SIP (2015), vol. 4, e13, page 1 of 19 © The Authors, 2015.

doi:10.1017/ATSIP.2015.12

ORIGINAL PAPER

Mesh-based piecewise planar motion compensation and optical flow clustering for ROI coding

HOLGER MEUEL, MARCO MUNDERLOH, MATTHIAS RESO AND JÖRN OSTERMANN

For the transmission of aerial surveillance videos taken from unmanned aerial vehicles *(UAVs),* region of interest *(ROI)-based coding systems are of growing interest in order to cope with the limited channel capacities available. We present a fully automatic detection and coding system which is capable of transmitting high-resolution aerial surveillance videos at very low bit rates. Our coding system is based on the* transmission *of ROI areas only. We assume two different kinds of ROIs: in order to limit the transmission bit rate while simultaneously retaining a high-quality view of the ground, we only transmit new emerging areas (ROI-NA) for each frame instead of the entire frame. At the decoder side, the surface of the earth is reconstructed from transmitted ROI-NA by means of global motion compensation (GMC). In order to retain the movement of moving objects not conforming with the motion of the ground (like moving cars and their previously occluded ground), we additionally consider regions containing such objects as interesting (ROI-MO). Finally, both ROIs are used as input to an externally controlled video encoder. While we use GMC for the reconstruction of the ground from ROI-NA, we use meshed-based motion compensation in order to generate the pelwise difference in the luminance channel (difference image) between the mesh-based motion compensated and the current input image to detect the ROI-MO. High spots of energy within this difference image are used as seeds to select corresponding superpixels from an independent (temporally consistent) superpixel segmentation of the input image in order to obtain accurate shape information of ROI-MO. For a* false positive *detection rate (regions falsely classified as containing local motion) of less than 2% we detect more than 97%* true positives *(correctly detected ROI-MOs) in challenging scenarios. Furthermore, we propose to use a modified high-efficiency video coding (HEVC) video encoder. Retaining full HDTV video resolution at 30 fps and subjectively high quality we achieve bit rates of about 0.6–0.9 Mbit/s, which is a bit rate saving of about 90% compared to an unmodified HEVC encoder.*

Keywords: Region of interest ROI coding, Mesh-based motion compensation, Superpixel segmentation, Low bit rate HDTV video coding, Moving object detection

Received 26 June 2014; Revised 14 August 2015; Accepted 14 August 2015

I. INTRODUCTION IN REGION OF INTEREST (ROI) CODING

For aerial surveillance tasks, e.g. for disaster area monitoring as well as for police surveillance operations, *unmanned aerial vehicles* (UAVs) become more prevalent nowadays. One of the main challenges hereby is the transmission of high-resolution image data recorded on-board the UAV over channels with only limited capacities. Taking into account the high resolutions of today's and upcoming camera sensors (4 K and above), and the demand for multiple or multi-view video streams, efficient data compression is of growing interest. In this work, we aim at providing high image quality for the transmission of high-resolution video sequences (HDTV) at low bit rates.

Institut für Informationsverarbeitung (TNT), Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany. Phone: +49 511 762-19585

Corresponding author:
H. Meuel
Email: meuel@tnt.uni-hannover.de

A) Related work

Trivial approaches transmit high-resolution video data from UAVs to the ground by simply using broader channels like Wi-Fi or using a high-image compression ratio targeting low bit rates. This either results in the disadvantage of very limited range of operation (in the case of e.g. Wi-Fi) or results in poor image quality and the possible loss of interesting image content [1].

1) ROI CODING

In order to reduce the bit rate after encoding while maintaining interesting image content, ROIs coding is commonly applied, spatially dividing each frame of a video sequence into ROIs and non-ROIs. Hereby, the quality of ROIs is left untouched. Non-ROI areas of a frame could be

blurred in a preprocessing step prior to actual video encoding or coarsely quantized within the video encoder itself to reduce the overall bit rate [2–4]. A modified or externally controllable block-based hybrid video coder like *Advanced Video Coding* (AVC) [5] or *High-Efficiency Video Coding* (HEVC) [6] is employed in [7–10] and [11, 12], respectively, in order to apply different QPs for the coding of ROI and non-ROI blocks.

The drawback of ROI detection and coding approaches discussed above is the degradation of non-ROI areas that cannot be reconstructed at full quality at the decoder. The modular ROI detection and coding system introduced and extended in our earlier publications [13, 14] and [15] exploits the characteristic of aerial video sequences on a planar landscape to overcome this drawback and maintains full resolution and high-quality video over the entire frame at low bit rates. It relies on the transmission of only new emerging image content (*New Areas*, ROI-NA) for each of the frames, which are stitched together in a mosaicking post-processing step at the decoder to reconstruct the static parts of the scene (background) by means of *global motion compensation* (GMC) [13, 14]. Since only small parts of each frame have to be transmitted, this ROI detection and coding system is capable of providing high image quality at low bit rates. Since other approaches introduced here are not able to reconstruct the background in full quality at bit rates considerably lower than those of a common state-of-the-art video encoder, we use this system as a basis. It is described in detail in Section II.

2) ROI DETECTION

Although, theoretically, ROIs can be arbitrarily defined, e.g. in the center of the image or by detecting skin color in a teleconferencing system like in [16], more context-sensitive approaches are desirable. For aerial surveillance scenarios, *moving objects* (MOs) are often considered as ROI, further on referred to as ROI-MO. Most of the recent work in the field of surveillance video processing, especially for automatic MO detection, relies on a static (non-moving) camera, e.g. [17–20] and consequently cannot deal with camera ego-motion. Hence, those approaches are not suitable for aerial surveillance tasks with the camera attached to a UAV.

Popular approaches rely on the GMC of the background pixels due to the camera movement prior to calculate the pelwise image differences (difference image) between two frames of the image sequence or between the current frame and a reconstructed background reference image [21–24]. More efficient detectors can also handle non-perfect conditions like parallax effects by employing the epipolar geometry [25]. Other approaches are based on optical flow analysis in order to detect MO [26]. In [27], image features are classified as stationary or moving on the ground, clustered and finally tracked over several frames. The clustering itself is based on morphological filtering of a binarized difference image [28, 29] provide an extensive overview on recent publications in the field of aerial surveillance systems from moving cameras.

However, since the image signal itself is not considered in any of the above MO detection strategies, the shape of MOs cannot be detected accurately, especially in homogeneous areas within MO (e.g., car roofs). Moreover, since MO are often detected on a frame-by-frame basis, missing detections in single frames lead to entire ROIs not being detected.

To overcome these limitations of the MO detection results and the image signal [30], uses the difference image-based MO detection results as seeds for a mean shift clustering [31] in order to accurately determine the shapes of MOs. In [32], MO detection is performed by processing the difference image between an affine motion compensation and an optical flow estimation. Based on the detected blobs, *GraphCut* is employed as a signal-dependent segmentation method to determine the shapes of MOs in the input video frame. In our previous work [15], we showed that a superpixel segmentation [33] is able to outperform a *GraphCut* method in a MO detector. By additionally exploiting the temporal consistency of these superpixels, we were able to handle the problem of temporally missing detections of MOs in single frames. Based on [15], we further reduced falsely as static classified MOs due to motion parallax in [34] by replacing the background motion compensation within the MO detector by a mesh-based motion compensation and a clustering of displacement vectors [35].

Thus, we decided to use our MO detection and ROI-based coding system from [34] as a basis for extension with an efficient modified HEVC video encoder.

The contributions of this work are:

(1) We summarize our previous work [13, 15, 34, 36] and thoroughly describe the complete MO detection and coding system for the low bit rate transmission of high-resolution aerial video sequences in detail.
(2) We review the mesh-based motion compensation and the mesh-based cluster filter for the reduction of non-MOs, falsely classified as moving from [34], and present previously unpublished details of the cluster filter in Section IV-A.
(3) We show a more detailed evaluation of the MO detector including *receiver operation characteristics* (ROCs) for different test sequences, also considering the publicly available VIRAT test data set [37, 38].
(4) We propose to integrate a modified HEVC video encoder in the coding system and evaluate the performance compared to the AVC-based video encoder employed in our previous work as well as to an unmodified HEVC encoder. In order to analyze the maximum coding performance of the proposed video encoder, we use more test sequences containing no MOs in a second test set.
(5) Finally, we present a run-time analysis for each component in order to underline the suitability of the proposed system for usage on-board an UAV.

The remainder of this work is organized as follows: In Section II, we review the ROI-based coding system for

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

low bit rate transmission of aerial video and introduce our adaption of HEVC as video encoder (Fig. 1: brown) to exploit the improved coding efficiency of HEVC compared to AVC. In Section III, we describe the integration of superpixels in the system. In Section IV, we explain our mesh-based MO detector in detail, employing our cluster filter approach for reliably distinguishing non-moving and MOs and a mesh-based motion compensation for the compensation of non-planar structures. In the experimental results in Section V, we present results of the improved MO detection system for an extended test set (compared to our previous work [34]) in terms of detection accuracy (Section V-B), coding efficiency (Section V-B) as well as run-time (Section V-C). In order to demonstrate the maximum coding efficiency of the modified HEVC video encoder, we use a second, publicly available test set [39], containing self-recorded high-resolution aerial sequences without MOs. Section VI concludes this work.

II. OVERVIEW OF THE PROPOSED ROI CODING SYSTEM FOR AERIAL SURVEILLANCE VIDEO

The entire block diagram of the ROI coding system for aerial surveillance video sequences including all proposed improvements is depicted in Fig. 1 (based on [34]). In order to visualize each processing step, we also integrated preview images into the block diagram. We will introduce all components and explain the entire pre-processing procedure needed prior to the actual video encoding within this subsection at appropriate positions.

Assuming a planar landscape, the camera motion between the recorded frames at the encoder on-board the UAV can be estimated. This estimated motion is transmitted as projective transformation parameters to the decoder at the ground station. Assuming a first, regularly coded frame by the encoder, these parameters are used to predict the current frame from already known video frames by applying a GMC of the background. Since the background can be reconstructed at full quality by means of GMC, no additional transmission cost is necessary for already transmitted background for any predicted frames. Image content which is not reconstructed by the global motion model, such as newly visible background (*New Area*, ROI-NA) or MOs (ROI-MO), is transmitted using an externally controlled arbitrary video codec, e.g. AVC (also known as H.264 or MPEG-4 part 10), or HEVC. Compared to a block-based motion compensation, we can reconstruct a high-quality image without blocking artifacts with our GMC approach.

As a basis for further processing, we derive the global motion out of the video frames as follows: A pel $\vec{p} = (x, y)^\top$ in frame k can be mapped on a corresponding coordinate $\vec{p}' = (x', y')^\top$ in the preceding frame $k-1$ using the projective transformation $F(\vec{p}, \vec{a_k})$ (equation (2)) with the projective transformation parameter set $\vec{a_k}$ (equation (1)).

$$a_k = (a_{1,k}, a_{2,k}, \ldots, a_{8,k})^\top, \tag{1}$$

$$F(\vec{p}, \vec{a_k}) = \left(\frac{a_{1,k} \cdot x + a_{2,k} \cdot y + a_{3,k}}{a_{7,k} \cdot x + a_{8,k} \cdot y + 1}, \frac{a_{4,k} \cdot x + a_{5,k} \cdot y + a_{6,k}}{a_{7,k} \cdot x + a_{8,k} \cdot y + 1} \right)^\top. \tag{2}$$

Thus, one plane, i.e. one frame, can be mapped into another with the projective transformation, whereas the parameters a_3 and a_6 express translational movement in direction of x and y. The parameter set is embedded into the bit stream of the video encoder as *supplemental enhancement information* (SEI). Since only nine floating point numbers per frame are required, the additional bit rate is negligible.

To estimate the global motion, first a *Harris Corner Detector* [40] detects corner features in the current frame k. Secondly, a sparse optical flow (Fig. 1: white) is calculated by a *Kanade–Lucas–Tomasi* (KLT) feature tracker from frame k to the previous frame $k-1$ [41, 42]. By employing a projective transformation motion model (equation (2)), *Random Sample Consensus* (RANSAC) is able to estimate a set of projective transformation parameters for the mapping of all pixels from frame $k-1$ to frame k while removing the outliers [43] (Fig. 1: green). Using this parameter set, *New Area* is computed as the image regions not contained in the frame $k-1$ but in the current frame k (on a pelwise basis). These regions are marked for video encoding in a map of pixels to be coded, further on called the *coding mask*. For the detection of MOs, the pelwise difference in the luminance channel (Y) between the current frame k and the globally motion compensated prediction $\hat{k}$ is computed (further referred to as *difference image*) and spots of high energy are marked as MOs in an *activation mask*.

A) Increase of true positive detections of MOs by integrating temporally consistent superpixel in the MO detector

Such difference image-based MO detectors lack accuracy when it comes to unstructured, homogeneous regions within the MOs – e.g. car roofs – as for those areas where the pixel differences between the current and the motion compensated frame are relatively small [19]. Figure 2 illustrates occurring problems: if parts of a MO (original in Fig. 2(a)) are detected as ROI whereas other parts of the same MO are not recognized (Fig. 2(b)), reconstruction errors might occur since the motion compensated ground (background) and foreground (ROI) might not match exactly, leading to errors in the reconstructed video (Fig. 2(c)) [15]. We identify MO areas more accurately by combining an independently calculated superpixel segmentation with the difference image-based detector result (Fig. 3, middle and bottom row): the results from the difference image-based detector are used as seeds to automatically activate only those superpixels containing MOs. Additionally, by using a temporally consistent superpixel (TCS) segmentation our system is able to bridge temporal detection gaps, thus reducing the amount of missed detections per frame (see also Fig. 4 in Section III-A for illustration). As shown in [15], the TCSs

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

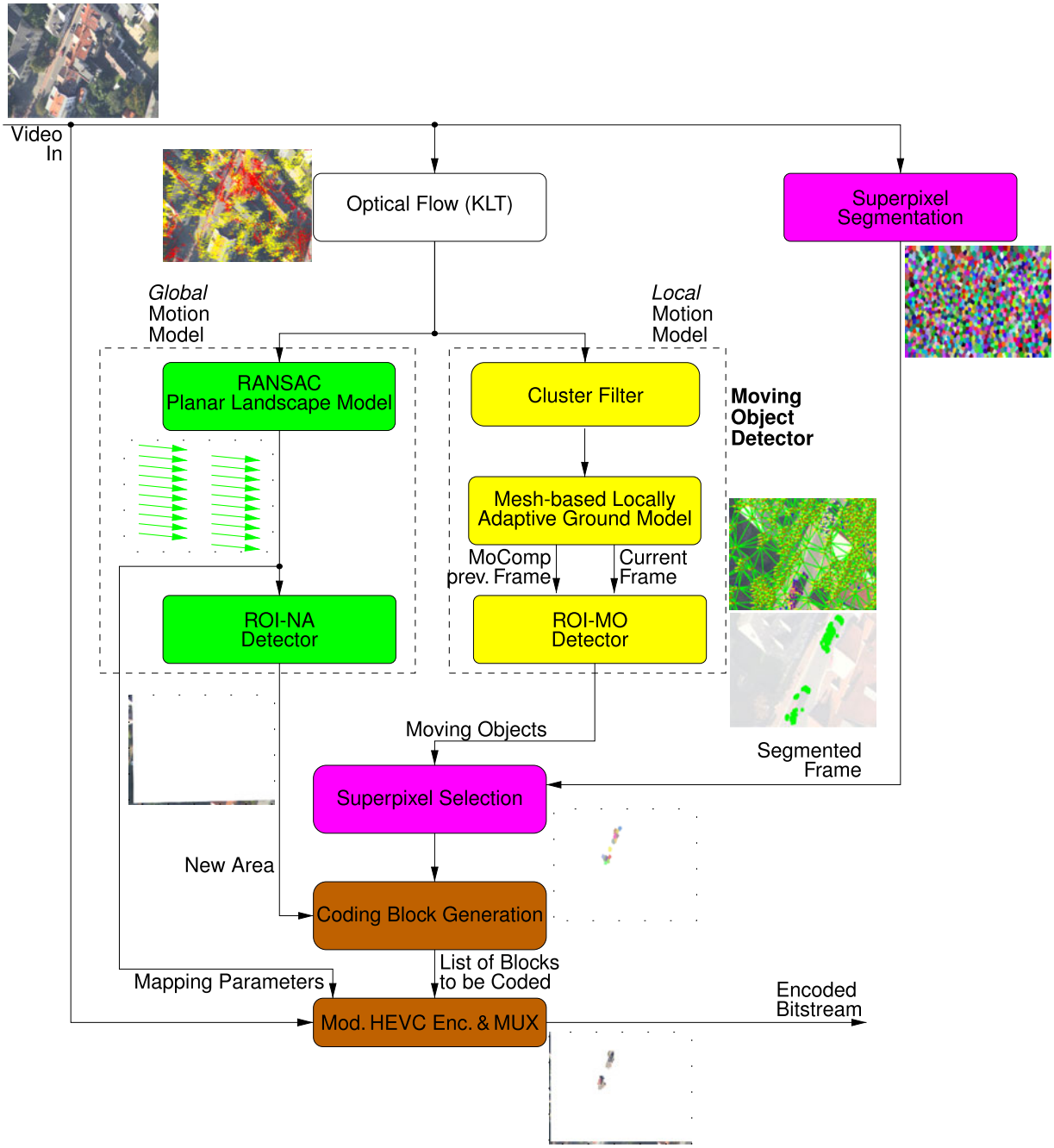

Fig. 1. Block diagram of ROI detection and coding system: Bold framed block: proposed cluster filter to eliminate *false positive* (FP) detections; white: optical flow; yellow: mesh-based motion estimation/compensation incl. ROI detector; magenta: superpixel segmentation and selection; green: global motion estimation and new area detector; brown: block generation, video coder and muxing (based on [34]).

are able to outperform other state-of-the-art segmentation methods like an efficient *GraphCut*-based *SlimCut* implementation [44]. The TCS segmentation itself [33] and the integration into the detection system [15] (Fig. 1: magenta) are described in Section III.

B) Reduction of FP detections of MOs by integrating a mesh-based MO detector

Given the use of the projective transformation, we must assume a planar ground which is (prevalently) true for sequences recorded at high flight altitudes. This assumption is not suitable for non-planar ground structures like buildings or trees. These lead to image regions falsely detected as MO *false positive* (FP) detections resulting in an increased ROI area. Consequently an increased number of superpixels is selected for encoding. For the MO detection, we propose to replace the planar GMC by a mesh-based motion estimation and compensation [36]. Instead of one global plane for the full frame, multiple smaller planes are used to enable the motion compensated image to adapt to non-planar scene geometry (Fig. 1: yellow) [34]. We describe

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

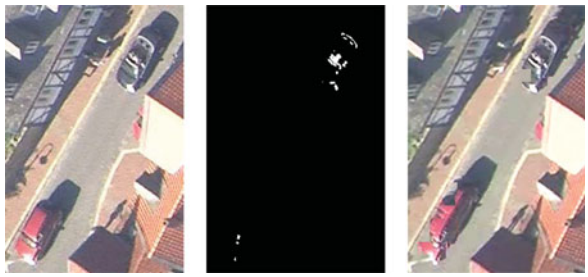

Fig. 2. Original outtake (a) and reconstructed image after ROI encoding and decoding (c) with inaccurate MO detection due to homogeneous, unstructured regions on the car roof. Missing detections (b) of the rear part of the red car as ROI lead to reconstruction errors since the front part of the car (ROI) does not match the reconstructed background [15].

the mesh-based local motion estimation and compensation as well as a locally adaptive outlier detector (Fig. 1: yellow) for MO detection to deal with non-planar areas in Section IV [34].

C) Further reduction of the bit rate by introducing an HEVC video encoder

Whereas the reference system in [34] employs a modified AVC encoder, we propose to replace the video encoder by a recent HEVC encoder in order to gain from the increased coding efficiency of HEVC compared to AVC. The approach to determine which image areas finally have to be encoded (ROI, non-skip mode) and which not (non-ROI, skip mode), remains the same: the pelwise information of ROI-NA as well as ROI-MO is extended to a fixed block grid (*macroblock*/*Coding Unit* level). If at least one ROI-NA or ROI-MO pel is located in a 16×16 block, this block is marked for encoding in non-skip mode in a *Final block coding mask* as shown in Fig. 3 which is used to control the video encoder externally. Since common video coding standards like AVC and HEVC only define the decoding, our encoder control does not affect the (HEVC) standard compliance of the bit stream. However, an additional post-processing is necessary as described above to reconstruct non-ROI areas (static background) of the scene [13, 14].

The coding gain of our system compared to the encoding of entire frames with an unmodified video encoder depends on the amount of ROI to be encoded. As an upper limit we have to encode the entire frame (e.g. if MOs are all over the frame). In this case the system falls back to encode and to transmit the full frame, resulting in a coding efficiency of the unmodified video coder (anchor). As a lower limit we can encode the entire frame in skip mode, if no UAV motion is prevalent and no MOs are detected within the scene. However, for typical scenarios only a few percent of each frame have to be encoded and transmitted.

We would like to emphasize that single components (e.g., the video encoder, the image segmentation or the MO detector) could be exchanged by similar components without loss of generality or loss of functionality of the entire system.

III. SUPERPIXEL-SEGMENTATION

In order to improve the detection accuracy of non-textured MOs without decreasing the precision, it was proposed in [15] to use superpixels for the context-adaptive enlargement of the activation mask. Superpixel algorithms as initially proposed by Ren and Malik in [45] group spatially coherent pixels sharing the same color or which are part of the same texture into segments of approximately same size and shape. Their boundaries should comply with appearance boundaries of objects present in the scene. In this work, we use the superpixel segmentation framework proposed by [33] which provides segmentation masks for all frames of the input video. Superpixels occupying the same image region in different frames share the same label establishing a temporal connection between these superpixels.

The framework produces superpixels by clustering pixels using their five-dimensional (5D) feature vector $[labxy]$ containing the three color values in CIE-Lab color space and

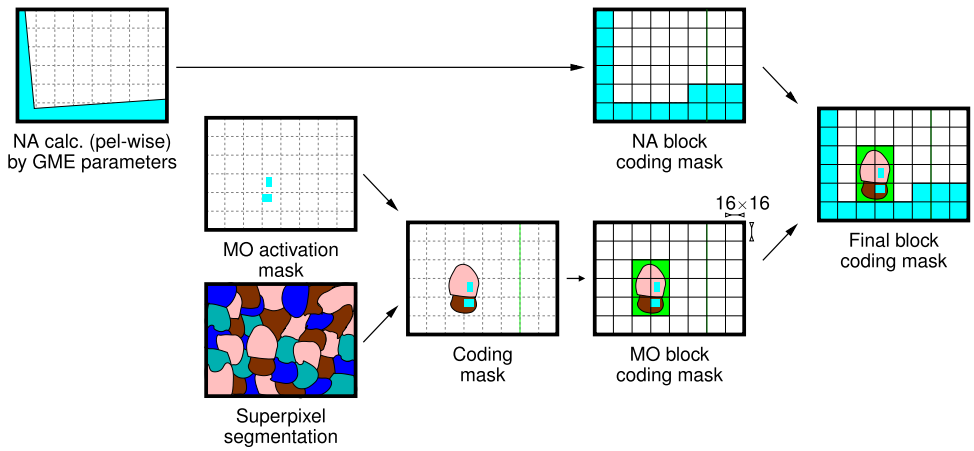

Fig. 3. Coding mask generation for new area (top row) and MOs. The MO *activation mask* from the difference image calculation is overlaid with an independent *Superpixel segmentation* in order to get accurate shape information of the MOs. The *Coding mask* is adapted to a coding block pattern (*MO block coding mask*) and combined with the *NA block coding mask* to the *Final block coding mask*. Cyan and green blocks in the latter will be encoded as ROI.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

the pixel's xy-coordinates. To capture the temporal connections between superpixels in different frames, the clustering is performed over an observation window spanning multiple frames. Pixels in different frames being part of the same superpixel should share the same color but not necessarily their position over multiple frames. Therefore, each cluster center (representing one superpixel) consists of one color center and multiple spatial centers (one for each frame in the observation window). In order to represent the image content adequately by a superpixel segmentation, an optimal set of cluster centers Θ_{opt} as well as a mapping $\sigma_{i,k}$ of pixels i in frame k to these cluster centers have to be obtained. A cost function (equation (3)) is defined which sums up all distances of the pixels to their assigned cluster center

$$D_{total} = \sum_{k} \sum_{i} (1-\alpha) D_c(i_k, \sigma_{i,k}) + \alpha D_s(i_k, \sigma_{i,k}), \quad (3)$$

where $D_c(i_k, \sigma_{i,k})$ and $D_s(i_k, \sigma_{i,k})$ denote the Euclidean distance of pixel i_k to the cluster center $\sigma_{i,k}$ in color space and image plane. The spatial distance is normalized by the average superpixel size which depends on the frame resolution and the number of superpixels chosen by the user. The trade-off between color-sensitivity and spatial compactness can be controlled by the weighting factor α. If α is set to 1 no color information is used resulting in Voronoi cells which only depend on the initial positions of the superpixels' spatial centers. On the other hand, a low α leads to less compact superpixels which vary more in their size and have irregular shapes (for our experiments we set α to 0.96 which was empirically determined in [33]). An approximation of the optimal set $\hat{\Theta}_{opt}$ and a corresponding mapping $\hat{\sigma}_{i,k}$ is obtained by applying an alternating *expectation–maximization* (EM) scheme. In the expectation-step an optimal mapping for the three latest frames in the observation window is obtained by minimizing equation (3). This is done by assigning each pixel to the cluster center for which the weighted sum of color and spatial distances is minimal. In the maximization-step the cluster centers are updated by calculating the mean color and spatial values of the assigned pixels. The expectation- and maximization-steps are alternated five times before the observation window is shifted one frame forward. The connectivity of the superpixels is ensured by a post-processing step. The initialization is done by subsequently filling the observation window with frames while performing several iterations (five in our experiments) of the expectation- and maximization-step after adding a frame. The first frame is initialized by distributing the cluster centers uniformly on the frame. After the observation window finally spans an amount of 15 frames, new frames are inserted into the window. Simultaneously the oldest frame is removed which results in a shift of the observation window.

With the integration of TCSs as described, we are able to accurately segment shapes of MOs. However, in case of missing activations of superpixels due to missing detections from the difference-image-based MO detector, e.g. for slow moving or shortly occluded MOs, those MOs might still not be detected and thus wrongly reconstructed.

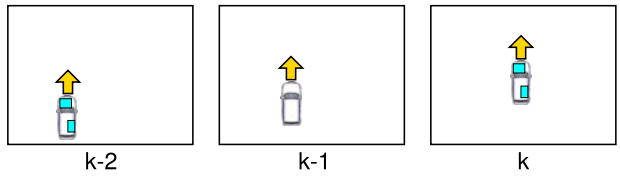

Fig. 4. *Temporally consistent superpixels (TCSs)* are used to bridge *false negative* detections of the ROI-MO: if no MO (white car) is detected by the MO detector (cyan), the MO in frame $k-1$ would not be selected for coding. Due to the temporal consistency of the superpixels the position of the car can also be predicted in frame $k-1$ and thus correct processing and transmission of the car in all frames can be guaranteed.

A) Bridging short-time missing MO detections by utilizing temporal consistency

In order to activate blocks containing those slow MOs in the coding mask, i.e. to detect MOs which erroneously were not detected in a single frame, but in the surrounding frames, we employ a sliding window approach (Fig. 4): an active superpixel in the current frame within a *sliding window width* (SWW) will also activate the past and next SWW/2 temporally associated superpixels. SWW = 1 represents no superpixel activation propagation, "3" specifies a lookback and a lookahead of one frame each. Thus, besides the propagation of activations into homogeneously colored areas this TCS enhanced system guarantees the accurate detection of MOs in case of short-time missing detections caused by e.g. very slow object movement.

Although the integration of superpixels in the coding system offers several benefits, wrongly detected MOs were erroneously also enlarged by the corresponding superpixels. Thus, we have to ensure a *FP* detection rate being as low as possible.

IV. REDUCTION OF FP DETECTIONS OF MOS BY MESH-BASED MOTION COMPENSATION AND CLUSTER FILTERING

The GMC uses a projective transformation (homography) to model the movement of the background pixels between the frames originated by the ego-motion of the camera during recording. This model assumes all scene points to lie on the surface of a single plane in three-dimensional (3D) space, i.e. the surface of the earth. This approximation is only valid if the surface of the earth is completely planar or if the distance of the camera to the earth is high and the focal length of the camera is chosen small. Violation of this assumptions e.g. by low flight altitudes, large focal lengths, or non-planar ground structures like buildings or trees result in falsely detected MOs due to the motion parallax effect. The effect describes the difference in displacement of projected pixels between frames of a moving

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

camera and depends on the distance of the scene points to the camera center.

The homography based GMC is only capable of compensating the displacement of projected scene points which are positioned on the surface of a single plane at a specific distance. All scene points not placed on the surface of the plane might result in spots of high energy in the difference images due to their displacement not being perfectly compensated. This results in lots of FP detections of MOs and consequently leads to an unnecessary high bit rate in the ROI coding system. Assuming that several small planes can be better fitted to a non-planar landscape than one plane per frame, we replaced the single plane GMC by a locally adaptive mesh-based motion compensation [14, 36, 46] *for the MO detector only*. We approximate the earth surface with a mesh of connected (small) planar triangles, which are called patches further on. Each patch is assumed to be planar and has an individual slope and orientation. This allows the mesh surface to better adapt to non-planar ground structures and parallax effects.

Since the KLT features are designated as nodes of the triangles, the feature points have to be pruned by outliers, such as feature points on MOs.

A) Mesh-based cluster filtering for outlier removal

As with multiple small planes there is no single global homography parameter set to optimize, the RANSAC outlier removal has to be replaced by evaluating the motion of features in a local neighborhood surrounding each of the patches. For this purpose, we have designed a filter which is based on the clustering of motion vectors, gained from the KLT feature tracker, using a region growing approach. Since KLT has to be performed anyway for global motion estimation, no extra effort is necessary to generate these motion vectors. We assume a smooth optical flow field: small changes between adjacent motion vectors suggest them to be part of the same object while discontinuities indicate objects with differing motion. This filter we call *cluster filter* (CF), as it clusters the optical flow into regions of similar motion [34]. To follow small changes in the vector field, the region growing approach assumes clusters to be defined by the motion vectors on their boundaries only: if the spatial distance (equation (4)) of an unclustered motion vector $\vec{v_k}$ in frame k to the closest border motion vector $\vec{c_k}$ of an already existing cluster in the same frame is smaller than a threshold t_{d1} and if furthermore the difference in their displacements $\vec{d_v} = \vec{v_k} - \overrightarrow{v_{k-1}}$ and $\vec{d_c} = \vec{c_k} - \overrightarrow{c_{k-1}}$ between the frames k and $k-1$ (equation (5)) is also smaller than a threshold t_{d2}, both vectors are considered similar and the unclustered vector is added to the cluster. The displacement similarity t_{d2} is hereby scaled by the distance to force a higher similarity for nearby motion vectors:

$$||\vec{v_k} - \vec{c_k}|| < t_{d1}, \tag{4}$$

$$||\vec{d_v} - \vec{d_c}|| < t_{d2} \cdot \frac{||\vec{v_k} - \vec{c_k}||}{t_{d1}}. \tag{5}$$

If no further unclustered vector fulfills the similarity condition according to equations (4) and (5) for any cluster, a new cluster has to be founded. The process repeats until every vector is assigned to a cluster. A MO is defined by a common motion and hard discontinuities in the vector field at its borders. Therefore, a MO forms an individual cluster. The displacement vectors on non-planar structures, however, only change slightly and continuously. These changes are relatively small compared to those of real MOs. Therefore, the cluster filter is capable of assigning high objects which protrude from the ground plane into the background motion cluster by simply following the small changes in displacement from the bottom up to the top. As an example imagine a church spire: the ground plane of the church spire will have no displacement caused by motion parallax due to the moving camera. In contrast to that, the maximum displacement at the church top will be very high. Considering only pairwise neighbored motion vectors, starting from the bottom up to the top, the displacement will increase slightly and continuously. The background motion cluster is finally defined as being the largest one in the scene (brown dots in Fig. 5). Only the background cluster is used for motion compensation whereas small clusters are further processed as MO candidates (Fig. 5, blue crosses with purple and white dots). Clusters containing less motion vectors than a threshold t_f are considered to be outliers and have to be removed (Fig. 5, blue crosses).

B) Mesh-based motion compensation

To define the piecewise planar patches of the mesh from the background motion vector field, a triangle mesh is generated using the feature point coordinates of the background cluster of the frame k given by the cluster filter as nodes for the mesh (see Fig. 6, based on [36]). As the changes in perspective between the frames are relatively small, an affine transformation is greatly sufficient to compensate the content of each patch. We create a triangle mesh with triangles t_i and the feature points as vertices, employing a Delaunay triangulation [47] using the Guibas–Stolfi divide and conquer algorithm from [48]. The displacement vectors point

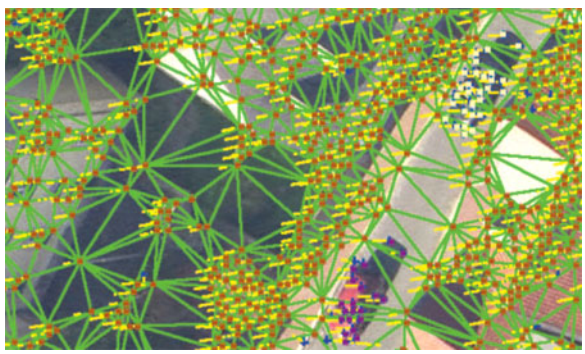

Fig. 5. Triangulated mesh (green triangles) between detected features (brown dots: background features, blue crosses: motion candidates including outlier, purple and white dots: detected MOs after cluster filtering) and trajectories (yellow lines) in the motion compensated destination frame [34]. Best viewed in color.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

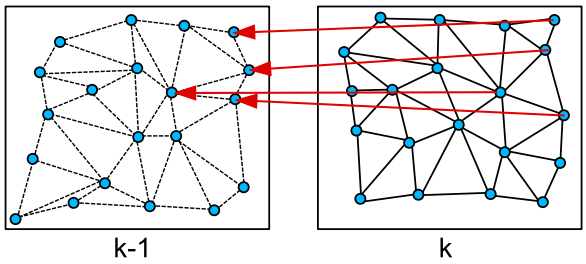

Fig. 6. The Delaunay triangulation of the feature point cloud in frame k creates the mesh. The triangulation is performed in the frame k (right). The displacement vectors point to the frame $k-1$ (left) and define the mesh in the frame $k-1$ (based on [36]).

into frame $k-1$ and define the mesh in that frame. To compensate the motion inside the patches, an individual affine transformation H_i is determined for each of the triangles using the three mesh nodes defining the triangle:

$$\mathrm{H}_i = \mathrm{T}_{i,k-1} \cdot \mathrm{T}_{i,k}^{-1}, \tag{6}$$

wherein $\mathrm{T}_{i,k}$ and $\mathrm{T}_{i,k-1}$ are matrices containing the coordinates of the three mesh nodes of the triangle t_i in the frames k and $k-1$ as homogeneous coordinates in column form. The affine transformation H_i is then applied to each of the pixels of the triangle t_i resulting in a motion compensated frame. Due to the locally adapted motion parameters the difference images contain less falsely detected spots of high energy as MO candidates. Hence, the activation mask is cleared by lots of FPs. As only the motion compensation of the MO detector is modified, no additional information has to be signaled to the decoder.

Since non-planar structures like buildings (with motion parallax) are correctly motion compensated by the mesh-based motion compensation as background, FP detections are largely decreased leading to less blocks to be coded and an increased coding efficiency.

V. EXPERIMENTS

We present detection results of the proposed MO detector as well as coding results for the proposed HEVC-based video encoder in this section.

We define two different test sets. The first set (*Test Set 1*) is used for the evaluation of the proposed MO detector, whereby bit rates are additionally provided. It consists of two self-recorded publicly available video sequences in full HDTV resolution (named after the flight height they were recorded at) [34, 39] and a low resolution, interlaced aerial video sequence with relatively low image quality from the publicly available VIRAT data set [37, 38]. Example frames are printed in Fig. 7. The self-recorded 750 m sequence (Fig. 7(a)) contains lots of houses and cars, most of them are parking, two are moving. An accurate detection and segmentation of the MOs including their shadows as well as previously covered ground is very challenging. The other

Fig. 7. Example frames of the test set used to evaluate the MO detection and ROI coding framework.(*Test Set 1*). (a) MOs (black and red car with shadows) in the 750 m sequence, HDTV resolution, ground resolution: 21 pel/m [34, 39]. (b) MO (white car in the middle) in the 350 m sequence, HDTV resolution, ground resolution: 43 pel/m [34, 39]. (c) MOs (white and red car in the middle) in the *VIRAT* test data set, original resolution: 720 × 480, interlaced [37, 38].

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

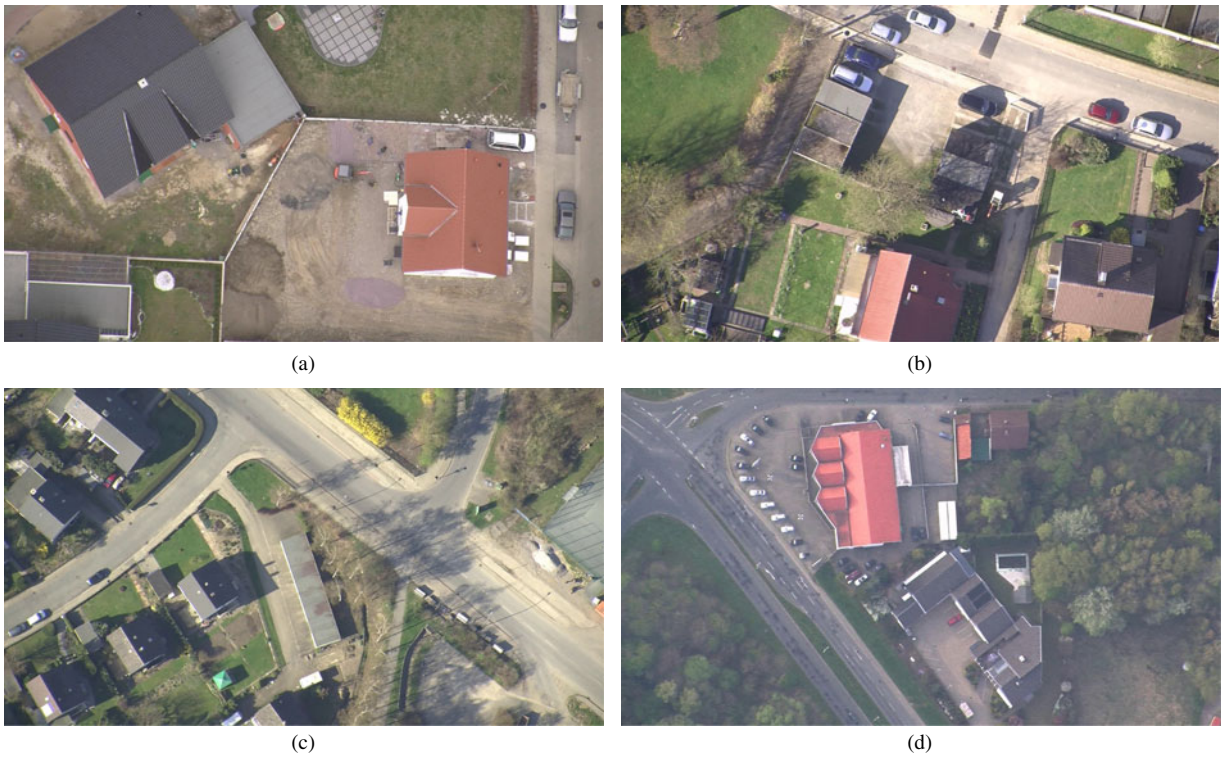

Fig. 8. Test sequences (self-recorded) for coding (*Test Set 2*) [39]. (a) Frame of the 350 m sequence, HDTV resolution, ground resolution: 43 pel/m. (b) Frame of the 500 m sequence, HDTV resolution, ground resolution: 30 pel/m. (c) Frame of the 1000 m sequence, HDTV resolution, ground resolution: 15 pel/m. (d) Frame of the 1500 m sequence, HDTV resolution, ground resolution: 10 pel/m.

self-recorded sequence (350 m sequence, Fig. 7(b)) is much easier to segment since the moving car on the street has a high contrast against the background. Compared to the 750 m sequence only a small number of objects is contained in the 350 m sequence and the latter sequence is less textured overall. We additionally considered the publicly available VIRAT data set (Fig. 7(c)) [37, 38] in order to show that our algorithms also work on low resolution, interlaced video sequences with relatively low overall image quality.

In order to evaluate the maximum coding performance of the video encoder and in absence of more high-resolution aerial video sequences containing MOs, we define a second, publicly available *Test Set 2* [39]. It contains four self-recorded HDTV resolution aerial video sequences with a frame rate of 30 fps. Likewise, the sequences are named after the flight height they were recorded at and each sequence contains between 821 and 1571 frames (Fig. 8).

A) Classification results

To show the performance of our proposed MO detector we will give qualitative and quantitative results for the test sequences from *Test Set 1* (Fig. 7). We use fixed thresholds for our experiments, which were empirically optimized. For the generation of the TCSs we set the compactness weighting factor α to 0.96 (equation (3)) and use five iterations of the expectation- and maximization-step after adding a frame to the observation window with a total length of 15 frames. For the mesh-based cluster filter we set $t_{d1} = 80$ and $t_{d2} = 3.6$ (equations (4) and (5) and a minimum of $t_f = 3$ motion vectors for a motion cluster in the cluster filter) for HDTV resolution and typical flight speeds, whereas we linearly downscaled the thresholds for sequences with smaller resolution (e.g. the VIRAT test sequence). For the MO detector we define true positive (TP) detections as the (pelwise) correct classification of MOs as such compared to a manually labeled reference (ground truth). Similarly, *FP* detections are image pels belonging to static objects falsely classified as moving. In Fig. 9, an example of the 750 m sequence is shown [34]. Figure 9(a) shows a cropped region of the original frame containing a MO, whereas Figs 9(b) and 9(c) depict the corresponding decoded frame (cropped region and whole frame) using the cluster filter and mesh-based motion compensation. The activation masks for the planar GMC-based MO detector including many FP detections at the gable of the building can be seen in Fig. 9(d). Figure 9(e) shows the result of the MO detector improved by the mesh-based cluster filter which removed almost all false detections (we used 3000 feature points as a maximum). The resulting coding masks for ROI-MOs after the superpixel enhancement are shown in Figs 9(f) and 9(g). The reduced FP detections (missing white regions in the left part of the image) for the proposed MO detector approach compared to the GMC MO detection approach assuming only one planar ground (Figs 9(d) and 9(e)) lead to a greatly improved coding mask after superpixel enhancement (Figs 9(f) and 9(g)) for the coding system. The TP detection rate for the moving car including the shadows stays almost the

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

same. Since the entire car (MO) is detected as one, it can be properly reconstructed without errors, which is confirmed in informal subjective viewings. Moreover, nearly no non-moving areas (FPs) are marked for video encoding, resulting in an improved detection accuracy and thereby in a reduced bit rate.

For an objective evaluation of our system we generated *Receiver Operating Characteristics* (ROCs) [49] using the manually labeled ground truth data. To generate the curves we employed different SWWs. This parameter controls to which extent a temporal gap between single MO detections can be bridged by the system and thus has an impacts on the TP and FP rate. As a baseline we included a non context-adaptive approach into the ROC, based on simple thresholding and trivial dilation operations (3×3 structuring element). The activation masks (result of pelwise difference between the images) for every frame were dilated n-times (accordingly labeled as $n\times$ in the ROC curves) before they are used as coding mask. No superpixel enhancement was performed for the baseline case. Consequently, $0 \times$ dilation represents the MO detection rates just for the difference image similar to the detection method from [13, 14] (750 m sequence: TP = 7%, FP = 0%; 350 m sequence: TP = 36.4%, FP = 0%).

The ROC curves are shown in Fig. 10. For the 750 m sequence and a reasonable operation point with SWW = 3 we achieve a FP rate of 1.8% at a very high TP detection rate of about 97.9%. With increasing the SWW, the slope of the ROC curve gets flatter, resulting in a small increased TP detection rate at the cost of an unintentional highly increased FP detection rate. Without the mesh and cluster filter but with superpixel enhancement, i.e. with a MO detector like in [15], the system still achieves reasonably good FP rate of 2.2%. For our proposed system (MO detection from [34]) and with a SWW of nine frames a FP detection rate of 2.8% and a simultaneously increased TP detection rate of 98.6% is achieved (results for the superpixel enhanced system but without mesh and cluster filter are: FP about 3.4%, TP about 97.2%). Note that the SWW is used only for MO detection and thus is completely independent from encoding. We did not investigate longer SWWs since the FP rate would increase dramatically. The segmentation results of the MOs are better for any operating point for the proposed system, consequently the detection accuracy according to equation (7) is increased (from 97.2 up to 98.9% for SWW = 9 and SWW = 1, respectively) in the fully automatic system. Since only relatively small parts (<5%) of one frame are actually MOs, this is a noticeable achievement in terms of bit rate saving.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}, \tag{7}$$

where TP is the number of *True Positive* detections, TN is the number of *True Negative* detections, FP is the number of *False Positive* detections and FN is the number of *False Negative* detections.

For the 350 m sequence both the TP as well as the FP detection rates were highly increased compared to a simple dilation approach. Since after the superpixel integration, but without the mesh-based locally adaptive motion model, the TP detection rate already was between 99.3% (SWW = 1) and 100% for a SWW greater than 1 and no FPs caused by model violations were detected (Fig. 10(b)), no improvement in terms of detection accuracy was possible by introducing the mesh/cluster filter.

For the VIRAT sequence (Fig. 7(c)) both systems compared (GMC+dilation, GMC+SP) mainly fail to segment

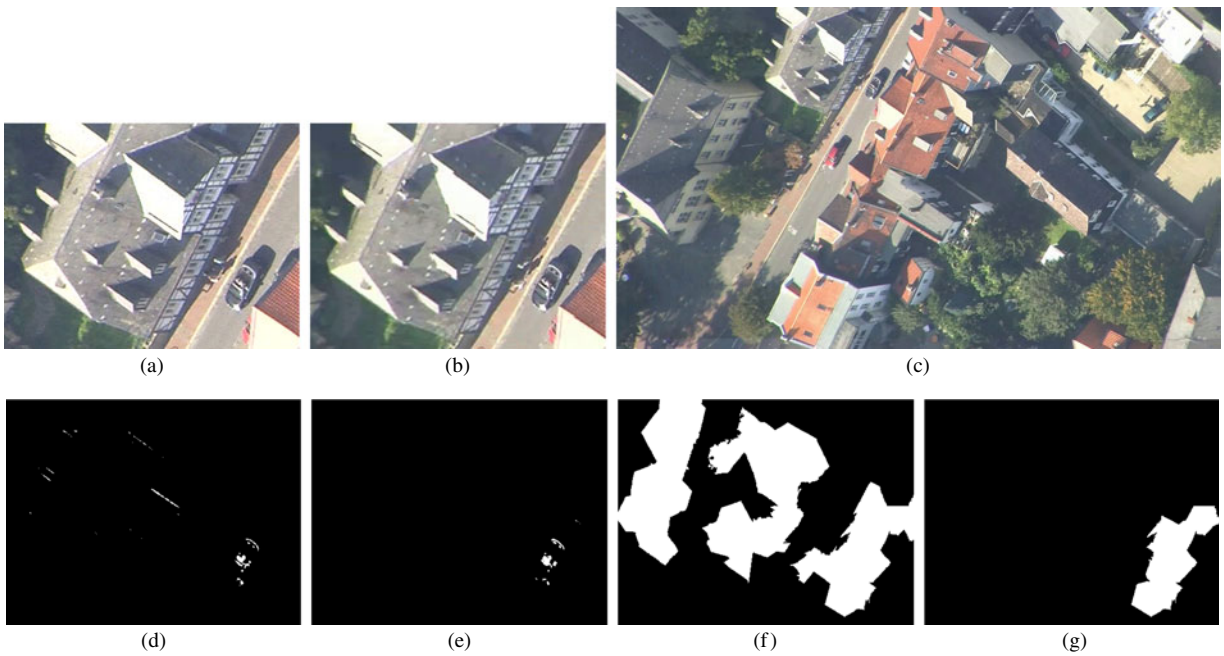

Fig. 9. MO detections (d, e) and coding masks including superpixel (SP) enhancement (f, g) for the GMC-based (d, f) and the CF-based (e, g) MO detector. Panel (a) shows the original frame and (b,c) the decoded result [34]. (a) Original frame (cropped). (b) Decoded (CF+Mesh+SP, cropped). (c) Decoded (CF+Mesh+SP, whole frame). (d) GMC activation mask. (e) CF activation mask. (f) GMC+SP coding mask. (g) CF+SP coding mask.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

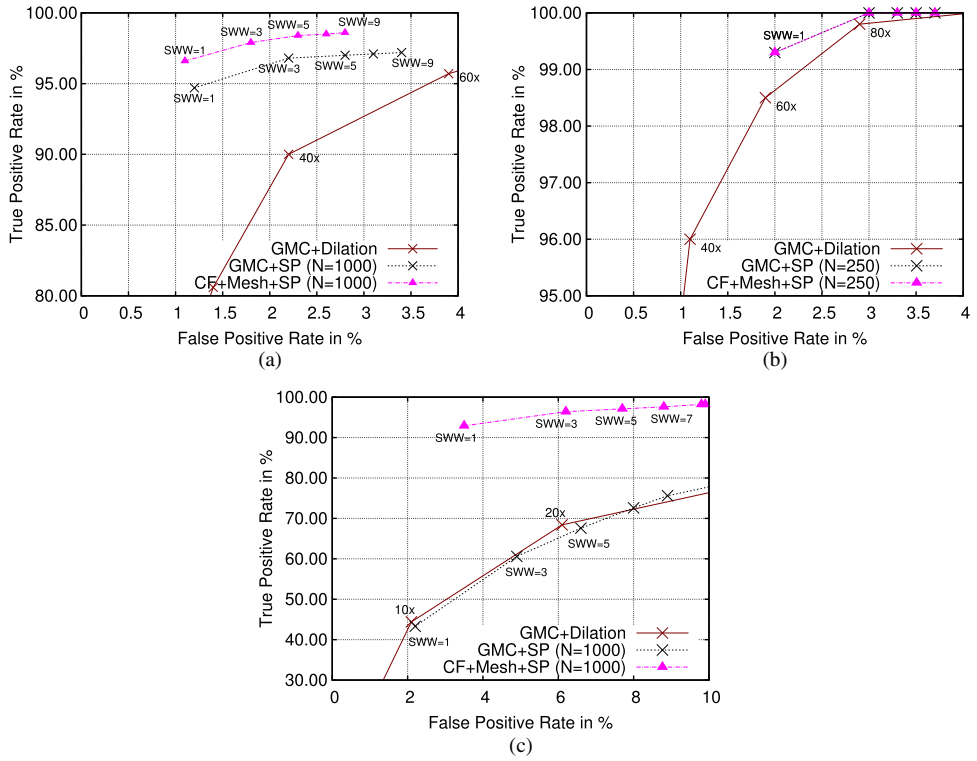

Fig. 10. Receiver Operating Characteristics (ROCs) for (a) 750 m sequence, (b) 350 m sequence, (c) VIRAT test scene (TP rate calculated pel-wise, SP = Superpixel, CF = Cluster Filter, SWW = Sliding Window Width, N = No. of superpixels used for image segmentation, Difference Image only is 0× Dilation (not in the figures): for 750 m sequence: TP = 7%, FP = 0%; for 350 m sequence: TP = 36.4%, FP = 0%; for VIRAT: TP = 6%, FP = 0%).

MOs at a reasonable low FP detection rate (less than 80% TP detection rate at approximately 10% FP rate). It can be seen in Fig. 10(c) that our system including cluster filtering/mesh-based motion compensation and superpixel enhancement performs best with 92.9% TP detection rate at 3.5% FP detection rate. This even holds for the case when no temporally consistency of the superpixel segmentation is exploited (operating point SWW = 1). For a SWW = 3 we reach a TP detection rate of 96.4% at a FP detection rate of 6.2%.

B) Coding results and image quality evaluation

We used a modified *x*264 (v0.78) [50] AVC-encoder [5] at *High Profile* (Level 4.0) – further on referred to as AVC-skip – as a reference video coder using a fixed *quantization parameter* (QP). *x*264 has a coding performance similar to the JM reference software [51, 52] at similar settings but is significantly faster due to software optimizations, e.g. its multi-threading capability. The QP setting itself influences the image quality of the ROI blocks and consequently the resulting image quality after the decoding.

According to informal subjective tests (ten persons, using *Mean Opinion Scores*, MOSs, 0 = worst, 5 = best), the perceived image quality after the video decoding and reconstruction remains very high over the entire image as expected. For the MOS evaluation, the original frame as recorded by the camera was used as the hidden reference and common AVC was used as a low-quality anchor. Using our proposed ROI detection and coding system with a modified HEVC video codec ("ROI HEVC") instead of a common unmodified HEVC codec, the MOS values are significantly increased by 0.7 and 0.8 up to 3.6 and 3.8 for 300 and 500 kbit/s, respectively. As can be seen in the magnifications in Fig. 11, common AVC is not able to produce a high image quality for bit rates equal to or lower than 500 kbit/s (Fig. 11(d)). Especially for very low bit rates below 500 kbit/s, our ROI-based system (Fig. 11(e)) retains much more high-frequency details (e.g. at a bit rate of 150 kbit/s like shown) resulting in a perceptively higher image quality compared to HEVC (Fig. 11(f)). However, due to the GMC of the background, small discontinuities at non-planar structures reconstructed from different *New Areas* might occur as can be seen, e.g. in Fig. 9(b) at the gable or in Fig. 11(b) at the upper right house roof. Although our test sequences were recorded in hilly terrain, which violates the planarity assumption, we were always able to reliably estimate the global motion of the scene and thus to reconstruct the video sequences by means of GMC.

For an objective evaluation, we compare the results of AVC-skip to our proposed HEVC-skip implementation based on HM 10.0 [53] – called HEVC-skip – at *low delay-* (LD), *Low Delay-P-* (LD-P), and *random access-* (RA) profile-based settings with modified maximum block sizes (*Coding Tree Units*, CTUs, formerly known as *argest Coding Units*, LCUs) of 16 × 16 or 64 × 64 and smallest block sizes of 4 × 4 each. The generated bit streams are

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

Fig. 11. Subjective image quality comparison for different video codecs and different very low bit rates (350 m sequence, 150–500 kbit/s). (f) ROI HEVC is proposed. Best viewed in pdf. (a) Original frame (whole frame). (b) ROI HEVC en- and decoded (whole frame, 300 kbit/s). (c) Original. (d) AVC 500 kbit/s. (e) HEVC 150 kbit/s. (f) ROI HEVC 150 kbit/s.

decodable with the HEVC compliant reference decoder HM 16.2. Apart from a modified maximum block size (and the corresponding partition depth resulting in 4×4 blocks) we applied the settings defined in the default HM configuration files `encoder_lowdelay_main.cfg`, `encoder_lowdelay_P_main.cfg` and `encoder_randomaccess_main.cfg`. Our configuration details are listed in Table 1. For the RA profile, an intra (I) frame period of 32 was selected, whereas for the LD-P (only using predicted (P) frames as inter frames) and LD (containing bi-predicted/B frames as inter frames) profile only the first frame of a sequence was encoded in intra mode.

However, since with the new area of each frame, which is often intra coded anyway, there is a kind of "rolling intra frame". Thus, theoretically, there is no need for the transmission of any intra frame at all since the decoder just has to wait for the next intra blocks within the new areas in order to continue decoding. Consequently, the highly efficient LD profile might be a good choice for scenarios with a demand of highest coding performance.

For our modified HM implementation (HEVC-skip) the *skip mode* was forced for non-ROI areas, whereas *non*-skip mode was forced for ROI areas (i.e. intra/any other inter mode than skip, PCM prohibited by configuration).

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

Table 1. Configuration settings of the HEVC encoder for *Low Delay* (LD), *Low Delay-P* (LD-P) and *Random Access* (RA), based on the common HM configuration files `encoder_lowdelay_main.cfg`, `encoder_lowdelay_P_main.cfg` and `encoder_randomaccess_main.cfg`.

	LD	LD-P	RA
Unit definition			
Max CU width	64 × 64 or 16 × 16	64 × 64	64 × 64 or 16 × 16
Max CU Height	64 × 64 or 16 × 16	64 × 64	64 × 64 or 16 × 16
Max partition depth	4 or 2	4	4 or 2
Log2 of maximum transform size quadtree-based TU coding	5 or 3	4	5 or 3
Log2 of minimum transform size quadtree-based TU coding	2	2	2
Quadtree TU max depth inter	3	3	3
Quadtree TU max depth intra	3	3	3
Coding structure			
Intra period	Only first	Only first	32
Decoding refresh type	0	0	1
GOP size	4	4	8
Motion search			
Fast search	TZ search	TZ search	TZ search
Search range	64	64	64
Bipred search range	4	4	4
Hadamard ME	1	1	1
Fast encoder decision	1	1	1
Fast decision for merge RD cost	1	1	1
Quantization			
QP	24–35	24–35	24–35
Max delta QP	0	0	0
Max Cu DQP depth	0	0	0
Delta Qp RD	0	0	0
RDOQ	1	1	1
RDOQTS	1	1	1
Coding tool			
SAO	1	1	1
AMP	1	1	1
Transform skip	1	1	1
Transform skip fast	1	1	1
SAOLcuBoundary	0	0	0
Misc.			
Deblocking	On (defaults)	On (defaults)	On (defaults)
Internal bit depth	8	8	8
Slices	0	0	0
PCM	0	0	0
Tiles	0	0	0
Wave front	0	0	0
Quant. Matrix – scaling list	0	0	0
Lossless	0 (all)	0 (all)	0 (all)
Rate control	0	0	0

Rate distortion (RD) plots are printed for different encoders (AVC, AVC-skip, HEVC, and HEVC-skip) and a maximum coding block size of 16 × 16 in Fig. 12. For the PSNR calculation we only considered luminance values within ROI areas. Similar evaluations can be found, e.g. in [54, 55]. Errors introduced by GMC, e.g. caused by parallax, are assumed to be irrelevant as they influence the perceptional quality of the background only marginally and much less than a coarse quantization over the entire image. For the AVC-skip encoder at QP = 33 – corresponding to a reconstructed Y-PSNR "video quality" of about 35 and 32 dB for the 350 m sequence and the 750 m sequence, respectively – we see a bit rate saving of about 80% compared to the unmodified non-ROI AVC coder which can be found in the same magnitude all over the RD plot. Similar findings hold true for HEVC-skip compared to the unmodified HEVC. The red arrows in the RD plots emphasize similar Y-PSNR quality levels comparing an unmodified HEVC encoder to the HEVC-skip system. Employing the HEVC-skip encoder, additional coding gains (Bjøntegaard delta, BD, BD-rate and BD-PSNR, cubic interpolation, QP range: 24–35 [56, 57]) according to Table 2 can be achieved, e.g. for inter frames up to 33.2% for the high-resolution sequences and about 65% for the low-resolution sequence, corresponding to BD-PSNR gains of up to 1.7 dB (high-resolution sequences) and 3.46 dB (low-resolution sequence), respectively. Subjectively sufficient quality can be provided at a bit rate below 2 Mbit/s for each of these sequences, especially including the HDTV resolution sequences. It is noteworthy, that the actual coding gains of the inter predicted frames (which basically were examined and improved in this work) is about 28.8 and 33.2% for the 750 m sequence and the

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

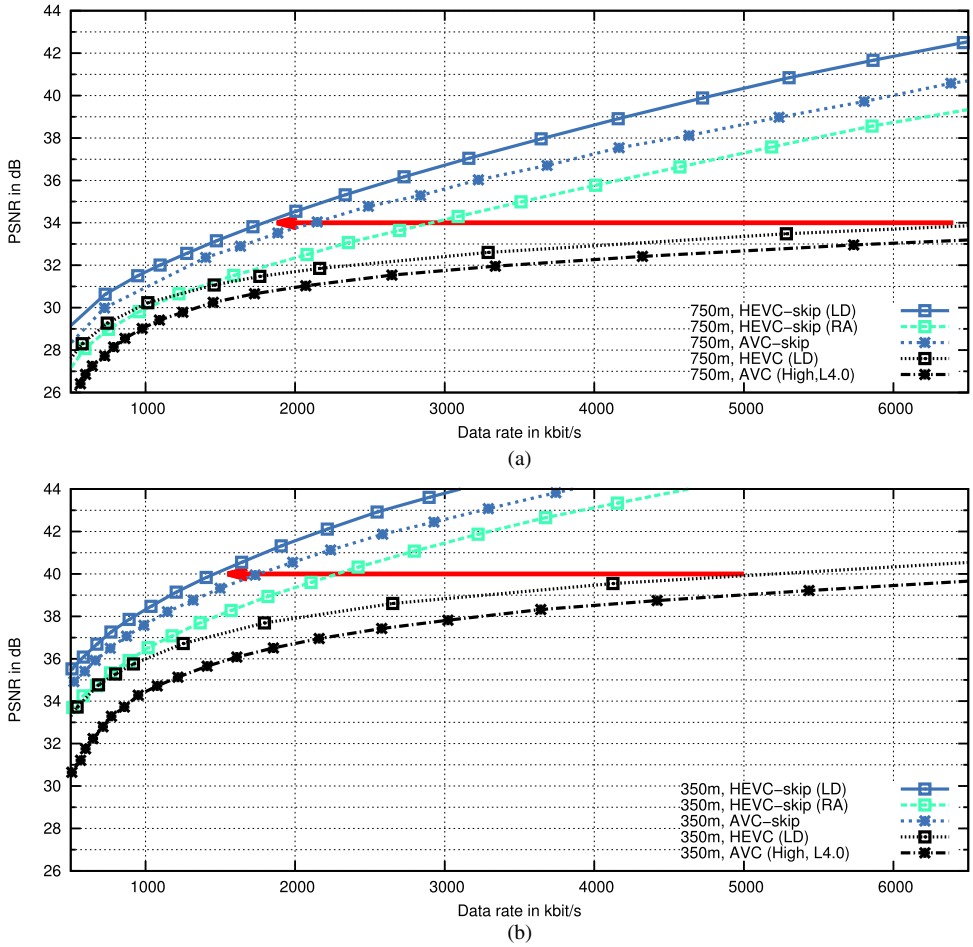

Fig. 12. RD diagrams for two test sequences from *Test Set 1* (stars: AVC/AVC-skip, squares: HEVC/HEVC-skip; black: common, unmodified encoders, green: RA-alike, blue: LD-alike, maximum block sizes: 16 × 16 each, minimum block size: 4 × 4 each), red arrows emphasize bit rate saving of HEVC-skip system compared to unmodified HEVC coding. (a) 750 m sequence. (b) 350 m sequence.

Table 2. Bjøntegaard delta (BD, BD rate, cubic, QP range: 24–35 and BD-PSNR) [56, 57] for *Test Set 1*, negative BD-rate numbers represent coding gains of the proposed HEVC-skip coding system over the AVC-skip ROI coding system. "All" represents total gains over the entire sequence, whereas "Inter only" represents BD gains only for inter predicted frames, based on 16 × 16 (CTU16) and 64 × 64 (CTU64) largest coding block size for HEVC.

	750 m sequence		**350 m sequence**		**VIRAT test sequence**	
	All	Inter only	All	Inter only	All	Inter only
BD-rate, CTU16 (%)	−17.2	−23.3	−16.1	−19.3	−47.6	−65.0
BD-PSNR, CTU16 (dB)	0.90	1.14	0.74	0.91	2.22	3.59
BD-rate, CTU64 (%)	−20.2	−28.8	−26.6	−33.2	−48.1	−64.8
BD-PSNR, CTU64 (dB)	1.06	1.7	1.25	1.6	2.25	3.46

350 m sequence, respectively. Whereas the latter sequence contains low noise and is easy to encode, for the former sequence neither an unmodified AVC nor the unmodified HEVC encoder will reach more than 35 dB Y-PSNR for bit rates smaller than 6500 kbit/s. Although the BD gains of the modified HM 10.0 encoder compared to the *x*264 encoder are smaller than reported in the literature [58] they approximately reflect the coding gains of 20–30% from an unmodified HEVC compared to an unmodified AVC.

For the very low resolution VIRAT test sequence we need a bit rate of less than 500 kbit/s with our system compared to 1760 and 1580 kbit/s (at approximately 40 dB) for AVC and HEVC encoding, respectively.

For a reasonable operation point at SWW = 3 we reduce the bit rate for the transmission of detected MOs by more than 24% compared to a MO detector relying on GMC for full HDTV video sequences at 30 fps. Including *New Areas*, the bit rate decreases only by 4%.

Coding results for the second test set containing no MOs (Fig. 8) using the modified HEVC video encoder (HM-skip) are provided in Table 3. Using the AVC-skip bit rates and the corresponding ROI Y-PSNR "qualities" as the anchor, we adjusted the QP for the competitors to match the quality as closely as possible. Finally, we interpolated the bit rate linearly in order to match the desired PSNR exactly. This linear interpolation is justified when looking at the

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

Table 3. Coding gains (negative numbers) for *Test Set 2* of proposed HEVC-based over AVC-based ROI coding system compared to the reference (Ref.) as marked in the table column by column. AVC and HEVC bit rates without ROI coding are additionally given (LD configurations-based with modified block-size according to the table, minimum MB/CU size = 4 × 4).

Coder	MB/CTU size (pel)	Bit rate (kbit/s)	Diff. (%)	Diff. (%)	Diff.(%)
(a) 350 m sequence, 43 pel/m, 821 frames, PSNR ≈ 38.9 dB					
AVC	16 × 16	9287	Ref.	–	–
HEVC (LD)	16 × 16	6489	−30.1	–	–
HEVC (LD-P)	64 × 64	7443	−19.9	–	–
HEVC (LD)	64 × 64	5568	−40.0	Ref.	–
AVC-skip	16 × 16	943	−89.9	−83.1	Ref.
HEVC-skip (LD)	16 × 16	715	−92.3	−87.2	−24.2
HEVC-skip (LD-P)	64 × 64	590	−93.6	−89.4	−37.4
HEVC-skip (LD)	64 × 64	**579**	−93.8	**−89.6**	−38.6
(b) 500 m sequence, 30 pel/m, 1121 frames, PSNR ≈ 37.2 dB					
AVC	16 × 16	11491	Ref.	–	–
HEVC (LD)	16 × 16	8973	−21.9	–	–
HEVC (LD-P)	64 × 64	11194	−2.6	–	–
HEVC (LD)	64 × 64	7947	−30.8	Ref.	–
AVC-skip	16 × 16	1423	−87.6	−82.1	Ref.
HEVC-skip (LD)	16 × 16	1020	−91.1	−87.2	−28.3
HEVC-skip (LD-P)	64 × 64	885	−92.3	−88.9	−37.8
HEVC-skip (LD)	64 × 64	**872**	−92.4	**−89.0**	−38.7
(c) 1000 m sequence, 15 pel/m, 1166 frames, PSNR ≈ 37.7 dB					
AVC	16 × 16	10337	Ref.	–	–
HEVC (LD)	16 × 16	7243	−29.9	–	–
HEVC (LD-P)	64 × 64	10028	−3.0	–	–
HEVC (LD)	64 × 64	5849	−43.4	Ref.	–
AVC-skip	16 × 16	1153	−88.8	−80.3	Ref.
HEVC-skip (LD)	16 × 16	865	−91.6	−85.2	−25.0
HEVC-skip (LD-P)	64 × 64	796	−92.3	−86.4	−31.0
HEVC-skip (LD)	64 × 64	**762**	−92.6	**−87.0**	−33.9
(d) 1500 m sequence, 10 pel/m, 1571 frames, PSNR ≈ 37.6 dB					
AVC	16 × 16	13560	Ref.	–	–
HEVC (LD)	16 × 16	11942	−11.9	–	–
HEVC (LD-P)	64 × 64	12470	−8.0	–	–
HEVC (LD)	64 × 64	11901	−12.2	Ref.	–
AVC-skip	16 × 16	967	−92.9	−91.9	Ref.
HEVC-skip (LD)	16 × 16	743	−94.5	−93.8	−23.2
HEVC-skip (LD-P)	64 × 64	644	−95.3	−94.6	−33.4
HEVC-skip (LD)	64 × 64	**634**	−95.3	**−94.7**	−34.4

RD curves in Fig. 12 at around 37–39 dB. Results are provided for the LD-P and LD profile at different block sizes. Bit rates for the LD-P profile are less than 10% lower for our test set than the achieved AVC bit rates, already considering the larger maximum coding block size of 64 × 64 for HEVC compared to 16 × 16 for AVC. In contrast to that for our skip-implementations the LD-P coding gains of more than 34% are comparably high. To achieve the best coding efficiency, we recommend the LD profile including bi-prediction for inter frames with a CTU size of 64 × 64. With this profile and the proposed HM-skip video encoder we are able to provide bit rates between 579 kbit/s (350 m sequence, 38.9 dB) and 872 kbit/s (500 m sequence, 37.2 dB), depending on the sequence characteristic, which is a bit rate saving of 35.1% compared to AVC-skip or 90.0% compared to common HEVC without any modifications. As already mentioned, the bit rate in a real system depends on the amount (and distribution) of ROIs to be encoded, which is typically 5–10% of a frame in our tests.

C) Run-time considerations

Since our proposed detection and coding system aims at real-time processing on-board an UAV, we consider the run-time for full HDTV resolution sequences recorded at 30 fps of our non-optimized C/ C++ and Matlab code on a typical modern desktop PC with an Intel Core i7-3770K CPU at a clock rate of 3.5 GHz running Linux. Except for parts of the superpixel segmentation all run-times are measured in single-thread processing, albeit parallelization of the components is able to decrease the run-time significantly. No hardware acceleration like GPU processing was applied yet.

The run-times for each component are listed in Table 4. It is obvious that the HEVC-skip video encoder consumes by far most of the time. Our experiments were carried out using a modified version of the reference software HM, which is used in the standardization process. A hardware HEVC encoder will even be capable of real-time processing of full HDTV bit streams at low power consumption and low cost.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

Table 4. Run-times per frame, non-optimized components written in C/C++, single thread execution on CPU (no GPU or other hardware acceleration), PC with an Intel Core i7-3770K CPU, clock rate of 3.5 GHz.

Component	Run-time per frame (ms)
KLT	61.0
RANSAC	27.5
Cluster filter	37.5
Mesh-based motion compensation	280.0
Difference image calculation	10.0
Superpixel segmentation	900.0
MO detection + block selection	25.0
Video encoding	8415.9
Total	9756.9

The Matlab-based superpixel segmentation consumes second-most of the run-time in the entire processing chain. Our Matlab implementation can segment a full HDTV resolution image into 1000 superpixels on our PC in about 900 ms. In [59] a much faster C++ implementation was proposed which could be additionally sped-up by employing the GPU (using the NVIDIA CUDA framework and a NVIDIA GTX460 graphic card) by a factor of 20 compared to the sequential algorithm run on an Intel Core i7-2600 (3.6 GHz) CPU. Aiming at a real-time application, we also prepare an optimized C/ C++ implementation.

The third most computational burden is generated by the mesh-based motion compensation which can be easily parallelized on a triangle basis. Consequently, it can be realized in real-time on a (small) GPU.

The OpenCV [60] KLT implementation as the fourth most consumer of run-time is able to process full HDTV content by tracking 3000 features as a maximum in a single thread at about 16 fps which means two CPU cores can easily process the video sequence for 30 fps sequences in real-time. More efficient KLT implementations like proposed in [61] might reach an even higher computational efficiency and thus shorter run-time.

Our cluster filter can process HDTV video sequences in (nearly) real-time. Whereas the run-time of RANSAC depends on the percentage of outliers and the number of iterations until the consensus is reached [62], it only consumes about 2.8% of the entire processing time in our tests and thus is real-time capable. The run-times of the remaining components, like the new area calculation based on the projective transformation parameters, are negligible in the entire processing chain.

For the proposed system we need a total run-time of nearly 10 s for the processing of each frame in full HDTV resolution on a single CPU core which equals a processing with about 0.1 fps. Assuming the usage of an HEVC IP core and the sped-up superpixels we should be able to already process at least 2 fps in software. Further algorithmic optimizations might include the usage of the sparse optical flow from KLT as well for the superpixel segmentation and the usage of background feature points for global motion estimation directly from the cluster filter while omitting RANSAC completely (which is a valid simplification for predominantly planar scenes). Taking into account the above software optimizations and the usage of parallel processing or even dedicated hardware like FPGAs or HEVC encoders, our proposed system can easily become real-time capable. Since power consumption and form factor restrictions apply on-board an UAV, the usage of dedicated hardware is advisable anyway.

VI. CONCLUSIONS

We present an aerial surveillance video coding system which can provide very low bit rates maintaining full image quality over the entire image. GMC is employed to reconstruct the background at the decoder side from already transmitted images. New areas contained in the current but not in the previous frame as well as MOs and previously covered background are transmitted. In order to limit the bit rate, it is crucial – especially for surveillance applications – to accurately detect new area and MOs. Therefore, non-moving regions falsely detected as moving have to be avoided to keep the bit rate as low as possible.

To decrease the *FP* detection rate we propose to replace the GMC by a mesh-based locally adaptive multiplanar approach within the MO detector. The mesh-based approach is capable of modeling distinct 3D structures more precisely. A cluster filter is introduced to distinguish between background motion and MOs based on an optical flow analysis. The reduced model aberrations lead to a decreased FP detection rate.

Since the MO detector is not able to accurately detect the shapes of MOs leading to reconstruction errors when not entire MOs are transmitted, we use an independently calculated, context-adaptive, TCS segmentation to increase the *TP* detection rate of the system.

Combining the superpixel segmentation and the mesh-based motion compensation, we are able to achieve a FP detection rate of only 1.8% while simultaneously increasing the TP detection rate to 97.9% (for a reasonable operating point) for challenging sequences. For the interlaced, low-resolution test sequence from the publicly available VIRAT data set we are able to detect 96.4% TPs at a FP detection rate of 6.2%.

Our final contribution is the integration of a modified HEVC encoder (employing the skip-mode by external control) into the coding system. In order to make the entire processing chain real-time capable for on-board usage at small and mid-size UAVs, optimized and hardware-accelerated algorithms are in preparation. Compared to a similarly modified AVC video encoder we gain an additional 30% (BD rate) or an equivalent of 1.65 dB (BD PSNR) for inter frames with the proposed HEVC-skip encoder for high-quality HDTV resolution aerial sequences (30 fps) and even more for lower resolution sequences. Typical aerial sequences containing MOs can be encoded at bit rates far below 2 Mbit/s. Compared to an unmodified HEVC encoder, we achieve a much higher image quality for very low bit rates (150–500 kbit/s).

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

REFERENCES

[1] Ciubotaru, B.; Muntean, G.; Ghinea, G.: Objective assessment of region of interest-aware adaptive multimedia streaming quality. IEEE Trans. Broadcast., **55** (2) (2009), 202–212.

[2] Karlsson, L.; Sjöström, M.; Olsson, R.: Spatio-temporal filter for ROI video coding, in *Proc. of the 14th European Signal Processing Conf. (EUSIPCO)*, September 2006, 1–5.

[3] Doulamis, N.; Doulamis, A.; Kalogeras, D.; Kollias, S.: Low bit-rate coding of image sequences using adaptive regions of interest. IEEE Trans. Circuits Syst. Video Technol., **8** (8) (1998), 928 –934.

[4] Chen, M.-J.; Chi, M.-C.; Hsu, C.-T.; Chen, J.-W.: ROI video coding based on H.263+ with robust skin-color detection technique. IEEE Trans. Consum. Electron., **49** (3) (2003), 724–730.

[5] AVC: Recommendation ITU-T H.264 and ISO/IEC 14496-10 (MPEG-4 Part 10): Advanced Video Coding (AVC), 3rd ed., *ISO/IEC and ITU-T*, Geneva, Switzerland, 2004.

[6] HEVC: ITU-T Recommendation H.265/ ISO/IEC 23008-2:2013 MPEG-H Part 2/: High Efficiency Video Coding (HEVC), 2013.

[7] Liu, Y.; Li, Z.G.; Soh, Y.C.: Region-of-interest based resource allocation for conversational video communication of H.264/AVC. IEEE Trans. Circuits Syst. Video Technol., **18** (1) (2008), 134–139.

[8] Wu, C.-Y.; Su, P.-C.; Yeh, C.-H.; Hsu, H.-C.: A joint content adaptive rate-quantization model and region of interest intra coding of H.264/AVC, in *IEEE International Conf. on Multimedia and Expo (ICME)*, July 2014, 1–6.

[9] Liu, Y.; Li, Z.; Soh, Y.; Loke, M.: Conversational video communication of H.264/AVC with region-of-interest concern, in *IEEE Int. Conf. on Image Processing (ICIP)*, October 2006, 3129–3132.

[10] Wu, C.-Y.; Su, P.-C.: A region of interest rate-control scheme for encoding traffic surveillance videos, in *Fifth Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, September 2009, 194–197.

[11] Xing, P.; Tian, Y.; Huang, T.; Gao, W.: Surveillance video coding with quadtree partition based ROI extraction, in *Proc. of the IEEE Picture Coding Symposium (PCS)*, December 2013, 157–160.

[12] Meddeb, M.; Cagnazzo, M.; and Pesquet-Popescu, B.: Region-of-interest based rate control scheme for high efficiency video coding, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, 7338–7342.

[13] Meuel, H.; Munderloh, M.; Ostermann, J.: Low bit rate ROI based video coding for HDTV aerial surveillance video sequences, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition – Workshops (CVPRW)*, June 2011, 13 –20.

[14] Meuel, H.; Schmidt, J.; Munderloh, M.; Ostermann, J.: Advanced Video Coding for Next-Generation Multimedia Services – Chapter 3: Region of Interest Coding for Aerial Video Sequences Using Landscape Models, *Intech*, 2013. [Online]. Available at: http://www.intechopen.com/books/advanced-video-coding-for-next-generation-multimedia-services/region-of-interest-coding-for-aerial-video-sequences-using-landscape-models.

[15] Meuel, H.; Reso, M.; Jachalsky, J.; Ostermann, J.: Superpixel-based segmentation of moving objects for low-complexity surveillance systems, in *Proc. of the 10th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, August 2013, 395–400.

[16] Terrillon, J.-C.; David, M.; Akamatsu, S.: Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments, in *Proc. of the Third IEEE Int. Conf. on Automatic Face and Gesture Recognition*, April 1998, 112 –117.

[17] Sakaino, H.: Video-based tracking, learning, and recognition method for multiple moving objects. IEEE Trans. Circuits Syst. Video Technol., **23** (10) (2013), 1661–1674.

[18] Dey, B.; Kundu, M.: Robust background subtraction for network surveillance in H.264 streaming video. IEEE Trans. Circuits Syst. Video Technol., **23** (10) (2013), 1695–1703.

[19] Bang, J.; Kim, D.; Eom, H.: Motion object and regional detection method using block-based background difference video frames, in *Proc. of the 18th IEEE Int. Conf. on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, August 2012, 350 –357.

[20] Zhang, X.; Tian, Y.; Huang, T.; Dong, S.; Gao, W.: Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling. IEEE Trans. Image Process., **23**, (10) (2014), 4511–4526.

[21] Jones, R.; Ristic, B.; Redding, N.; Booth, D.: Moving target indication and tracking from moving sensors, in *Proc. of Digital Image Comput.: Techniques and Application (DICTA)*, December 2005, 46.

[22] Shastry. A.; Schowengerdt, R.: Airborne video registration and traffic-flow parameter estimation. IEEE Trans. Intell. Transp. Syst., **6** (4) (2005), 391–405.

[23] Cao, X.; Lan, J.; Yan, P.; and Li, X.: KLT feature based vehicle detection and tracking in airborne videos, in *Sixth Int. Conf. on Image and Graphics (ICIG)*, August 2011, 673–678.

[24] Ibrahim, A.; Ching, P.W.; Seet, G.; Lau, W.; Czajewski, W.: Moving objects detection and tracking framework for UAV-based surveillance, in *Fourth Pacific-Rim Symp. on Image and Video Technology (PSIVT)*, November 2010, 456–461.

[25] Kang, J.; Cohen, I.; Medioni, G.; Yuan, C.: Detection and tracking of moving objects from a moving platform in presence of strong parallax, in *Tenth IEEE Int. Conf. on Computer Vision (ICCV)*, vol. **1**, October 2005, 10–17.

[26] Yalcin, H.; Hebert, M.; Collins, R.; Black, M.: A flow-based approach to vehicle detection and background mosaicking in airborne video, in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. **2**, June 2005, 1202.

[27] Teutsch, M.; Kruger, W.: Detection, segmentation, and tracking of moving objects in UAV videos, in *Proc. of the IEEE Ninth Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, September 2012, 313–318.

[28] Kumar, R. *et al.*: Aerial video surveillance and exploitation. Proc. IEEE, **89** (10) (2001), 1518–1539.

[29] Teutsch, M.: Moving Object Detection and Segmentation for Remote Aerial Video Surveillance. Ph.D. dissertation, Karlsruhe Institute of Technology (KIT), Germany, 2014.

[30] Mundhenk, T.N.; Ni, K.-Y.; Chen, Y.; Kim, K.; Owechko, Y.: Detection of unknown targets from aerial camera and extraction of simple object fingerprints for the purpose of target reacquisition, in *Proc. SPIE*, vol. **8301**, 2012, 83 010H–83 010H–14. [Online]. Available at: http://dx.doi.org/10.1117/12.906491.

[31] Comaniciu, D.; Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell., **24** (5) (2002), 603–619.

[32] Xiao, J.; Cheng, H.; Feng, H.; Yang, C.: Object tracking and classification in aerial videos, in *Proc. of the SPIE Automatic Target Recognition XVIII*, vol. **6967**, 2008, 696 711–696 711–9. [Online]. Available at: http://dx.doi.org/10.1117/12.777827.

[33] Reso, M.; Jachalsky, J.; Rosenhahn, B.; Ostermann, J.: Temporally consistent superpixels, in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, December 2013, 385–392.

[34] Meuel, H.; Munderloh, M.; Reso, M.; Ostermann, J.: Optical flow cluster filtering for ROI coding, in *Proc. of the Picture Coding Symp. (PCS)*, December 2013, 129–132.

[35] Munderloh, M.: Detection of Moving Objects for Aerial Surveillance of Arbitrary Terrain. Ph.D. dissertation, Leibniz Universität Hannover, Germany, 2015.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

[36] Munderloh, M.; Meuel, H.; Ostermann, J.: Mesh-based global motion compensation for robust mosaicking and detection of moving objects in aerial surveillance, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2011, 1–6.

[37] Defense Advanced Research Projects Agency (DARPA): VIRAT Video Dataset, 2009. [Online]. Available at: http://www.viratdata.org/.

[38] Oh, S. *et al.*: A large-scale benchmark dataset for event recognition in surveillance video, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2011, 3153–3160.

[39] Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover: TNT Aerial Video Testset (TAVT), 2010–2014. [Online]. Available at: https://www.tnt.uni-hannover.de/project/TNT_Aerial_Video_Testset/.

[40] Harris, C.; Stephens, M.: A combined corner and edge detection, in *Proc. of the Fourth Alvey Vision Conf.*, 1988, 147–151.

[41] Tomasi, C.; Kanade, T.: Detection and Tracking of Point Features, Carnegie Mellon University, Technical Report, CMU-CS-91-132, April 1991.

[42] Shi, J.; Tomasi, C.: Good features to track, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, June 1994, 593–600.

[43] Fischler, M.A.; Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, **24** (6) (1981), 381–395. [Online]. Available at: http://dx.doi.org/10.1145/358669.358692.

[44] Scheuermann, B.; Rosenhahn, B.: SlimCuts: graphcuts for high resolution images using graph reduction, in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, ser. Lecture Notes in Computer Science (Y. Boykov, F. Kahl, V. Lempitsky, F.R. Schmidt, eds), 219–232, *Springer*, Berlin, Heidelberg, vol. **6819**, 2011. [Online]. Available at: http://dx.doi.org/10.1007/978-3-642-23094-3_16.

[45] Ren, X.; Malik, J.: Learning a classification model for segmentation, in *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2003, 10–17.

[46] Munderloh, M.; Klomp, S.; Ostermann, J.: Mesh-based decoder-side motion estimation, in *Proc. of the IEEE Int. Conf. on Image Processing (ICIP)*, September 2010, 2049–2052.

[47] Dwyer, R.A.: A faster divide-and-conquer algorithm for constructing delaunay triangulations. Algorithmica, **2** (1–4) (1987), 137–151.

[48] Guibas, L.J.; Stolfi, J.: Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams, in *Proc. of the 15th Annual ACM Symp. on Theory of Computing*, ser. STOC, *ACM*, New York, NY, USA, 1983, 221–234. [Online]. Available at: http://doi.acm.org/10.1145/800061.808751.

[49] Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett., **27** (8) (2006), 861–874. [Online]. Available at: http://dx.doi.org/10.1016/j.patrec.2005.10.010.

[50] VideoLAN Organization: *x264*, 2009. [Online]. Available at: http://www.videolan.org/developers/x264.html.

[51] Joint Video Team (JVT) of ISO/IEC MPEG & ITU: H.264/14496-10 AVC Reference Software (JM), 2009. [Online]. Available at: http://iphome.hhi.de/suehring/tml/.

[52] Tourapis, A.M.; Leontaris, A.; Sühring, K.; Sullivan, G.: H.264/14496-10 AVC reference software manual, in *Joint Video Team Doc. JVT-AE010, 31th Meeting*, London, UK, July 2009.

[53] Kim, I.-K.; McCann, K.; Sugimoto, K.; Bross, B.; Han, W.-J.: High effic. video coding (HEVC) test model 10 (HM10) encoder description, in *JCT-VC Doc. JCTVC-L1002*, Geneva, Switzerland, January 2013.

[54] Grois, D.; Hadar, O.: Complexity-aware adaptive spatial preprocessing for ROI scalable video coding with dynamic transition region, in *Proc. of the 18th IEEE Int. Conf. on Image Processing (ICIP)*, September 2011, 741–744.

[55] Gorur, P.; Amrutur, B.: Skip decision and reference frame selection for low-complexity H.264/AVC surveillance video coding. IEEE Trans. Circuits Syst. Video Technol., **24** (7) (2014), 1156–1169.

[56] Bjøntegaard, G.: Calculation of average PSNR differences between RD curves, in *ITU-T SG16/Q6 Output Document VCEG-M33*, Austin, Texas, April 2001. [Online]. Available at: http://www.wftp3.itu.int/av-arch/video-site/0104_Aus/VCEG-M33.doc.

[57] Bjøntegaard, G.: AI11: improvements of the BD-PSNR model. ITU-T Study Group 16 Question 6. *35th Meeting in ITU-T SG16 Q*, Berlin, Germany, 2008.

[58] Sullivan, G.; Ohm, J.; Han, W.-J.; Wiegand, T.: Overview of the high efficiency video coding (HEVC) standard. IEEE Trans. Circuits Syst. Video Technol., **22** (12) (2012), 1649–1668.

[59] Ren, C.Y.; Reid, I.: gSLIC: A Real-Time Implementation of SLIC Superpixel Segmentation, University of Oxford, Department of Engineering Science, Technical Report, 2011.

[60] Bradski, G.: OpenCV Library. Dr. Dobb's Journal of Software Tools (2000). [Online]. Available at: http://code.opencv.org/projects/opencv/wiki/CiteOpenCV.

[61] Mainali, P.; Yang, Q.; Lafruit, G.; Van Gool, L.; Lauwereins, R.: Robust low complexity corner detector. IEEE Trans. Circuits Syst. Video Technol., **21** (4) (2011), 435–445.

[62] Chum, O.; Matas, J.: Optimal randomized RANSAC. IEEE Trans. Pattern Anal. Mach. Intell., **30** (8) (2008), 1472–1482.

Holger Meuel Holger Meuel studied Electrical Engineering at the *Technische Universität (TU) Braunschweig* with a focus on signal processing and communication techniques. He received his Dipl.-Ing. degree from the "Institute for Communications Technology, Department of Electronic Media: System Theory and Technology" of the *TU Braunschweig*, Germany, in 2010. After graduation he joined the *Institut für Informationsverarbeitung* (TNT) of *Leibniz Universität Hannover* as a Research and Teaching Assistant. He became the senior engineer end of 2010. Holger attended several standardization meetings for the video coding standard *High Efficiency Video Coding* (HEVC) of the MPEG and VCEG *Joint Collaborative Team on Video Coding* (JCT-VC). In that context, he also dealt with radial camera lens distortion compensation, scalable video coding, and screen content coding. His research interests are video coding with special focus on low bit rate video coding for aerial surveillance applications. Currently he is working towards his Dr.-Ing. degree.

Marco Munderloh Marco Munderloh achieved his Dipl.-Ing. degree in Computer Engineering with an emphasis on multimedia information and communication systems from the *Technical University of Ilmenau*, Germany, in 2004. His diploma thesis at the *Fraunhofer Institute for Digital Media Technology* dealt with holographic sound reproduction, the so-called wave field synthesis (WFS) where he helds a patent. During his work at the Fraunhofer Institute he was involved in the development of the first WFS-enabled movie theater. At the *Institut für Informationsverarbeitung* of *Leibniz Universität Hannover*, Marco Munderloh wrote his thesis with a focus on motion detection in scenes with non-static cameras for aerial surveillance applications and received his Dr.-Ing. degree in 2015.

Matthias Reso Matthias Reso studied Information Technology with an emphasis on Communication Technology and

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press

Microelectronics, at the *Universität Paderborn*. He wrote his diploma thesis at the *Fachgebiet Nachrichtentechnik* about blind source separation and achieved his Dipl.-Ing. degree in December 2011. In January 2012, he joined the *Institut für Informationsverarbeitung* at the *Leibniz Universität Hannover* as a research assistant. Since then he is working towards his Dr.-Ing. degree. His research interests are image and video segmentation with an emphasis on the topic of temporally consistent superpixel segmentation.

Jörn Ostermann Jörn Ostermann studied Electrical Engineering and Communications Engineering at the *University of Hannover* and *Imperial College London*. He received Dipl.-Ing. and Dr.-Ing. from the *University of Hannover* in 1988 and 1994, respectively. In 1994, he joined *AT&T Bell Labs*. From 1996 to 2003 he was with *AT&T Labs – Research*. Since 2003 he is Full Professor and Head of the *Institut für Informationsverarbeitung* at the *Leibniz Universität Hannover*, Germany. Since 2008, Jörn is the Chair of the Requirements Group of MPEG (ISO/IEC JTC1 SC29 WG11). Jörn received several international awards and is a Fellow of the IEEE. He published more than 100 research papers and book chapters. He is coauthor of a graduate level text book on video communications. He holds more than 30 patents. His current research interests are video coding and streaming, computer vision, 3D modeling, face animation, and computer–human interfaces.

https://doi.org/10.1017/ATSIP.2015.12 Published online by Cambridge University Press