Multimed Tools Appl (2010) 48:411–436
DOI 10.1007/s11042-009-0315-7

Successive refinement of side information for multi-view distributed video coding

Lucian Ciobanu · Luís Côrte-Real

Published online: 4 July 2009

Abstract Inter-camera registration in multi-view systems with overlapped views has a particularly long and sophisticated research history within the computer vision community. Moreover, when applied to Distributed Video Coding, in systems with at least one moving camera it represents a real challenge due to the necessary data at decoder for generating the side information without any a priori knowledge of each instant camera position. This paper proposes a solution to this problem based on successive multi-view registration and motion compensated extrapolation for on-the-fly re-correlation of two views at decoder. This novel technique for side information generation is codec-independent, robust and flexible with regard to any free motion of the cameras. Furthermore, it doesn't require any additional information from encoders nor communication between cameras or offline training stage. We also propose a metric for an objective assessment of the multi-view correlation performance.

Keywords Distributed video coding (DVC) · Wyner-Ziv (WZ) · Side information generation · Multi-view registration (MVR) · Overlapped views · Motion compensated extrapolation (MCE) · Scale-Invariant feature transform (SIFT)

1 Introduction

Distributed Video Coding (DVC), as a particular paradigm of Distributed Source Coding (DSC), has been one of the most active research areas in the signal processing community in the last years, providing a revolutionary new perspective over the

L. Ciobanu (✉) · L. Côrte-Real
Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
e-mail: lciobanu@inescporto.pt

L. Ciobanu · L. Côrte-Real
INESC, Porto, Portugal

L. Côrte-Real
e-mail: lreal@inescporto.pt

conventional video compression (e.g., the MPEGx, H.26x families). It has arisen in the favourable context of the ever increasing use of distributed architectures due to the emerging technical advances in the last decade, thus making it possible to deploy cheap, low-power sensing devices widely spread over large areas, from hand-held digital cameras to the omnipresent multimedia cellular phones.

The roots of Distributed Source Coding date back to the 1970s when the information-theoretic results of Slepian-Wolf in 1973 for lossless coding with side information at decoder side [28], and then in 1976 extended by Wyner-Ziv for lossy coding [31], have shown the conceptual importance of this distributed paradigm. As stated in theory, the Distributed Source Coding enables the same coding efficiency for architectures with independent encoders, having no communication between each other, and joint decoding as in the case the encoders are jointly encoding. As a consequence, when applied to Distributed Video Coding, it is an essentially reversed paradigm that enables the shift of the bulk of computation from encoder to decoder, as opposed to the conventional (e.g., H.26x, MPEGx) non-distributed coding.

Among the many challenging topics of Distributed Video Coding is the generation of side information, usually based on an elaborated correlation model that is essentially more complex when it comes to multi-view architectures [13, 19, 25, 26, 32]. There are currently two main techniques for generating the side information: trajectory-based motion interpolation (TMI) and hash-based motion estimation (HME) [5]. The former uses past and future reference frames, usually conventionally coded, in order to determine the motion vectors and finally to interpolate the frame in between. It proves to be highly reliable as long as the motion is sufficiently smooth and predictable, overwise other methods are needed in addition. As for works implementing the latter technique [1, 2, 5, 27], the hash-based motion estimation, the encoder sends a signature or hash in order to feed the decoder with the necessary data for the motion estimation process, at the cost of an extra-load on encoder and communication channel. Some works use in particular Cyclic Redundancy Check (CRC), as in [27], or high pass filters in the DCT domain [1] in order to generate the hash [5].

In the multi-view setting the above mentioned trajectory-based motion interpolation is called intra-camera interpolation. Additionally, it is usually performed an inter-camera interpolation (spatial interpolation) between neighbour cameras. Most works use a fusion of both, intra-camera and inter-camera interpolation, or use mode decisions in order to achieve optimal results on side information generation [4, 25, 26].

Most DVC schemes use a successive (iterative) refinement of the side information for each frame to decode, up to an acceptable achieved quality or equivalently, an acceptable small error-probability. For most codecs the iterative process is inherently performed by the inner turbo decoding, for others is part of a particular elaborated method dedicated to side information refinement [12]. In the latter case it employs multi-stage encoders and decoders where each frame decoding exploits the generated information from previous stages jointly with the side information, which may vary from stage to stage.

Camera registration has been recently an important topic of research in computer vision, as fundamental task in distributed video processing applications. A few solutions towards automatic, generic and flexible camera registration were proposed, usually assuming some context-specific knowledge. Szlavik et al. [29] and Benedek

et al. [6] use co-motion statistics to achieve a flexible camera registration of partially overlapped camera-views, i.e., by detecting any concurrent motion of various objects present in the scene and establishing the corresponding point-pairs. A general multi-view registration technique is proposed in [7] by reducing prior registration errors, e.g., from calibrated acquisition setup or a crude manual alignment, between all pairs in a set of range views.

Some still-image based techniques rely on image registration and attempt to detect static features in images like edges, corners, contours, shape, color, areas, etc. They are often applied on image pairs having small differences (e.g., stereo images) so that the difference between features is insignificant [20].

A comprehensive survey on image registration was presented in [9], and more recently in [33]. The latter reviews the existing techniques according with their nature (*area-based* and *feature-based*) as well as by four basic steps involved in the image registration process: *feature detection*, *feature matching*, *mapping function design* and, *image transformation and resampling*. The rapid development of image acquisition devices has led to intense research efforts on automatic image registration, and as [33] concludes, this remains an open issue. The recent approaches on automatic image registration are usually focused on specific fields like remote sensing (e.g., satellite imagery [8, 14, 16, 21], aerial imagery [18, 30]) and medicine [3, 22].

As stated in [33], given the variety of images to be registered as well as the diversity of possible discrepancies between the acquired images, it is impossible to design an adequate universal method for all registration tasks. Furthermore, within the area of camera registration for multi-view Distributed Video Coding it seems unrealistic to achieve an automatic, generic and constraint-free method applicable on most real-life scenarios (e.g., surveillance networks, mobile monitoring of large areas, visual tracking of humans, objects, events, sports, etc.), resilient to partial occlusions present in some views, insensitive to object movement in the scene or camera movement, etc. The lack of such methods motivated the authors to propose a novel technique, adequate for these scenarios. The key lies in a successive refinement of inter-camera correlation. See Sections 2 and 3 for more details.

This paper is focused on both the multi-view scenario and iterative decoding, describing a novel technique for generating the side information in a simplified two-view scenario (see Fig. 1) with one moving camera (the *target camera*). The *reference camera* performs conventional encoding (e.g., H.26x, MPEGx) of its perceived view (the "scene" view) and provides the side information, while the target camera conventionally encodes only the 1^{st} frame and applies Wyner-Ziv encoding for the remaining frames. Here the two views are considered to be completely overlapped and the proposed technique assumes that in a multi-camera environment (e.g., comprising dozens of cameras) there can be allocated at least one reference camera to capture the entire scene. Since the 2D representation of the 3D scene provided by the reference camera can result in a poor side information due to the different view angles of the reference and target cameras, we considered the use of multiple reference cameras. This approach allows the system improve the performance on complex scenes avoiding the use of complex 3D models. Section 5 contains partial results for a two reference camera scenario with fused side information.

The proposed technique is based on a fusion approach combining successive multi-view registration (MVR) and motion compensated extrapolation (MCE) for on-the-fly re-correlation of the two views at decoder, it doesn't require any additional

information from encoders nor communication between cameras or offline training stage. Unlike other works that use fusion techniques for generating the side information in the multi-view setting (see [4, 25, 26]), usually using 2 side cameras (left and right) situated at a close distance from the target (Wyner-Ziv) camera, this technique uses different methods so that it does not require any specific camera arrangement.

In Section 2 are presented an overview of the Scale-Invariant Feature Transform (SIFT) [23] and the proposed technique for multi-view registration of the cameras that is based on SIFT. Section 3 contains a detailed description of our technique for side information generation. In Section 4 we also propose a metric for evaluation of the produced multi-view correlation (based on MVR and MCE), as part of the side information generation. The achieved results are presented in Section 5. Finally, in Section 6 the conclusions and future work are drawn.

2 SIFT-based multi-view registration (SMVR)

2.1 Scale-Invariant Feature Transform (SIFT)

The SIFT algorithm, published by David Lowe in 1999 [23], enables tracking and detection specific applications to easily find highly distinctive key-points in images, with low-probability of mismatch and effortless one to one matching. Considering the proposed two-view scenario (see Fig. 1) and given a pair of images, each captured from one of the cameras, the problem of finding the point to point correlations between them is reduced to finding the one to one matches between the 2 sets of SIFT-generated key-points, one set for each image (see Fig. 2). Moreover, in this context we considered as few as 3 such matches between 2 images to be sufficient in order to compute its approximate view location and pose relative to the other image. Thus arises the SIFT's importance for automatic determining of point to point matches in two correspondent frames captured from two video sources having overlapped views. More references on SIFT can found in [10, 11, 15, 17, 24].

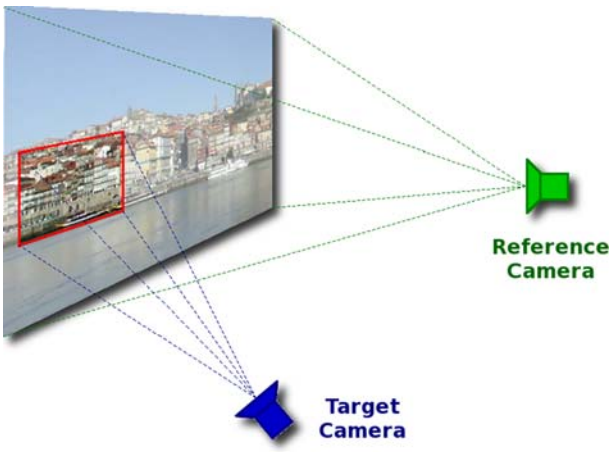

Fig. 1 The considered 2-camera scenario with complete-overlapped views

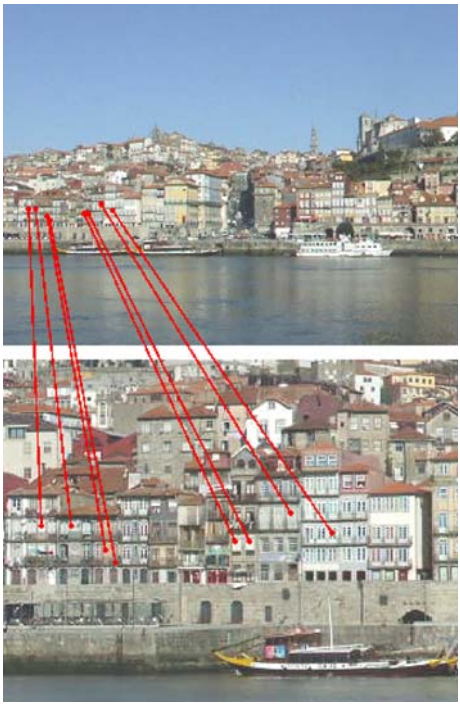

Fig. 2 Point to point correlations (*red lines*) as generated by the SIFT algorithm

The SIFT-generated key-points are further on exploited for achieving the proposed multi-view registration in a 2-step process: first for eliminating the false positives among matches and the second for building the actual multi-view correlation.

Throughout this paper it was used the "Porto" raw sequence (320×240, 100 frames, 5 fps) for illustration and evaluation purposes.

2.1.1 Eliminating the false positives

The accuracy of the results depends on the image resolution and quality, scene content, amount of details, etc., as well as the sensitivity set internally to SIFT for application specific purposes. There are however some drawbacks relative to the key-point false positives (outliers). Although the SIFT algorithm is highly performant just a few outliers can be sometimes fatal when trying to subsequently perform the multi-view registration (see Subsection 2.1.2). For this reason we developed a set of basic yet effective filter methods for eliminating the outliers, as follows: 1) *common end-points*, 2) *match crossings* and 3) *match direction* (*angle*) *coherence*; these are seen as a post-processing phase after generating the matches. The filters are applied on the SIFT generated matches as identified by their geometric features (illustrated in Fig. 2): line segment (red line) bounded by 2 end-points (tiny red circles) in a two-dimensional space.

The first filter (common end-points) is used to eliminate the duplicate matches as well as matches with one common end-point (in the latter case they are all removed). The match crossings filter eliminates those matches that cross many of the others (experimentally determined: more than 2), they usually are outliers. Then, the match

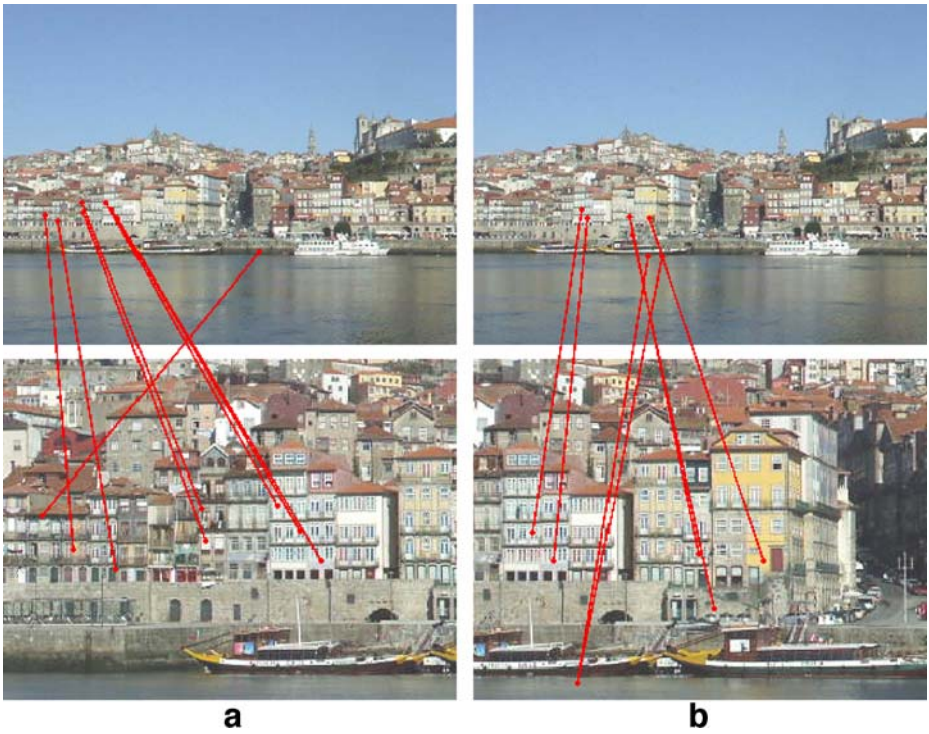

Fig. 3 Examples of outliers (**a**, **b**)

direction coherence takes into account the angle of each match. As they usually are almost parallel one to another (see Fig. 2) there was empirically determined a threshold (e.g., 0.1 radians) to determine the angle disparity between neighbour matches, thus detecting the lack of angle coherence that may appear among them. An usual cause is the presence of an outlier. See Fig. 3 for two examples of outliers.

2.1.2 Building the multi-view correlation

As suggested in the previous section, we use at least 3 matches between 2 images (correspondent frames), each one captured from one of the views, in order to roughly determine the approximate overlapping area between them (see Fig. 1). Normally, the more matches the more accurate is the estimation. For most scenarios it's imperative to have a rough approximation than not having anything at all. Additional validations are performed in order to prevent erroneous approximation when there are very few matches, as detailed below. However, for more robustness in certain scenes the minimum required number of matches can be set higher.

We call this area a "window" and it is basically the goal of our developed SIFT-based multi-view registration (SMVR) algorithm applied between 2 cameras, as illustrated in Fig. 4. More exactly, there's one determined window for each pair of frames, as shown in Fig. 5. Actually, a window is identified by a precise geometric quadrilateral (usually a rectangle) in a two-dimensional space, indicating

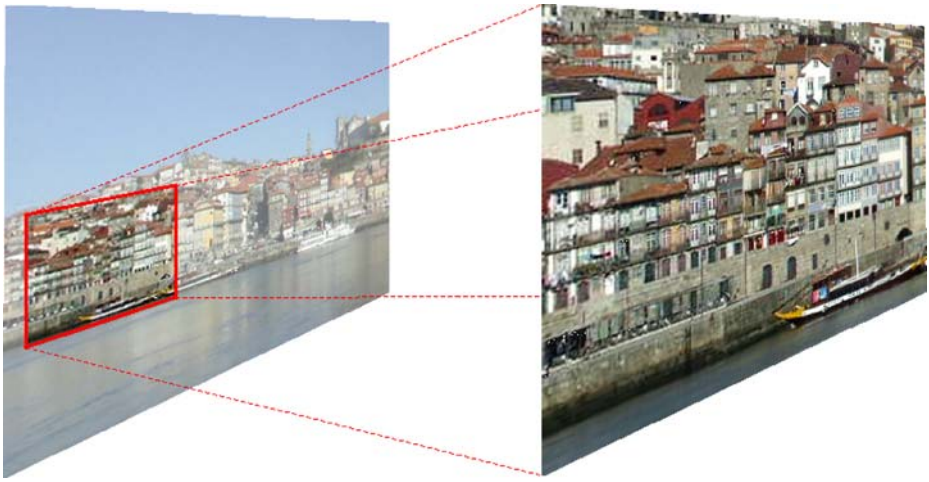

Fig. 4 Overlapping area (window) as determined by SMVR

the approximate position, angle and size of one view (target view) relative to the other (reference view), as depicted in Fig. 4.

The following are the steps for the actual estimation of the window:

1) select the most distant 3 matches (the extremes); then, for each corner-point of the target frame execute the following steps:
2) pick each pair of 2 matches from the previous subset of 3; for each such pair apply a linear projection of the corner-point in order to determine its correspondence into the reference frame (see Fig. 6); finally, there will be determined 3 such points (candidates)
3) consider the average of the 3 candidates as an approximation of the current window corner.

Nevertheless, each window may still suffer from the improper filtering of the previous mentioned outliers (among SIFT generated matches). As a consequence we developed two methods for validation purposes: *window area coherence* and *window angle coherence* that validate the current window with respect to the area (size) and angle of the previous found valid window. These are applied, along with the above mentioned set of filters for outliers elimination, as part of a post-processing phase before running our SMVR algorithm (see Section 3).

As expected, any failure in finding enough valid matches as generated by SIFT (the minimum is 3) or a valid window at the end is regarded as a failure of the entire SMVR process. In this case a complementary technique will be applied as presented in Section 3.

3 Proposed technique for side information generation

In this paper we propose a novel technique for successive generation of side information at decoder and implicitly, iterative decoding. More exactly, it involves

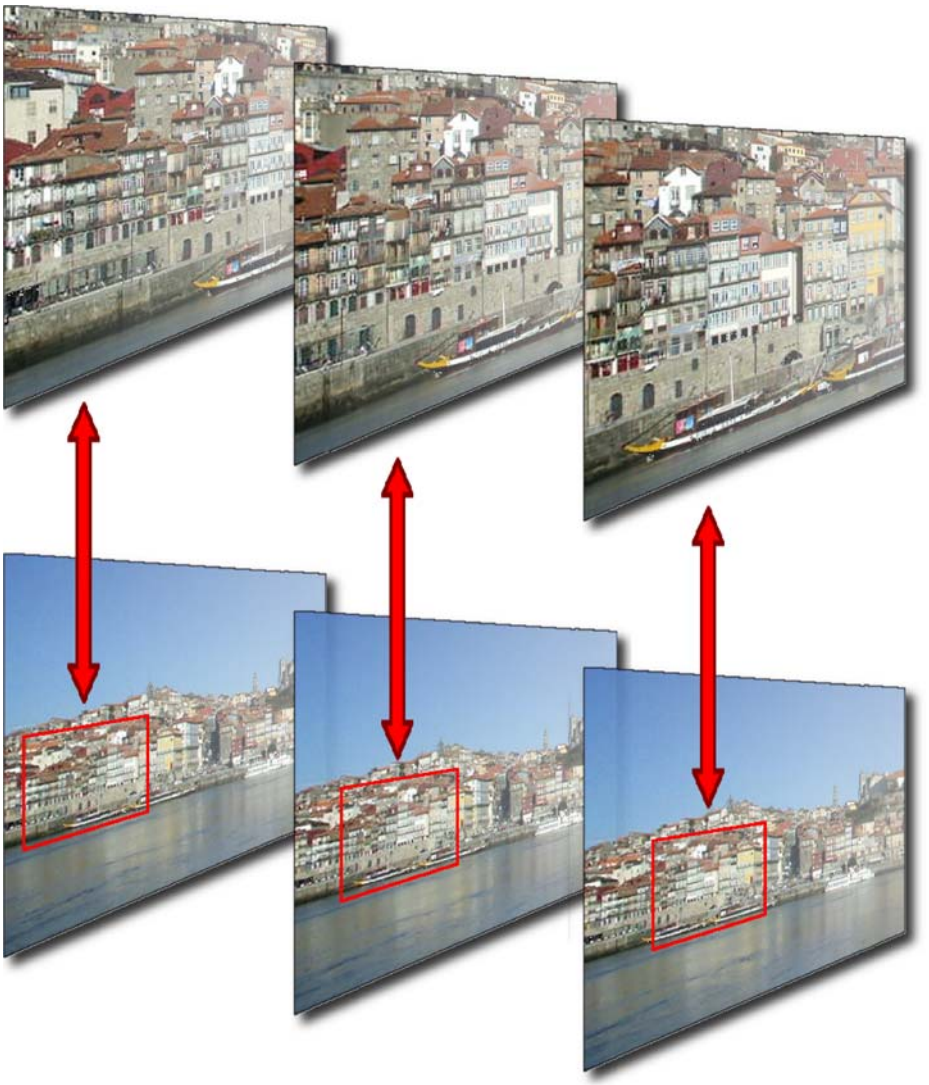

Fig. 5 Frame by frame correspondence between the 2 views as determined by the SMVR process (the *red quadrilaterals*)

successive refinement of side information performed in various iterations per each decoded frame. It's based on periodic Multi-View Registration (MVR) based on the previously described SIFT algorithm and Motion Compensated Extrapolation (MCE) for on-the-fly re-correlation of the 2 video sources at decoder, meaning that either for one method or the other there's always determined a window indicating the current correlation between the 2 views (as in Fig. 5).

Moreover, the new technique is codec-independent, that is, it relies only on the previous decoded frames and the side video source (from the reference view, see

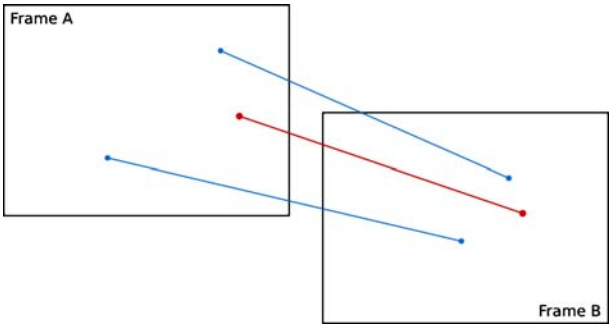

Fig. 6 Linear projection of one (*red*) point based on other two (*blue*) point projections

Fig. 1) for generating the proper side information for the current decoded frame, without any additional data from encoders or knowledge of the architecture.

We propose a simplified distributed system with two video cameras having complete-overlapped views as illustrated in Fig. 1. It is assumed a considerable distance and angle between cameras in order to independently capture partially different 3D details of the same scene. However, the proposed technique does not require any specific camera arrangement and its main goal is to provide the optimal side information for every given scenario. A generic architecture for the used scenario is presented in Fig. 7.

Without reducing the generality of the technique, several considerations are assumed as the basis of the used scenario: one of the cameras performs conventional encoding (e.g., H.26x, MPEGx) of the captured frames, we call it the *reference camera*. The other, the *target camera*, performs Wyner-Ziv (WZ) encoding of its perceived view, except for the first frame that is conventionally encoded. As requirement of the presented technique, it is considered that the target view (TV) will be always totally overlapped over (seen as included into) the reference view (RV) (at different scale, angle or rotation though), the latter acting as the "scene view" (see Fig. 1). Nevertheless, there can be employed more than one TV (every TV being completely overlapped over the RV) and a single joint decoder for all the encoders (RV + TVs).

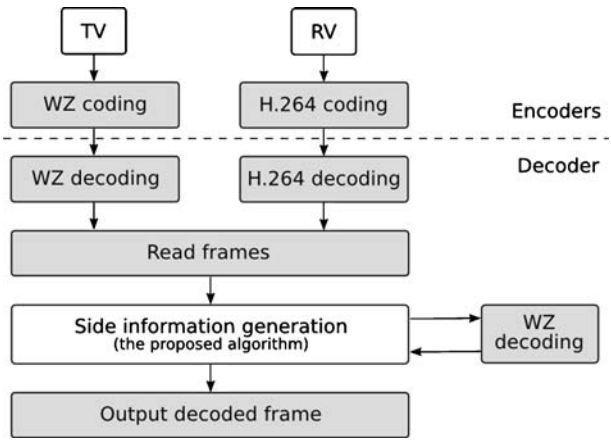

Fig. 7 Generic architecture of the multi-view codec embedding the proposed algorithm

At a glance, the algorithm of the proposed technique performs several iterations for each decoded frame (see Fig. 8). On each iteration it generates a temporary side information and one temporary image decoding based on it, until the algorithm decides that the generated side information is accurate enough and represents the best estimate for the current frame captured from the target camera, and therefore the final decoded image is supposed to have the finest quality. For that reason it is considered the current decoded frame. On each iteration it is used the same WZ bitstream, as provided by the target camera, for decoding the current frame. Also, the side information gets better estimated with each iteration and it is concluded one more step towards the optimal side information for that respective frame.

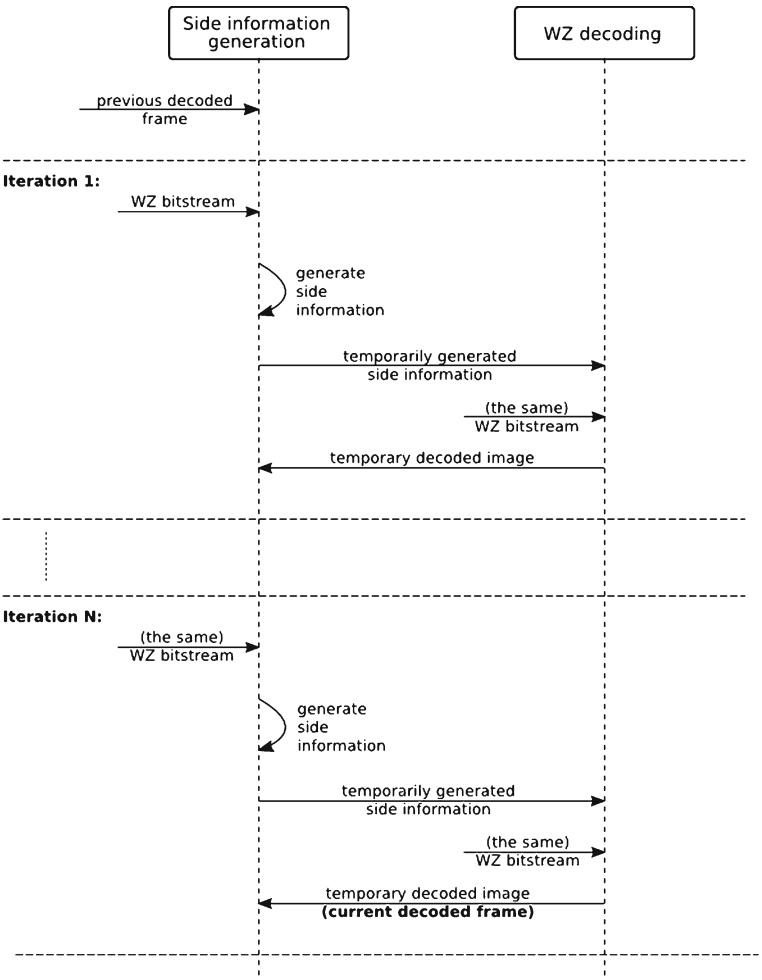

Fig. 8 Iteration diagram: the sequence of steps performed on each iteration in order to decode the current frame

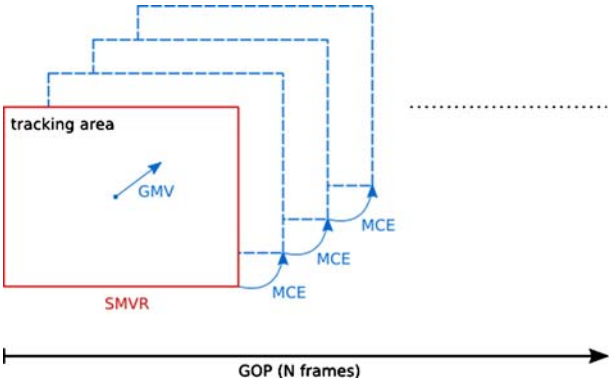

Fig. 9 GOP representation: the SMVR method is re-applied periodically (e.g. from 10 in 10 frames), the MCE is applied in between. The GMV (Global Motion Vector) is the MCE's generated motion vector between successive frames

The SIFT-based multi-view registration (SMVR) of the two views is an essential part of generating the side information in this multi-view architecture. As mentioned in Section 2, its goal is to determine, the best it can, which portion of the reference frame corresponds to the entire view captured from the target camera, as illustrated in Fig. 4. The SMVR is initially applied on the first pair of received frames that are conventionally decoded and then, for the subsequent frames is applied instead a motion compensated extrapolation (MCE) for tracking that respective area according to the target camera movement relative to the reference camera (see Fig. 9). Periodically (e.g., from 10 in 10 frames), the SMVR is re-applied for maintaining an updated fine-tuned correlation between the two cameras (Fig. 5). See Section 5 for more information on the benefits and drawbacks of each of the two methods.

For either correlation method, SMVR or MCE, a new window is generated being associated (relative) to the reference view (RV). Furthermore, bicubic interpolation is applied on that selected area for generating an estimation of the target view (TV), that is, an image with the same dimensions as the TV. This is what we call the generated (estimated) side information based on a window (pose), as illustrated in Fig. 10. Furthermore, the RV (see Fig. 1) is considered the "side" view since its main purpose is to serve for side information generation at decoder.

Fig. 10 Example of generated side information based on a window

In Figs. 11 and 12 are illustrated general flowcharts depicting the proposed algorithm for on-the-fly re-correlation of the 2 video sources. Figure 11 shows the initial phase that processes the first frames (conventionally decoded), one from each camera, it performs a SMVR method and achieves the first window that serves as reference for next frame to decode. Additionally, the Global Motion Vector (GMV) is set to null and also the retrieved conventionally decoded frame from the target camera is considered the first decoded frame. Then, Fig. 12 presents the core of the proposed algorithm: the successive side information generation and iterative decoding of each frame.

Due to the generality of the presented flowchart, there are several aspects of the algorithm that need to be detailed. First, both SMVR and MCE are used to estimate the pose of one view into the another. At this stage of development only the SMVR method is capable to fail the estimation, as indicated in Fig. 12 (not to be confused with the high success rate of the original SIFT algorithm). In this case either the MCE method is called instead or a new conventional encoded frame is requested from the target camera (e.g., I frame) and the decoding process is resumed. The latter case occurs only for the first decoded frame when SMVR fails and as a consequence there's no previous valid window serving as reference for MCE.

Regarding the periodicity of the applied SMVR, as stated previously, the algorithm tempts to re-apply this method after a predefined number of frames (e.g., from 10 to 10 frames). There are cases however where the SMVR cannot provide a valid window. In such situations, the algorithm simply performs yet another MCE instead and assigns the SMVR method the highest priority on the next iterations.

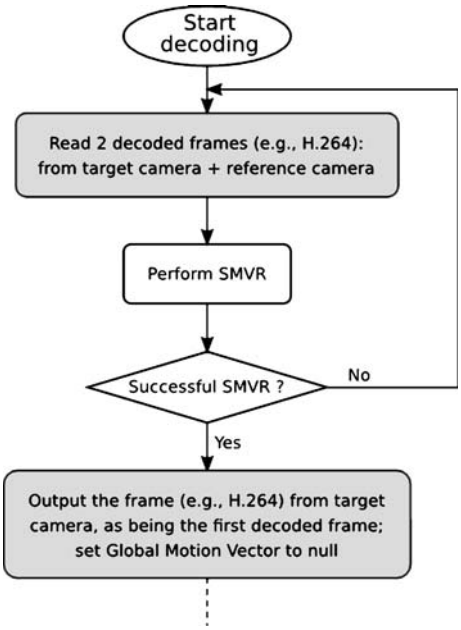

Fig. 11 Pseudocode algorithm for the first side information generation. A SMVR method is performed using the 2 conventionally decoded frames (e.g., H.264). The *grey blocks* are codec-specific

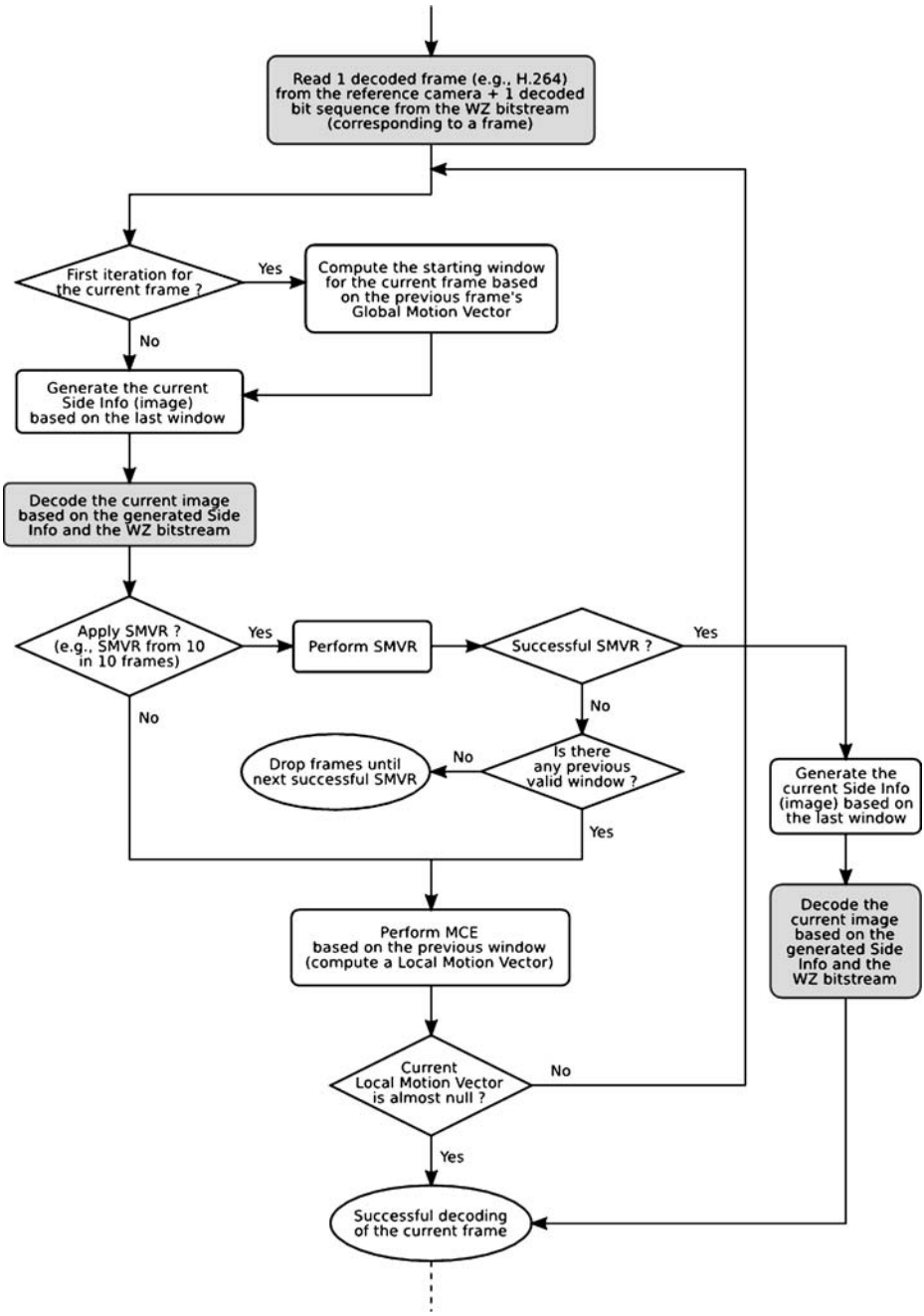

Fig. 12 Pseudocode algorithm for the iterative decoding of one single frame (the core of side information generation at decoder), based on SMVR and MCE. The *grey blocks* are codec-specific. The LMV (Local Motion Vector) is the MCE's generated motion vector between successive iterations

As depicted in the flowchart (Fig. 12), the iterations are successively performed while the MCE is still used and the Local Motion Vector (LMV) is not null (is above a predefined threshold, e.g., 0.1 pixels). The following are the steps for decoding one frame:

1) estimate the new location of the window for decoding the current frame based on the GMV from the last frame
2) generate the side information by bicubic interpolation based on the estimated window and the frame from the reference camera
3) perform WZ decoding of a new temporary image based on the generated side information
4) apply block-based Motion-Estimation (ME) based on the newly decoded image and the previously estimated window and determine the location difference, that is, a general motion vector that expresses how far the window has moved from one iteration to another; we call this a Local Motion Vector (LMV)
5) repeat the steps from 1 to 4 until the LMV gets almost null (below a predefined threshold, e.g., 0.1 pixels)
6) output the last decoded image (from last iteration) as being the decoded frame, store the Global Motion Vector for the current frame and move on to the next frame to decode.

See Fig. 13 for a more illustrative description of the performed steps from one frame to another.

On each iteration, the WZ bitstream used for decoding the current image remains unaltered. Also, for each ME process a LMV is calculated and it indicates the new position of the window relative to the one from the last iteration. Additionally, a GMV is continuously updated by adding all the LMVs from all iterations for the current decoded frame. When decoding the next frame, the GMV is used to initially estimate the starting window for that frame (the algorithm initially assumes that the camera movement is uniform). Concluding, the GMV dictates the general camera motion from one frame to another while the LMV indicates the refinement of the side information after each iteration.

When a LMV gets almost null it indicates that the window location has stabilised and it can produce the best side information for the current frame. Basically, in the 1^{st} iteration the side information is initially predicted based on the global camera movement between the 2 previous frames, then, starting with the 2^{nd} frame the side

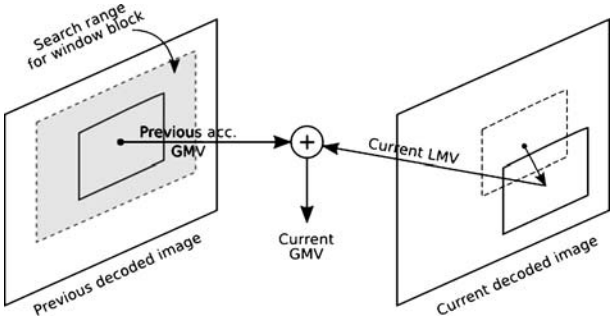

Fig. 13 Motion Estimation process on each iteration. A predefined search range is considered (e.g., 32×32 pixels). A Local Motion Vector (*LMV*) is determined and the previous accumulated Global Motion Vector (*GMV*) is then updated

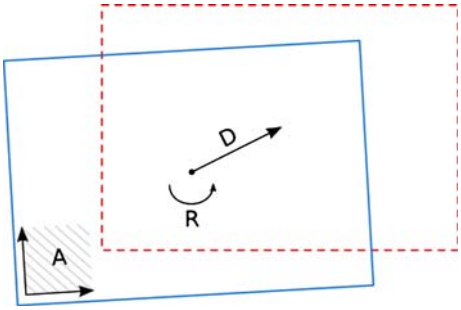

Fig. 14 The 3 metric criteria (window area - *A*, distance - *D* and rotation - *R*) for comparing the current generated window (*blue rectangle*) and the ground-truth window (*dashed red rectangle*)

information is continuously refined using the presented technique. According with the experiments, for most of the cases the algorithm uses 2 iterations in order to decode a frame. For unpredictable camera movement it may require more iterations (see Fig. 16). However, the iteration count is limited to a maximum of 3 as being sufficient in order to produce highly accurate side information.

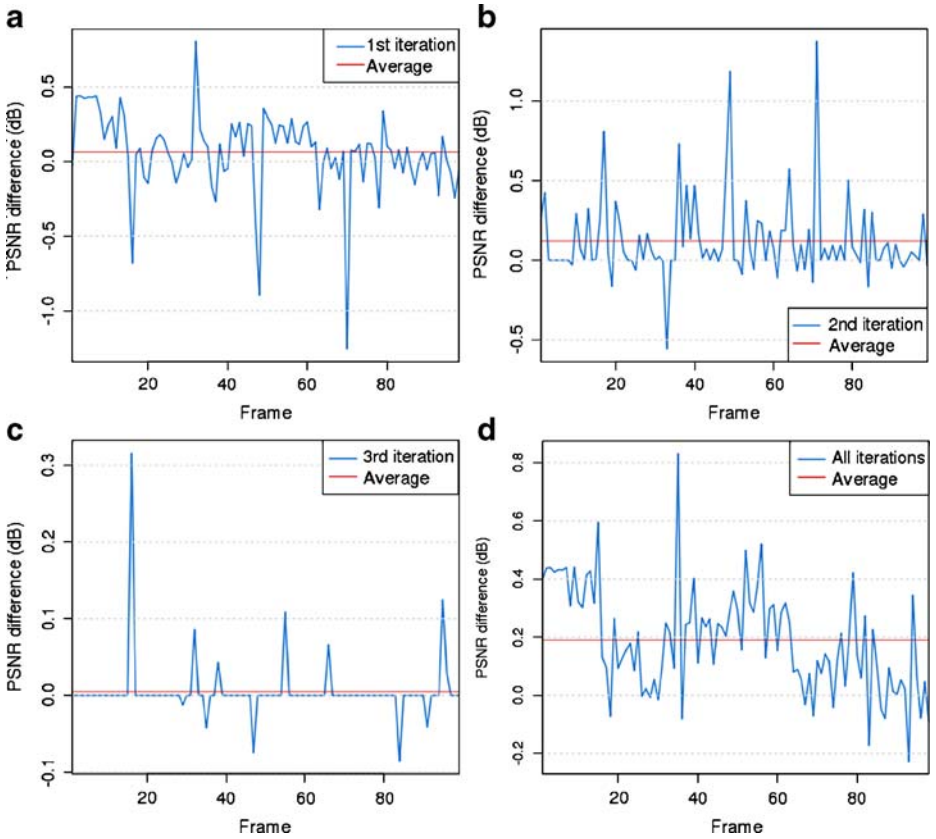

Fig. 15 Iteration-based performance evaluation of the generated side information for predictable target camera movement (**a**–**d**)

4 Proposed metric for multi-view correlation

Accurate multi-view correlation (MVC), obtained either by SMVR or MCE, plays a crucial role in the correct generation of side information at decoder and implicitly contributes to a high quality decoding (rate-distortion). However, due to the codec-independent nature of the proposed technique we are motivated to find additional evaluation methods that avoid codec-specific details/performance. Consequently, we focused on the problem of comparing the achieved results on MVC of two overlapped views (a set of windows, one for each pair of correspondent frames as illustrated in Fig. 5) with a ground-truth (another set of windows, manually indicated).

The metric we are proposing in this paper is based on 3 criteria (measurements), each one measures the difference between 2 compared windows with respect to one particular geometric feature: *area*, *location* and *general relative rotation angle* (as illustrated in Fig. 14 for the particular case of a rectangle-shaped window).

The measurement of the difference for each of the 3 features is then normalised taking into account the maximum achievable value for that respective difference. The

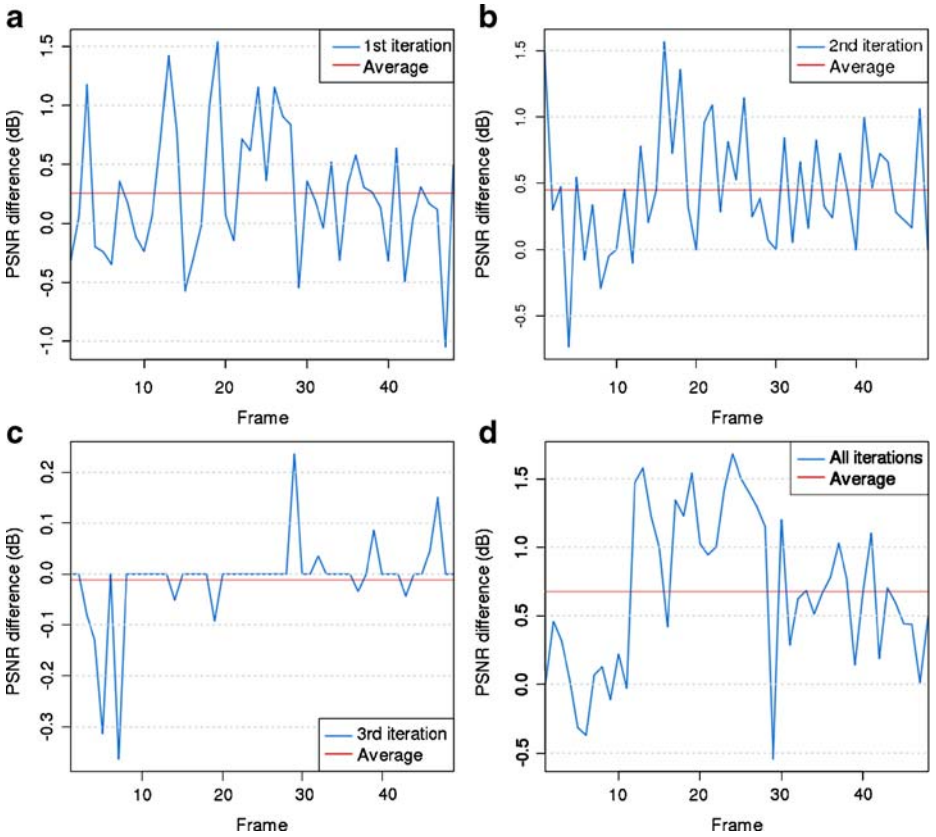

Fig. 16 Iteration-based performance evaluation of the generated side information for chaotic target camera movement (**a**–**d**)

goal is to attribute to each criterion (feature) an equal weight when computing the final value as the Mean Squared Error (MSE) of the 3 values (area difference - AD, location difference - LD and rotation difference - RD) as follows:

$$MSE_n = \frac{AD_n^2 + LD_n^2 + RD_n^2}{3} \tag{1}$$

the normalised values are indicated as "n".

Finally, the metric's result (M) is computed as an overall arithmetic mean calculated over the entire video sequence (N frames), based on each such local difference (MSE_n) between the current generated window and its corresponding ground-truth window, as a general evaluation of the entire multi-view correlation (MVC) process. The closer it gets to zero the better is the performance of the MVC.

$$M = \frac{\sum_{i=1}^{N} MSE_n(i)}{N} \tag{2}$$

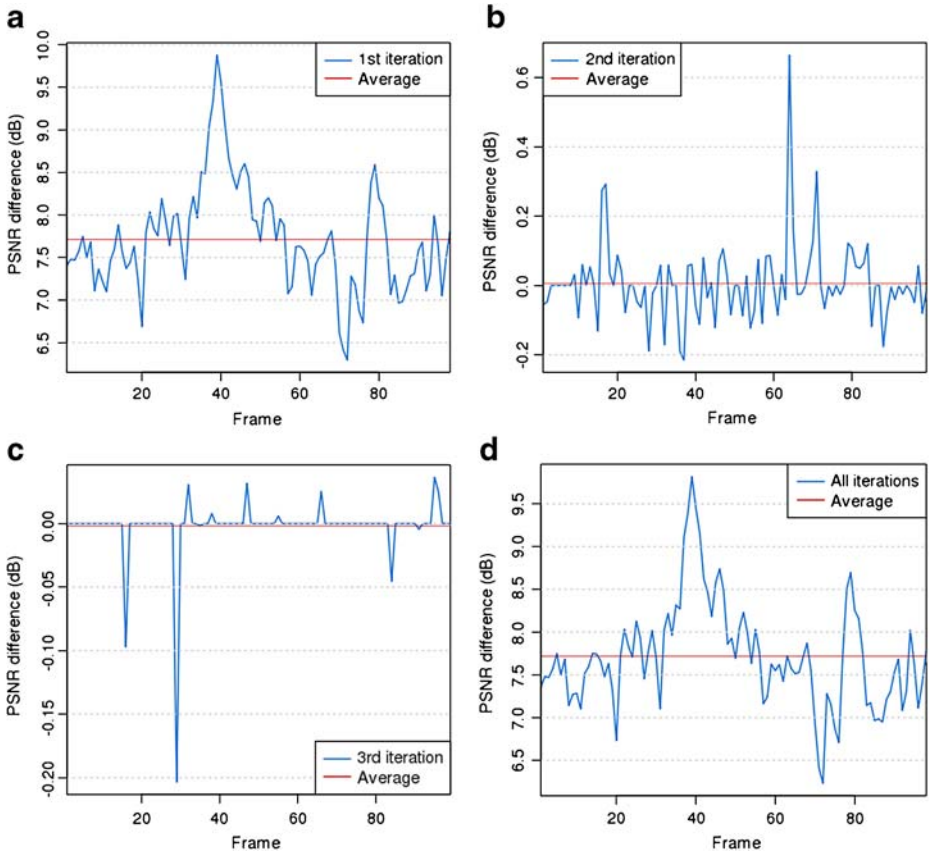

Fig. 17 Iteration-based performance evaluation of the decoded frames for predictable target camera movement (**a**–**d**)

Practical results are provided in Section 5 illustrating the sensitivity and objectivity of the proposed metric to window disparity (see Fig. 23).

5 Results

In this section we present the preliminary results of the side information generation performance focused on the decoding iterations, first for a predictable (smooth) target camera movement (see Fig. 15) using the "Porto" raw sequence (320 × 240, 100 frames, 5 fps) and then for a chaotic target camera movement (see Fig. 16) using the additional "Porto2" raw sequence (320 × 240, 50 frames, 15 fps). First it is determined the PSNR value between the decoder's internally generated side information frame at the end of each iteration and the original corresponding frame captured from the target camera. Then, Figs. 15 and 16 show only the differences between each two successive PSNR values, seen as an objective assessment of the enhancement of the side information from one iteration to another. Note the positive average values (well above zero) in the first graphs of Figs. 15 and 16.

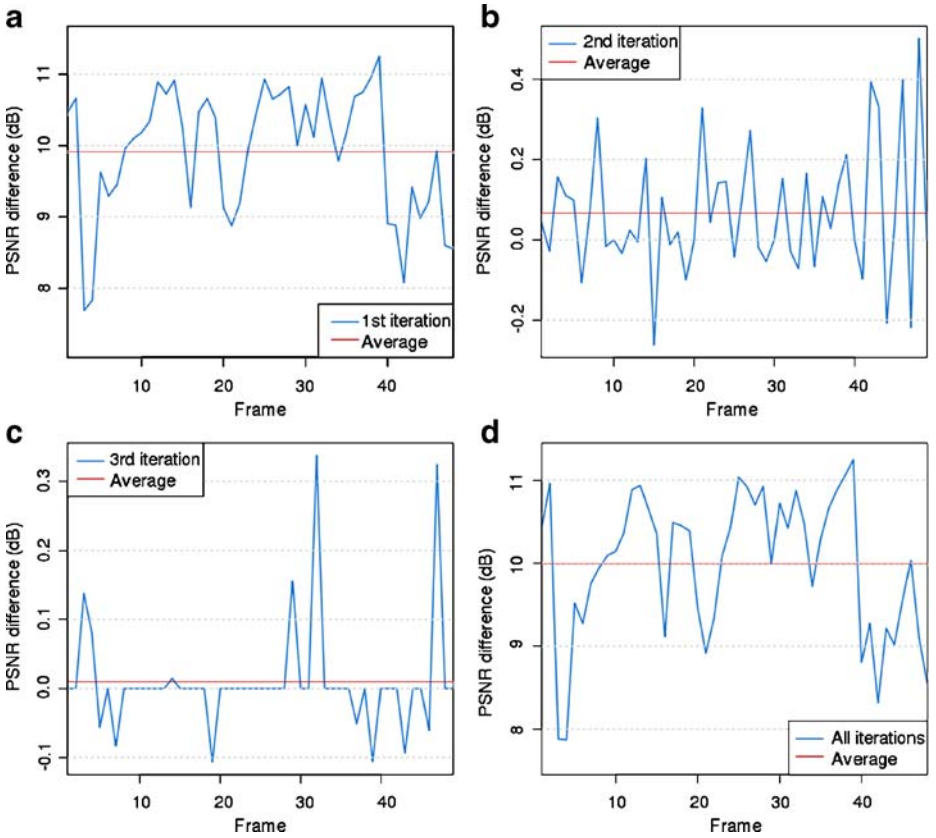

Fig. 18 Iteration-based performance evaluation of the decoded frames for chaotic target camera movement (**a**–**d**)

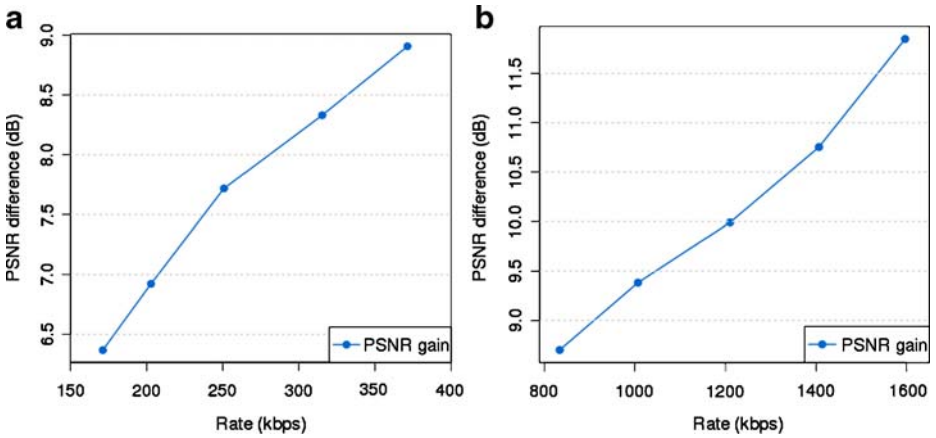

Fig. 19 PSNR gain in rate-distortion performance for both sequences, "Porto" raw sequence (320 × 240, 100 frames, 5 fps) and "Porto2" raw sequence (320 × 240, 50 frames, 15 fps). The middle point from each graph is the overall PSNR gain corresponding to the above results (**a**, **b**)

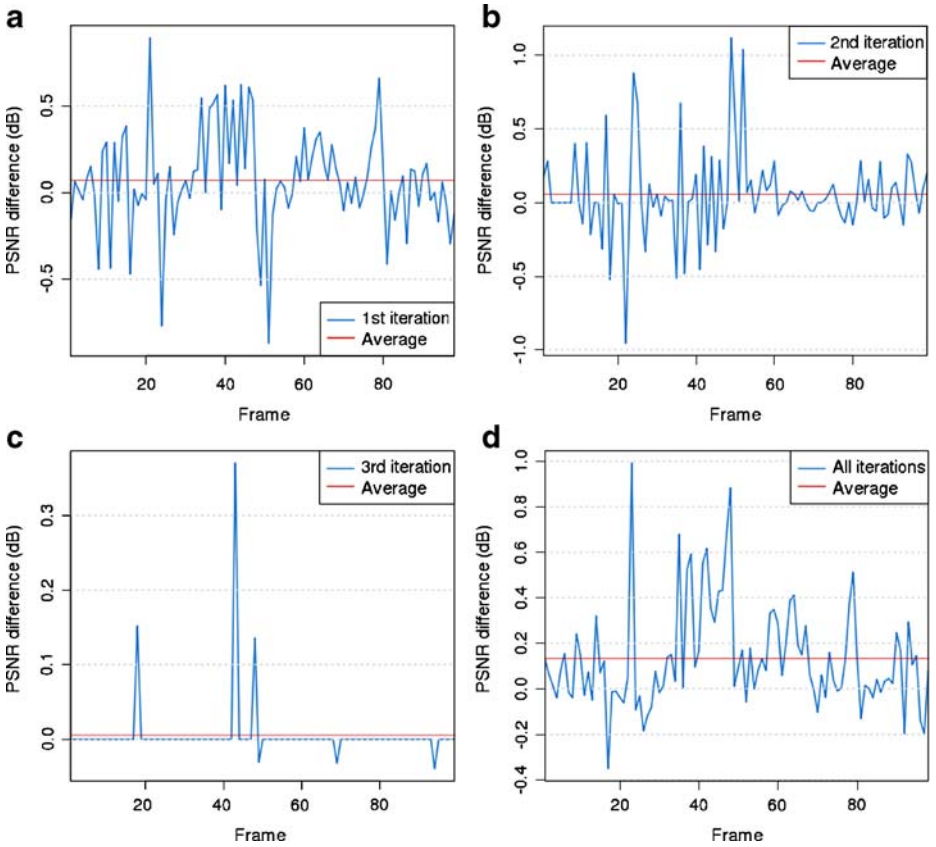

Fig. 20 Iteration-based performance evaluation of the generated side information for predictable target camera movement and high compression rate of the reference frames (99.09 kbps) (**a**–**d**)

As noticed in both cases (Figs. 15 and 16) the best performance is achieved in the second iteration (the side information is roughly estimated). As mentioned in Section 3, in the 1^{st} iteration is used the GMV determined in the previous decoded frame to estimate the side information, and in the remaining iterations the proposed technique is applied to effectively refine the side information, thus achieving the best performance in the second iteration (gaining approximately 0.12 dB in the first case and 0.45 dB in the second). As expected, in the 3^{rd} iteration the performance is significantly reduced corresponding to a yet a tiny refinement. Consequently, the best overall results were achieved in the latter case (chaotic movement): approximately 0.67 dB as shown in Fig. 16d.

The results presented in Figs. 15 and 16 were achieved for compressed reference sequences at 1021.23 kbps and 1478.71 kbps, respectively. Additionally, in Figs. 17, 18 and 19 are presented the overall results for both cases, provided only for reference as they are codec-specific. They were achieved by using an experimental multi-view Wyner-Ziv codec developed by the authors. It is now in an early stage of development and the codec's performance has been beyond the scope of the work presented in this paper.

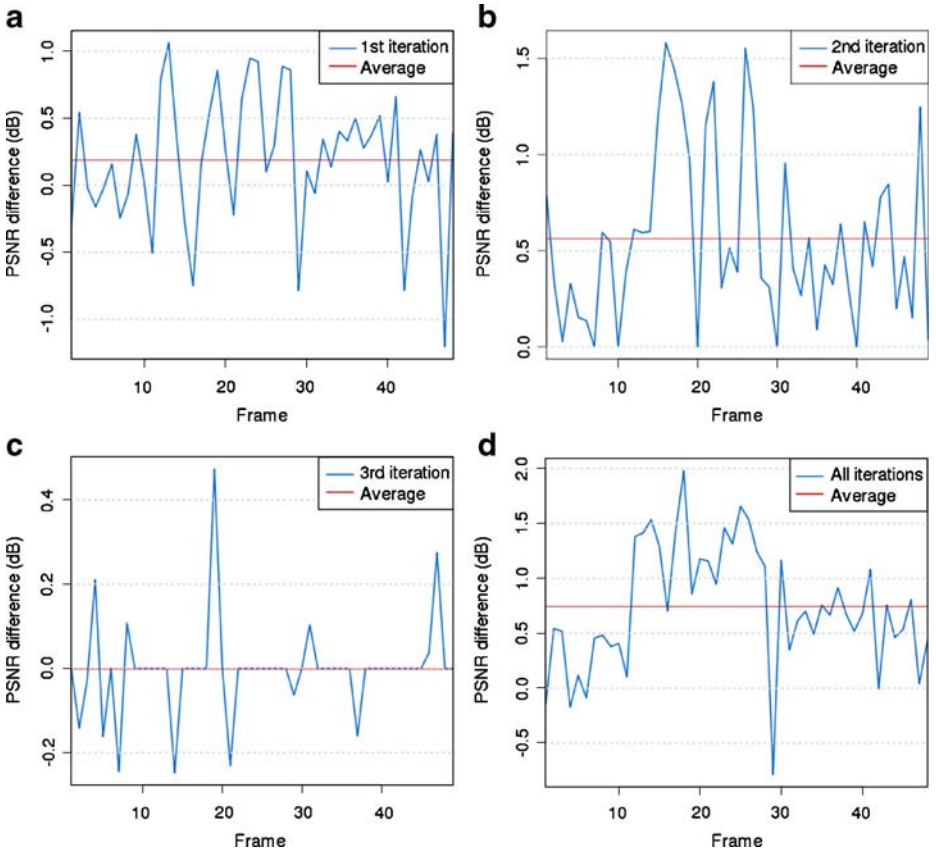

Fig. 21 Iteration-based performance evaluation of the generated side information for chaotic target camera movement and high compression rate of the reference frames (100.73 kbps) (**a**–**d**)

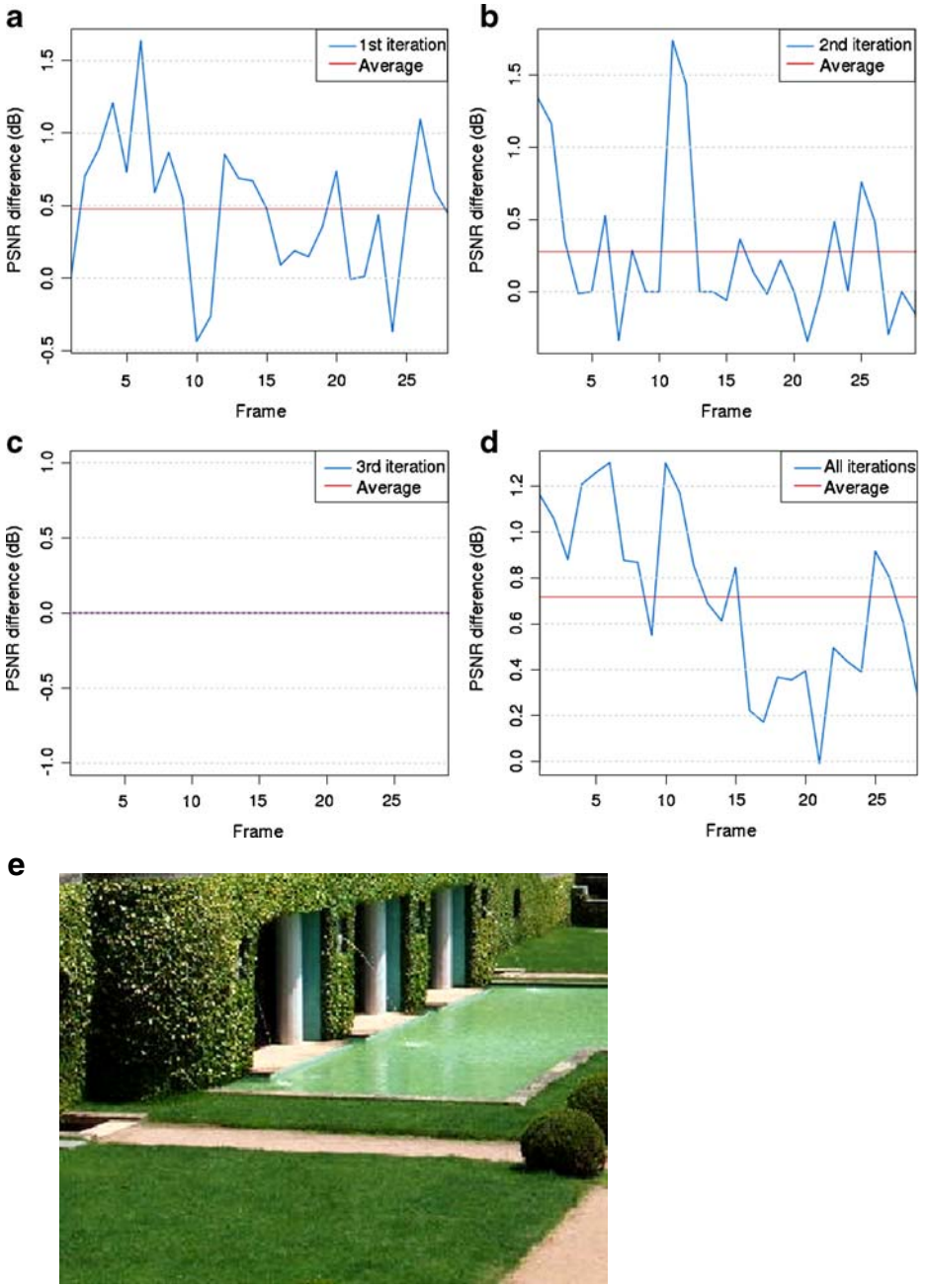

Fig. 22 Iteration-based performance evaluation of the generated side information for additional video sequence ("Garden") (**a**–**d**)

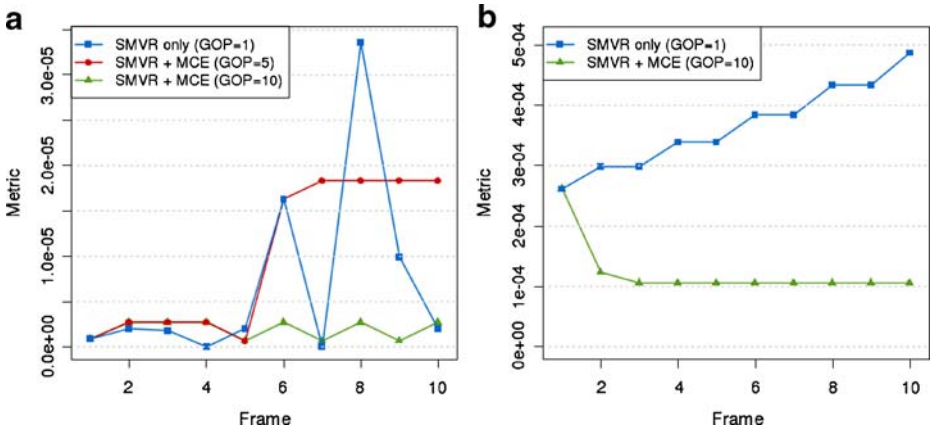

Fig. 23 SMVR vs. MCE-based MVC (**a**, **b**)

More results on both cases are presented in Figs. 20 and 21 showing the performance evaluation of side information generation for high compression rate of reference frames (99.09 kbps and 100.73 kbps, respectively). The overall PSNR gain is of approximately 0.13 dB on the first case (see Fig. 20d) and 0.74 dB on the second (see Fig. 21d). Concluding, the compression rate of reference frames has a limited influence on the overall results.

Additional performance results of side information generation are provided in Fig. 22 for other video sequence: "Garden", 320x240, 30 frames, 5 fps. It contains a predictable target camera movement. An example of target frame is also provided in Fig. 22e. As noticed, the fewer details wide spread over large areas make the overall algorithm more effective, i.e., there's a clear improvement from one iteration to another, thus the 3^{rd} iteration never occurred (see Fig. 22c).

Next are presented the performance results of the fusion technique for multi-view correlation (MVC) by using the proposed metric. We used for comparison our

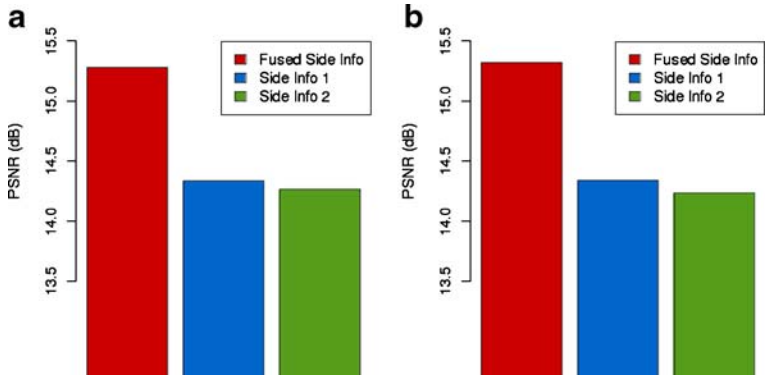

Fig. 24 Evaluation of the fusion between two individual side informations (the overall rate-distortion between side information frames and the original target frames), using the "Porto" raw sequence encoded at 508 kbps (**a**) and 797 kbps (**b**) respectively

manually indicated ground-truth (the lower are the presented values the closer are to ground-truth). There was taken into consideration two situations as being the most relevant for the proposed technique: for high and low success rate of SMVR on providing a valid window. Is was used 2 excerpts from the "Porto" raw sequence (10 frames each).

High success rate of SMVR: in the example depicted in Fig. 23a (a sequence of 10 frames), the *SMVR-only* (GOP=1) method has the most unpredictable behaviour, it provides both the best estimates and the worst. Starting with frame 6 there's a gap caused by the less accurate SMVR which is further on carried by MCE (*SMVR + MCE* method for GOP=5) until the next SMVR (scheduled at frame 11). On the other hand, MCE-based MVC brings the most constant results (e.g., for GOP=10). It does depend however on the periodic SMVR.

Low success rate of SMVR: in the scenario chosen in Fig. 23b the SMVR successfully provides a valid window only for the first frame and it fails in the rest. Therefore, it is shown a continuous increasing gap between the *SMVR-only* (GOP=1) method and the ground-truth due to the continuous failure to estimate a new window (in this case the SMVR always uses instead the last found valid window). On the other hand, the ME-based MVC is reducing the gap starting with frame 2, basically approaching the ground-truth.

In Fig. 24, additional simulation results are provided in order to emphasize the scalability of the proposed technique, in this case using two reference cameras and performing a fusion of the generated side informations on each iteration. This brings a significant improvement of the resulted side information to be used in the decoding process, rather than using each side information separately. The fusion is achieved by selecting the corresponding blocks, from each side information, that approximate better the ones from the decoded frame from the previous iteration. In this scenario the decoder performs as many iterations as necessary for achieving the best individual side information for each reference view (but no more than 3).

Note that each individual PSNR value is significantly influenced by the distance of the reference camera from the scene and projection angle. Figure 24 serves mainly for outlining the significantly improved quality of the fused side information when compared with each individual one.

6 Conclusions

In this paper we described an iterative method for successive refinement of the side information at decoder based on multi-view registration and motion compensated extrapolation. It aims to provide a general, almost constraint-free and yet robust solution for flexible multi-view environments where it can't be provided any extra-information from encoders or details about camera positions, angles, distances, etc.

The proposed technique can be furthermore expanded to scenarios with partial or no overlapping between views by using a feedback channel for this purpose. It indicates the encoder which portions (e.g., block by block) of the its perceived view to be Wyner-Ziv encoded (the overlapped ones), the remaining being coded conventionally (e.g., H.26x, MPEGx) at a low rate.

Due to experiments conducted in various scenes, we conclude that one reason to use SMVR would be in cases where there's a high amount of scene details thus

providing many key-points and implicitly an accurate estimation of the correlation window. In turn, the SMVR may have very poor success rate in certain circumstances: either due to a poor quality of the image(s), many repeated almost identical features which generate mismatches, significant difference between the two images that generates very few or no matches, or simply less details that generate few key-points.

Regarding the second method, the MCE-based multi-view correlation (MVC), it proves to be highly reliable. There's always a best estimate for a window that takes into account the match of the whole window-block, being very less sensitive to isolated mismatches. Nevertheless, it requires some guidance in the sense that a previous window has to be initially indicated in order to have a starting point for the window-block search.

At a glance it seems that each method acts as a complementary solution for the other's weakness. For this reason it was chosen a fusion of the two correlation methods as the best compromise.

As expected, the iterative nature of the algorithm clearly introduces some additional complexity to decoder. Nevertheless, the important achievement here is that there's no need for extra-processing on the encoder's side in order to achieve the presented results, thus maintaining the encoders as low-complexity as possible (as in the spirit of Distributed Video Coding) and also makes it codec-independent.

Further work towards a decision-taking enabled decoder for choosing on-the-fly between one method or the other depending on the circumstances seems very promising and can contribute to a more robust and accurate generation of side information. Also, the decision-taking could impose an automatically calculated, variable periodicity (as number of frames) for re-applying the SMVR depending on the global motion contained in the scene since the faster is the scene movement the more likely is for motion estimation to fail.

There are cases where both of the 2 proposed correlation methods are likely to fail due to scenes with less or almost no detail to be matched (e.g., river water, sky, empty walls, etc.), as illustrated in Fig. 3b—the two outliers. More methods and experiments will be studied for this purpose. More sophisticated filters can be used for the better elimination of the outliers produced by the SIFT algorithm, thus improving the overall performance of the MVC (e.g. using RANSAC followed by a probabilistic verification [10]).

Acknowledgement The first author acknowledges the *Fundação para a Ciência e a Tecnologia, Portugal*, for the financial support.

References

1. Aaron A, Girod B (2004) Wyner-ziv video coding with low encoder complexity. In: Proceedings international picture coding symposium, PCS'04, San Francisco
2. Aaron A, Rane S, Girod B (2004) Wyner-ziv video coding with hash-based motion-compensation at the receiver. In: Proceedings IEEE intl. conference on image processing, Singapore
3. Althof R, Wind M, Dobbins IJT (1997) A rapid and automatic image registration algorithm with subpixel accuracy. IEEE Trans Med Imag 16(3):308–316
4. Artigas X, Angeli E, Torres L (2006) Side information generation for multiview distributed video coding using a fusion approach. In: 7^{th} nordic signal processing symposium, NORSIG'06, Reykjavik, Iceland, 7–9 June 2006
5. Ascenso J, Pereira F (2007) Adaptive hash-based side information exploitation for efficient Wyner-Ziv video coding. In: International conference on image processing (ICIP)—2007, San Antonio

6. Benedek C, Havasi L, Sziranyi T, Szlavik Z (2005) Motion-based flexible camera registration. Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on, pp 439–444
7. Bergevin R, Soucy M, Gagnon H, Laurendeau D (1996) Towards a general multi-view registration technique. IEEE Trans Pattern Anal Mach Intell 18(50):540–547
8. Brivio PA, Della Ventura A, Rampini A, Schettini R (1992) Automatic selection of control-points from shadows structures. Int J Remote Sens 13(10):1853–1860
9. Brown LG (1992) A survey of image registration techniques. ACM Comput Surv 24(4):325–376
10. Brown M, Lowe DG (2003) Recognising panoramas. In: Proceedings IEEE international conference on computer vision, Nice, 13–16 October 2003
11. Brown M, Lowe DG (2007) Automatic panoramic image stitching using invariant features. Int J Comput Vis 74:59–73
12. Cheng S, Xiong Z (2005) Successive refinement for the Wyner-Ziv problem and Layered Code Design. IEEE Trans Signal Process 53:8
13. Dufaux F, Ouaret M, Ebrahimi T (2007) Recent advances in multiview distributed video coding. In: SPIE defense and security symposium (DSS 2007), Orlando, 9–13 April 2007
14. Fonseca LMG, Costa MHM (1997) Automatic registration of satellite images. In: Brazilian symposium on computer graphics and image processing, 10, Campos de Jordão, pp 219–226
15. Forssén P-E, Lowe DG (2007) Shape descriptors for maximally stable extremal regions. In: International conference on computer vision (ICCV), Rio de Janeiro
16. Goncalves H, Goncalves JA, Corte-Real L (2008) Automatic image registration based on correlation and hough transform. In: Bruzzone L, Notarnicola C, Posa F (eds) SPIE, vol 7109, no 1, p 71090J. http://link.aip.org/link/?PSI/7109/71090J/1
17. Gordon I, Lowe DG (2006) What and where: 3D object recognition with accurate pose. Toward Category-Level Object Recognition, pp 67–82
18. Growe S, Tönjes R (1997) A knowledge based approach to automatic image registration. In: International conference on image processing 97, pp 26–29
19. Guo X, Lu Y, Wu F, Gao W, Li S (2006) Distributed multi-view video coding. In: Proceedings of SPIE-IS&T electronic imaging, SPIE, vol 6077, San Jose, 15–19 January 2006
20. Izquierdo E (2003) Efficient and accurate image based camera registration. IEEE Trans Multimedia 5(3):293–302
21. Le Moigne J, Xia W, Chalermwat P, El-Ghazawi T, Mareboyana M, Netanyahu N, Tilton J, Campbell W, Cromp R (1998) First evaluation of automatic image registration methods. In: Geoscience and remote sensing symposium proceedings, 1998. IGARSS '98, vol 1. IEEE International, Piscataway, pp 315–317
22. Likar B, Pernus F (1999) Automatic extraction of corresponding points for the registration of medical images. Med Phys 26:1678–1686
23. Lowe D. Scale-Invariant Feature Transform (SIFT): matching with local invariant features. http://www.cs.ubc.ca/spider/lowe/research.html, http://www.cs.ubc.ca/~lowe/keypoints/
24. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
25. Ouaret M, Dufaux F, Ebrahimi T (2006) Fusion-based multiview distributed video coding. In: ACM international workshop on video surveillance and sensor networks, Santa Barbara, 27 October 2006
26. Ouaret M, Dufaux F, Ebrahimi T (2007) Multiview distributed video coding with encoder driven fusion. In: European conference on signal processing (EUSIPCO), Poznan, 3–7 September 2007
27. Puri R, Ramchandran K (2003) PRISM: a "reversed" multimedia coding paradigm. In: Proceedings international conference on image processing (ICIP), Barcelona
28. Slepian D, Wolf JK (1973) Noiseless coding of correlated information sources. IEEE Trans Inf Theory 19:471–480
29. Szlavik Z, Sziranyi T, Havasi L (2007) Video camera registration using accumulated co-motion maps. 61(5):298–306
30. Tipdecho T (2002) Automatic image registration between image and object spaces. In: Proceedings of the open source GIS—GRASS users conference 2002, Trento, 11–13 September 2002
31. Wyner D, Ziv J (1976) The rate-distortion function for source coding with side information at the decoder. IEEE Trans Inf Theory 22:1–10
32. Yeo C, Ramchandran K (2007) Robust distributed multi-view video compression for wireless camera networks. In: VCIP 2007, San Jose, 28 January–1 February 2007
33. Zitova B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21(11):977–1000. doi:10.1016/S0262-8856(03)00137-9

Lucian Ciobanu was born in Iasi, Romania, in 1978. He graduated in Software Engineering from the Faculty of Automatic Control and Computer Engineering, "Gh. Asachi" Technical University of Iasi, Romania. He is researcher at the Institute for Systems and Computer Engineering (INESC) Porto, Portugal, since 2002 and also PhD student of the Faculdade de Engenharia, Universidade do Porto, Portugal, since 2005. His research interests include image/video coding and processing.

Luís Côrte-Real was born in Vila do Conde, Portugal, in 1958. He graduated in Electrical Engineering from the Faculdade de Engenharia, Universidade do Porto, Portugal, in 1981. He received the M.Sc. degree in Electrical and Computers Engineering in 1986 from Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal and the Ph.D degree from the Faculdade de Engenharia, Universidade do Porto, in 1994. In 1984 he joined Universidade do Porto as a lecturer of telecommunications. He is currently associate professor at the Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto. He is researcher at the Institute for Systems and Computer Engineering (INESC) Porto, Portugal since 1985. His research interests include image/video coding and processing.