Superpixel-based Segmentation of Moving Objects for Low Bitrate ROI Coding Systems

Holger Meuel, Matthias Reso, Jörn Jachalsky*, Jörn Ostermann

Institut für Informationsverarbeitung
Universität Hannover, Germany
meuel@tnt.uni-hannover.de

* Technicolor Research & Innovation
Technicolor, Hannover, Germany
joern.jachalsky@technicolor.com

Abstract

*A major constraint for mobile surveillance systems (*e.g.* UAV-mounted) is the limited available bandwidth. Thus, ROI-based video coding, where only important image parts are transmitted with high data rate, seems to be a suitable solution for these applications. The challenging part in these systems is to reliably detect the regions of interest (ROI) and to select the corresponding macroblocks to be encoded in high quality. Recently, difference image-based moving object detectors have been used for this purpose. But as these pixel-based approaches fail for homogeneous parts of moving objects, which leads to artifacts in the decoded video stream, this paper proposes the use of a superpixel-based macroblock selection. It is shown that the proposed approach significantly reduces artifacts by raising the moving objects detection rate from 7 % to over 96 % while increasing the data rate by only 20 % to 1.5–4 Mbit/s (full HD resolution, 30 Hz) still meeting the given constraints. Moreover, other simple approaches (dilation) as well as state-of-the-art GraphCut-based approaches are outperformed.*

1. Introduction

One of the most common tasks of automatic surveillance systems is to detect moving objects. Classically, surveillance systems can be divided into fixed systems – *e.g.* for traffic monitoring at intersections – and non-fixed systems, such as mounted systems on *Unmanned Aerial Vehicles* (UAV). Since for the former circumstances are known and fixed regarding the angle of view, height above ground etc., it is relatively easy to design an appropriate moving object detector. Common approaches use *Frame Difference Method* (FDM)/*Difference Image-Based Detection* [1], employ background subtraction techniques to model and remove the background like shown in the overview paper [4], or use fixed cameras [7], sometimes supported by additional markers, *e.g.* to restrict the *Regions of Interest* (ROI) where moving objects can occur [9]. Mobile video-based surveillance systems like UAV-mounted systems have to deal with a high number of degrees of freedom like flight height, angle of view, or size of ground objects to be detected. Consequently, it is still a challenge to design a multipurpose surveillance system for moving objects without too many restrictions introduced by model assumptions.

Figure 1. Example for a fully automatic *moving object detection* utilizing a superpixel segmentation.

It is a further requirement for mobile platforms to consider the overall data rate since data transmission is crucial. For instance, an 8 bit *Pulse Code Modulation* (PCM) coded color video sequence in full HD resolution (1920×1080 pel, 30 Hz) results in a data rate of 622 Mbit/s. As shown in [10], for this input the popular video codec *Advanced Video Coding* (AVC) [8] is able to provide compressed data rates of 8–15 Mbit/s (employing *x264* software, v0.78 [18]) and also the latest ISO/MPEG video codec *High Efficiency Video Coding* (HEVC) [2] provides minimal compressed data rates of more than 6.5–12.5 Mbit/s for the test set at a subjectively "good" video quality [12]. A further data rate reduction using these codecs leads to subjectively annoying distortions and artifacts. However, for mobile channels an even lower channel capacity of only a few Mbit/s has to be expected. Consequently, specialized coding systems have to be designed.

One fundamental idea is to encode only the regions of interest in high quality while saving bandwidth on the others. An approach pursuing this idea employs a difference image-based moving object detector utilizing global motion compensation and is presented in [10]. However, as explained in [11, 1], difference image-based approaches fail at the detection of homogeneous regions like car tops. Thus, the detection rates of moving vehicles do not meet expectations. Another issue occurs if a moving object is not detected accurately in every single frame due to noise or occlusion. Both will lead to artifacts in the decoded image.

To encounter these shortcomings of difference image-based approaches while meeting the requirement of a very low data rate, this paper introduces an improved approach for moving object detection that combines the ideas of the difference image-based approach presented in [10] with an approach for temporally consistent superpixels [14].

The main contributions of this paper are twofold:

- A superpixel-based completion strategy to handle moving objects that were only partially or fragmentarily detected.
- A strategy to fill temporal gaps, which occur if a moving object is not detected in every frame, utilizing the spatio-temporal information provided by temporally consistent superpixels.

The results were evaluated in a video coding system similar to [10] that is capable of transmitting full HD video sequences at a bit rate of about 1.5–4 Mbit/s (at 30 Hz) at a subjectively high video quality.

The remainder of the paper is organized as follows: Section 2 takes a close look at the enhanced ROI coding system. The focus of this section is on superpixel-based macroblock selection extending activation masks generated by the difference image-based moving object detector. Section 3 shows the improved detection rate using the proposed superpixel-based macroblock selector, compares it to other macroblock selection approaches, and provides resulting coding bit rates. Section 4 finally concludes the paper.

2. Superpixel-based ROI Coding System

With regard to the constraints for UAV-mounted systems described in Section 1, this paper introduces a new ROI-based video coding system for UAVs, which is similar to [10] but utilizes temporally consistent superpixels [14] in order to significantly increase the detection rate for moving objects while still meeting the mentioned bandwidth constraints. A block diagram of the coding system is depicted in Figure 2. It consists of a modified AVC encoder that only encodes a subset of all macroblocks. The initial activations generated by the global motion compensation-based moving object detector (ROI-MO) (Section 2.1) are fed into a selection block (Section 2.3) utilizing the superpixel segmentation (Section 2.2) of the input video stream in order to generate the list of macroblocks to be encoded for the moving objects. In addition, the list of macroblocks for new emerging image regions at the image borders is generated by the ROI detector for new areas (ROI-NA) based on the global motion parameters (Section 2.1).

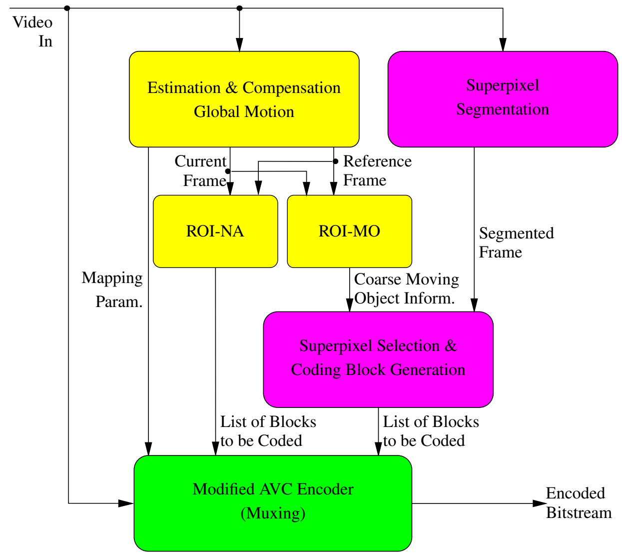


Figure 2. Block diagram of the enhanced superpixel-based ROI coding system (yellow (light): GMC/difference image-based moving object detector; magenta (dark): superpixel segmentation and selection; green (mid-tones): video encoder).

2.1. ROI Coding System

The (block-based) AVC video encoder is modified to use skip-mode only for macroblocks externally selected. In doing so, macroblocks containing moving objects (or new emerging image regions at the image borders) are forced to be encoded (non-skip mode) and all other macroblocks are forced to be not encoded (skip mode). To decode such a data stream, a special decoder is required, which is capable of mapping all non-skip macroblocks to their correct position and thus reconstruct the entire video sequence. A thorough explanation of different decoding possibilities can be found in [10].

For the *Global Motion Estimation and Compensation* (GME/GMC) a planar landscape is assumed. Thus, one frame $k-1$ can be morphed into the consecutive frame k employing a projective transformation. To determine its eight parameter set $\vec{a_k}$, all corners in the gray-scale representation of the frames $k-1$ and k are extracted by the two-dimensional gradient-based *Harris Corner Detector* [6]. Corners with high gradients are used as features by a *Kanade-Lucas-Tomasi* (KLT) feature tracker [17] to align the frames $k-1$ and k. *Random Sample Consensus* (RANSAC) [5] is then applied to eliminate outliers not matching the global motion. Since the camera movement over the landscape is slow in contrast to the frame rate of the camera ($30\ \frac{\text{frames}}{\text{second}}$), the translational offset of two consecutive frames is relatively small ($< 50\ \frac{\text{pel}}{\text{frame}}$) [10].

Given the global motion compensated previous frame $\hat{k}$, the moving object detector (ROI-MO) can compute the difference to the current image k indicating regions with motion different from the global motion. Basically, all pixels with unequal luminance values are considered to be moving objects. However, since noise and perspective aberration are not of interest in contrast to moving objects, a morphological filtering is applied afterwards. After that, only connected pixels are selected as moving objects and uncovered background. The output of the ROI-MO is a binary map with pixels representing either *moving object* ("1", *activated*) or *background* ("0", *not activated*). Such a binary map can be seen as initial activation mask for the macroblock selection. In such a scenario, each macroblock that contains an activated pixel is selected for encoding.

New emerging parts at the image borders are also considered to be "regions of interest". They are needed to properly reconstruct the new image areas at the decoder side. Based on the global motion parameters, those image areas, which are inside the current frame k but not in the motion compensated frame $\hat{k}$ can be easily calculated [10].

2.2. Superpixel Segmentation

Difference image-based moving object detectors aim at the detection of pixels not matching the global motion. In the case of homogeneous areas within a moving object such detectors may fail and will only segment the borders and textured areas of the moving objects [1]. When such detection results are used directly for the macroblock selection in the video encoder, macroblocks containing only homogeneous parts of moving objects will *not* be encoded. This leads to block errors (artifacts) in the final, reconstructed video frame.

In order to overcome this problem, this paper proposes the utilization of superpixels. The idea of superpixels was initially proposed in [13]. It is to group spatially coherent pixels sharing similar features, *e.g.* color or texture, into regions of almost homogeneous size and shape. This can be seen as a kind of oversegmentation. Superpixels are aligned with the main object boundaries. Thus, pixels within a superpixel in general belong to the same object.

Superpixels can be used to complement the pixel-based activations generated by the ROI-MO detector described in Section 2.1. It is achieved by broadcasting the pixel-based activations to all other pixels assigned to the same superpixel. In the proposed implementation this is done by the selection block from Figure 2, which selects all superpixels containing at least one activated pixel. For the generation of the superpixels, the temporally consistent superpixel approach (TCS) described in [14] was chosen. In contrast to common algorithms like mean-shift [3], which provide a superpixel segmentation for each video frame individually, TCS provides temporally consistent results with unique labels for each superpixel along the time line establishing a temporal connection between the superpixels of different frames. This finally allows to track the image region covered by each superpixel over time.

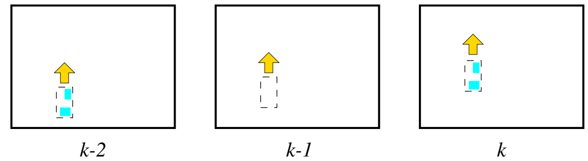


Figure 3. *Temporally Consistent Superpixels* are used to bridge *false negative* detections of the ROI-MO: if no moving object (dotted rectangle) is detected by the MO detector (cyan), the MO in frame $k-1$ would not be selected for coding. Due to TCS all blocks containing the car will be detected in all frames.

The algorithm creates superpixels by clustering pixels based on their five-dimensional feature-vector $[labxy]$ built of pixel color in the CIELAB color space and pixel position. The feature-space is separated into a global color subspace $[lab]$ and multiple local spatial subspaces $[xy]_k$ unique for each frame k. The clustering is done by minimizing an energy function using an iterative optimization framework. By utilizing an observation window that if shifted over the video volume along the time axis real-time capabilities are preserved. A more detailed description of the TCS algorithm can be found in [14].

2.3. Superpixel Selection

Superpixels are of almost homogeneous size and shape which is especially beneficial in the case of erroneous activations, when the ROI-MO gives *false positive* results, *i.e.* a moving object has been falsely detected. In those cases the utilization of unconstrained regions could lead to a wasteful use of the available bandwidth, as false pixel activations would be propagated to a potentially larger area.

In principle, a superpixel is selected if it contains an activated pixel. As a consequence an extended activation map is created, in which all pixels are activated that are comprised by the selected superpixel.

It can occur that a moving object is not accurately detected in every single frame due to either noise in the activations or occlusions. This leads to artifacts at the moving object in the reconstructed images. In those cases, the temporal connections between the temporally consistent superpixels of each frame can be utilized to fill these temporal gaps. Basically, in addition to the superpixels, which are selected in the current frame k, all superpixels that are temporally connected to them are also selected. Figure 3 illustrates this idea. Without the use of TCS the moving object (dotted rectangle) would be detected only in the frames $k-2$ and k based on the initial activations (cyan points) and the corresponding superpixel in frame $k-1$ would not be selected. To avoid the continuous tracking of *false positive* activations, *e.g.* edges of buildings falsely classified as moving objects, a *sliding window* with length L is introduced in the time domain and only superpixels that are temporally connected within this sliding window are selected. There is a trade-off between the bridging due to missing activations and the resulting coding efficiency. This trade-off can be adjusted altering L.

The selection of the right number of superpixels per frame is also a trade-off between *true positives* (TP, a moving object is classified correctly) and *false positives* (FP). Thus, it has a direct impact on the detection rate as well as the resulting coding efficiency. It is obvious that the optimal number of superpixels differs with flight height and camera optics, as both have an impact on the object size in the images. In order to normalize the number of superpixels per frame N, the ground resolution R – measured in pel/m – of the images was employed. An ordinary car with a length of 4.5 m and a width of 1.7 m (covering an area A) is assumed to be the average moving object. The number of superpixels covering one complete car C deliberately was selected to be approximately 2 to 3. Thus, for a full HD resolution the number of superpixels per frame can be calculated as follows:

$$N = \frac{C \times 1920 \times 1080}{R^2 \times A}. \tag{1}$$

3. Experiments

The coding system described in Section 2 was implemented using C++ and MATLAB. For the task of a fully automatic moving object detection and coding system, the proposed superpixel-based macroblock selection (SP) is compared to the following other macroblock selection methods:

- ROI-MO only, difference image: Diff
- ROI-MO with *dilation*: Dil
- ROI-MO with *GraphCut*-based segmentation: GC

The Diff-approach uses the simple difference image-based (binary) activation mask generated by the ROI-MO without any further processing. For the dilation-based approach (Dil) these binary activation masks were dilated several times until all moving objects are properly covered. For the *GraphCut*-based approach (GC), representing a state-of-the-art image segmentation approach, the efficient *SlimCut* implementation was used [16]. For the evaluation using SlimCut the weighting parameter between unary and pairwise potentials was set to $\gamma = 50$, which was proposed as a versatile setting for different images [15]. In subjective tests it proved to be a good choice also for the test sequences. White pixels from the binary activation masks were used as foreground initialization, black pixels eroded by 64 pixels were used as background initialization.

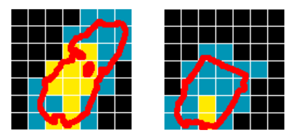
Several airborne HD sequences were recorded with a frame rate of 30 frames/s from different flight heights (see an example in Figure 4). For these recordings a consumer camcorder, the JVC GY-HM100E, was used at 390 mm (equivalent focal length in 35 mm film format) obtaining measured ground resolutions between 43–21 pel/m or an average superpixel diameter of 65–32 pel (see Section 2.3).

For the comparison of the different macroblock selection methods, an initial activation mask was created using

Figure 4. Frame 10 of the 750 m sequence (ground res. 21 pel/m).

(a) GT (b) Diff (c) Dil (d) GC (e) SP

Figure 5. Ground truth (GT) (a) and (extended) activation masks (b-e), (e) shows also the coding raster. White="moving object/uncovered background (ROI-MO)", black="background".



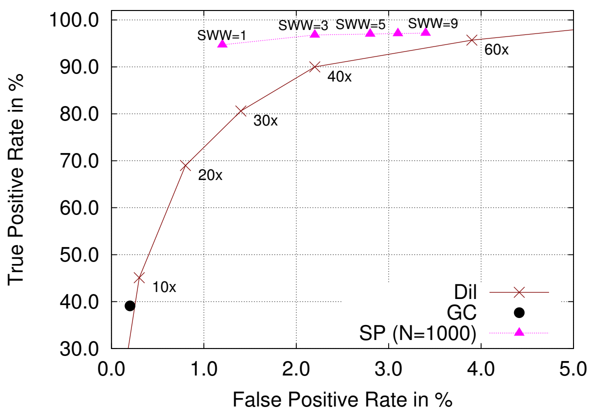
(a) Upper car (b) Lower car

Figure 6. Macroblock coding mask for both cars in the sequence. Yellow (light) blocks would be coded based on the result of Diff, blue (dark) blocks are additionally coded by SP.

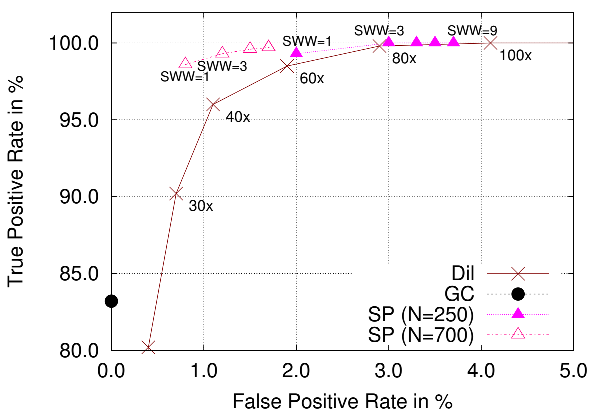
the output of the ROI-MO (Figure 5(b)). These initial masks were then used as input for the other selection methods in order to generate extended activation masks (Figure 5). Finally, these activation masks were turned into coding masks used during the encoding process. Figure 6 highlights the gain of the superpixel-based approach (SP) compared to the unmodified Diff proposed in [10], which already hints at an improved resulting image quality. Although the rear of the lower car is not entirely detected by the SP the detection accuracy beats any competitor.

However, in a coding system next to a high detection rate for moving objects also the resulting data rate has to be taken into account. Thus, the number of *false positives* (FP) needs to be as low as possible in order to efficiently use the available bandwidth and achieve a maximum image quality. Therefore, the ratio between *true positives* (TP) and FP is crucial. In order to assess the performance of the different approaches, results were generated for a number of test sequences. They are shown as *Receiver Operating Characteristics* (ROCs) in Figure 7.

As Diff was used to initialize all other methods, the results for Diff correspond with the beginnings of the curves for Dil, which is outside of the plots. The corresponding values are TP, FP: 7 %, 0 % for the low ground resolution (Figure 7(a)) and TP, FP: 36 %, 0 % for the high ground resolution (Figure 7(b)).



(a) ROC for 750 m sequence.



(b) ROC for 350 m sequence.

Figure 7. *Receiver Operating Characteristics* (ROCs). TP rate calculated pel-wise.

The Dil-approach was used as a baseline. The results for an increasing number of dilation iterations are depicted (labels on graphs in Figure 7). It is evident that the TP-rate increases with the number of iterations. GC achieves the lowest and thus the best FP-rate but at the cost of the lowest TP for all sequences (black dots). Hence, GC is considered not to be suitable for a surveillance system. Moreover, it can be stated that the results for GC strongly depend on the quality of the initialization for each moving object.

The results for SP were calculated with respect to Equation (1), once for the number of superpixels per frame $N = 1000$ for the low ground resolution (21 pel/m, Figure 7(a)) and once for $N = 250$ for the high ground resolution test sequence (43 pel/m, Figure 7(b)), respectively (purple filled triangles). Different sliding window lengths $l = [1 \ldots 11]$ have been tested to determine the optimal operating point for the selection task. For the former sequence, a SP number of $N = 1000$ (triangles) leads to an optimal segmentation. Also for the latter sequence the automatically derived $N = 250$ as suggested by Equation (1) already outperforms the dilation (dark red crosses). The optimal segmentation result is reached for $N = 700$ (light red triangles).

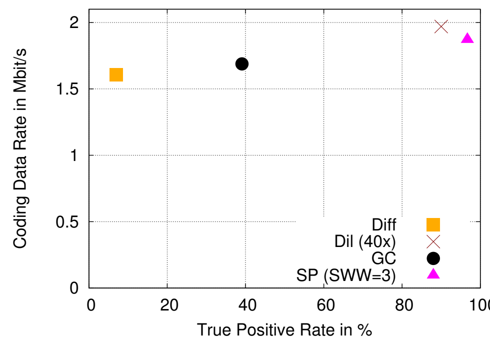
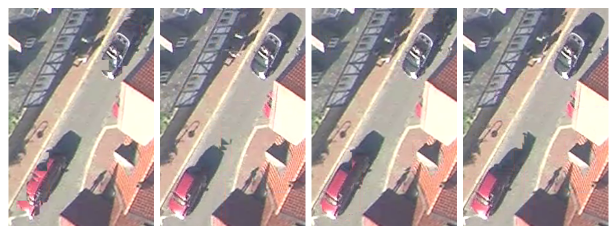


Figure 8. Data rate over TP-rate.



(a) Diff (b) Dil (c) GC (d) SP

Figure 9. Resulting image quality after decoding (frame 10). Especially note the high amount of artifacts in (a) and (c).

Figure 8 shows the resulting data rate over the (pel-wise calculated) TP-rate for fixed coding conditions (*Quantization Parameter* (QP) = 33). With an increase in data rate ($\approx$ 20 %) compared to Diff, both Dil (with 40 iterations) and SP achieve a significant higher detection rate, which is $\geq$ 90 %, whereas Diff has only a TP-rate of about 7 %. While consuming a lower data rate than Dil, SP achieves the highest TP-rate ($\approx$ 97 %) with a sliding window length (SWW) of three, which turned out to be the optimal value considering all test sequences. It should be noted that the data rate for Dil increases with the number of dilation iterations.

Finally, the resulting subjective image quality after ROI encoding and decoding is depicted in Figure 9. The image quality of Diff shown in Figure 9(a) indicates that this simple selection approach distorts moving objects. The Dil-approach is simple and works fine in terms of the resulting activation mask as well as in the subjective quality evaluation (Figure 9(b)). All *true positives* (moving objects/uncovered background) are (nearly) covered. Since at the same time also the *false positive*-rate is increased by selecting a lot of false macroblocks (see activation mask Figure 5(c)), the data rate is the highest in this test. Furthermore, a simple dilation of the activation masks requires manual control to set the correct dilation parameter (as done in this test) in order to make sure that all moving objects are covered correctly, especially if the initialization is of varying quality. With the given initialization, GC fails to properly and fully segment the moving objects for an appropriate macroblock selection. Thus, the resulting image quality is rather poor (Figure 9(c)). The given initialization is not sufficient, as the number of seed pixels required for a good result is too low. It is evident as stated above that GC is an unsuitable candidate for this non-interactive, automatic task. In contrast to that, the superpixel-based approach SP is able to segment all moving objects with the given initial activation masks without any further context information provided by the moving object detector. With the highest TP-rate and a lower data rate than the Dil-approach, it achieves a good subjective image quality (Figure 9(d)). Thus, this automatic approach is an ideal candidate to significantly improve the detection rate (and thus the image quality) with only a moderate increase for the resulting data rate compared to those approaches not meeting expectations.

4. Conclusion

In this paper a novel ROI-based video coding system for UAVs is introduced that combines a difference image-based moving object detector with a temporally consistent superpixel segmentation. It was shown that the utilization of superpixels to extend the activation masks of the object detector can raise the detection rate of moving objects from 7 % to over 96 % while increasing the data rate by only 20 % to 1.5–4 Mbit/s (full HD resolution, 30 Hz) still meeting the constraints for an UAV-mounted video coding system. The proposed approach was compared to a method utilizing a state-of-the-art *GraphCut*-based segmentation that achieves a significantly lower detection rate resulting in severe artifacts in the decoded video stream. Additionally, it was also shown that the proposed approach outperforms a dilation-based method in both the achievable detection rate as well as the resulting data rate. Moreover, the proposed approach allows for a fully automatic surveillance system, whereas the dilation approach requires at least an initial manual setting of the dilation parameters.

5. Acknowledgement

Special thanks goes to Marco Munderloh and Björn Scheuermann for their code contributions. Both are with the *Institut für Informationsverarbeitung* (TNT), Universität Hannover.

References

[1] J. Bang, D. Kim, and H. Eom. Motion Object and Regional Detection Method Using Block-Based Background Difference Video Frames. In *Embedded and Real-Time Computing Systems and Applications (RTCSA), 2012 IEEE 18th Internat. Conference on*, pages 350 –357, Aug. 2012. 1, 3

[2] B. Bross, W.-J. Han, G. J. Sullivan, J.-R. Ohm, and T. Wiegand. High Efficiency Video Coding (HEVC) text spec. draft 10 (for FDIS&Consent). In *JCT-VC Doc. JCTVC-L1003*, 12th Meeting: Geneva, Switzerland, Jan. 2013. 1

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603619, 2002. 3

[4] S. Y. Elhabian, K. M. El-Sayed, and S. H. Ahmed. Moving Object Detection in Spatial Domain using Background Removal Techniques - State-of-Art. *Recent Patents on Computer Science*, 1:32–34, 2008. 1

[5] 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*, 24(6):381–395, June 1981. 2

[6] C. Harris and M. Stephens. A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988. 2

[7] T. Hata, N. Kuwahara, T. Nozawa, D. Schwenke, and A. Vetro. Surveillance system with object-aware video transcoder. In *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, pages 1 –4, Nov. 2005. 1

[8] 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, July 2004. 1

[9] K. Kiratiratanapruk and S. Siddhichai. Practical application for vision-based traffic monitoring system. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009. 6th Int. Conf. on*, volume 02, pages 1138 –1141, may 2009. 1

[10] H. Meuel, M. Munderloh, and J. Ostermann. Low Bit Rate ROI Based Video Coding for HDTV Aerial Surveillance Video Sequences. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 13 –20, June 2011. 1, 2, 3, 4

[11] 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. 1

[12] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand. Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC). *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(12):1669 –1684, Dec. 2012. 1

[13] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, pages 10–17, 2003. 3

[14] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann. Temporally Consistent Superpixels. In *IEEE ICCV*, Submitted for review 2013. 2, 3

[15] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive Foreground Extraction using Iterated Graph Cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pages 309–314, New York, NY, USA, 2004. ACM. 4

[16] B. Scheuermann and B. Rosenhahn. Slimcuts: Graphcuts for high resolution images using graph reduction. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, Jul 2011. 4

[17] J. Shi and C. Tomasi. Good Features to Track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, June 1994. 2

[18] VideoLAN Organization. x264, http://www.videolan.org/developers/x264.html, 2009. 1