Optical Flow Cluster Filtering for ROI Coding

Holger Meuel, Marco Munderloh, Matthias Reso, Jörn Ostermann
Institut für Informationsverarbeitung, Gottfried Wilhelm Leibniz Universität Hannover
Hannover, Germany
Email: *{meuel, munderl, reso, office}*@tnt.uni-hannover.de

***Abstract*—Current *Moving Object Detectors* in airborne Region of Interest (ROI) coding systems for police surveillance applications used on-board of UAVs are often based on Global Motion Estimation (GME) techniques. Since in these scenarios the camera is moving, simple background removal approaches cannot be applied without a Global Motion Compensation (GMC). Common GMC algorithms assume the ground to be planar, allowing the pixels of the previous frame to be motion compensated into the current frame by applying a projective transformation. The difference image between the compensated frame and the current frame emphasis regions containing possible motion. Such moving object detectors are great in terms of run-time efficiency but are known to lack in terms of accuracy – especially for unstructured regions of moving objects – as well as the robustness against noise. Superpixel segmentation was recently proposed to overcome the issue of the imprecise region cuts given by the difference image. It provides a greatly improved *true positive* detection rate, but unintentionally also increases the area of *false positives*. This paper proposes the use of a mesh-based GME and GMC to detect the moving object regions wherein a cluster filter eliminates errors in the optical flow by assuming a smooth vector field as the global motion model. In doing so we improve the coding efficiency of the fully automatic ROI coding system by more than 24 % for moving object areas conserving the detection benefits of the integration of superpixel segmentation.**

Keywords: Global Motion Compensation, Moving Object Detection, Optical Flow, ROI Coding, Aerial Surveillance, Mesh, Superpixel, Low Bit Rate Video Coding

I. Introduction in Aerial Sequence ROI Coding

Lately, small *Unmanned Aerial Vehicles* (UAVs) became more prevalent for surveillance tasks, *e.g.* for police or disaster operations. One widely unsolved problem is the transmission of high resolution image data over channels with very limited bandwidths. Common approaches to transmit the video data are either to use broader channels like provided by WiFi or to highly compress the video data resulting in bad image quality [1]. Recent publications suggest a *Region of Interest* (ROI) detection and coding for aerial surveillance scenarios with moving camera [2], [3]. Basically, such system consists of a *Global Motion Compensation* (GMC) of the background and the insertion of ROI areas which are transmitted using a typical video codec, *e.g.* H.264/MPEG-4part10/AVC [4]. For real time applications, the ROI detectors as well as the coding system itself have to be highly computational efficient. Taking these limitations into account, a GMC/difference image-based approach for *Moving Object* (MO) detection often provides satisfactory detection results [5]. However, such difference image-based approaches lack accuracy when detecting unstructured, homogeneous areas within the MOs [6]. [7] proposes to use the difference image-based detector result only as initialization to select independently calculated superpixels which are designed to group pixels of homogeneous areas into connected regions [8]. Finally, all image areas covered by selected superpixels are treated as ROI. As the (implicit) planarity assumption of the *Global Motion Estimation and Compensation* (GME/GMC) system is not only violated by moving objects but also by high buildings and trees, these techniques generate lots of false positive detections. These false positive detections lead to an unnecessary increase in bandwidth usage which is exacerbated by the use of superpixels. As the superpixel enhancement helps to increase the resulting video quality it is crucial to decrease the false positive detection rate of the GME/GMC system. Therefore, we propose a mesh-based motion estimation and compensation employing a cluster filter to reduce false positive detections of moving objects.

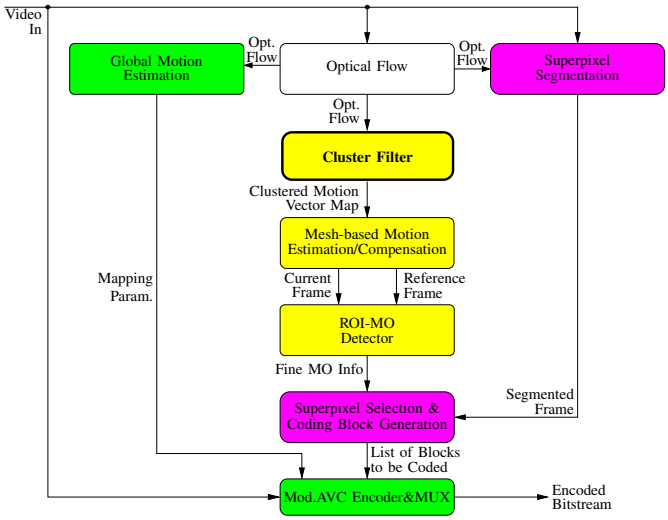


Figure 1. Block diagram of ROI detection and coding system (bold frame: proposed cluster filter to eliminate *false positive* detections, white: optical flow, yellow/light: mesh-based motion estimation/compensation incl. ROI detector, magenta/dark: superpixel segmentation and selection, green/mid-tones: global motion estimation and video coder).

The remainder of this paper is organized as follows: Section II describes the general ROI coding system whereas Section III gives details of the proposed improvement for the moving object detector. Section IV summarizes and analyzes experimental findings before Section V concludes the paper.

II. Superpixel Supported ROI Coding System

In [7], an efficient coding system was proposed for ROI coding. Basically, important image regions (ROI) are forced to

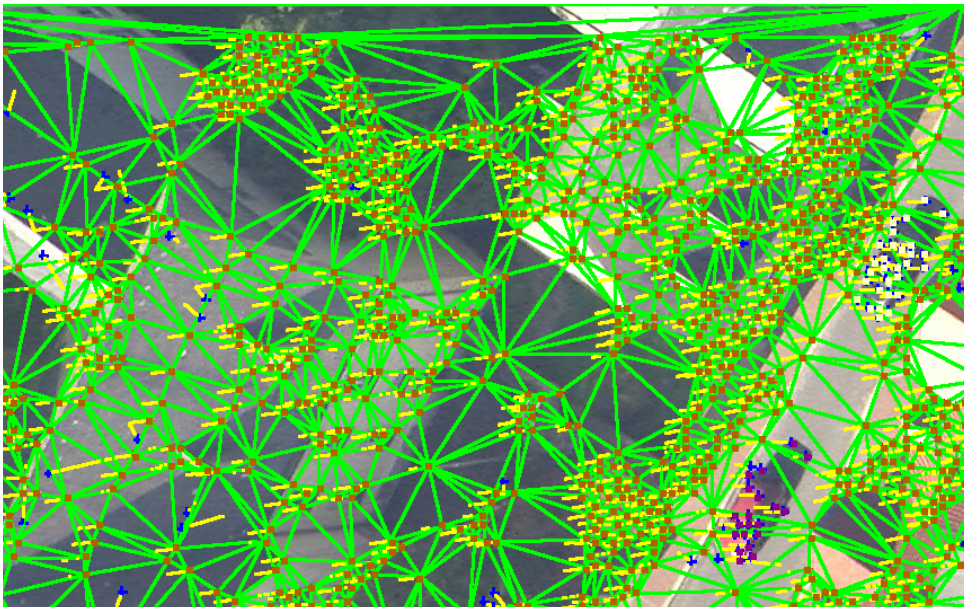


Figure 2. Triangulated mesh (green/mid-tones triangles) between detected features (brown/light dots: background features, blue/dark crosses: motion candidates including outlier, purple/dark and white dots: detected moving objects after cluster filtering) and trajectories (yellow/light lines) in the motion compensated destination frame. Best viewed in color.

be encoded (non-skip mode) whereas the remaining image is forced to be skipped (skip mode) by a modified AVC encoder. To reconstruct the video, a special decoder as presented in [5] is necessary.

The coding system assumes a planar landscape and consequently relies on GME/GMC and projective transformation. To estimate the global motion in the scene, firstly, a *Harris Corner Detector* [9] is employed to select features in the frame $k-1$. Secondly, a KLT feature tracker locates the position of the features in the consecutive frame k, resulting in a sparse set of trajectories [10]. By employing a projective transformation motion model, *Random Sample Consensus* (RANSAC) is able to calculate a set of projective transformation parameters for the mapping of all pixels from frame $k-1$ to frame k while removing the outlier trajectories [11].

Employing this parameter set, a *New Area Detector* is used to compute the image regions not contained in frame $k-1$ but in the current frame k. The regions are marked for video encoding in a coding mask [3]. Since in such a system moving objects remain frozen at the position of their first occurrence, a *Moving Object Detector* (MOD) is employed to detect movements of objects (with a predefined minimum size) not matching the global motion. For this, the difference image between the motion compensated frame $k-1$ and the current frame k is calculated and spots of high energy were marked as MOs in an activation mask.

In order to improve the segmentation recall of moving objects with unstructured texture without decreasing the segmentation precision, an independently calculated temporally consistent superpixel segmentation is created from the input video frames [12]. Using the activation mask, the coding mask is extended by inserting the areas of each superpixel which is covered by at least one marked pixel of the activation mask. This enables a correct processing of entire MOs, including homogeneous parts. Additionally, MOs not detected in single frames are encoded using the temporal connections of the superpixels between different frames using a sliding window approach: An active superpixel in the current frame within the *Sliding Window Width* (SWW) will also activate the past and next $\text{SWW}/2$ temporally associated superpixels. $\text{SWW} = 1$ represents no superpixel activation propagation, "3" specifies a lookback and a lookahead of 1 frame each. This superpixel enhanced system guarantees the accurate detection of moving objects in case of short time missing detections caused by *e.g.* occlusion or too slow object movement. Finally, all image regions referenced by the coding mask are video encoded and transmitted. While [7] introduces a great advantage in terms of *true positive* (TP) detections (MOs), also the *false positive* (FP) rate (non-moving objects falsely classified as moving) was increased unintentionally. Our proposed approach will address this issue.

III. Proposed Reduction of False Positives

To overcome this fundamental problem of false positive detections, we propose to replace the RANSAC-based GMC in the MOD by a mesh-based motion compensation using a *cluster filter* (CF) as outlier detector [13]. The cluster filter is based on the assumption that real MOs are characterized by discontinuities in the optical flow (KLT trajectories), whereas monotonic changes indicate surface modeling errors, *e.g.* by perspective degradations at high buildings. Due to the small image patches which are motion compensated individually in the mesh-based approach, perspective degradations due to model violations of the planarity assumption remain small in contrast to the previously described MOD. Consequently, in the difference image MO candidates are more emphasized and

(a) Overview frame #013 of the 750 m sequence.

(b) Overview frame #010 of the 350 m sequence.

Figure 3. Test sequences example frames.

are more reliable to detect and hence the activation mask is cleared by lots of false positives. As only the MOD is modified, no additional information has to be signaled to the decoder.

A. Mesh-based Cluster Filtering

The cluster filter is necessary to remove false trajectories and to separate FP from TP. It works as a region growing approach based on the smoothness of the motion vector field [14]. We consider a cluster as an image patch represented by the motion vectors on its boundary only. In order to reliably detect discontinuities in the vector field, all motion vectors $\overrightarrow{n_k}(x,y)$ are clustered by their similarity either into one existing cluster or into a new one. If the spatial distance of a motion vector to its neighbors (1) and its displacement difference to the border of a cluster (2) are smaller than a threshold, the motion vector is assigned to that cluster.

$$||\overrightarrow{r_k}(x,y) - \overrightarrow{n_k}(x,y)|| \quad < \quad t_{d1} \qquad (1)$$

$$||\vec{d}_{\overrightarrow{r_k}}(x,y) - \vec{d}_{\overrightarrow{n_k}}(x,y)|| \quad < \quad t_{d2} \qquad (2)$$

$\overrightarrow{r_k}(x,y)$ is the position of the closest motion vector of an already classified cluster to the yet unclassified motion vector $\overrightarrow{n_k}(x,y)$ at position (x,y) in frame k. The displacement vectors $\vec{d}_{\overrightarrow{r_k}}$ and $\vec{d}_{\overrightarrow{n_k}}$ of the motion vectors ($\overrightarrow{r}$ and $\overrightarrow{n}$) in frame k point to their positions in frame $k-1$. We used fixed thresholds because methods with dynamic thresholds like *e.g. Markov Random Fields* tend to under-segment the image [15]. Clusters containing less motion vectors than a threshold t_f (here $t_f = 3$) are considered as outliers and are removed (Fig. 2, blue/dark crosses).

The largest cluster is defined as background (brown/light dots in Fig. 2) and is used for the mesh-based motion compensation. Small clusters are correctly detected as being inconsistent with the motion model and treated as MO candidates (Fig. 2, blue/dark crosses with purple/dark and white dots). Since non-planar objects are correctly motion compensated by the mesh-based motion compensation, FP detections are largely decreased leading to less blocks to be coded and an increased coding efficiency. The computational complexity of the cluster filter is comparable to that of RANSAC.

IV. Experiments

Results of the proposed improved detection system are presented in this section. We used airborne sequences with very different characteristics (Fig. 3). One sequence contains lots of houses and cars, most of them are parking, two are moving ("750 m sequence", Fig. 3a). As these moving objects are very small compared to the entire frame it is very difficult to detect and segment them accurately. The other sequence is much easier to segment since the moving car on the street is one of the main biggest elements in the scene and has a high contrast against the background ("350 m sequence", Fig. 3b).

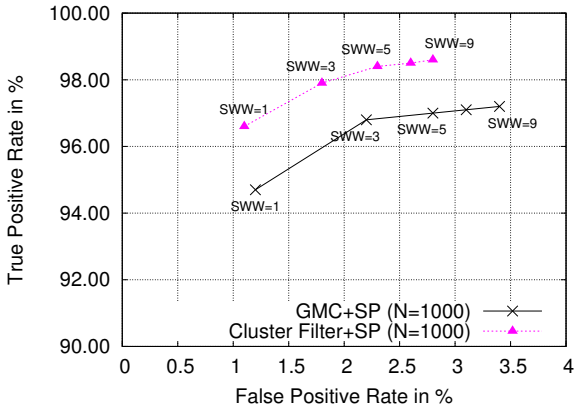
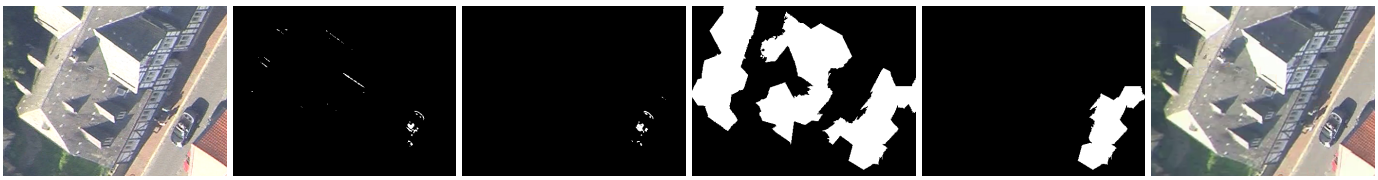


Figure 5. *Receiver Operating Characteristic* (ROC curve) for 750 m sequence. (TP rate calculated pel-wise, SWW =*Sliding Window Width*, N = number of superpixels.)

The example in Fig. 4 shows a magnification from the 750 m sequence (Fig. 4a), the activation masks for the planar GMC-based moving object detector, including lots of false positive detections at the building's gable (Fig. 4b), and the mesh-based cluster filter detector result without those false detections (Fig. 4c). The resulting coding masks after the superpixel enhancement are printed in Figs. 4d and 4e, and the decoded result in Fig. 4f. The reduced FPs (white regions in the left part of the image) for the proposed MOD approach compared to the GMC approach assuming a planar ground (Figs. 4b, 4c) lead to a greatly improved coding mask after superpixel enhancement (Figs. 4d, 4e) for the coding system. The TP detections for the moving car stay (mostly) the same. Since the entire car (moving object) is detected as one, it can be properly reconstructed without errors. Moreover, nearly no non-moving areas (FPs) are marked for video encoding, resulting in a reduced coding data rate as well as an improved detection accuracy.



(a) Orig. frame (outtake). (b) GMC activation mask. (c) CF activation mask. (d) GMC+SP coding mask. (e) CF+SP coding mask. (f) Decoded frame.

Figure 4. Moving object detections (b,c) and coding masks (d,e) for the GMC-based (b,d), the cluster filter-based (CF) system (c,e) and decoded result (f).

A. Classification Results

To create the *Receiver Operating Characteristic* (ROC) shown in Fig. 5 for the 750 m sequence, we employed different sliding window widths to visualize TP over FP. Depending on the SWW, we achieve FP reductions of up to 18.2 % (2.2 % to 1.8 % FP rate reduction in the ROC) at simultaneously increased TP detections. Since only relatively small parts (<5 %) of one frame are actually MOs, this is a noticeable achievement in terms of bit rate saving. At a reasonable operation point (SWW = 3) we are able to reduce the FP rate from 2.2 % to less than 1.8 % and for the worst case (SWW = 9) from 3.4 % to <2.8 %. The segmentation result of the moving objects is better for any operating point, consequently the detection accuracy according to (3) is increased (from 96.6 % up to 98.9 %) in the fully automatic system.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

For the 350 m sequence the TP rate already was between 99.3 % (SWW = 1) and 100 % for SWW > 1 and no FPs caused by model violations were detected. Thus no improvement was possible.

B. Coding Results

We used a modified *x264* (v0.78) [16] AVC-encoder as coding backend at a fixed *Quantization Parameter* QP = 33 at *High Profile*. Only the first frame is encoded as intra frame (I), leading to a coding data rate of less than 1.5 Mbit/s for inter-predicted frames (P) (overall data rate <2.0 Mbit/s) for high-textured sequences with two moving cars – which is typical for suburban environments with not much traffic. This is a data rate saving of about 80 % compared to the unmodified non-ROI coder. For the former operation point (SWW = 3) we reduce the data rate for the transmission of the detected moving objects more than 24 % (4 % smaller overall coding data rate, including new arising areas at image borders) compared to a planar GMC system (full HDTV video sequences, 30 fps).

V. Results and Conclusions

To reduce *false positive* detections leading to unnecessarily high data rates in fully automatic ROI coding systems for aerial video sequences we propose a new moving object detector scheme consisting of mesh-based local motion estimation and compensation. It employs a cluster filter approach to distinguish real moving objects (TP) from false ones based on an optical flow analysis. We are able to reduce the FP detection rate up to 18.2 % while simultaneously increasing the TP detection rate and providing coding data rates of less than 2 Mbit/s. We save more than 24 % data rate for the moving object areas (4 % overall data rate including new areas, respectively) for a full HDTV resolution sequence at 30 fps at the same quality level compared to a planar GMC-based moving object detector.

References

[1] B. Ciubotaru, G. Muntean, and G. Ghinea, "Objective Assessment of Region of Interest-Aware Adaptive Multimedia Streaming Quality," *Broadcasting, IEEE Transactions on*, vol. 55, no. 2, pp. 202–212, 2009.

[2] H. Cheng and J. Wus, "Adaptive region of interest estimation for aerial surveillance video," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 3, Sept. 2005, pp. III – 860–3.

[3] H. Meuel, M. Munderloh, and J. Ostermann, "Low Bit Rate ROI Based Video Coding for HDTV Aerial Surveillance Video Sequences," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2011, pp. 13 –20.

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

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

[6] J. Bang, D. Kim, and H. Eom, "Motion Object and Regional Detection Method Using Block-Based Background Difference Video Frames," in *IEEE 18th Internat. Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, Aug. 2012, pp. 350 –357.

[7] H. Meuel, M. Reso, J. Jachalsky, and J. Ostermann, "Superpixel-based Segmentation of Moving Objects for Low-Complexity Surveillance Systems," in *Tenth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, August 2013.

[8] X. Ren and J. Malik, "Learning a classification model for segmentation," in *ICCV*, 2003, pp. 10–17.

[9] C. Harris and M. Stephens, "A Combined Corner and Edge Detection," in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[10] J. Shi and C. Tomasi, "Good Features to Track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, June 1994.

[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, June 1981. [Online]. Available: http://dx.doi.org/10.1145/358669.358692

[12] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann, "Temporally Consistent Superpixels," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013.

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

[14] M. Munderloh, S. Klomp, and J. Ostermann, "Mesh-based Decoder-Side Motion Estimation," in *Proc. of IEEE International Conference on Image Processing*, Sept. 2010, pp. 2049–2052.

[15] F. Z. B. Kettaf and J. P. D. Asselin de Beauville, "A comparison study of image segmentation by clustering techniques," in *3rd International Conference on Signal Processing*, vol. 2, Beijing, 1996.

[16] VideoLAN Organization. (2009) x264. [Online]. Available: http://www.videolan.org/developers/x264.html