On the Adoption of Multiview Video Coding in Wireless Multimedia Sensor Networks

S. Colonnese*, F. Cuomo*, O. Damiano†, V. De Pascalis† and T. Melodia‡
* University of Rome, Sapienza, DIET, Via Eudossiana 18, 00184 Rome Italy
e-mail: (stefania.colonnese, francesca.cuomo)@uniroma1.it
† University of Rome, Sapienza, DIS, Via Ariosto 25, 00185 Rome Italy
e-mail: (depascalis.valeria, ori.damian)@gmail.com
‡ Department of Electrical Engineering, State University of New York at Buffalo, NY 14260, USA
e-mail: tmelodia@eng.buffalo.edu

***Abstract*—This article explores the potential performance gains achievable by applying the *multiview video coding* paradigm in wireless multimedia sensor networks (WMSN). Recent studies have illustrated how significant performance gains (in terms of energy savings and consequently of network lifetime) can be obtained by leveraging the spatial correlation among partially overlapped fields of view of multiple video cameras observing the same scene. A crucial challenge is then how to describe the correlation among different views of the same scene with accurate yet simple metrics. In this article, we first experimentally assess the performance gains of multiview video coding as a function of metrics capturing the correlation among different views. We compare their effectiveness in predicting the correlation among different views and consequently assess the potential performance gains of multiview video coding in WMSNs. In particular, we show that, in addition to geometric information, occlusions and movement need to be considered to fully take advantage of multiview video coding.**

***Index Terms*—Multiview video coding; wireless multimedia sensor networks; bandwidth efficiency.**

I. Introduction

The H.264 Multiview Video Coding (MVC) is a new emerging standard defined by the Joint Video Team (JVT) of the ISO/IEC Moving Pictures Experts Group (MPEG) and the ITU-T Video Coding Experts Group (VCEG). MVC provides a compact representation for multiple views representing the same video scene. It is an extension of the single view H.264/MPEG-4 AVC standard [1] and it includes a number of new techniques to improve the coding efficiency, reduce the coding complexity, as well as new functionalities for multiview operations. The availability of a MVC standard can benefit several applications, including free-viewpoint video, 3D Television (TV) and immersive teleconferencing. The work in [2] describes these application scenarios: a multiview video is first captured and then encoded by a MVC encoder. Then, multiple copies of the coded bitstream are transmitted to different clients, each running a different application. In free-viewpoint video the user can interactively select the preferred viewpoint (target view) in a 3D space among several candidates, each with a different viewing angle. Obviously, the target view needs to be efficiently extracted from the bitstream and decoded every time the viewpoint changes. 3D TV is an extension of traditional 2D TV and its key feature is that more than one view is decoded and displayed simultaneously. Last, immersive teleconferencing captures some aspects of both previous applications since it aims at making people perceptually feel immersed in a 3D environment.

In this paper, we analyze the expected benefits of applying the MVC technology in wireless multimedia sensor networks (WMSN) [3] [4]. MVC outperforms single view video coding considerably in terms of compression efficiency and therefore enables the support of bandwidth demanding multimedia services on WMSN. In this way WMSN may provide advanced services such as video-surveillance, multi-camera moving object tracking, ambience intelligence, to cite a few. Intuitively, the effective compression gain due to multiview encoding varies with the camera geometric setup as well as with the characteristics of the captured video. In particular, the multiview gain depends on the correlation between different views, and it is related to the angular displacement of the cameras. A refined spatial correlation model is presented in [5], where the authors provide an analysis of the correlation characteristics of visual information observed by cameras with overlapped fields of view. Therein, the authors propose to leverage correlation in selecting the most relevant camera views. In more detail, geometrical parameters such as the location of the camera, the sensing radius, the sensing direction and the offset angle, are combined to evaluate the inter-view correlation. In [6], the authors build on the model in [5] to propose a network dependency graph that connects sensors exhibiting the most correlated camera views. Then, a distributed multi-cluster coding protocol is proposed to enable efficient network scalability. In the simulation scenarios analyzed in the paper, joint encoding of the views performed by suitably clustered sensors reduces the bandwidth required to transmit all the available WMSN views.

In this paper, we make the following contributions:

- We first experimentally assess the performance gains of multiview video coding as a function of metrics capturing the correlation among different views. In particular, we discuss the relationship between parameters describing camera geometry and the effective gain provided by MVC as compared to single view encoding techniques.
- We compare the effectiveness of different metrics in

predicting the correlation among different views and consequently assess the potential performance gains of multiview video coding in WMSNs based on experiments. When the difference in sensing direction between two cameras is small (i.e., cameras are close and their fields of view are overlapped "enough"), MVC outperforms AVC considerably in terms of compression efficiency and quality of the decoded sequence.

- We show that when the camera views exhibit differences that cannot be explained in terms of camera geometry only, the model of [5] can be extended to take into account specific issues that do not depend only on the geometry of the camera positions, including occlusions and movement.

The rest of the paper is structured as follows. After a brief description of the MVC standard, in Section II we introduce the relation between MVC coding gain and geometric camera settings; besides, assuming that MVC joint coding of multiple views is adopted, we analyze the criteria for selection of the reference view among the available ones. Section III introduces our numerical experiments setup while Section IV reports the performance evaluation in case of the selected video sequences. Finally Section V discusses our results and future work.

II. The Multiview Encoding Technology and its Adoption in WMSN

MVC compresses videos by leveraging the temporal and spatial redundancy introduced by cameras that capture the same scene from overlapping viewpoints. Video frames are then not only predicted from temporal references, but also from inter-view references[1]. In the recently defined H.264 MVC standard, one view is mandatorily encoded as *base*, single-view, bitstream, to provide backward compatibility with the H.264 AVC standard. The other views are encoded and encapsulated in suitable Network Adaptation Layer (NAL) units[2], intended to be recognized only by decoders conforming to one of the MVC profiles [7]. H.264 MVC also introduces a new type of picture, called *anchor*. The anchor picture provides random access to multi-view video sequence decoding, in a more efficient way than classical Intra-coded frame. [3]

The afore-described H.264 MVC standard was originally developed for high-quality video services, mainly referring to high definition and standard definition TV video sequence formats, such as those employed in 3D TV as well as free view point TV. In these services, multiple views are generated under the encoder control, so that the views are certainly correlated and the compression efficiency gain of MVC on single view coding is straightforwardly assessed.

The herein presented study is motivated by the fact that, in WMSN, the spatial allocation of the sensor nodes may be chosen out of the encoder control, on the basis of different criteria. For instance, sensor locations may be assigned based on requirements of network topology considerations, including effective radio coverage and link availability. Therefore, the resulting camera geometry may or may not be viable for MVC, and the benefits expected from the adoption of MVC need to be assessed.

In [5], the authors introduce a spatial correlation model that detects similarities between different views by exchanging only parameters associated to the geometry of cameras. These geometrical measurements are used to define a spatial correlation coefficient that captures inter-views dependencies.

As a first contribution, in this paper we assess the compression efficiency gain of MVC over single view coding on reduced resolution sequences, such as those envisaged in resource limited communication services provided by WMSN. In [5] and [6], the authors develop a relation between the camera geometry and the MVC gain over single view video coding. Here, we extend the analysis in [5], by conducting several numerical experiments taking into account also the objective quality of the decoded sequences. Besides, we emphasize the applicability limits of a geometry-based MVC performance analysis, envisaging relevant, content-dependent video scene characteristics that can affect the MVC performances.

As a second contribution, we have evaluated the impact of the *reference view* selection on the overall bit rate generated by the multiview codec. In fact, the selection of reference view to be used in MVC, that is the selection of the view that is first encoded in accordance to a single view paradigm, affects the overall compression efficiency of the MVC system. The herein provided experimental results show that suitable cameras clustering is not the only factor to be taken into account in order to minimize the employed bandwidth for an assigned decoded video quality. In fact, selection of the reference view itself brings advantages in terms of the overall employed bandwidth.

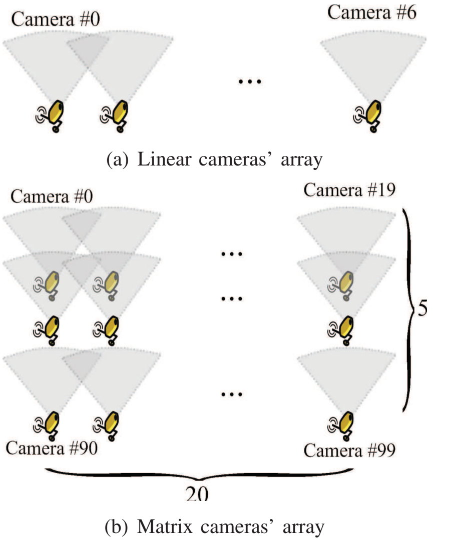
III. Numerical Experiments Setup

Our experimental results are obtained by using Joint Multiview Video Coding (JMVC), that is the reference software for MVC, on a selected set of multiview video test sequences, namely the sequences *Akko & Kayo*, *Kendo*, and *Balloons*. The sequences were acquired in accordance to different acquisition geometries [8] for 3D TV study purposes. Here, the sequences were resampled to a QCIF format, corresponding to a 176×144 image, that best matches the limited resources typical of WMSN. The frame rate is 30 frames per second for all the sequences. In all the experiments, the compression of the reference single view sequence is realized using a Quantization Parameter (QP) of 32, whereas the second and possibly the third views are encoded using a QP of 35 and 36, respectively. Such values lead to decoded sequences

[1] Inter-view prediction is allowed only between frames corresponding to the same temporal instant.

[2] Apart from the adoption of a NAL unit type extension, MVC involves a few high level syntax changes for signaling MVC specific information, mainly affecting sequence related information and supplemental enhancement information.

[3] When an Intra coded frame is introduced in the so called reference view, an anchor picture is introduced in each of the other views. These anchor pictures are predicted only from the said Intra coded frame, and therefore provide random access points in all the other views.



(a) Linear cameras' array

(b) Matrix cameras' array

Fig. 1. Examples of different geometries of multiple cameras to be considered in WMSNs.

TABLE I
EXPERIMENTAL PARAMETERS

Video Sequence	*Kendo*	*Balloons*	*Akko & Kayo*
Encoded Frames	400	500	30
Symbol Mode	CABAC	CABAC	CAVLC
GOP Size	4	4	15

whose average Peak Signal to Noise Ratio (PSNR) ranges in between 32 to 35dB, which corresponds to typical video telephony quality. Additional JMVC configuration parameters[4] are summarized in Table I.

The *Kendo* and *Balloons* sequences were both acquired by 7 cameras with 5 cm spacing, as shown in Figure 1(a). This configuration is representative of a linear array multimedia sensors. *Kendo* and *Balloons* sequences are characterized by medium to high movement. Figures 2(a) and 2(b) show the first frame of the views corresponding to cameras 0 and 6 of the *Kendo* sequence, respectively. Again, Figures 2(c) and 2(d) show the first frame of the views corresponding to cameras 0 and 5 of the *Balloons* sequence, respectively.

A more complex camera configuration is represented in Figure 1(b), i.e., a multimedia sensor grid. The *Akko & Kayo* multiview video sequence is originally acquired by 100 cameras organized in a 5×20 matrix structure, with 5 cm horizontal spacing and 20 cm vertical spacing. The herein reported experimental results have been obtained using a subset of 9 over 100 camera views; the subset corresponds to a 3×3 camera matrix with about 50 cm horizontal spacing and 40 cm vertical spacing. The first frame corresponding to each of the selected cameras is shown in Fig. 3.

The selected multiview sequences present different characteristics, in terms of camera angular displacement, floor-to-camera height, observed scene dynamics, and percentage of common areas in the framed images.

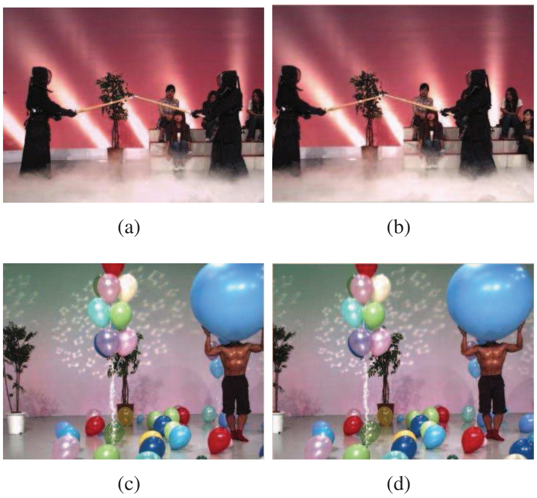


Fig. 2. Snapshots of the views corresponding to camera 0 (a) and 6 (b) of the *Kendo* sequence and to cameras 0 (c) and 5 (d) of the *Balloons* sequence.

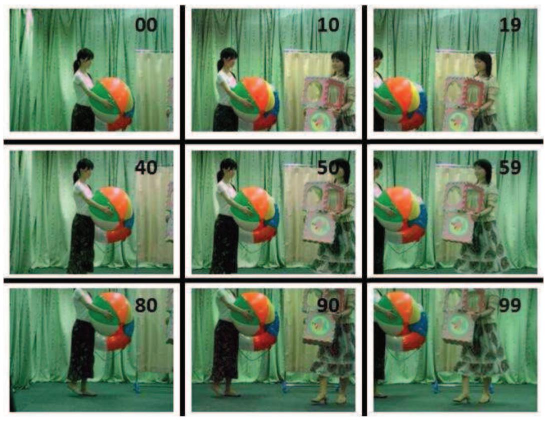


Fig. 3. Snapshots of the views corresponding to a selected set of cameras extracted from the *Akko & Kayo* multiview sequence (First row: camera 0, 10, 19; second row camera 40, 50, 59; third row camera 80, 90, 99).

IV. NUMERICAL EXPERIMENTS RESULTS

In this Section, we analyze the performance of H.264 MVC on multiple views. First, we assess the performance of joint multiview coding as a function of camera displacement. To this end, for each sequence, we encoded several pairs of views, always using the first view as the reference. In Fig. 4, we show the average rate in $kbit/s$ generated by the codec (gray bars) for the reference view (#0) and for the second view, as a function of the second view index. This latter, in turn, is related to the angular displacement between the corresponding cameras. As a reference, we also report the average rate (black bars) for each view using single view encoding, namely H.264 AVC. From Figs. 4 and 5 it is clear that the coding cost of the second view increases as the index increases.

[4] The group of pictures (GOP) used for encoding the video sequences consists of an INTRA or anchor frame(s), respectively introduced in the reference view and in the predicted view(s), followed by several hierarchically coded B and P frames.

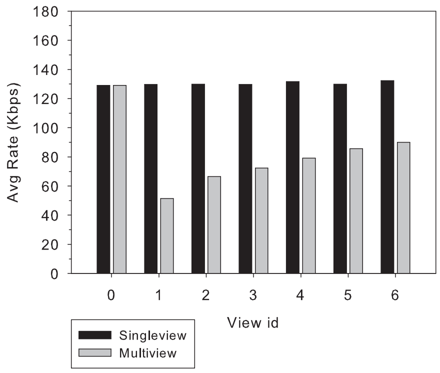


Fig. 4. Measured average bit rate in case of *Kendo* sequence

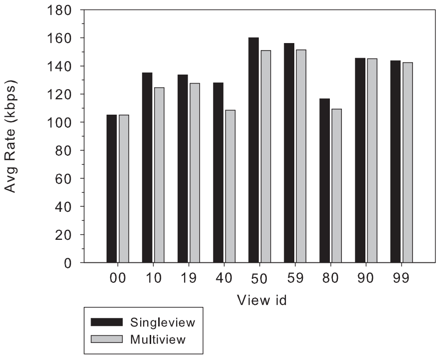


Fig. 6. Measured average bit rate in case of *Akko & Kayo* sequence

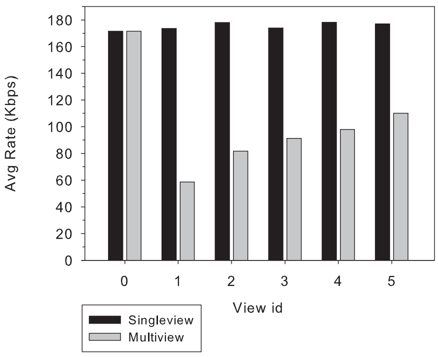


Fig. 5. Measured average bit rate in case of *Balloons* sequence

Then, the effectiveness of MVC encoding decreases with the angular distance of the cameras. Besides, we evaluated the MVC efficiency in case of camera views taken under the same viewing angle but at different camera heights, and the more general case of different heights and different angles. The encoding results obtained on the *Akko & Kayo* sequence are summarized in Fig. 6. All pairs $0-10, \cdots 0-99$ have been encoded assuming the view #0 as the reference view. As aforementioned, the relative gain of MVC encoding decreases as the angular displacement increases. This can be seen for instance from the increasing coding costs of view 10 and 19 both predicted from view $\#0$. The advantages of MVC encoding with respect to single view coding also reduces when the effective overlapped area between different views decreases, as occurs for the view pairs 0-40 and 0-60. When the angular displacement is large and the common image area is small the MVC gain over single view encoding becomes negligible.

An alternative parameter for evaluating the improvement due to MVC encoding with respect to single view encoding is the joint coding efficiency computed as:

$$\eta = 1 - r_{MVC}(0,i)/(r_0 + r_i) \tag{1}$$

where $r_{MVC}(0,i)$ denotes the total rate for H.264 MVC encoding of views 0 and i, and r_i the rate for H.264 AVC encoding of view i. In Figs. 10 and 11 we observe that, on the *Kendo* and *Balloons* sequences, η decreases as the sensing direction difference increases. Besides, on the *Akko & Kayo* sequence, reported in Fig. 12, the efficiency reflects the afore discussed considerations about overlapping areas between the views; interestingly enough, the efficiency peak is found for the closest camera pair, with no angular displacement (00-40). Moreover, among the three possible directions in the sensor grid of Figure 1(b) (horizontal, vertical and diagonal) the best gain is achieved for the vertical one. This shows that: (i) the angular displacement is not the only factor that influences the MVC gain; indeed pairs (00-40) and (00-80) where view are acquired under the same angle present different efficiencies; (ii) a major factor affecting the MVC gain is the percentage of overlapping areas; (iii) remarkably, the overlapping image area not only depends on the fields of view of the cameras, but also on the depth and movement of the framed objects.

To sum up, MVC brings advantages in terms of compression efficiency for views with large overlapped area as well as limited angular displacement. By observing these results we can suggest that, as a rule of thumb, a minimal percentage of 50% area overlap is required for exploiting the MVC efficiency. Besides, the cameras angular displacements should not exceed 10-15 degrees with respect to the real world objects to have MVC outperforming AVC.

A further issue addressed by the performance analysis is the impact of the *reference view* selection among an assigned set of views. To clarify this, we have considered triplets of views belonging to the above introduced test sequences. Then, we have considered different dependencies during the encoding phase. Specifically, we have encoded each triplet in accordance to the three different dependency orders in Fig. 13. Although the encoding cost of each view varies in the different cases, the overall triplets encoding costs exhibit a clear behavior. Table II reports the different total average bit rate on the various considered triplets. From Table II, it is clear that order 3 always presents the higher coding

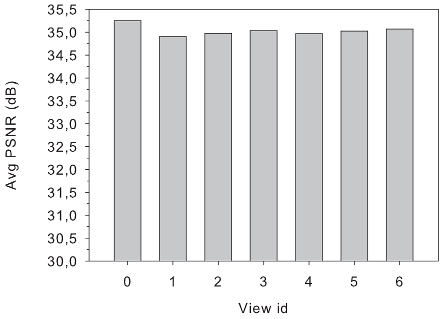


Fig. 7. Measured average PSNR in case of *Kendo* sequence

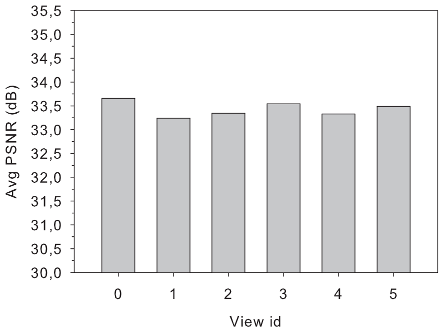


Fig. 8. Measured average PSNR in case of *Balloons* sequence

cost, whereas order 1 and 2 present quite similar results. To recap, the encoding dependency order affects the MVC efficiency; besides, establishing a pairwise dependency among the most correlated view systematically turns out to be the most effective choice. This kind of result should be taken into account when a clustering policy is adopted in a WMSN.

V. Discussion and concluding remarks

In this paper, we presented a preliminary analysis on the benefits of adopting multiview video coding techniques in wireless multimedia sensor networks. As a first contribution, we assessed the expected performance gain in adopting H.264 multiview video coding rather than multiple single view encoding. Specifically, the multiview video coding compression efficiency can be coarsely predicted on the basis of geometric camera angular displacement, as well as by using the more refined inter-view correlation model introduced in [5]. A few criteria for application of multiview video coding in wireless multimedia sensor networks were established. Finally, the experimental results on multiview video coding efficiency points out that state-of-the-art inter-view correlation models need to be generalized to take into account the effective overlapped image view areas. The latter depend on both acquisition geometry settings, i.e., the fields of view of the cameras, and on the acquired scene dynamics, such as the presence of moving objects at different depths. Let us remark that multiview video coding efficiency varies with the mentioned video sequences features that in turn may vary with time. Hence, the feasibility of multiview video coding encoding in a wireless multimedia sensor networks is expected to change as well, and the exchanging of suitable parameters for control and setting of multiview video coding shall be repeated to track such changes.

TABLE II
Multiview Triplet Encoding Cost

Video Sequence	Views	Order 1	Order 2	Order 3
Kendo	0-1-2	234.57	230.76	246.99
	0-3-6	274.40	272.79	291.33
	0-1-6	265.10	263.41	270.33
Balloons	0-1-2	286.50	277.57	311.89
	0-2-4	322.06	316.61	351.13
	0-1-4	310.58	303.11	328.14
Akko & Kayo	0-10-19	338.86	341.24	356.26
	0-40-80	322.16	326.14	322.90
	0-50-99	392.80	396.00	398.41

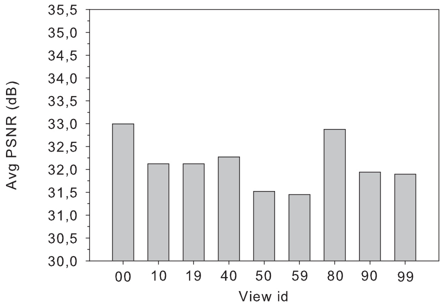


Fig. 9. Measured average PSNR in case of *Akko & Kayo* sequence

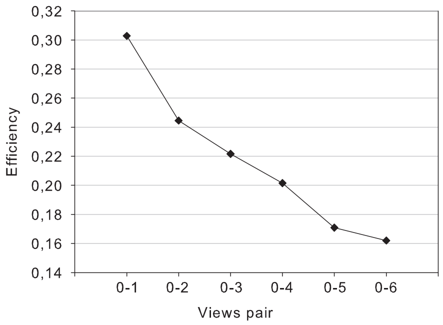


Fig. 10. Measured efficiency in case of *Kendo* sequence

References

[1] A. Puri, X. Chen, and A. Luthra, "Video coding using the H.264/MPEG-4 AVC compression standard," *Signal Processing: Image Communication*, vol. 19, no. 9, pp. 793 – 849, 2004.

[2] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," *EURASIP J. Appl. Signal Process.*, vol. 2009, pp. 8:1–8:13, January 2009.

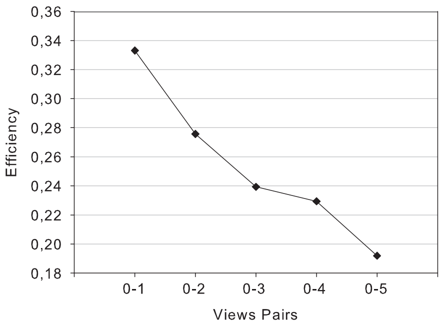


Fig. 11. Measured efficiency in case of *Balloons* sequence

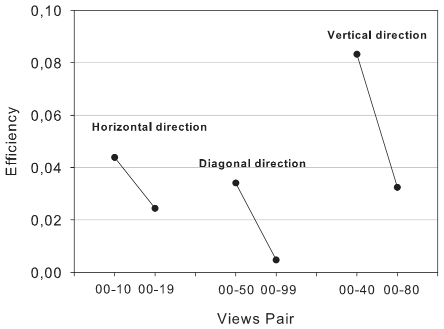


Fig. 12. Measured efficiency in case of *Akko & Kayo* sequence

[3] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, “A survey on wireless multimedia sensor networks,” *Comput. Netw.*, vol. 51, pp. 921–960, March 2007.

[4] R. Puri, A. Majumdar, P. Ishwar, and K. Ramchandran, “Distributed video coding in wireless sensor networks,” *Signal Processing Magazine, IEEE*, vol. 23, no. 4, pp. 94 –106, July 2006.

[5] R. Dai and I. F. Akyildiz, “A spatial correlation model for visual information in wireless multimedia sensor networks,” *Multimedia, IEEE Transactions on*, vol. 11, pp. 1148–1159, October 2009.

[6] P. Wang, R. Dai, and I. Akyildiz, “A Spatial Correlation-Based Image Compression Framework for Wireless Multimedia Sensor Networks,” *Multimedia, IEEE Transactions on*, vol. 13, no. 2, pp. 388 –401, april 2011.

[7] A. Vetro, T. Wiegand, and G. Sullivan, “Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard,” *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626 –642, April 2011.

[8] A. Smolic, G. Tech, and H. Brust, “Report on generation of stereo video data base, v2.0,” Tech. Rep. Mobile3dtv, July 2009.

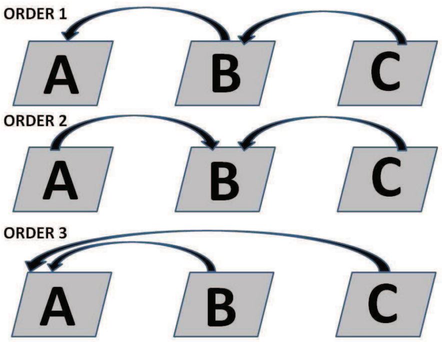


Fig. 13. Different dependency orders for coding triplets