A sensor aided H.264 encoder tested on aerial imagery for SFM

A SENSOR AIDED H.264 ENCODER TESTED ON AERIAL IMAGERY FOR SFM

C.V. Angelino, L. Cicala, G. Persechino

CIRA, the Italian Aerospace Research Center
Payload Sensors & HMI
Capua, Italy

E. Baccaglini, M. Gavelli, N. Raimondo

Istituto Superiore Mario Boella
Multi-Layer Wireless Solutions
Torino, Italy

ABSTRACT

Standard video coding systems currently employed in UAV (Unmanned Aerial Vehicle) and aerial drone applications do not rely on some peculiarities in terms of scene 3D model and correlation among successive frames. In particular, the observed scene is static, i.e. the camera movement is dominant, and it can often be well approximated with a plane. Moreover, camera position and orientation can be obtained from the navigation system. Therefore, correspondent points on two video frames are linked by a simple homography. This paper presents novel results obtained by a low-complexity sensor aided H.264 encoder, recently developed at CIRA and yet tested on simulated data. The proposed encoder employs a new motion estimation scheme which make use of the global motion information provided by the onboard navigation system. The homography is used in order to initialize the block matching algorithm allowing a more robust motion estimation and a smaller search window, and hence reducing the complexity. The tests are made coding real aerial imagery, captured to be used for 3D scene reconstruction. The images are acquired by an high resolution camera mounted on a small drone, flying at low altitude.

Index Terms— motion imagery coding, h.264, UAV, aerial drones, SFM

1. INTRODUCTION

Sensor aided video coding for UAV and aerial drone applications gained attention in the last few years. Indeed the use of external sensors is suitable for low complexity global motion estimation. Rodriguez et.al [1] use the available global motion information to simplify block ME in a MPEG-4 encoder. Although their approach reduces the complexity of a standard video encoder, transmitting the global motion information instead of the motion vectors derived from it might be more efficient. Video compression schemes suited for UAV applications that exploit the available global motion information have also been proposed. Gong et.al [2] use a homography to model the global motion, merge the first intra frame and subsequent inter frame residues in a frame group into a single big image, and code it using JPEG2000. M. Bhaskaranand and J. D. Gibson [3] proposed a low-complexity encoder whose distinctive attributes are no block-level motion estimation, global motion compensated prediction with global motion parameters input from the camera mount system, and spectral entropy-based coefficient selection and bit allocation. Comparing the performance of the proposed encoder with that of a H.264, they showed that for videos typical of UAV reconnaissance, the proposed encoder achieves better R-D performance at lower bit rates and lower variation of quality across frames. They also demonstrated that the proposed encoder requires fewer memory accesses and computations.

In [4] GME is performed using external position and orientation sensors. The observed scene is supposed to be approximated with a plane. Based on this simple hypothesis it is possible to obtain a prediction of the global motion. If motion vector predictors and refinement stage (as described in [5]) can receive additional information from sensors about the motion flow, it could be of great benefit for the whole encoding stage. The proposed work allows to choice the prediction method (inter/intra) for each macroblock and initializing the motion with predictors derived from data delivered by UAV navigation sensors.

In [4] the focus was on typical Medium Alititude Long Endurance UAV missions, in which the planar approximation of the overflight landscape fits well, and the novel encoder was validated on simulated videos. Experiment results showed better performances of the proposed encoder w.r.t. standard H.264 encoder.

In this work, instead, we present results of the same encoder in a low altitude scenario and on real aerial imagery. The low altitude scenario, in which the altitude relative to the ground is comparable with the depth of the 3D scene, represents the worst case for the planar approximation hypothesis. For this reason, the rate-distortion performance of the encoder, in such situation, can be considered the lowest limit of the performance envelopment. In our tests, the image acquisitions are made by a small aerial drone from different points of view, in order to have a picture set to make a 3D reconstruction of the overflight scene. In this scenario the available bandwidth (or storage memory) can be used for transmitting (or storing) a high definition video at a low framerate instead

of a standard definition full motion video. Moreover, it is important to keep low the encoding complexity in order to increase the battery life and hence the drone mission duration. Our experiments clearly show that, even if the planar approximation is unsatisfied, the proposed sensor aided encoder is less computation demanding than the reference h.264 encoder.

2. PROPOSED METHOD

The relationship between world coordinates P_c, where the subscript c stands for camera centered reference system, and homogeneous image coordinates p, expressed in pixels, is described by the linear equation

$$p \simeq CP_c, \tag{1}$$

where the symbol $\simeq$ means that the equality is up to a scale factor and C is the calibration matrix at time sample k. Let us consider the observed scene as a 2-D plane Π in 3-D space with normal n_t and distance h from the camera center O, where the subscript t refers to the tangent reference frame centered at O.

The equation linking two image points $p[k-1]$ and $p[k]$ which are the projection of the same point of Π in two successive video frames is:

$$p[k-1] = C[k-1] \cdot H[k-1,k] \cdot C^{-1}[k]p[k], \tag{2}$$

where the matrix H is the sum of two components H_R and H_T, related to the rotation and translation of the camera respectively, with the following expression [6, p.131]

$$H_R[k-1,k] = R_{t\to c[k-1]} \cdot R^T_{t\to c[k]}, \tag{3}$$

$$H_T[k-1,k] = \frac{1}{h} \cdot R_{t\to c[k-1]} \cdot (O_t[k-1] - O_t[k]) \cdot n_t^T \cdot R^T_{t\to c[k]}. \tag{4}$$

Here, $R_{t\to c}$ represents the direction cosine matrix for the reference change from the tangent (t) to camera (c) frame. The knowledge of the homography matrix from sensor data allows to initialize the motion field as explained below.

This model has been employed in the H.264 motion estimation system through its open source implementation x264 mainly modifuing two specific aspects:

- the decision on the frame type that has to be encoded (Intra, Predicted, Bi-Predicted) and the correspondent best frame reference;
- the initialization of the motion field to generate the predictors which will be adopted in the subsequent macroblock analysis.

The above coding steps are improved, with respect to the x264 implementation, considering the *metadata* (the position and the orientation of the flying platform, as delivered by the navigation sensors).

The module developed to perform these operations is called *initialization module*. The proposed changes allow to interface this module with the x264 library. The above tasks are performed by two main functions. The *frametype_decide* function is called to decide the type and the proper reference of the frame associated with the metadata available in an auxiliary buffer. The function receives as input two lists of metadata information: *aub_ref* and *aub*. In the first list we store metadata that can be used as a reference by those present in the *aub* list. The *aub* list contains metadata associated with the frame for which the decision has yet to be made. The current implementation defines a set of functions that generate fixed GOP structures. The proper function is selected by the user according to an a-priori evaluation of the length of the motion vectors. In the current setup, the output GOP structure reflects the one generated by the original x264 decision function. The *initialize_me* function is used to initialize the motion field and it provides MV predictors for each macroblock of size 4×4. Each predictor is obtained multiplying the spatial coordinates of the central pixel of the macroblock by the homographic matrix calculated with respect to the reference frame. This function is called for P- and B- frame type only. A careful analysis of the original x264 libraries allows these modification to fit into the original standardized H.264/AVC framework. The preliminary stage of this proposed encoding system allows to initialize crucial parameters (frame type, MV, ...) with the available sensor information in order to speed up the whole encoding stage.

As described above, the proposed approach allows to generate a H.264/AVC-compliant video stream which can be later decompressed and reconstructed by any compatible decoder.Further details are reported in [4].

3. EXPERIMENTS

3.1. Scenario and data description

The input of the presented experiments is a sequence of high resolution images recorded by the drone while overflying a high voltage pylon in a rural areas (see Figure 3). The objective of the acquisition is to get images in order to make a 3D reconstruction of the pylon. To obtain a good reconstruction, many points of view of the camera (102 frames) and high resolution images (3000x2000 pixels) are needed. The drones moves at an average velocity of 2 m/s and at an average altitude of 80 m. The horizontal field of view of the camera is 74 degrees. In order to obtain the desired image overlap, the snapshots are taken at the frequency of 1 Hz.

The camera parameters are known after an off-line calibration. The position and orientation of the camera are available from the on board sensors, the GNSS (Global Navigation Satellite System) and the AHRS (Attitude and Heading

	PSNR (dB)		FPS (Hz)	
Coder	@1 fps	@0.5 fps	@1 fps	@0.5 fps
sa264	**35.32**	**32.44**	**4.23**	**4.29**
x264	34.48	31.78	3.36	3.28

Table 1. PSNR and FPS results for the test sequence @3250 kbps.

Reference Systems). A subsampled sequence (0.5 Hz) of 51 frames has also been considered, in order to further reduce the overlap between consecutive frames.

The experiments are performed using the preset *"medium"* of x264. Experimentally, in fact, we found that for this kind of video sequences, the use of more complex preset increases the computational burden without significantly improving the rate-distortion performance.

3.2. R-D performance and coding speed

The performances of the coded sequences at 3250 kbps in terms of quality (PSNR - Peak Signal to Noise Ratio) and speed (FPS - encoded Frames Per Second) are shown in Table 1 for both 0.5 and 1 Hz.

The R-D curves for the subsampled sequence are shown in Figure 1. The gain in terms of PSNR for the sensor aided encoder (sa264) is low at low bitrates (about 1000 kbps) where the video quality is low. It increases up to 0.6 dB for higher bitrates. The same happens for the non subsampled sequence, where the gain is up to 0.9 dB.

Figure 2 shows that sa264 outperforms x264 also in terms of coding speed. Tests were conducted on a PC with a 8 core i7-2710QE cpu @2.1 GHz and 4 Gb of RAM. The proposed encoder is over 25 percent faster. In fact, x264 can make a lot of computation before finding motion vectors, because consecutive frames have a very low overlap with respect of the case of full motion videos and the motion vectors are very long. Moreover the transformation between the frames can be quite prospective and the affine motion model can fail in many cases. Strategies that do not consider very small blocks for the motion estimation, often are not the best. The encoder sa264, instead, considers always 4x4 blocks in the prediction model, without expensive calculations.

3.3. 3D reconstruction quality

In order to evaluate the video compression effect on the 3D reconstruction, we extracted a point cloud from the original image sequence with the software Agisoft Photoscan. This software extracts point clouds in the 3D space from multiple images of the same scene by using standard techniques of SFM (Structure From Motion). In our test, SFM took advantage of sensor data for position and orientation measurements. Sensor data were suitably filtered for robust estimation [7].

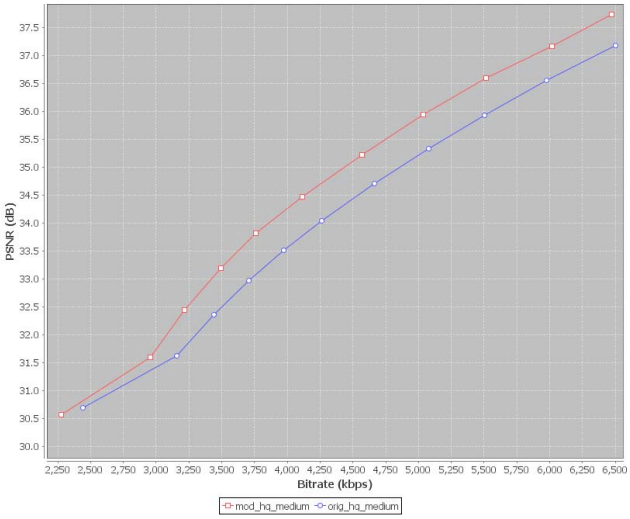

Fig. 1. PSNR vs bitrate @0.5 fps (blue curve = x264, red curve = sa264).

With the same workflow we extracted point clouds from the x264 and sa264 sequences. A comparison analysis was made directly on the point clouds instead of their relative mesh surfaces, using the Hausdorff distance as error measure and the original point cloud as reference. Using the same standard configuration of the SFM algorithm, with a bitrate of 6000 kbps, the sa264 coded sequence outperforms the x264. Indeed, x264 has an error with mean of 4.2 cm and standard deviation of 14.8 cm, while for sa264 the error has mean of 3.0 cm and standard deviation of 12.3 cm. The reconstructed 3D model is shown in Figure 3.

3.4. Results with simulated data

Because, in the previous experiments, there is an important 3D part in the scene (the high voltage pylon), is interesting to understand if there is a lack of rate-distortion performance due to the fact that the planar approximation is unsatisfied in many regions of the image. For this reason, we proposed the same trajectory to generate simulated data, taking in account the orography of the overflight terrain but removing the most relevant 3D object from the simulated scene, the pylon.

The image generator we used is not able to reach the same digital resolution of the camera, but it can generate images at resolution of 1056x704 pixels, with an horizontal field of view of 60 degrees.

In Tab. 2 the results of the simulated aerial imagery are reported. In this case we have a better improvement in coding speed, and a much more relevant improvement in rate-distortion performance w.r.t. original real data. These experiments suggest that probably the R-D performance of the sensor aided encoder increases with the relative altitude with respect the scene depth.

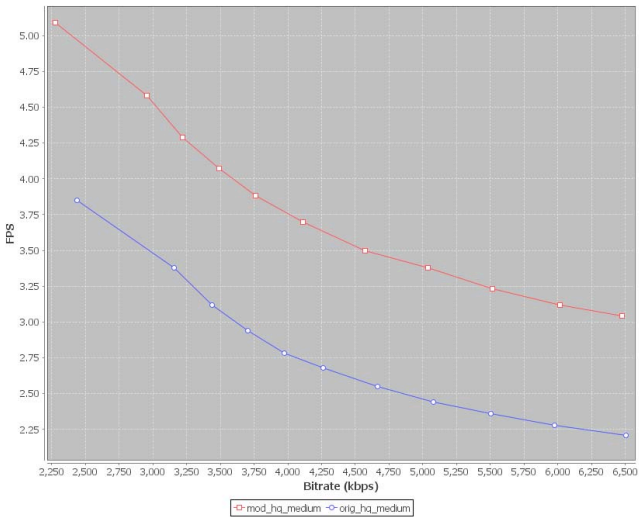

Fig. 2. Coded FPS vs bitrate at @0.5 fps (blue curve = x264, red curve = sa264).

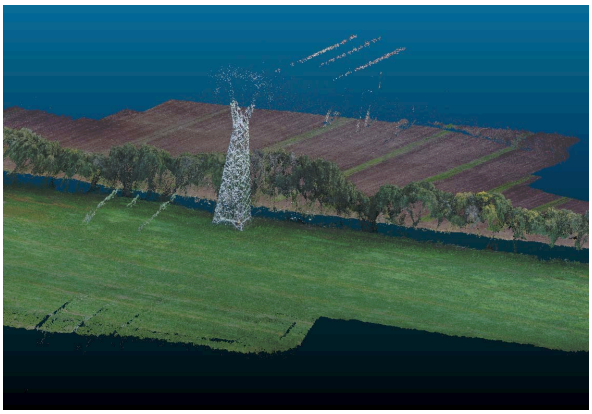

Fig. 3. 3D reconstruction of the overflight area from the compressed video sequence @0.5 fps and @6000 kbps.

Other experiments on simulated data generated with the same trajectory of the drone, but on urban areas (with textured 3D buildings), showed similar results. These experiments suggest that probably the R-D performance improves with the details in the overflight landscape, because the relevance of the inter-frame coding with respect to the intra-frame coding is higher. These considerations are supported by the analysis of the motion fields.

4. CONCLUSION

This paper presented the application of the Sensor Aided Video Coding technique introduced in [4], in the context of SFM 3D reconstruction from aerial motion imagery. Working conditions differ from [4] because the drone flies at a relative low altitude. Therefore, there is a strong perspective effect and the planar approximation is less valid.

	PSNR (dB)		FPS (Hz)	
Coder	@1 fps	@0.5 fps	@1 fps	@0.5 fps
sa264	**44.83**	**42.63**	**41.58**	**41.85**
x264	41.85	38.44	32.08	32.64

Table 2. PSNR and FPS results for the simulated sequence @2000 kbps.

Experiments are promising even if more real sequences are needed. Test on simulated sequences show that better results are achieved with higher altitude and with more scene details (e.g., urban areas). Such a condition is difficult to test with a small drone because of the Italian restrictions in terms of permit to fly. However, the case study clearly shows a speed up of computations which preserves battery life and hence the mission duration. Results in terms of R-D curves and the quality of 3D reconstruction, show that this technique brings advantages also in terms of data quality at same bitrate.

5. REFERENCES

[1] A. Rodriguez, B. Ready, and C. Taylor, *Using Telemetry Data for Video Compression on Unmanned Air Vehicles*, American Institute of Aeronautics and Astronautics, 2014/02/15 2006.

[2] J. Gong, C. Zheng, J. Tian, and D. Wu, "An image-sequence compressing algorithm based on homography transformation for unmanned aerial vehicle," *Intelligence Information Processing and Trusted Computing, International Symposium on*, vol. 0, pp. 37–40, 2010.

[3] M. Bhaskaranand and J.D. Gibson, "Low-complexity video encoding for UAV reconnaissance and surveillance," in *MILCOM, Military Communications Conference*, 2011, pp. 1633–1638.

[4] C. V. Angelino, L. Cicala, M. De Mizio, P. Leoncini, E. Baccaglini, M. Gavelli, N. Raimondo, and R. Scopigno, "Sensor aided h.264 video encoder for uav applications," in *Proceedings of te 30th Picture Coding Symposium, PCS, San Jose, CA, USA*, December 2013.

[5] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, July 2003.

[6] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*, SpringerVerlag, 2003.

[7] C. V. Angelino, V. R. Baraniello, and L. Cicala, "Uav position and attitude estimation using imu, gnss and camera," in *15th International Conference on Information Fusion, Singapore*, July 2012, pp. 735–742.