Sensor Aided H.264 Video Encoder for UAV Applications

C.V. Angelino, L. Cicala, M. De Mizio, P. Leoncini
Payload Sensors & HMI Lab
C.I.R.A., Centro Italiano Ricerche Aerospaziali
Capua, Italy
{c.angelino, l.cicala, m.demizio, p.leoncini}@cira.it

E. Baccaglini, M. Gavelli, N. Raimondo, R. Scopigno
Multi-Layer Wireless Solutions
Istituto Superiore Mario Boella
Torino, Italy
{baccaglini, gavelli, raimondo, scopigno}@ismb.it

Abstract—This paper presents a new low-complexity H.264 encoder for Unmanned Aerial Vehicles (UAV) applications. Standard video coding systems currently employed in UAV 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. The 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. Experimental results show that the proposed scheme outperforms standard H.264 in terms of PSNR and throughput. The results are relevant in low frame rate video coding, which is a typical scenario in UAV behind line-of-sight (BLOS) missions. Experiments open new directions in developing new sensor aided video coding standards.

I. INTRODUCTION AND RELATED WORK

Global motions in a video sequence are caused by camera motion, which can be modeled by parametric transforms. The process of estimating the transform parameters is called global motion estimation (GME) and compensation (GMC).

Usually, GME would require extra computations, because all the pixels are involved in the estimation of global motion parameters. There is a considerable amount of low-complexity GME methods proposed, e.g., [3] that only use a small subset of pixels in motion estimation to calculate motion model parameters; [4] allows fast estimation of model parameters from coarsely sampled motion vector field. Moreover, there are some other methods employed GMEC for predictive motion estimation, e.g., [5]. Tools for the GME are available in the MPEG4 part 2 standard. However, H.264/AVC (MPEG4 part 10) has not adopted GME in the standard due to its suboptimal rate-distortion performance and increased complexity.

In order to reduce the computational burden of GME, several works on the use of external motion sensors have been recently proposed. In [10] the authors studied the application of GME techniques that process sensor data (gyroscopes, accelerometers and magnetometers) to video coding. The focus is on mobile devices and the complexity reduction of motion estimation reflects into increasing battery life. The use of external sensors, thus, is suitable for low complexity global motion estimation. To show the effectiveness of this approach, the authors showed that SaVE has a strong noise-resistant capability when employed with particular predictor insertion strategies. The proposed method has been also patented [8] by the authors and the Rice University in 2010. A similar approach has been also patented by R. Collin and Nokia [6] in 2009. M. Stephan and Harman Becker Automotive Systems GmbH [7] patented in 2008 a sensor assisted video compression system that uses also position sensors (e.g. GPS receiver), together with orientation sensors, to estimate the global motion.

Video coding systems currently employed in NATO military UAV applications meet the STANAG 4609 standard [14], which refers to MPEG4 and Motion JPEG2000 standards for motion imagery coding. However, these standards are for general purpose applications and do not rely on the particular features of aerial video sequences. Indeed, the latters present some peculiarities in terms of observed scene geometry and point of view. In particular, the observed scene is almost static (the camera movement is dominant) and generally easy to model (usually approximated with a plane). Moreover, the point of view (camera position and orientation) is far from the scene and approximatively known since it can be obtained from the navigation system. UAV transmitted video has two main applications:

1) remote piloting: the video helps the remote pilot during take-off, landing and critical situations;
2) surveillance: for both military and civil applications such as the remote monitoring of hostile area, coasts and boundaries surveillance, disaster management (earthquakes, floodings, fires, etc.).

In the first application, the video transmission must respect the
real-time constraint even at the cost of decoded video quality, while the the second application needs high defined images under a more relaxed real-time constraint. Furthermore, an UAV video coding system should meet both bandwidth (BLOS transmission) and computational (battery life) constraints.

Sensor aided video coding for UAV application gained attention in the last few years. Rodriguez et.al [11] 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 [12] 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 [13] 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 this paper 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, Section II shows how to obtain a prediction of the global motion. The focus is on low frame rate video coding, which is a typical scenario in BLOS missions. If motion vector predictors and refinement stage (as described in [15]) 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 choose the prediction method (inter/intra) for each macroblock and initializing the motion vectors by substituting the non-standard x264 predictors with predictors derived from external data. Furthermore, we also adress the case of different refinement policies to improve the overall motion estimation process.

II. SENSOR AIDED GLOBAL MOTION ESTIMATION

This Section describes the model used in this work to predict the global motion from sensors data. Then, it shows how this model has been employed in the H.264 motion estimation system through its open source implementation x264 (version 0.128.30M 5d25de2).

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

\[
p \simeq CP_c,
\]

where the symbol \( \simeq \) means that the equality is up to a scale factor. The calibration matrix \( C \) at time sample \( k \) is defined as

\[
C[k] = \begin{pmatrix} -f[k] & 0 & \frac{c}{h} \\ 0 & -f[k] & \frac{c}{w} \\ 0 & 0 & 1 \end{pmatrix}.
\]

Here \( f \) represents the focal length in pixel and \( c \times r \) is the camera frame digital resolution (rows \( \times \) columns).

Let us consider the observed scene as a 2-D plane \( \Pi \) in 3-D space described by the equation

\[
n_t^T \cdot P_t = h,
\]

where \( n_t \) is the normal to the plane and the subscript \( t \) refers to the tangent reference frame centered at camera center \( O \).

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

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

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 [9, p.131]

\[
H_R[k-1,k] = R_{t \rightarrow c}[k-1] \cdot R_{t \rightarrow c}^T[k],
\]

\[
H_T[k-1,k] = \frac{1}{h} \cdot R_{t \rightarrow c}[k-1] \cdot (O_t[k-1] - O_t[k]) \cdot n_t^T \cdot R_{t \rightarrow c}^T[k].
\]

Here, \( R_{t \rightarrow c} \) 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.

The proposed approach wants to assess whether there are possibilities for improvement, in terms of time and quality, of the encoding process of the x264 libraries when they are integrated with external data provided for instance by UAV sensors. In fact, especially for aerial sequences, a strong global motion flow is present in each frame and for each macroblock. The possibility of providing the encoder with some preferential motion directions and significant frames to be used as reference can thus leverage the encoding process. Our work mainly addresses two specific points:

- 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 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 \texttt{frametype\_decide} function is called to decide the type and the proper reference of the frame associated with the actual available sensor data (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 4x4. 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. In fact, the proposed pre-encoding process exploits external data in the sense that it initializes the MV predictors and the frame type decision methods leaving unaltered the subsequent stages.

III. PERFORMANCE EVALUATION

Test sequences have been generated by means of a computer graphic simulation tool for image acquisition over photorealistic and 3-dimensional scenarios. Acquisition camera is installed under the UAV in a looking-down setup. Simulated data reflects non-ideality characteristics of sensors (accelerometers, gyros, magnetometer, GPS) in terms of sampling rate, delay, noise, available bandwidth. The sequence generated for encoding comparison in the current work is made of 300 fps, flying altitude of approx 200 meters over a city\(^1\). Similar consideration holds for another tested sequence, not reported here for brevity, made of 480 frames, 2.5 fps, flying altitude of approx 400 meters over a rural scenario.

All tests have been executed at constant bitrate with different runs in the range \([100 – 1500]\) kbps, frame size 990 \(\times\) 704 and reflects BLOS (beyond line of sight) scenarios. We report the results for low spatial overlapping sequences (lower than 60\%) as they allow to achieve the most significant benefits in the proposed scenario. It is worth noting that the proposed encoder can also be adopted in case of full motion video which can be transmitted over a Line-of-Sight datalink. All sequences have been encoded at Main Profile @ 3.1 to allow the support for B slices and entropy coding. Reference frames are set to 1 and MB partitioning is set to \(i8\times8, i4\times4, b8\times8, p8\times8\). Motion estimation refinement has been set to HEX and UMH. No exhaustive search and no \(p4\times4\) partitioning have been adopted due to the fact that encoding time is significantly increased with a very little quality gain.

In the reported results, the curves labeled ORIG refer to the original x264 libraries in which the GOP structure is left to default optimization policies. Both UMH and HEX refinement have been tested for the original encoder. In general, for the tested sequences, short GOP are created so a significant number of Intra-coded picture are generated. The curve labeled INTRA refers to the intra-code mode i.e. all frames are I frames. Curves labeled MOD refer to the proposed libraries which exploit data from sensors to initialize motion fields. MOD-1 refers to the GOP structure \(I, B_{ref}, B, I\) and no P frames are coded. This configuration allows to bi-predict the B frames both from the previous and the next I frames. For this scenario, UMH refinement is adopted. MOD-2 refers to the GOP structure \(IPP\ldots\) i.e. a I frame followed by 250 P frames. In this scenario, both UMH and HEX refinement have been tested. In Fig.1 we report the results in terms of output quality (Peak Signal-to-Noise Ratio expressed in dB) versus the encoding bitrate (in kbps) and it is possible to highlight the benefits of the proposed approach. The proposed algorithm outperforms traditional H.264/AVC coding because the motion estimation fails when the frame overlapping is very low. The GOP structure \(I, B_{ref}, B, I\) (MOD-1) outperforms traditional coding and the gain is even higher if the encoder adopts long sequences of P frames (MOD-2). Around 500 kbps, for instance, when using data received from sensors to initialize MV predictors, the gain can be approximately 1.4 dB when comparing the approach MOD-2 vs the original x264 with hexagonal motion refinement. In case of UMH refinement (which is usually not adopted in real-time application) the gain reduces to 0.6 dB. These results are coherent with the fact that the subsequent frames overlapping is low and then the traditional x264 motion estimation algorithms fail when

\(^1\)The sequence can be downloaded at www.telematica.polito.it/sas-ipl/Baccaglini/video/test.avi
looking for similar MB in previous frames. For this reason, x264 rate-distortion policies decide to code frames in Intra-mode and this requires higher bandwidth. As a performance bound, in the figure we also report the encoder performance in Intra-mode. As can be seen, the output quality is lower than the other schemes because the bandwidth required to code each frame as an I frame is significant. Concerning encoding time, it is worth noting that motion estimation is significantly faster when it is possible to include in the video compressor system data acquired from sensors. As shown in Fig. 2, it is possible to notice that there is a significant reduction in the encoding time. For a total output bitrate of 400 kbps, both for HEX and UMH refinement, a 25% speed-up in encoding time is achieved and this value slightly when the total output rate increases. The Intra-code scheme is the fastest as no prediction is performed but, as seen before, this is the worst scenario in terms of quality vs bitrate. UMH refinement requires higher computational time with respect to HEX search. We can note that MOD-2-UMH encoding time is approximately the same as the original x264 with HEX search (ORIG-HEX).

Finally, Fig. 3 shows the image quality comparison between the standard ITU H.264 and the proposed method. As one can see, the proposed method better preserves the image details which are very important for UAV surveillance and monitoring missions.

IV. Discussion

In this paper we proposed a new low complexity H.264 compliant video encoder for UAV applications. The encoder employs a new motion estimation strategy which initialize the motion vector prediction by means of navigation system data. Experiment results show better performances of the proposed encoder w.r.t. standard H.264 encoder in low frame rate coding scenario, which is typical for UAV BLOS mission. Future works will address a more general model for the observed scene.

REFERENCES