution would not affect inter-view prediction since both frames have a similar degree of information loss. This is demonstrated through the amounts of inter-view prediction in both prediction directions; it is in range of 4.4–31.93% when full spatial-resolution frames are used in the base view, while it is in range of 0.1–5.1% when low spatial-resolution frames are used in the base view.

Table 3 Settings for the quantisation parameters

Multi-view video	High quality-QP_L	Medium quality-QP_M	Low quality-QP_H
Break-dancers	22	26	31
Ballroom	29	31	34
Exit	26	29	31
Race1	24	26	28
Akko & Kayo	24	29	36
Rena	23	28	33

These results are not consistent with the findings of Brust and co-workers [16]. However, it should be pointed out that their study used asymmetric quality in conjunction with mixed spatial-resolution stereoscopic video coding. They reported that both prediction directions provide similar coding efficiency for mixed spatial-resolution stereoscopic video coding. In order to understand the effect of the asymmetric quality on inter-view prediction, a similar experiment using asymmetric quality was conducted in the work reported herein. The amount of inter-view prediction was analyzed using different settings for delta quantisation (ΔQP) among frames with mixed spatial-resolution, which was set in the range of (0, 10). Based on a regression analysis, using the six multi-view videos, the relationship between inter-view prediction (IVP) and ΔQP was found to fit the equation

$$\mathrm{IVP} = 1.492 + 1.096\,\Delta\mathrm{QP}$$

This would explain the finding of Brust and co-workers. From rate-distortion curves which were reported in their study (at low bitrates), ΔQP was set to a value ranging from 2 to 3 when full spatial-resolution frames were used, while it was in the range from 7 to 9 when low spatial-resolution frames were used in the base view. Although applying asymmetric quality for MSR-MVC would improve the coding efficiency, it is very critical from the point of view of suppression theory, since full spatial-resolution frames (which contain the high frequency components) are highly quantized.

There are several outcomes from the study presented in this section. First, the mixed spatial-resolution frames should use a different reference frame selection. This is due to the dissimilar effect of inter-view prediction during the coding of full spatial-resolution and low spatial-resolution frames. Also, reference frame ordering for full spatial-resolution frames should index full spatial-resolution frames prior to low spatial-resolution

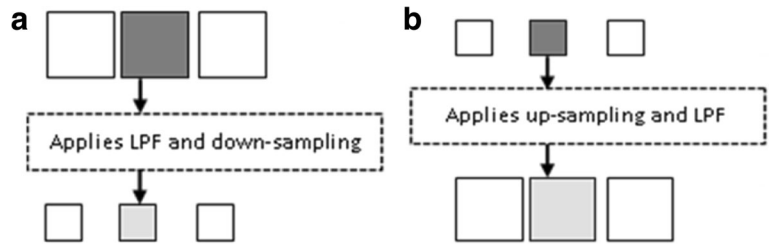

Fig. 2 Different scenarios for inter-view prediction direction when (**a**) full spatial-resolution (FR) frames and (**b**) low spatial-resolution (LR) frames are used in the base view

reference frames. This is due to the lower inter-view prediction efficiency resulting from using low spatial-resolution reference frames in predicting frames with higher spatial-resolution.

3.2 Evaluation of decimation and interpolation methods

A comparative study among different decimation and interpolation methods in terms of computational complexity and coding efficiency is presented in this section. Since H.264/AVC enables inter-picture prediction at a level of quarter-pixels, each reference frame is represented by 16 samples which include: one integer sample; three Half-Pixel (H-Pel) samples; and twelve Quarter-Pixel (Q-Pel) samples.

From the literature survey, there are two methods for decimating full spatial-resolution frames; they are the conventional decimation method and the high-performance decimation method. The conventional method applies the decimation separately on each sample which belongs to a full spatial-resolution reference frame [3]. The high-performance method filters and down-samples first the integer sample followed by obtaining the remaining samples from the decimated integer sample [17]. Therefore, this method filters a lower amount of samples compared to the conventional method since it is only applied for the integer sample. This is due to applying an MPEG filter; 13-tap (or AVC filter; 6-tap) for decimating (or interpolating) a single reference frame rather than applying it for sixteen frames as in the conventional method. Figure 5 sketches these methods, where a downwards and an upwards arrows refer to sub-sampling and up-sampling, respectively.

The conventional and high-performance decimation methods were assessed. The views which are described in the previous section were coded by the prediction architecture depicted in Fig. 3a. The coding performance and the time needed by each decimation method were compared. The measurements reported here are based on the quantisation setting for coding each stereoscopic video at low bitrate (Table 3; QP_H). Based on rate-distortion results, the conventional and high-performance decimation methods gave similar coding efficiency, where the high-performance method achieved slightly better coding efficiency than the conventional method by saving the bitrate by 0.88 Kbps. With regards to total decimation time, the high-performance method decreased decimation time by 24% compared to the conventional method.

Different interpolation methods were also examined. The conventional method applies interpolation for each sample separately. On the contrary, the high-performance method interpolates the integer sample first by the AVC 6-tap filter, while the remaining sub-pel samples are generated using the interpolated integer sample. Similar experiments were conducted using the prediction architecture depicted in Fig. 3b. Based on rate-distortion results, the conventional decimation method and the high-performance decimation method gave the same coding efficiency. However, the latter method reduced the amount of time needed for interpolation significantly, by up to 56% compared to the time needed by the conventional interpolation method.

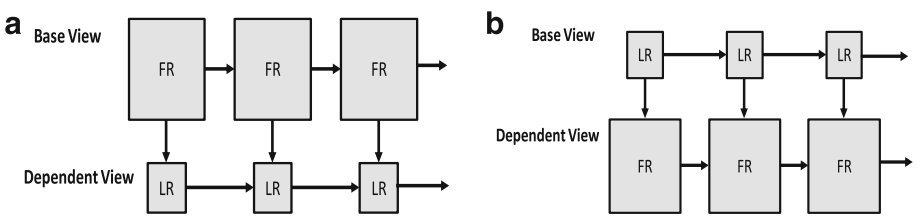

Fig. 3 Sequential view prediction structure for (**a**) first and (**b**) second inter-view prediction direction. FR and LR stand for full spatial-resolution frames and low spatial-resolution frames, respectively

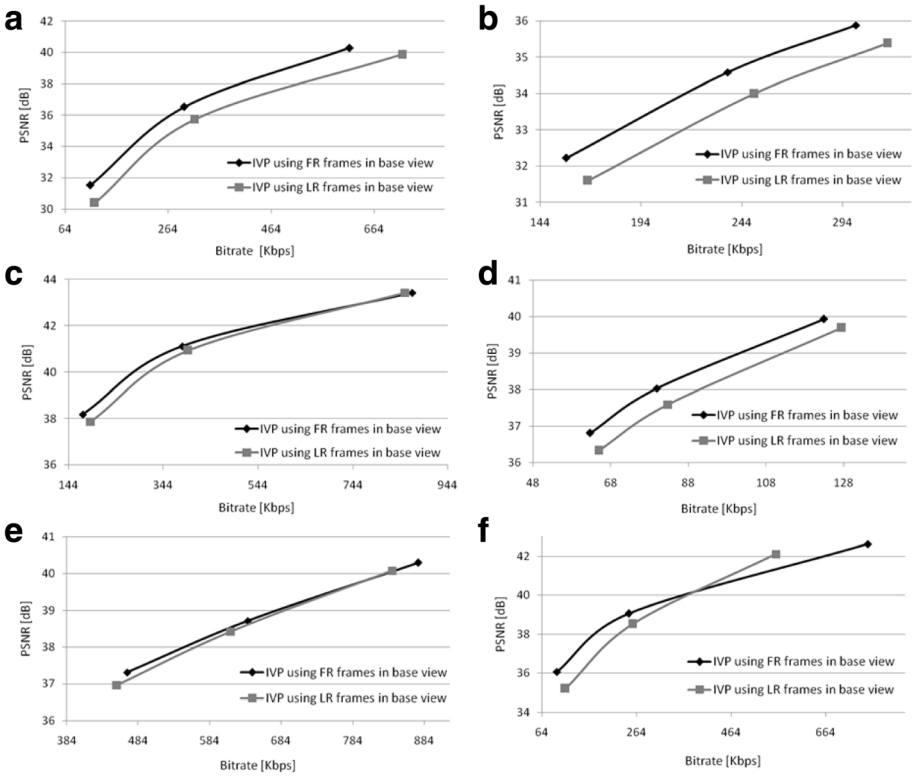

Fig. 4 a–f Rate-distortion curves for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena videos. IVP, FR and LR stand for inter-view prediction direction, full spatial-resolution frames and low spatial-resolution frames, respectively

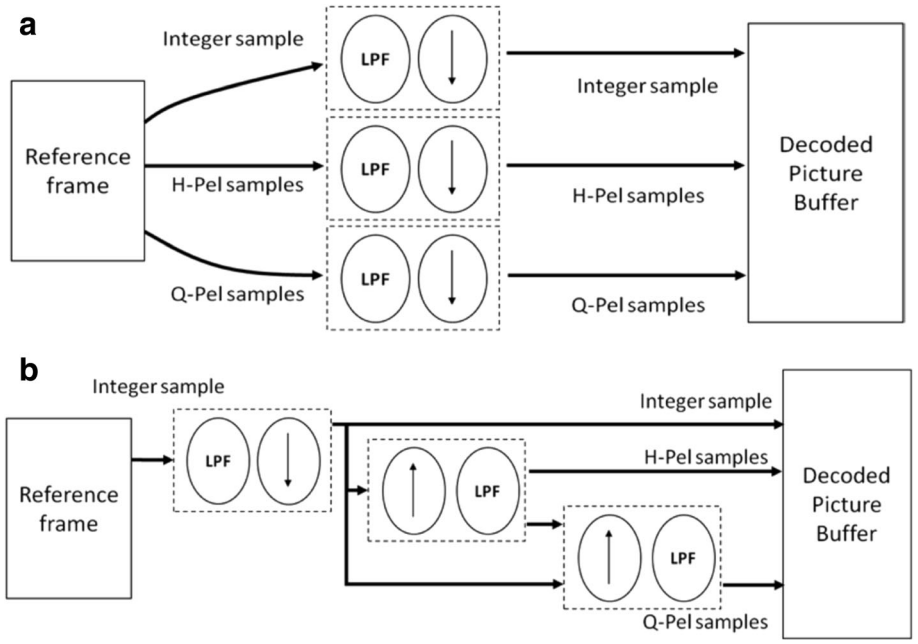

Fig. 5 Decimation methods. **a** Conventional method. **b** High-performance method

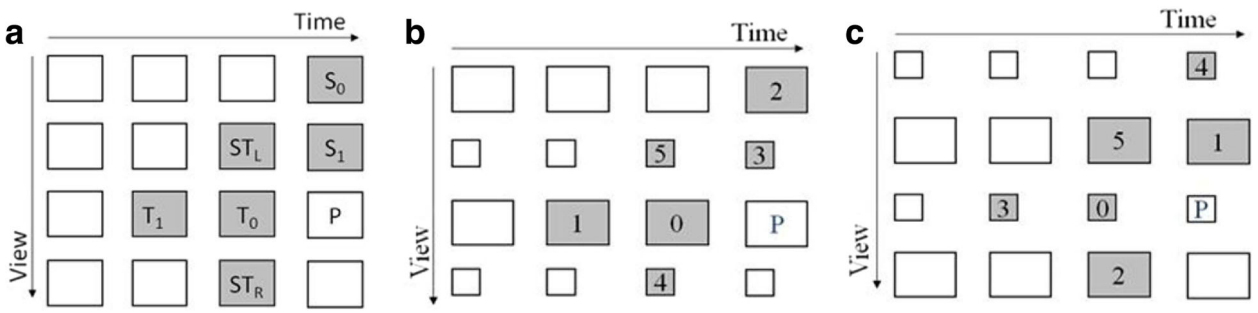

Fig. 6 Prediction architectures. **a** Symmetric spatial-resolution frame. **b** Full spatial-resolution frames. **c** Low spatial-resolution frames. *T*, *S* and *ST* denote temporal, spatial and Spatio-temporal reference frame, respectively

Based on the comparative study, it is clear that deploying high-performance methods for decimating and interpolating reference frames would be a preferred choice in terms of coding efficiency and computational complexity.

4 Prediction architecture for mixed spatial-resolution MVC based on block matching statistics

This section discusses the main investigations towards proposing a prediction architecture for MSR-MVC. These investigations start with analyzing block matching among frames with mixed spatial-resolution in order to define the reference frames which play a most significant role in block matching. Since videos have diverse characteristics, another level of block matching analysis is conducted to find a key for how to skip reference frames which play an insignificant role in block matching (dynamic reference frame selection). Lastly, the adaptive reference frame ordering algorithm is integrated with the proposed prediction architecture to code videos with hard scene changes efficiently [24].

Block matching statistics among reference frames were computed. The Break-dancers dataset was used in the analysis because it has balanced amounts of temporal and inter-view correlations [27]. Based on the outcomes reported in section 3.1, two experiments were conducted in order to define the reference frame selection for full spatial-resolution and low spatial-resolution frames. Four-view videos were used in each experiment; full spatial-resolution frames and low spatial-resolution frames were used in the base view for the first and the second experiments, respectively. Since the dataset contains eight views, five different sequences were obtained, where the first sequence contains View 0 up to View 3, while the last sequence contains View 4 up to View 7. Both experiments were conducted for these sequences, where the average block matching was computed using a preliminary prediction architecture (which was previously proposed for symmetric multi-view coding) as shown in Fig. 6a [23]. All frames were predicted using the same reference frame selection method in both experiments. Figures 6b, c depicts the prediction architectures for both experiments, where the shaded blocks are for reference frame selection while numbers inside these blocks indicate the reference frame ordering. Based on the results presented in section 3.2, the high performance decimation and interpolation methods were applied to enable inter-view prediction among mixed spatial-resolution frames. Table 4 shows the analysis results, where the significant reference frames for predicting full spatial-resolution frames are T_0 and S_0. These frames contribute by 91.1% while T_0 and S_1 have a significant role in block matching for predicting low spatial-resolution frames (on average 92.2%). The most challenging part in MSR-MVC is coding full spatial-resolution frames which belong to dependent views. This is due to a lower reliability of inter-view prediction for S_1, as shown in Table 4. The second temporal frame is therefore included during the prediction of full spatial-resolution frames which belong to a dependent view.

Predicting full spatial-resolution frames is a major source for computational complexity, since each frame is four times bigger than a low spatial-resolution frame (when it is decimated by a factor of two horizontally and vertically). Since multi-view videos have a variety of scene characteristics, the reference frame selection for full spatial-resolution frames is adaptive, where the spatial

Table 4 Average block matching statistics for full and low spatial-resolution frames

Statistical analysis results (%)	T_0	T_1	S_0	S_1	ST_R	ST_L
Full-resolution frame	79.294	3.959	11.8068	4.0044	0.5794	0.3564
Low resolution frame	60.2724	1.7838	0.866	31.967	4.1662	0.9446

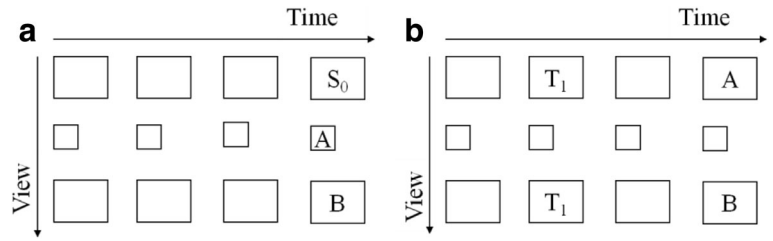

Fig. 7 Correlation among frames; predicted by **a** spatial frame and **b** second temporal frame

and second temporal reference frames are skipped when the expected amounts from their block matching are insignificant. Two statistical analyses were conducted to find a correlation among these frames with their nearby coded frames. The analysis results would provide a key for when to skip using these reference frames. Spatial reference frame S_0 is the source for inter-view prediction for A and B frames (both belong to depended views) as shown in Fig. 7a. The amounts of inter-view predicted blocks in A and B frames could be correlated. To validate this correlation, a statistical analysis was performed to compute the number of inter-view predicted blocks for A and B frames using the same reference frame S_0. The average inter-view prediction correlation, based on the six videos, was 0.44. This indicates a moderate positive relationship between the number of inter-view predicted blocks, when coding low spatial-resolution frames and full spatial-resolution frames. The number of inter-view predicted blocks for low spatial-resolution frames (A frame) was therefore analyzed. When this number is less than the threshold (discussed at the end of this section), then reference frame S_0 is skipped during the coding of a full spatial-resolution frame (B frame). Similarly, a statistical analysis was conducted in order to validate the correlation among temporal-predicted blocks in both frames (A and B frames), as depicted in Fig. 7b. The figure shows that a similar relationship exists (with a correlation coefficient measured to be 0.42) among second temporal reference frames during the coding of A and B frames. The T_1 temporal frame is therefore skipped during the coding of B frame when the amount of block matching during the coding of A frame (by second temporal reference frame) is less than the threshold.

To set the threshold value, six videos were coded via H.264/AVC-based MVC, where different thresholds were used (0, 2.5, 4, 6, 12 and 20). Each value for the threshold represents the amount of block matching as a percentage. According to the literature, block matching in the range from 5 to 6, is described as relatively low, and it is described as significantly high when it is greater or equal to 12 [20, 21]. Increasing the threshold value reduces the amount of time needed to encode a multi-view video, through skipping more reference frames at the expense of increasing the average bitrate, compared to the same codec which does not apply the threshold. Figure 8 shows the effect of using different threshold values upon the increase of the bitrate; setting the threshold to 2.5 results in a small bitrate increase (0.12 Kbps) compared to setting it to 12 (which causes a significant bitrate increase by 12.3 Kbps). With regards to deploying the same multi-view coding technique without using the threshold, the results show that the savings in average coding time, when thresholds are set to 2.5 and 12, are 9 and 31.5%, respectively.

A prediction architecture is thus proposed, based on the block matching statistics given in the foregoing. Figure 9 presents the proposed prediction architecture, where the group-of-picture size was set to 8. The prediction architecture deploys low spatial-resolution frames in the middle view. Dashed arrows are reference frames which are used when conditions A and B (as described below) are true. When the number of inter-view prediction blocks for a low spatial-resolution frame is higher than the threshold, then condition A is true. Similarly, when temporal predicted blocks for a frame, which belongs to the base view is higher than the threshold, then condition B is true. The threshold was set to 2.5%, which indicates an insignificant number of matching blocks. For full spatial-resolution frames, which belong to the third view, there are four possible cases for reference frame selection as illustrated in Table 5. They represent all combinations of reference frame selections for full spatial-resolution frames.

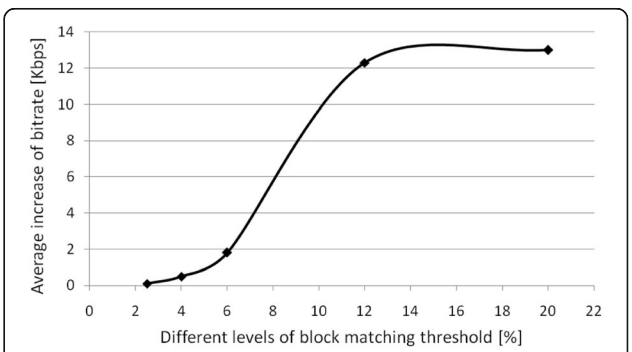

Fig. 8 Effect of the block matching threshold on the average bitrate

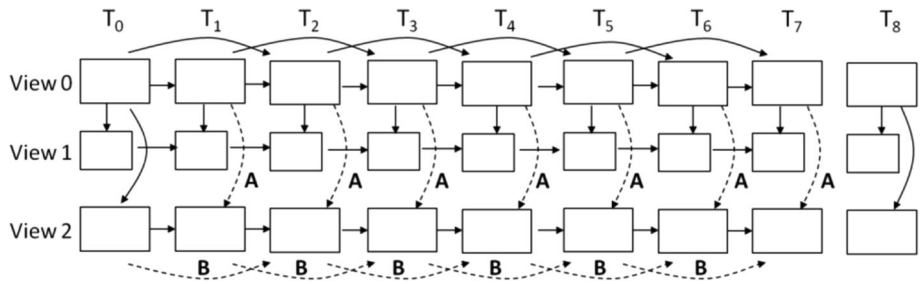

Fig. 9 Proposed prediction architecture for mixed spatial-resolution MVC

The adaptive reference frame ordering algorithm, reported previously [24], is integrated with the proposed prediction architecture. The algorithm is independent of reference frame selection and it offers an efficient mechanism for reordering frame indices which is vital when coding multi-view videos which contain multiple scenes. Coding a frame which belongs to a new scene would change the reference frame ordering, where the most significant reference frame becomes the nearest spatial frame instead of the recent temporal frame. Therefore, the algorithm first detects scene changes by analyzing the amount of intra-prediction for frames which belong to dependent views, then it alters reference frame ordering accordingly so that the spatial frames are indexed prior to temporal frames [24]. The new reference frame ordering is applied for the following frames which belong to neighbouring views.

5 Evaluation of the performance of the prediction architecture

The proposed prediction architecture was evaluated against other architectures in terms of coding efficiency, computational complexity and memory consumption. The hierarchical B-picture and an extended architecture based on 3D-DMB were used in the comparison. Three-view videos were coded by H.264/AVC using these prediction architectures, where the middle view uses low spatial-resolution frames and the group-of-picture size was set to 8. The comparison was performed on two coding scenarios which include coding videos depicting a single scene and coding a video which shows different scenes.

Table 5 Four cases for reference frame selection during the coding of full spatial-resolution frames

Condition A	Condition B	1st REF	2nd REF	3rd REF
False	False	T_0	n/a	n/a
True	False	T_0	S_0	n/a
False	True	T_0	T_1	n/a
True	True	T_0	S_0	T_1

N/A not applicable

In the context of the first scenario, H.264/AVC using the proposed prediction architecture reduced the amount of memory by 31.6 and 51.9% while it speeded-up coding by on average of 14 and 57%, compared to the same codec deploying an extended architecture based on 3D-DMB and hierarchical B-picture, respectively. It was found that the proposed prediction architecture needs less bitrate for transmitting mixed spatial-resolution videos, compared to the extended architecture based on 3D-DMB, by on average 13.1 Kbps. HBP was found to be more coding efficient than the proposed prediction architecture, where HBP obtained better quality by on average 0.78 dB while requiring less bitrate by on average 24.9 Kbps. Figure 10 shows rate-distortion curves for the codec that uses these prediction architectures; HBP, the proposed prediction architecture, and the extended architecture based on 3D-DMB. From these results, it is clear that the proposed prediction architecture is a better choice than 3D-DMB when coding videos which contain a single scene, while it gives inferior coding efficiency, it has less computational complexity and less memory complexity compared to the HBP architecture.

In the context of the second scenario, a multi-view video with hard scene changes is coded using H.264/AVC multi-view video coding. Figure 11 shows rate-distortion curves obtained when coding the video using the three prediction architectures. The proposed prediction architecture integrated with the adaptive reference frame ordering algorithm saved on average 28.7 and 35.4 Kbps compared to the HBP architecture and to the extended architecture based on 3D-DMB, respectively. It was seen to give similar quality for multi-view video coded with the extended architecture based on 3D-DMB. HBP achieved better quality by on average 0.38 dB compared to the corresponding video that was coded by the proposed prediction architecture. The proposed prediction architecture accelerates coding time by on average 64 and 33%, compared respectively to HBP and to the extended 3D-DMB architectures.

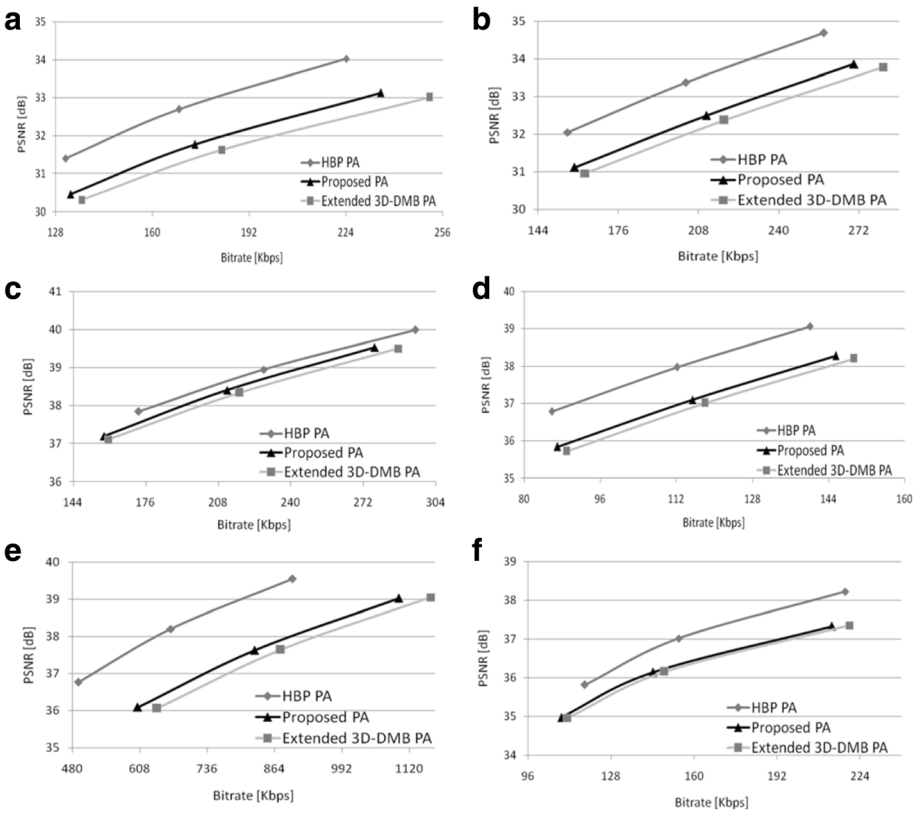

Fig. 10 a–f Rate-distortion curves for coding, by different prediction architectures (PAs), the multi-view videos known as Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena, respectively

6 Conclusions

This paper presents investigations of mixed spatial-resolution multi-view video coding, and it proposes a new prediction architecture. The investigations which underpinned the development of the proposed prediction architecture include: exploring the effect of inter-view prediction direction upon the efficiency of multi-view video coding; comparing different methods for the decimation and interpolation of reference frames; and conducting statistical analyses of block matching. Based on the outcomes from these studies, a prediction architecture is proposed, and it is integrated with the adaptive reference frame ordering algorithm, to provide an efficient coding solution for videos with hard scenes change.

The effect of different inter-view prediction directions on the coding efficiency of mixed spatial-resolution stereoscopic video coding is discussed. At low bitrates, mixed spatial-resolution stereoscopic video coding provides superior coding efficiency, when using full spatial-resolution frames rather than low spatial-resolution frames in the

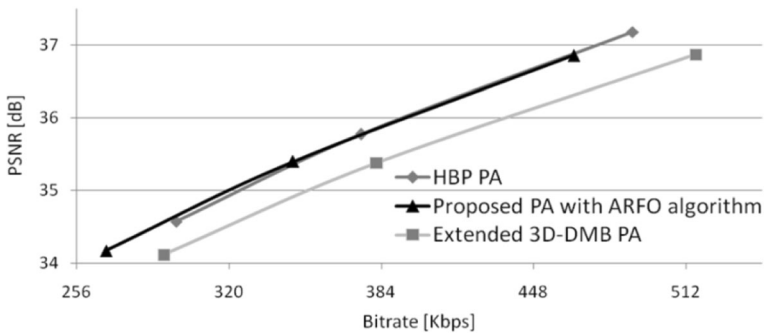

Fig. 11 Rate-distortion curves for coding, by different prediction architectures (PAs), a multi-view video that has hard scene change

base view. This implies that full spatial-resolution and low spatial-resolution frames should use different reference frame selection and reference frame ordering processes. A comparison of different decimation and interpolation methods showed that the high-performance methods reduce the amount of time needed for both processes through filtering fewer samples than the conventional methods. The high-performance methods for decimation and interpolation were therefore used when computing block matching statistics and in the comparisons reported in Section 5.

Based on the outcomes of the investigation of interview prediction and of the investigation of the decimation and interpolation of reference frames, in addition to results from statistical analyses of block matching, a prediction architecture is proposed. In this prediction architecture, nearest temporal and spatial reference frames are selected during the coding of a low spatial-resolution frame. A full resolution frame which belongs to the dependent view uses two temporal frames and a neighbouring full spatial-resolution reference frame. Temporal and spatial reference frames are dynamically skipped when their expected numbers of matching blocks are insignificant. The proposed prediction architecture is integrated with the adaptive reference frame ordering algorithm, to dynamically adapt the reference frame ordering when coding a video which depicts hard scene changes.

The proposed prediction architecture is compared to the extended architecture based on 3D-DMB and hierarchical B-picture prediction architectures in terms of computational complexity, memory consumption and coding efficiency. From the results, the proposed prediction architecture is shown to have less computational complexity (by on average from 14 to 57%) and less memory consumption (by on average from 31.6 to 52%) compared to the other architectures. Its coding efficiency is superior to a corresponding codec, which deploys the extended architecture based on 3D-DMB by demanding less bitrate by on average 13.1 Kbps, while HBP provides the best coding efficiency among other architectures when coding videos, which depict a single scene. The proposed prediction architecture integrated with the adaptive reference frame ordering algorithm provides better coding solution among other architectures when coding multi-view video which depicts several scene changes. It requires less bitrate by on average from 28.7 to 35.4 Kbps, less computational complexity (by on average from 33 to 64%) compared to a codec which deploys the extended architecture based on 3D-DMB or HBP prediction architectures.

Abbreviations

3D-DMB: 3D digital multimedia broadcast; ARFO: Adaptive reference frame ordering; HBP: Hierarchical B-picture; H-Pel: Half-pixel; MSR: Mixed spatial-resolution; MVC: Multi-view video coding; Q-Pel: Quarter-pixel

Acknowledgements

I would like to acknowledge Staffordshire University for PhD scholarship to carry out the research titled "Low bitrate multi-view video coding based on H.264/AVC".

Funding

This research project is funded by Staffordshire University through partial scholarship.

Authors' contributions

HS carried out the studies reported in the manuscript in addition to preparing manuscript draft. MM conceived of the study and participated in its design and coordination and helped to draft and review the manuscript. CC helped to draft and review the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]College of Engineering, Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt. [2]School of Architecture, Computing and Engineering, University of East London, London, UK. [3]Faculty of Computing, Engineering and Sciences, Staffordshire University, Stoke-on-Trent, UK.

Received: 21 December 2015 Accepted: 23 January 2017
Published online: 13 February 2017

References

1. H Brust, A Smolic, K Mueller, G Tech, T Wiegand, *Mixed Resolution Coding of Stereoscopic Video for Mobile Devices. Paper presented at the true vision - capture, transmission and display of 3D video*, 2009, pp. 1–4
2. P Aflaki, MM Hannuksela, M Gabbouj, Subjective Quality Assessment of Asymmetric Stereoscopic 3D Video. SIViP, **9**(2), 331–345 (2013)
3. C Fehn, P Kauff, S Cho, H Kwon, N Hur, J Kim, *Asymmetric Coding of Stereoscopic Video for Transmission over T-DMB. Paper presented at the true vision - capture, transmission and display of 3D video*, 2007, pp. 1–4
4. G Miao, N Himayat, Y Li, A Swami, Cross-layer optimization for energy-efficient wireless communications: a survey. Wirel. Commun. Mob. Comput. **9**, 529–542 (2009)
5. M Paul, G Sorwar, *Encoding and decoding techniques for medical video signal transmission and viewing. Paper presented at the 6th IEEE/ACIS international conference on computer and information science*, 2007, pp. 750–756
6. F Dufaux, B Pesquet-Popescu, M Cagnazzo, *Emerging technologies for 3D video* (John Wiley & Sons, Ltd, Chichester, 2013), p. 147
7. V De Silva, HK Arachchi, E Ekmekcioglu, A Fernando, S Dogan, A Kondoz, S Savas, *Psycho-physical limits of interocular blur suppression and its application to asymmetric stereoscopic video delivery. Paper presented at the international packet video workshop*, 2012, pp. 184–189
8. L Stelmach, Wa James Tam, D Meegan, A Vincent, Stereo image quality: effects of mixed spatio-temporal resolution. IEEE Trans. Circuits Syst. Video Technol. **10**(2), 188–193 (2000)
9. G Saygili, CG Gurler, AM Tekalp, *Quality assessment of asymmetric stereo video coding. Paper presented at the IEEE international conference on image processing*, 2010, pp. 3–6
10. MT Pourazad, P Nasiopoulos, RK Ward, *A New Prediction Structure for Multiview Video Coding. Paper presented at the international conference on digital signal processing*, 2009, pp. 1–5
11. S-H Jung, W-J Park, T-Y Kim, *Fast reference frame selection with adaptive motion search using rd cost. paper presented at the spring congress on engineering and technologyconference*, 2012, pp. 1–4
12. European Broadcasting Union, Digital audio broadcasting; digital multimedia broadcasting video service; user application specification (2005), http://www.etsi.org/deliver/etsi_ts/102400_102499/102428/01.01.01_60/ts_102428v010101p.pdf. Accessed 1 Feb 2011
13. H Yang, M Yu, G Jiang, *Decoding and Up-sampling Optimization for Asymmetric Coding of Mobile 3DTV. Paper presented at the TENCON 2009 IEEE region 10 conference*, 2009, pp. 1–4
14. M Yu, H Yang, S Fu, F Li, R Fu, G Jiang, *New Sampling Strategy in Asymmetric Stereoscopic Video Coding for Mobile Devices. Paper presented at the international conference on E-Product E-Service and E-Entertainment*, 2010, pp. 1–4
15. Y Chen, S Liu, Y Wang, MM Hannuksela, H Li, M Gabbouj, *Low-complexity Asymmetric Multiview Video Coding. Paper presented at the IEEE international conference on multimedia and expo*, 2008, pp. 773–776

16. H Brust, G Tech, K Mueller, T Wiegand, *Mixed resolution coding with inter view prediction for mobile 3DTV. Paper presented at the true vision - capture, transmission and display of 3D video conference*, 2010, pp. 1–4
17. P Aflaki, W Su, M Joachimiak, D Rusanovskyy, MM Hannuksela, *Coding of mixed-resolution multiview video in 3D video application. Paper presented at the international conference of image processing*, 2013, pp. 1704–1708
18. E Ekmekcioglu, ST Worrall, AM Kondoz, *Bit-rate adaptive downsampling for the coding of multi-view video with depth information. Paper presented at the true vision - capture, transmission and display of 3D video conference*, 2008, pp. 137–140
19. Y Chen, Y-K Wang, K Ugur, MM Hannuksela, J Lainema, M Gabbouj, The emerging MVC standard for 3D video services. EURASIP J. Adv. Signal Process. (1), 1–13 (2009)
20. P Merkle, A Smolic, K Muller, T Wiegand, Efficient Prediction Structures for Multiview Video Coding. IEEE Trans. Circuits Syst. Video Technol. **17** (11), 1461–1473 (2007)
21. A Kaup, U Fecker, *Analysis of Multi-Reference Block Matching for Multi-View Video Coding. Paper presented at the 7th workshop digital broadcasting*, 2006, pp. 33–39
22. Y Zhang, S Kwong, G Jiang, H Wang, Efficient multi-reference frame selection algorithm for hierarchical B pictures in multiview video coding. IEEE Trans. Broadcast. **57** (1), 15–23 (2011)
23. H Said, A Sheikh Akbari, *H.264/AVC Based multi-view video codec using the statistics of block matching. paper presented at the 55th international symposium ELMAR*, 2013, pp. 97–100
24. H Said, A Sheikh Akbari, M Moniri, *An adaptive reference frame re-ordering algorithm for H.264/AVC based multi-view video codec. Paper presented at the international conference EUSIPCO*, 2013, pp. 1–5
25. Y Su, A Vetro, A Smolic, *Common test conditions for multiview video coding, JVT Doc. JVT-T207*, 2006
26. K Sühring, JM reference software version 18.0 (2011), iphome.hhi.de/suehring/tml/download/old_jm/. Accessed 1 Jan 2011
27. Y Zhang, G Yi Jiang, M Yu, YS. Ho, Adaptive multiview video coding scheme based on spatiotemporal correlation analyses. ETRI J. **31**(2), 151–161 (2009)