Moving Object Tracking for Aerial Video Coding using Linear Motion Prediction and Block Matching

Holger Meuel, Luis Angerstein, Roberto Henschel, Bodo Rosenhahn, Jörn Ostermann
Institut für Informationsverarbeitung
Gottfried Wilhelm Leibniz Universität Hannover
Hannover, Germany

Email: {meuel, angerste, henschel, rosenhahn, office}@tnt.uni-hannover.de

Abstract—Region of Interest (ROI) coding is a common method for data reduction in scenarios where bandwidth is crucial like in aerial video surveillance from Unmanned Aerial Vehicles (UAVs). In order to save bits, non-ROI areas are typically reduced in quality or not transmitted at all and thus, an accurate ROI classification is mandatory. Moving objects (MOs) are often considered as ROIs and consequently have to be accurately detected on-board. However, common detection approaches either rely on computationally demanding processing which is not available at board. While not detected MOs lead to a degraded representation at the decoder, erroneously detected MOs lead to an unnecessary high bit rate. We tackle all these issues utilizing an efficient object proposal computation.

Based on a dual-thresholding approach applied to image differences, we propose a linear prediction-supported block matcher. Compared to a simple thresholding approach, it shows superior performance and is robust to threshold tuning. By integrating superpixels into the framework, we further recover the complete shape of the MOs. Finally, an efficient tracking-by-detection system is employed to produce accurate detections from the proposals, thereby recovering missed MOs and denying wrong proposals, making the coding more efficient.

We achieve an improved detection precision of up to 76% compared to a simple difference image-based approach. By using a general ROI coding framework we reduce the bit rate of our test set by 70% compared to common HEVC.

I. Introduction

The high resolution Pulse Code Modulation (PCM) video data rate of 622 Mbit/s for a color video sequence with full High Definition Television (HDTV) resolution (1920 × 1080) can be typically compressed to 4–13 Mbit/s at a reasonable image quality [1], [2] using standardized hybrid video coding like High Efficiency Video Coding (HEVC) [3]. Since bandwidth is of crucial importance for scenarios like aerial surveillance from Unmanned Aerial Vehicles (UAVs) and taking into account even higher camera resolutions or multi-camera setups, the bit rate often has to be much further reduced, e.g. by using Region of Interest (ROI) coding. It typically relies on quality reduction of non-ROI areas, resulting in a reduced bit rate for non-ROI areas and thus for the entire video. For aerial surveillance systems, moving objects (MOs) are typically of high interest and thus are considered as ROI. In order to achieve the optimal coding performance and to provide a high quality for all MOs, the latter have to be accurately detected on-board. Therefore, we propose to employ a model-less, three-stages moving object detector (Fig. 1). We use a dual-threshold, block matching supported (pel-wise) difference image first (stage 1) to select corresponding superpixels from an independent superpixel segmentation [4] (stage 2). To cope with non-perfect detections from the simple detectors, e.g. introduced by noise or motion parallax, a tracking-by-detection [5], [6] approach is used to eliminate outliers and also link single detections (stage 3). The resulting coding mask is used for the control of a highly efficient ROI-based coding framework for aerial videos [7].

A. Related Work

Most ROI coding systems provide the highest image quality only for predefined regions whereas non-ROI areas are degraded in quality. For instance, blurring or a coarse quantization of a frame either as preprocessing prior to actual video encoding or within the video encoder itself can be applied [8]–[10], resulting in reduced bit rates. In contrast to that a video coding system relying on reconstruction of non-ROI areas by means of Global Motion Compensation (GMC) was proposed in [7], [11], [12]. By using GMC, a subjectively high image quality could be provided over the entire frame at very low bit rates of 0.8–2.5 Mbit/s for full HDTV resolution video sequences [2]. The ROI coding systems defines two different ROIs: the first one contains only newly emerging areas for each frame (ROI-NA) which were not contained in the previous frame. These ROI-NAS are stitched together by means of GMC into a background panorama image at the decoder, using a projective transform homography $H_n$ estimated on-board the UAV and transmitted as side-information. From the panorama image, the video frames are reconstructed [13]. In order to also reconstruct objects with a local motion (moving objects), moving objects are detected by a moving object detector which calculates the pel-wise luminance difference (difference image) between the motion compensated frame $\hat{n}$ and the current frame $n$. Spots of high energy in this difference image are additionally defined as ROI (ROI-MO). Like other simple approaches, this difference image-based moving object detector is highly computational efficient and provide...
satisfactory detection results in several applications [14]–[17]. However, since the detection performance typically depends on single thresholds, e.g. for noise filtering, these approaches tend to be non-robust for different scenarios. As a consequence, they often lack accuracy for small objects with very small amplitudes in the difference image as well as for objects with low contrast compared to the background. Thus, these detectors cannot fulfill high detection precision demands. More efficient detectors were proposed, which exploit parallax effects [18] or use block matching motion vectors [19] or an optical flow analysis in order to detect moving objects [20], [21], [22] finally tracks clustered image features over several frames in order to improve the detection accuracy. All these approaches have in common that the actual moving frames in order to improve the detection accuracy. 

In contrast to those approaches, specialized detectors cannot consider the image content after the feature derivation object detection is performed only based on features but do not consider the image content after the feature derivation anymore. In contrast to those approaches, specialized detectors may be utilized for the detection of certain objects, e.g. cars, by sophisticated classification or machine learning algorithms (SVM, HOG, SIFT, CNNs etc.). However, since in a surveillance system it is more important to detect any local motion instead of only predefined objects, these methods cannot be used. Thus we propose to combine a model-less dual-threshold difference image-based moving object detector combined with a (modified) block matcher (mask matcher) in order to find and track moving objects also in subsequent frames. The result of the difference image-based analysis is used as seed for a mask matcher. By this approach we find corresponding areas to any detection in subsequent frames and thereby can highly improve the precision of the detector. Since small and highly optimized, low energy block matchers are available from common video coding (e.g. in mobile phones), no additional efforts must be undertaken for the energy-efficient usage on-board of UAVs. For accurate shape retrieval especially for homogeneous, unstructured areas of moving objects, a superpixel segmentation is employed before a tracking-by-detection framework is used to link moving objects over long temporal distances. Using the enhanced moving object detections, a ROI-based coding is performed.

The remainder is organized as follows: In Section II our proposed block matching-assisted difference image-based moving object detector is described in detail. We present experimental results in Section III before Section IV concludes the paper.

II. BLOCK MATCHING SUPPORTED MOVING OBJECT TRACKING

We use the ROI detection and coding system from [7] as a basis, since it provides a subjectively high image quality over the entire frame by encoding only newly emerging areas for each frame (ROI-NA) and locally moving objects (ROI-MO) es explained at the beginning of Section I-A. To overcome the weaknesses of its difference image-based detector, we propose to combine this simple method with block matching to exploit similarities of the shape of moving objects over neighboring frames. Basically, we apply a block matching between two video frames \( n - 1 \) and \( n \) for every spot of high energy in the difference image between both frames. If a proposal trajectory can be built up for one object over several frames, we assume it to be a moving object, otherwise the proposal is discarded. The proposals of the remaining trajectories are extended to the MS full shape using superpixels as proposed in [23] in order to obtain accurate shape information. At the same time, this circumvents splitting of an object into multiple proposals.

Finally an efficient tracking-by-detection framework [6] is employed to reliably link detections and to overcome the limits of the difference image-based tracker, e.g. missing detections as well as partial and long term occlusions (Fig. 1).

A. Proposed Moving Object Detector

The block diagram of the proposed moving object detector is shown in Fig. 2. We use a dual-threshold strategy (also known as hysteresis thresholding) to filter as many false positive detections (static objects falsely detected as moving) on the one hand, and to retrieve the rough shape from the difference image on the other hand: Firstly, we calculate the pel-wise luminance differences between the global motion compensated (block “GMC” in Fig. 2) frame \( \hat{n} \) and the camera-captured frame \( n \) (diffImage output of block “Calc. pel-wise abs. difference” in Fig. 2). Secondly, we generate two different binarized versions (diff2Image and diff2LNFimage) out of the same difference image (diffImage) with different noise filter settings, one with a high \( (t_{\text{high}}) \) and the other with a very low \( (t_{\text{low}}) \) noise filter threshold, respectively. For each pixel in the
In order to find new objects which are not tracked yet, the tracked objects are compared with the set of objects found in the difference image. A new Object is added for each untracked object. As the results of the mask matcher may not be the true motion of the object, a tracked object may be moved to a wrong position leading to non-ROI parts erroneously being marked as ROI. Therefore an overlay check is performed that compares all already tracked and all newly added Objects with the improvedObjects list. Every tracked object that does not match with a improvedObject will be marked as an virtual object. Based on the assumption that real moving objects (true positive detections) reappear at similar positions in subsequent frames, we can distinguish them from false detections: if a virtual Object is not recovered for $x$ frames, it is deleted. Otherwise, in the case of a successful overlay check, the estimated motion is confirmed and the mask of the Object is refined by merging its mask with the mask of the corresponding improvedObject. All pixels from a mask of a Object have a corresponding weight between 0 and 1, indicating the probability that it belongs to the Object. The weight of points in the mask of the Object that do not appear in the mask of the corresponding improvedObject mask is reduced in order to further refine the mask of the objects in subsequent frames and may further be used for detection and tracking confidence. The final ROI-MO activation mask (Fig. 3) is used for further processing.

### B. Accurate shape retrieval by Superpixel Segmentation

In order to retrieve accurate shape information of moving object proposals with unstructured texture, we employ the independently calculated superpixel segmentation of the input video frames. Using the activation mask, a tracking mask (Fig. 2) is generated by inserting the areas of each superpixel which is covered by at least one marked pixel of the activation mask. By combining multiple detections of parts of the same moving object proposal to one proposal in the tracking mask, a correct processing of entire MOTs including homogeneous parts is realized. The tracking mask finally is passed to the object tracker for a temporal linking of the single detections.

### C. Tracking-by-Detection

Using the proposed block matcher we obtain good initial objects proposals. Thereby superpixel help to improve the segmentation quality and to reduce the false positive rate by merging proposals belonging to the same object. However, the proposals are still prone to miss objects in cases of (partial) occlusion or deformation as they are based on image information. To recover objects also in cases where the appearance assumptions fail, we employ the robust and efficient multi object tracker (MCA-Tracker), using only position information.

The hierarchical MCA-Tracker is a tracking-by-detection approach that seeks for a minimum cost arborescence in the detection association graph and obtains the optimal linking of the current association graph in linear time in the number of detections. It is thus suitable for the application in UAVs. Moreover, their proposed tree tracklet trajectory model allows it to reliably reject wrong detections and recover missing ones.

Finally, the computed detections are added as ROI to the coding mask.

### III. Experiments

We used the 350m sequence from the TNT Aerial Video Testset (TAVT) as well as the self-recorded aerial HD video sequence Parking Lot containing lots of moving objects on a parking lot and its nearby streets (overview frame in Fig. 4(e)). Whereas fast and large objects (e.g., the big white car in the 350m sequence) can be satisfactorily detected for a subsequent processing by the unmodified difference image-based moving object detector from [7], small and slow objects like the pedestrian in the 350m sequence or the cars in the Parking Lot sequence are still challenging. Since they may still be missing after tracking-by-detection was applied on the simple difference image (Table 1), those objects may lead to erroneous results. Since our ROI coding system relies on background reconstruction (i.e., GMC), not detected MOTs get lost and are replaced by corresponding background which is not meaningful for a surveillance system. To this end, a high detection recall rate in the first stage (Fig. 1) is desirable, which is achieved by our robust dual-threshold strategy. At
the same time a high precision rate is required to keep the required bandwidth low. We accomplish this by the mask matcher at stage 1 together with the tracking-by-detection system. To obtain quantitative evidence of the effectiveness of our approach, we evaluate the system using the object based true positive (TP) and false positive (FP) rates and, respectively, and recall as well as precision measure as in \([25]\) against manually created ground truth data:

\[
\text{Recall} = \frac{\#TP}{\#GT} \quad \text{Precision} = \frac{\#TP}{\#TP + \#FP} \quad (1)
\]

Table I shows that the precision of simple difference image-based moving object detectors is low. Whereas the integration of superpixels into the system improves the recall, the detection precision still remains low in absolute terms (below 37 \%). By using our proposed mask matcher before the superpixel enhancement, we increase the precision to more than 70 \%. With the improved detections, we were able to additionally increase the tracking accuracy of the tracking-by-detection framework \([6]\) by up to 13 \%

In the Parking Lot sequence (Fig. 4) we observed that the detection of small and slow moving objects becomes possible only with our proposed system because they are often not contained in the common difference image (Fig. 4(a)). Using the low noise filter threshold alone leads to high false detections at the static background, \(e. g.\) due to motion parallax effects (Fig. 4(b) detected building structures top left). Since the latter detections do not have a directed movement but behave similar to noise, we can remove them with our proposed moving object detector without impairing the TP detections of the real moving objects (Fig. 4(c)). Accurate shape information can be retained by utilizing independently calculated superpixels (Fig. 4(d)). Fig. 4(e) finally shows the decoded image after reconstruction by means of GMC.

In order to demonstrate the robustness of our system we show that the detection results are mostly insensitive to the used threshold \(t_{\text{high}}\). The noise filter threshold \(t_{\text{high}}\) basically defines the number of objects which are processed by our mask matcher and can be set \(e. g.\) to values between 300 and 450 (we used \(t_{\text{high}} = 350\)) without impacting recall or precision much (Fig. 5). Since the low filter threshold \(t_{\text{low}}\) is only used to find a rough object shape as initialization for the mask matcher, the detection results are mainly unaffected as long as the threshold is set to a small value allowing a lot of energy in the diff2LNImage (also including noise), \(e. g.\) \(t_{\text{low}} = 200\).

Using the general ROI coding framework from \([7]\) with the x265 video encoder software \((\text{Lavc} 57.48.101 \text{libx265}) [26]\), we can reduce the bit rates by 76.2 \% and 69.7 \% for the Parking Lot and the 350 m sequence, respectively, compared to common HEVC encoding at similar quality levels (37 \(\text{dB}\) and 41 \(\text{dB}\)) \([2], [27]\). We would like to emphasize that a subjectively high image quality is provided over all entire frames due to the reconstruction of non-ROI by global motion compensation of previously transmitted ROI New Area.

### IV. Conclusion

We present a reliable detection system for small and slow moving objects in aerial video sequences which can be employed \(e. g.\) in ROI-based detection and coding systems onboard of (small) UAVs. By utilizing a dual-threshold strategy, our system becomes robust against parameter tweaking without impairing the detection precision much. After applying our proposed modified block matcher (mask matcher), we integrate independently calculated superpixels to cluster pixel-wise detections to compact objects and thus retrieve improved object shapes. Due to the decreased false positive and false negative detections, a tracking-by-detection framework can be utilized to efficiently eliminating remaining false detections, whereas a reliable tracking is not possible without our proposed preprocessing due to too many false detections.

We showed that the precision of the object-based detection is highly increased by the integration of the mask matcher from 12 \% to more than 80 \% in a fully automatic process. Employing the improved detections in a ROI coding framework we reduce the bit rates by about 70 \% compared to common HEVC coding for the encoding of full HDTV resolution sequences (@30 fps) with a subjectively high image quality over
the entire frame. At the same bit rate, we retain more details with our ROI coding framework than with common HEVC.

REFERENCES


Figure 4. Detection precision improvement for self recorded parking lot sequence (a–d: magnifications); a) Diff2image: Binarized difference image from [7]: lots of MOs are not or not entirely detected in every frame; b) Diff2LNFCimage: all MOs but also lots of false positive detections (e.g. building upper left) are detected; c) Diff2 images and applied mask matching: all MOs are detected, false positive detections are nearly entirely removed; d) Mask matching result (from c)) with superpixel enhancement: shapes of MOs are retrieved; e) Decoded and reconstructed entire video frame.