

Word Confidence Using Duration Models

Original

Word Confidence Using Duration Models / Scanzio, S; Laface, Pietro; Colibro, D; Gemello, R.. - (2009), pp. 1207-1210. (Interspeech 2009 Brighton 6-10/9/2009).

Availability:

This version is available at: 11583/2280631 since:

Publisher:

ISCA

Published

DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

An H.264 sensor aided encoder for aerial video sequences with in-the-loop metadata enhancement

Luca Cicala^{a*}, Cesario Vincenzo Angelino^a, Nadir Raimondo^b, Enrico Baccaglini^b, Marco Gavelli^b

^a CIRA, the Italian Aerospace Research Centre, 81043 Capua, Italy
(c.angelino, l.cicala)@cira.it

^b Istituto Superiore Mario Boella, Torino, Italy
(raimondo, baccaglini, gavelli)@ismb.it

Abstract. Unmanned Aerial Vehicles (UAVs) are often employed to capture high resolution images in order to perform image mosaicking and/or 3D reconstruction. Images are usually stored on-board or sent to the ground using still image or video data compression. Still image encoders are preferred when low frame rates are involved, because video coding systems are based on motion estimation and compensation algorithms which fail when the motion vectors are significantly long. The latter is the case of low frame rate videos, in which the overlapping between subsequent frames is very small.

In this scenario, UAVs attitude and position metadata from the Inertial Navigation System (INS) can be employed to estimate global motion parameters without video analysis. However, a low complexity analysis can refine the motion field estimated using only the metadata.

In this work, we propose to use this refinement step in order to improve the position and attitude estimation produced by the navigation system with the aim of maximizing the encoder performance. Experiments on both simulated and real world video sequences confirm the effectiveness of the proposed approach.

1 Introduction

Unmanned Aerial Vehicles (UAV) are mainly employed in order to collect data [4]. Often this task is achieved using a set of on-board digital video cameras. Typical constraints of UAV missions are related to limited bandwidth, *e.g.*, when they operate Behind Line of Sight (BLOS), as well as battery life. In the first case, is unlikely to achieve a high frame rate acquisition especially because additional data gathered by the other payload sensors share the same data link and further reduce the bandwidth available to the video stream. On the other hand, when UAVs are used in order to acquire high resolution images for mosaicking and/or 3D reconstruction, there is no need to transmit the video stream and data

* Corresponding author

are stored on-board. In this situation, the main mission constraint is the duration of the battery that supplies the vehicle.

It can be desirable to optimize the available resources (bandwidth, power supply) in order to improve the mission performance (more data, more flight time). When high frame rate videos are not a desiderata of the mission, one solution can be the reduction of the acquisition frame rate. In such a scenario, the video sequences are sent/stored at few frames per second (fps) and hence the overlap between two consecutive frames is lower than standard video streams. Usually at low frame rates the commercial video encoders fail in performing a good motion estimation/compensation, due to the length of the motion vectors (MVs) and to the prospective changes among frames that make hard the MV prediction. In such situations a still image encoder can be more or equally performing.

However, the motion of the UAV camera can be derived by the position and orientation data delivered by the on-board navigation systems. Moreover, the geometry of the overflight scene is approximately known and can be estimated using, for example, a laser altimeter, or the GPS (Global Positioning System) position and a Digital Terrain Model (DTM). With such an information, a global motion in the image plane can be inferred without video analysis. In [3], the authors investigate a low complexity encoder with GM based frame prediction and no block Motion Estimation (ME). For fly-over videos, it is shown that the encoder can achieve a 40% bit rate savings over a H.264 encoder with ME block size restricted to 8x8 and at lower complexity. In [10] and [12], global motion parameters are used to compensate frames that are used as reference for block ME using GM within standard MPEG-4 and H.264 codecs. In [5], the authors propose a framework tailored for UAV applications that uses the GM information and a homography model to code the stream using JPEG2000. In [11] and [2], the authors present modifications of the H.264/AVC encoder to initialize the MVs using the camera motion information from UAV sensors. These latter approaches perform block ME at a lower complexity, and transmit the derived block MVs. Both approaches guarantee the generation of a standard-compliant H.264/AVC bitstream thus no changes at the decoder side are required.

In this paper, we propose a sensor aided video encoder to be used at high resolution and low frame rates on aerial video sequences, following the studies reported in [2] and [1]. The encoder is obtained by modifying the open source implementation of H.264/AVC video coding standard (ISO/IEC, 2006) x264 [9] and fully compliant with H.264. As opposed to the previous works, here the problems of video coding and of the metadata correction are tackled in the same integrated design. This paper is focused on the improvements in terms of rate-distortion performance. Moreover, further aspects about a sensor aided encoder design, unpublished results and consideration, are reported.

The paper is organized as follows. Section 2 introduces the proposed sensor aided coding scheme with in-the-loop metadata correction. Section 3 describes how to improve the navigation data using the optical flow calculated by the proposed video encoder. In Section 4 experimental results are presented and in Section 5 conclusions and future work are discussed.

2 Sensor aided motion imagery coding

The structure of the proposed encoder is shown in Figure 1. A common H.264 encoding scheme is modified in order to take account of meta-data (position and orientation) coming from the navigation system of the UAV. The camera is supposed to be internally and externally calibrated with respect to the navigation system. A Global ME (GME) is performed using metadata and a rough planar representation of the overflight scene (*i.e.*, assuming the ground to be an horizontal plane and using an altimeter to determine the distance of the aerial platform from the ground). A further MV refinement is performed by block matching, as proposed in the original version of x264, but starting from a more accurate initial estimate of the MVs, as provided by the GME module. Further, in addition to the scheme presented in the cited work, the proposed solution uses the estimated motion field as optical flow estimation for a state-of-the-art camera egomotion algorithm based on RANSAC homography model estimation and algebraic motion data extraction. The camera egomotion is used in loop with an unscented Kalman filter in order to refine the position and orientation data provided by the navigation system. Such use of the motion field will be discussed in the Subsection 3.

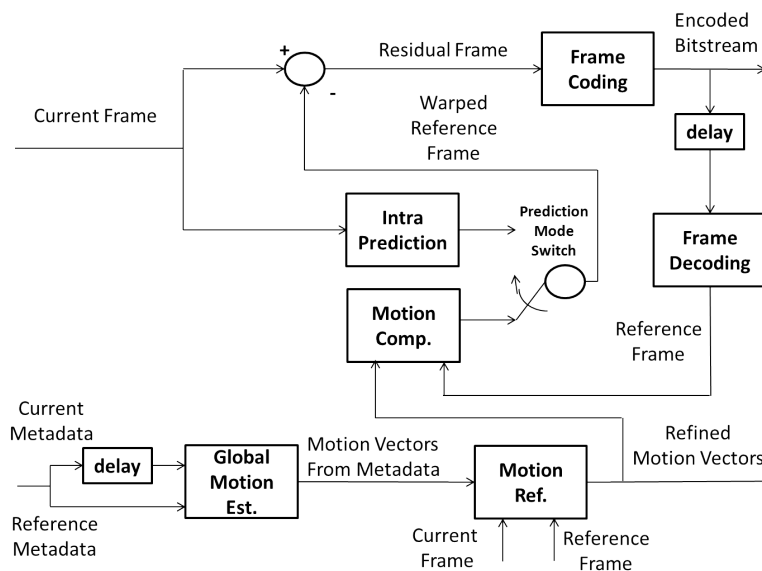


Fig. 1. Sensor aided video encoder scheme.

Figure 2 show a situation in which the ME algorithm of x264 fails while the proposed ME process, initialized with the sensor based GME, performs with success. The vectors in overlay represent the MVs found by the ME process. When an appropriate MV cannot be estimated, Intra prediction is performed instead.

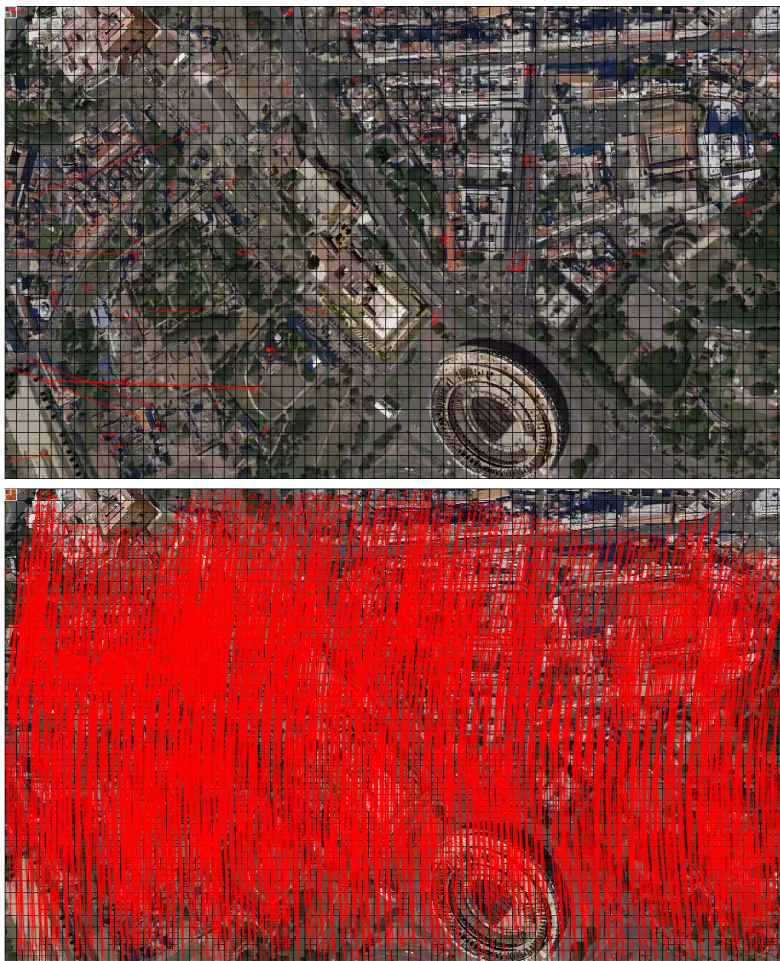


Fig. 2. Top) x264 ME algorithm at low frame rates (0.5 fps). Down) Sensor aided ME. Motion vectors are represented with red arrows.

3 Metadata enhancement

The overall data fusion architecture is sketched in Figure 3, where the sensor fusion block implements the Kalman filtering of the data provided by the Navigation System and the camera egomotion data from the video processing system. The camera egomotion module is based on the homography matrix, which relates homologous points in two different views of the same scene. In this work the correspondences are given by the refined MV provided by the encoder. Obtained a set of correspondences of points for a couple of successive frames, the homography matrix is estimated and then decomposed into his motion and structure parameters. The estimation and decomposition procedure is behind the scope of this work and will be omitted. The interested reader may refer to [6] and the reference therein.

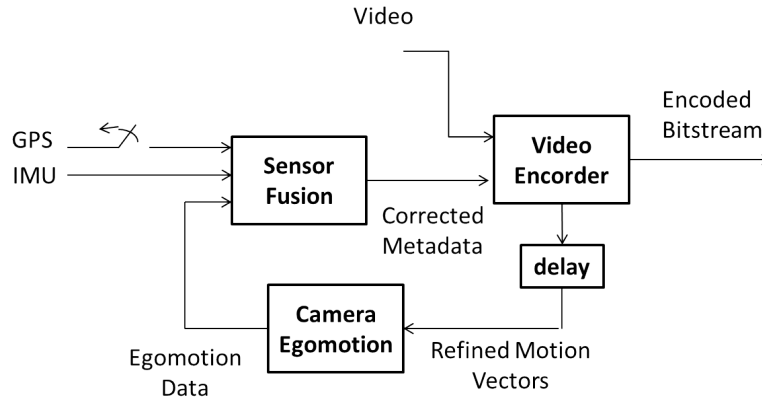


Fig. 3. Metadata improvement by sensor fusion.

The final purpose of the proposed sensor fusion algorithm is the estimation of the position and attitude of the camera, supposed internally and externally calibrated. The data fusion algorithm is based on the Unscented Kalman Filter (UKF) [13],[8], because dynamic and observation equations are non-linear in their original form. Like all the Kalman Filters, an UKF consists of two steps: model forecast and data assimilation. Sigma points are used to represent the current state distribution and to propagate the distribution to the next state and to the output. Mean and covariance of the transformed sigma points can be used to calculate the Kalman gain and to update the state prediction. Often such a filter has been used to estimate the pose of an UAV. In particular, in this work, we adopt the same solution proposed in [1]. Angular velocities and linear accelerations provided by the Inertial Measurement Unit (IMU)

are used in the Kalman prediction step. GPS position and speed as well as camera egomotion parameters are used in the Kalman update step, in order to correct the position and the orientation drift due to the integration of the IMU data. A magnetometer is used also in order to correct the heading.

4 Experimental results

4.1 Test data

Three different aerial sequences [7] have been encoded and then their motion data have been processed. We considered low frame rate sequences (0.5 - 1 fps) and relative long MVs as this is often the case for UAV acquired high resolution video sequences. The characteristics of the three sequences are reported in Table 1.

Video Seq.	FR [fps]	Res [pix x pix]	h-FOV [deg]	Speed [km/h]	Alt [m]
Cape Pend.	1	1088x672	60	250	800
Rome	0.5	1088x672	60	250	800
Brezza	1	3000x2000	73.7	2	80

Table 1. Aerial video sequences characteristics.

The sequences "Cape Pendleton" and "Rome" have been generated using Google Earth. In the "Cape Pendleton" sequence, the overflight region is a military base and the surrounding areas. The area is substantially homogeneous and with few details. The "Rome" sequence refers to a flight over the city of Rome, rich of details. For the simulated sequences the horizontal Field Of View (FOV) is 60 degrees, the frame resolution is 1088x672 pixels. The flying altitude is 800 m for both the simulated video sequences. The "Brezza" sequence is part of a video recorded using a real multi-rotorcraft mini-UAV over a rural region poor of details (grass with some trees and only a few of man-made structures). The frame resolution is 3000x2000 pixels. The flying altitude (80 m) is much lower than the simulated video sequences. The horizontal FOV is 73.7 degrees.

Ground truth metadata are provided by the image generator for the synthetic video sequences, while for the "Brezza" sequence, they are estimated from multiple views by a bundle adjustment technique. However, in the experiments a noisy version of these metadata has been generated according to the sensor model described in [1]. The parameters of the sensor model, as reported in the cited paper, are extracted by the datasheet of a well known commercial GPS aided Attitude and Heading Reference System commonly employed in aeronautical applications.

4.2 Encoder settings

The x264 library offers several presets. Each preset is a collection of parameters which are set in order to get a good trade-off between quality and coding time for different application scenarios. The "medium" preset is general purpose, and is compatible with low computationally demanding scenarios. Because in this scenario the video frame rate is very low (i.e. 0.5 fps), a time demanding preset, i.e. the "slower", can be also considered, in order to reach better encoding quality. These preset options for the proposed modified x264 encoder are labeled in the figures as "medium" and "slower", instead the same configurations in the original x264 encoder are labeled with the prefix "x264", and are respectively "x264_medium" and "x264_slower".

The sensor aided encoder often uses only one reference frame in the GOP, because the low overlap among the frames. For this reason, the comparison with the reference x264 encoder with only one reference frame in the GOP is presented. In this case only the "medium" preset is reported. The corresponding label is "x264_medium_ref1".

Two other coding option are specifically presented for the proposed sensor aided encoder. A first option excludes the refinement step of the ME through video analysis. This option is labeled as "medium_nors", where the word "medium" indicates the used preset and the acronym "nors" is for NO Refinement Search.

A further option is added in order to force the sensor aided encoder to perform the ME also when Intra coding is possible. This processing step can be useful in order to produce more accurate motion field that can be used by the sensor fusion module. These experiments, reported for the preset "medium", are labeled as "medium_uem" or "medium_nors_uem", where the latter acronym is for Use Estimated MVs.

4.3 R-D performance

In the following experiments the Rate-Distortion (R-D) performance of the proposed sensor aided encoder, using corrected metadata, is compared to that of the x264 implementation of H.264. Eight rate-distortion curves are plotted in Figures 4-6 for each tested video sequence. On the x-label the encoding bitrate is reported, while on the y-label the PSNR (Peak Signal to Noise Ratio), that is a commonly used objective video quality measure.

A first observation is that the proposed sensor aided encoder performs better than the reference x264 encoder both with the medium and the slower preset. For example, for the "Cape Pendleton" video sequence, at 400 kbps, the PSNR of the proposed encoder is 35.41 dB versus 34.76 dB of the reference with the medium preset, and 36.24 dB versus 34.31 dB, with the slower preset. For the sequence "Rome", at the same bitrate of 400 kbps, the PSNR of the sensor aided encoder is 33.93 dB versus 32.74 dB of the reference with the medium preset, and is 35.18 dB versus 32.50 dB for the slower preset.

The proposed sensor aided encoder has a similar behaviour on real video sequences also. On the sequence "Brezza", for example, at 3250 kbps, the PSNR is 35.12 dB versus 34.48 dB of x264.

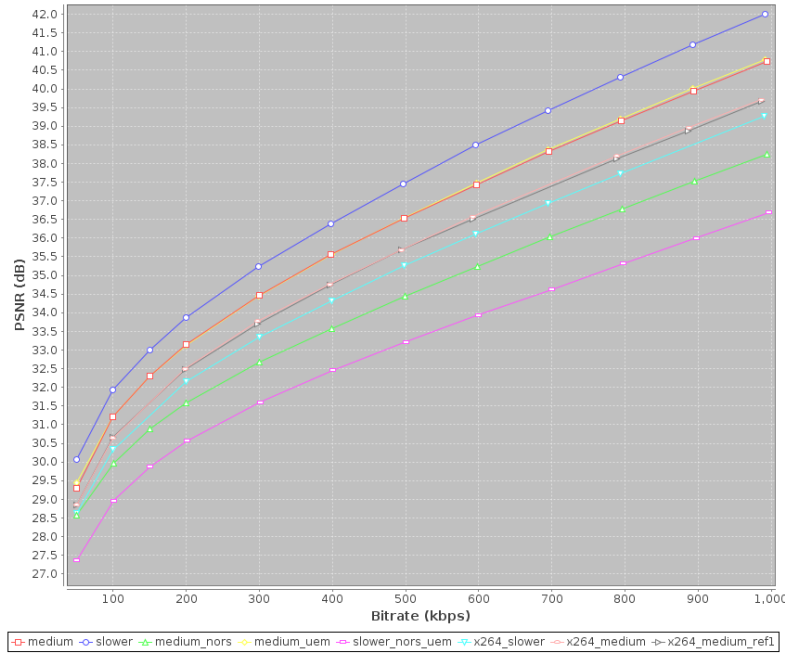


Fig. 4. R-D Curves for the "Cape Pendleton" sequence, with a resolution of 1088x672 pixels and a frame rate of 1 fps.

It is worth to notice that the reference x264 encoder uses a complex GOP analysis in order to optimize the use of the I, P and B frames. The proposed sensor aided implementation instead, at the current stage of development, uses a more simple strategy based on only one I frame per GOP and all P frames (this strategy is reasonable, due to the continuity of the camera motion). For this reason it is more correct to compare the "medium" curve with the "x264_medium_1Ref" curve, instead of the "x264_medium" curve. Comparing these couples of curves, the proposed solution can be further appreciated.

From the figures it is also possible to note that the "slower" preset has lower quality than "medium" preset for the reference x264 encoder. This is due to the large number of B-frames selected by the x264 GOP decision algorithm. In the considered scenarios, in which there is low overlap among successive frames, the use of B frames has bad effects on the output quality. The proposed sensor aided implementation, instead, uses the same GOP structure for the two different presets, that is similar to the best option selected by x264.

Analyzing only the sensor aided implementations, other considerations can be made. Comparing the "medium" and "medium_nors" curves, it is clear that the motion search refinement step, based on video analysis,

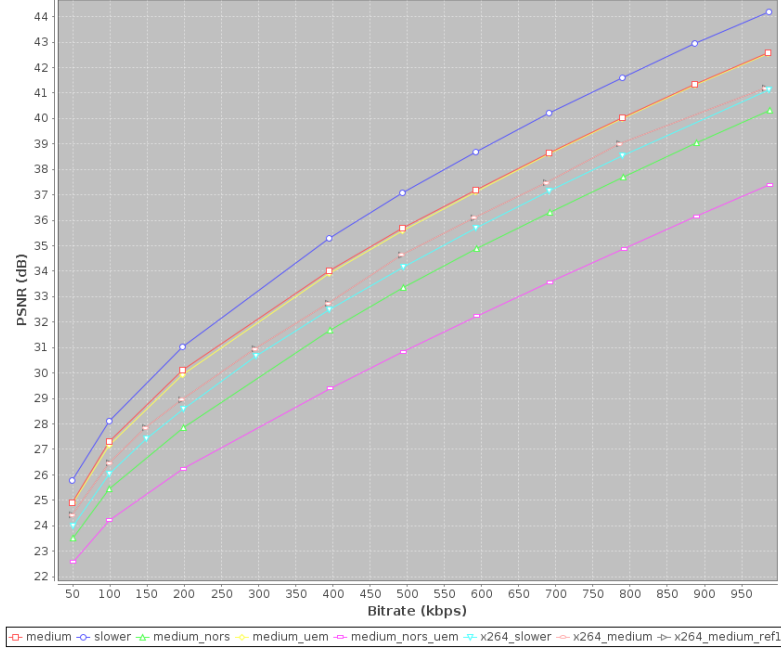


Fig. 5. R-D curves for the "Rome" sequence, with a resolution of 1088x672 pixels and a frame rate of 0.5 fps.

is essential to reach high rate distortion performance. Instead, comparing the "medium_nors" and "medium_nors_uem" curves, it is possible to conclude that a ME pure approach cannot be preferred to a combined strategy approach, based both on Intra and Inter block prediction, at least in the case in which video analysis is not used to refine the MVs. The comparison between the "medium" and the "medium_uem" curves, however show that, using the video analysis for MVs refinement, also don't considering the Intra option, it is possible to reach a performance near to the combined approach (both Intra and Inter blocks). In the case of Cape Pendleton and Rome, the gap is negligible.

In general, the metadata correction performance is generally very good (the standard deviation of the orientation estimation error results less than 0.2 degrees for all the video sequences), so that the obtained encoding results are very similar to that obtained using the ground truth (difference in PSNR less than 0.15 dB).

5 Conclusions

In this work we proposed an integrated solution of sensor aided video encoder, able to process corrected metadata in order to estimate the

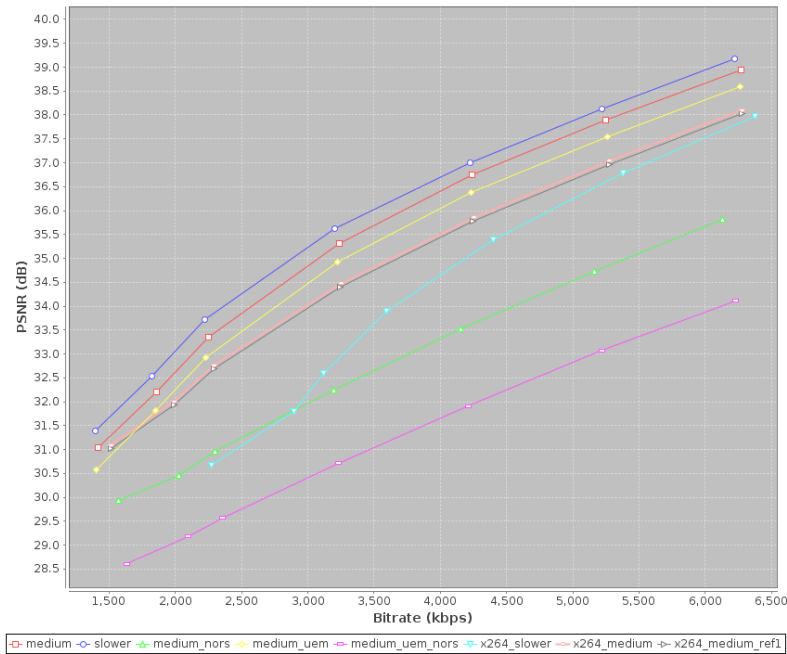


Fig. 6. R-D curves for the "Brezza" sequence, with a resolution of 3000x2000 pixels and a frame rate of 1 fps.

global motion in an aerial video sequences, strongly reducing the need of video analysis. A novel encoder architecture is presented and a fully H.264 implementation is proposed and tested, on simulated and real video sequences. The experimental results show the effectiveness of the proposed solution at high resolution and low frame rates. The suggested applications are to UAV imagery transmission and storage, under channel capacity or power consumption constraints. Future works will be focused on computational complexity aspects and on optimized solutions for high speed vision based metadata corrections.

References

1. Angelino, C.V., Baraniello, V.R., Cicala, L.: High altitude uav navigation using imu, gps and camera. In: Proceedings of the 16th International Conference on Information Fusion (FUSION). pp. 647–654. Istanbul, Turkey (July 2013)
2. Angelino, C.V., Cicala, L., De Mizio, M., Leoncini, P., Baccaglini, E., Gavelli, M., Raimondo, N., Scopigno, R.: Sensor aided h.264 video encoder for uav applications. In: Proceedings of the 30th Picture Coding Symposium (PCS). pp. 173–176 (Dec 2013)

3. Bhaskaranand, M., Gibson, J.: Global motion assisted low complexity video encoding for uav applications. *IEEE Journal of Selected Topics in Signal Processing* 9(1), 139–150 (Feb 2015)
4. Chen, X.l., Zhang, S.c., Liu, J.: Design of uav video compression system based on h.264 encoding algorithm. In: *Proceedings of the 1st International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT)*. vol. 5, pp. 2619–2622. Harbin, China (Aug 2011)
5. Gong, J., Zheng, C., Tian, J., Wu, D.: An image-sequence compressing algorithm based on homography transformation for unmanned aerial vehicle. In: *Proceedings of the 1st International Symposium on Intelligence Information Processing and Trusted Computing (IPTC)*. pp. 37–40. Huanggang, China (Oct 2010)
6. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
7. ISMB/CIRA: Test sequences [online] available at: http://www.ismb.it/mise_cira (2013)
8. Julier, S.J., Uhlmann, J.K.: Unscented filtering and nonlinear estimation. *Proceedings of the IEEE* 92(3), 401–422 (Mar 2004)
9. Merritt, L., Rahul, V.: X264: A high performance h.264/avc encoder [online] available at: http://neuron2.net/library/avc/overview_x264_v8_5.pdf (2006)
10. Morimoto, C., Burlina, P., Chellappa, R.: Video coding using hybrid motion compensation. In: *Proceedings of the 4th International Conference on Image Processing (ICIP)*. vol. 1, pp. 89–92. Santa Barbara, California, USA (Oct 1997)
11. Soares, P.H.F.T., Pinho, M.d.S.: Video compression for uav applications using a global motion estimation in the h.264 standard. In: *Proceedings of the 6th International Workshop on Telecommunications*. vol. 1. Santa Rita do Sapucaí, Brazil (May 2013)
12. Steinbach, E., Wiegand, T., Girod, B.: Using multiple global motion models for improved block-based video coding. In: *Proceedings of the 6th International Conference on Image Processing (ICIP)*. vol. 2, pp. 56–60. Kobe, Japan (Oct 1999)
13. Van Der Merwe, R., Wan, E.: Sigma-point kalman filters for probabilistic inference in dynamic state-space models. In: *Proceedings of the Workshop on Advances in Machine Learning*. Montreal, Canada (Jun 2003)