Detecting Moving Vehicles in Satellite Videos

Julian Wagner

Computer Vision Lab
Institute of Visual Computing & Human-Centered Technology
TU Wien

December 21, 2018

Supervisor: Roman Pflugfelder

We thank Planet Inc. for providing us the data in this work.

Contents

Abstract

Appearance based object detection algorithms are pushed to their boundaries when applied to small objects with little distinguishable features in satellite images. Recently satellite videos have become available and with them arise new opportunities and challenges for object detection. Challenges are low local contrast between targets and background, motion effects caused by the elevation angle of the image sensor, noise and targets only slightly larger than the resolution limit.

To overcome the limitations of appearance based methods an algorithm for the detection of moving vehicles is proposed that uses the temporal information.
After co-registering consecutive frames using the optical flow obtained fro, the Lucas-Kanade method, motion is detected using a Gaussian mixture background model. Components that are too small or too large to correspond to vehicles are removed. Motion artifacts resulting from stationary objects are detected and removed by analyzing the space-time trajectories of the remaining components using Local Principal Component Analysis.
The results of the proposed method with and without detection of motion artifacts caused by stationary objects are experimentally evaluated on three different regions from a satellite video. Using the detection of motion not caused by moving objects a F1 score of 0.61 is obtained in the best case and 0.12 in the worst case.

The results show that semi-obstructing entities like clouds lead to a larger number of false positives. Under-sampled targets that are reduced to single pixels are often mistaken for noise.

The detection of motion caused by stationary objects reduces the number of false positives for about 48%.

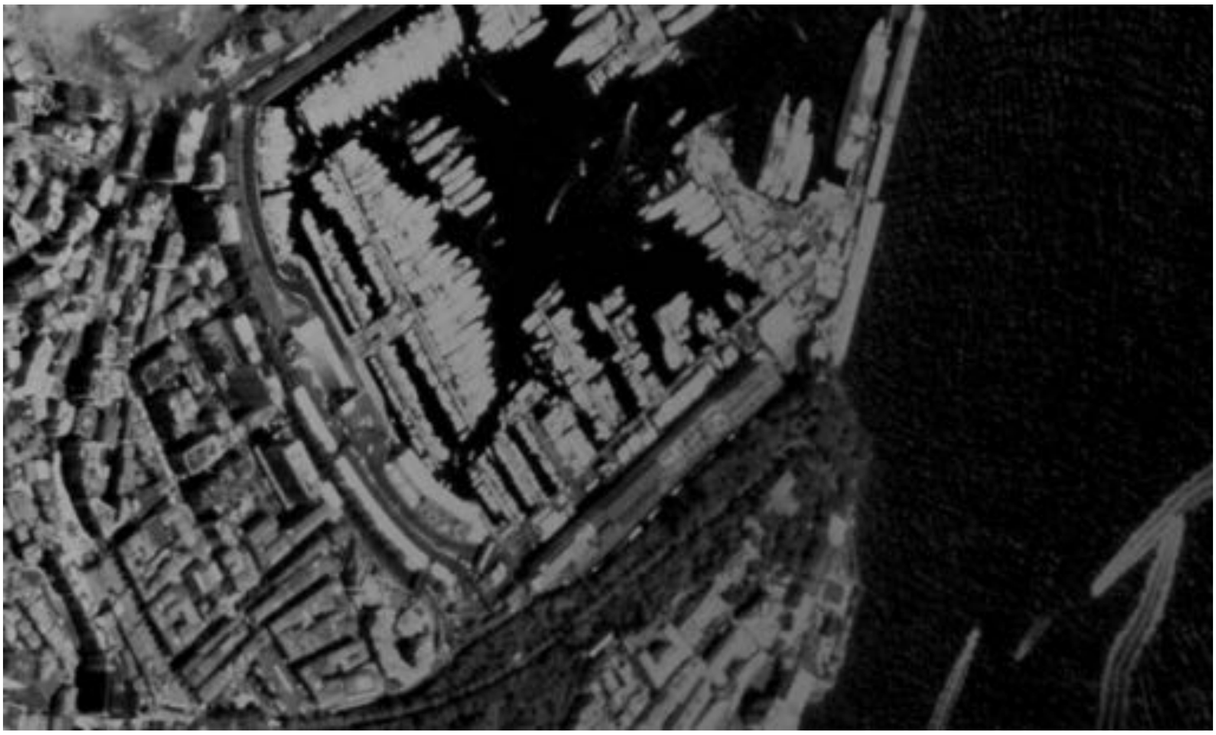


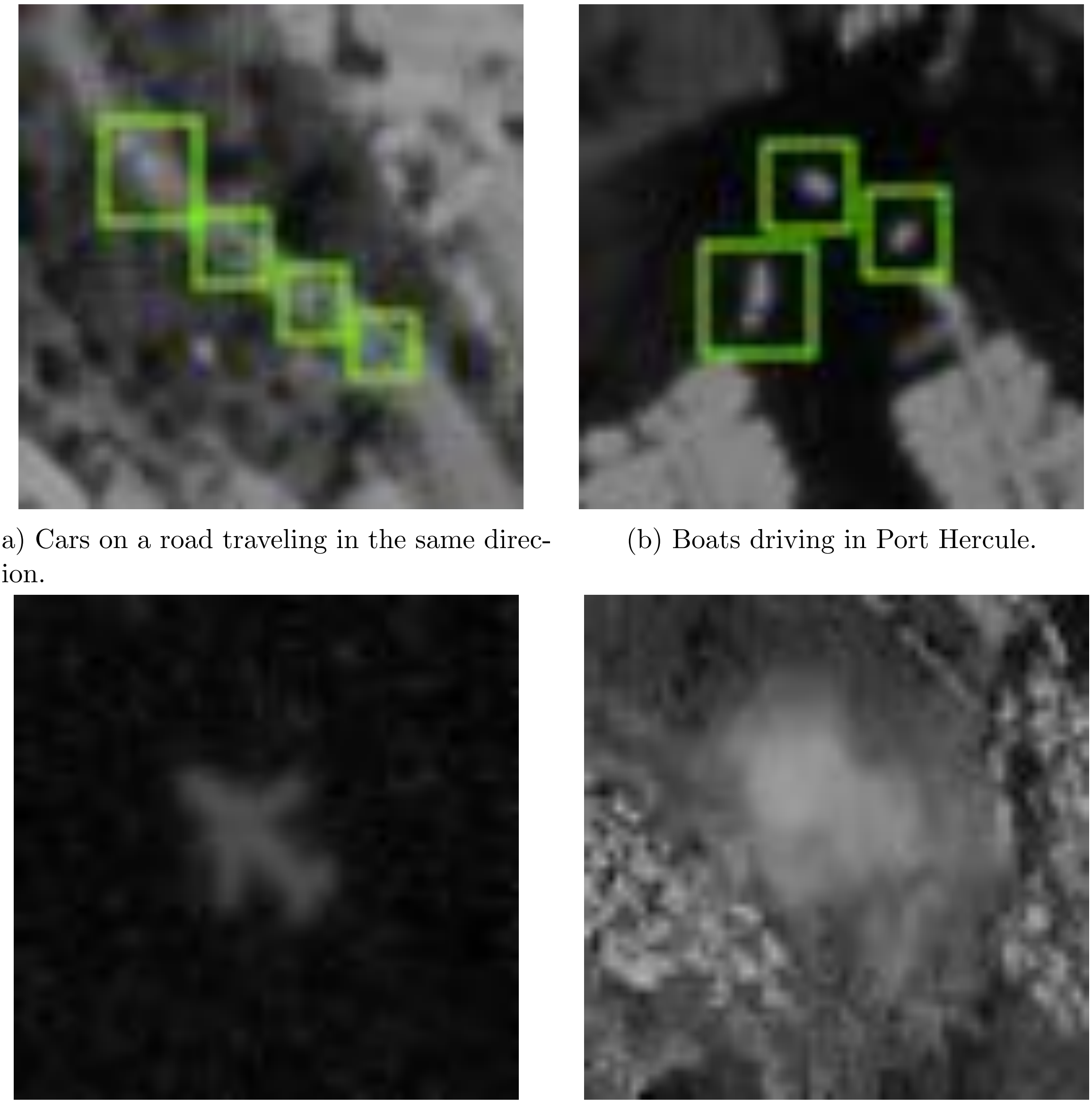
Figure 1: Satellite image of Monaco.

1 Introduction

Possible applications of satellite surveillance are numerous and reach from traffic monitoring, urban planning, road verification to border control and military reconnaissance. An important challenge in remote sensing and satellite image processing is the robust detection of larger objects like buildings and road networks, and of potentially small objects like ships, cars or trailers. A new generation of satellites, providing videos with frame-rates up to 30 frames per second and a ground sample distance of about one meter, offers new opportunities for robust object detection.

Detecting objects in satellite imagery leads to the challenge of having large search spaces and often small objects. Despite rapid progress in the last few years, the low resolution in current commercial satellite images still proposes a limitation. Robust and reliable detection of small objects like cars from space, that are only slightly larger than the resolution limit, is often hard using only spatial information. Appearance based object detection algorithms, successful with common imagery, often fail when applied to satellite images due to a lack of meaningful features [19].

This practicum is concerned with detecting moving vehicles in satellite motion imagery. The methods presented in this work aim to improve the detection of objects in satellite videos by using the available temporal resolution as additional information source. The presented approach aims at correlating moving objects from frame to frame in order to identify and classify movement.



(a) Cars on a road traveling in the same direction.

(b) Boats driving in Port Hercule.

(c) Plane captured by satellite.

(d) Clouds possible occluding objects of interest.

Figure 2: Examples of visible moving objects in given satellite video (see Figure 1).

2 Related Work

Due to the relatively new modality that is high resolution satellite video, little work has been published regarding object detection in satellite videos.

2.1 Satellite Videos

Meng et al. [22] propose an object tracking algorithm applied to high resolution multi spectral satellite images with multi-angular look capability. The frame rate at which their dataset has been captured is significantly lower than the frame rate of the dataset used in this practicum. The authors identify moving objects by creating an accumulative difference image and looking for pixels with large values. Target objects are then defined by extracting spatial and spectral features. For target matching they use a sliding window with a feature matching operator defined by three different metrics. The Bhattacharyya distance and histogram intersection are used to match spectral features and pixel count similarity is used to determine spatial matches.

Kopsiaftis et al. [18] use a background model based on an adaptive procedure where the background is dynamically estimated based on the mean intensity value of about 300 frames. They apply morphological operations and not nearer specified statistical analysis on the connected components to remove non vehicular components. Their method does not use motion compensation but uses an already stabilized dataset. They also apply prior knowledge of the exact road network.

Yang et al. [33] propose a two-step method in order to apply the detection only to regions where vehicular motion is to be expected. In the first step they use the VIBE background model to extract all motion including non vehicular motion causing false positives. They extract and analyze trajectories using a Hungarian algorithm. Trajectories that are to short or unstable are removed. A motion heat map is generated by utilizing a distance transform. The following procedures are only applied to hot areas in the motion heat map. In the second step they generate a local saliency map for each frame which accounts for low contrast between roads and vehicles. On the resulting map the authors again apply background subtraction using VIBE. The authors do not incorporate motion compensation because they used an already stabilized dataset.

Xu et al. [2] use optical flow and Shi-Tomasi features to co-register consecutive frames. They use a modified VIBE background model[3] to determine motion. To reduce artifacts from motion not caused by moving vehicles they use simple heuristics like the ratio of the bounding box enclosing the foreground candidate.

Du et al. [12] use Lucas-Kanade Optical Flow to track objects in satellite videos. Each velocity vector is transformed to a three dimensional RGB color using the HSV color space. By doing so, the authors can calculate the integral image for each band separately. The authors state that high optical flow velocities will produce a higher gray value in one band, while the other to will be lower. By finding the region near the target with the lowest integral result they aim to detect the location of the target.

2.2 Satellite Images

Chen et al. [11] apply a Hybrid Deep Convolutional Neural Network (HDNN) to satellite images obtained by Google Earth. By using a HDNN instead of a DNN they aim to extract variable-scale features.

Han et al. [14] perform object detection on satellite data where objects of interest have not been annotated. It is only known which image contains objects of interest. To overcome the lack of annotations, the authors combine weak supervised learning (WSL) with high-level feature learning.

2.3 Wide Area Motion Imagery

Liang et al. [20] combine histogram of oriented gradients (HOG) and Haar-like features in order to detect vehicles in wide area motion imagery (WAMI). Utilizing Generalized Multiple Kernel Learning they determine the trade-off between HOG and Haar-like features in form of coefficients and train the classifier. The authors evaluated their algorithm on the Columbus Large Image Format (CLIF) data set [1] and claim that it outperforms both Haar-like features and HOG when used separately. They do not use the temporal resolution at all.

LaLonde et al. [19] apply a two-stage, spatio-temporal convolutional neural network (CNN) to WAMI. The first stage (ClusterNet) aims to reduce the large search space that comes with WAMI data or with satellite images. Unless common Region Proposal Networks ClusterNet is provided with multiple video frames including the reference frame to incorporate temporal resolution. The resulting regions of objects of interest which can contain several hundred objects is sent to the second stage (FoveaNet). FoveaNet conducts high-resolution analysis on the given regions and predicts the centroids of objects of interest in the reference frame.

3 Problem Statement

The objects that we aim to detect are cars, an example of multiple cars driving in the same direction is given in Figure 2a. When examined using image manipulation tools, the cars consist of about 4 pixels. Other moving objects can be seen in Figures 2b 2c and 2d.

Cars, barely larger than the resolution limit of the satellites sensor, lack distinctive features required for appearance based solutions for object detection. Assuming that the temporal resolution of a satellite video is high enough to detect correlation of moving objects between frames, moving objects would then form distinct trajectories in a space-time domain like in Figure 3.

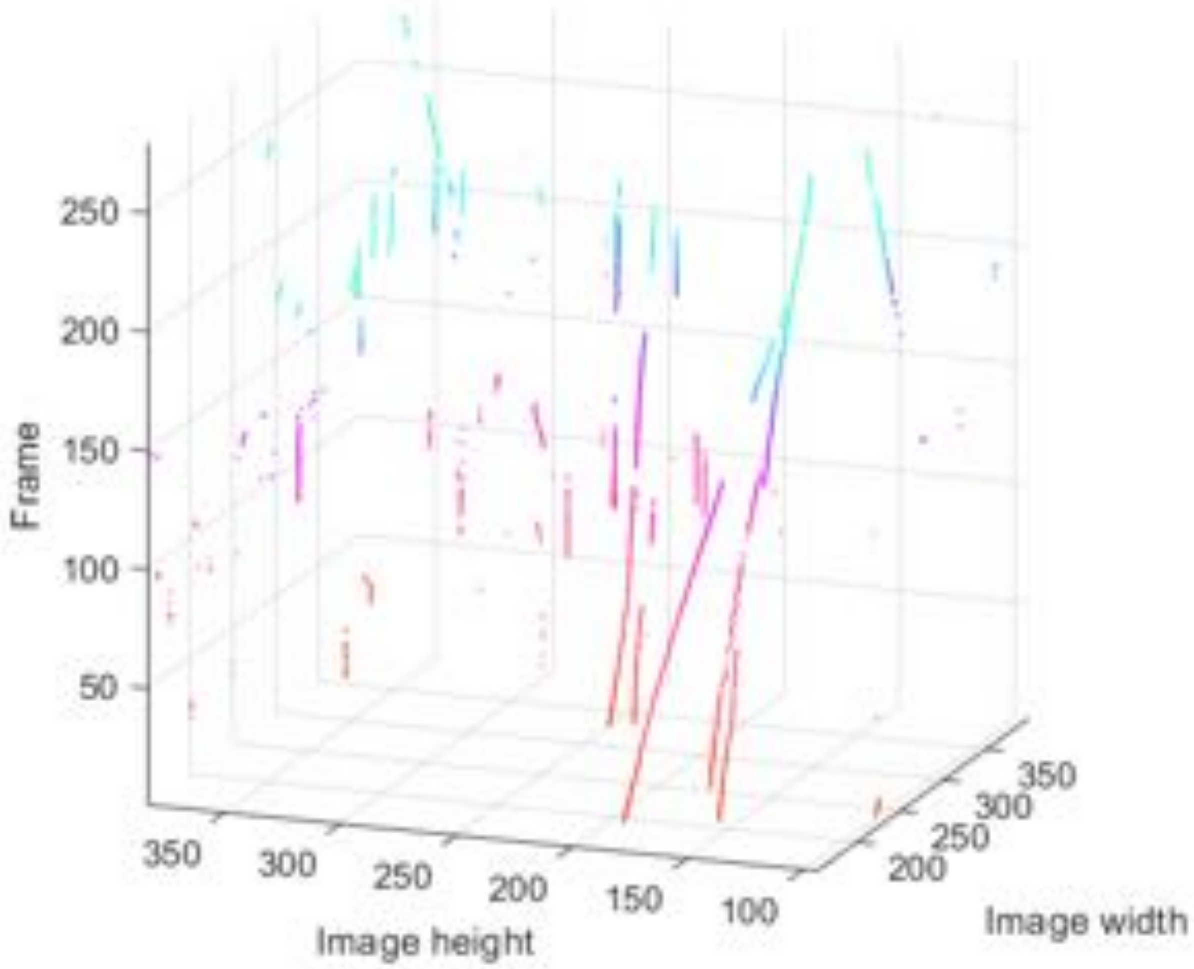


Figure 3: Different moving objects in a space-time domain.

A major problem is the movement of the satellite because the images are not co-registered. This leads to big differences between two frames. The global motion of the satellite, which changes the captured region of interest and local motion have to be discriminated. The local motion can further be sectioned into local motion mainly caused by moving vehicles and local motion not caused by moving objects.

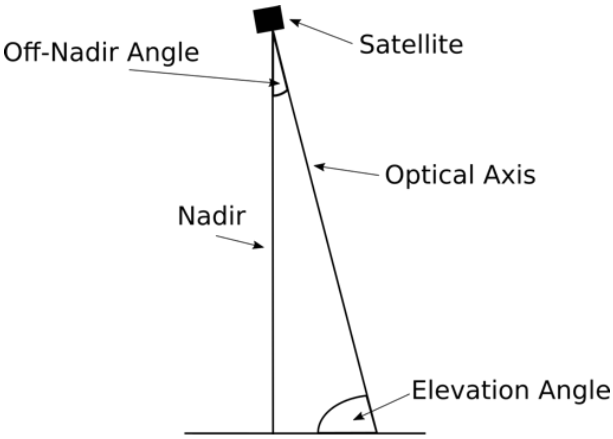


Figure 4: The elevation angle is defined by the optical axis of the satellite sensor and the horizon. The off-nadir angle is used interchangeable and is defined by the nadir and the optical axis.

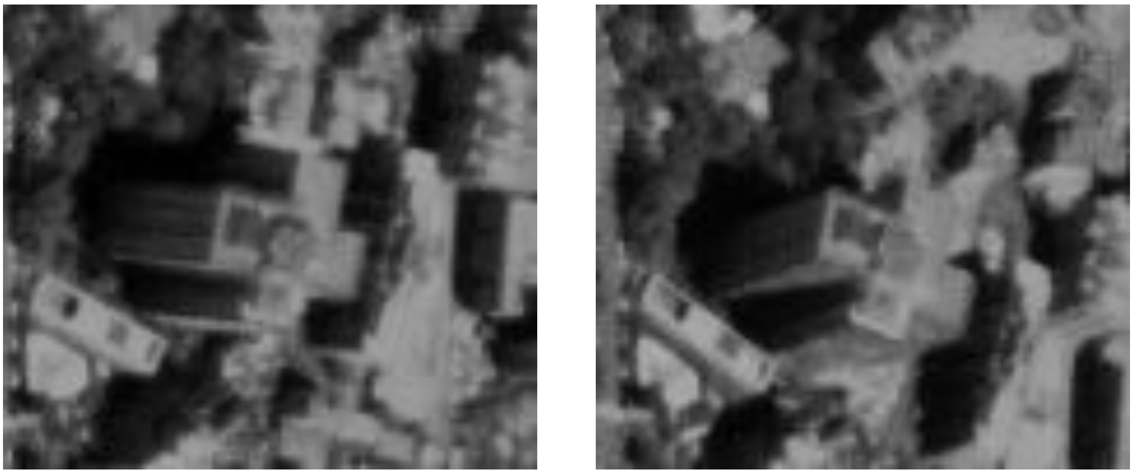


Figure 5: Image of high building in first frame (left) and last frame (right).

Motion not caused by moving objects is caused by the way the images are recorded.

Imagery captured by satellites only conform to an orthographic projection if the sensor is exactly positioned above the corresponding spot on the ground, i.e., along nadir. But satellite sensors usually capture at an angle, depicted as elevation angle or off-nadir angle, which allows for shorter revisit times and therefore improving efficiency (see Figure 4).

Consequences of capturing images off-nadir are perspective distortions that lead, among other problems, to the leaning-building effect where high objects seem to lean in a certain angle. This is especially problematic when detecting motion in satellite videos, because the angle at which the building is leaning changes with the sensor position which looks like high objects are moving. Motion detection algorithms for satellite imagery have to account for such motion effects in order to prevent false positives. The leaning-building effect over time can be seen in Figure 5.

4 Methodology

The frames of the given dataset are noisy and have low contrast. In a pre-processing step noise is reduced using a Gaussian filter kernel and contrast enhancement is performed using Histogram Equalization .

In order to detect local motion, we minimize global motion by co-registering consecutive frames. This is done finding Shi-Tomasi features [26] and tracking them along consecutive frames using optical flow [21]. After estimating the homography between each pair of frames the images are transformed to perform the video stabilization.

After extracting the foreground performing background subtraction [28], noise and larger components are discarded using simple heuristics. In order to detect motion caused by not moving objects we analyze the trajectories in space over time using Local Principal Component Analysis[6].

A flow chart of the proposed algorithm can be seen in Figure 6.

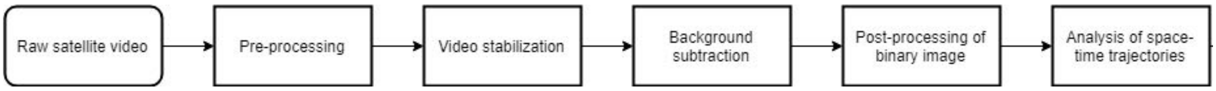


Figure 6: Flow chart of proposed algorithm.

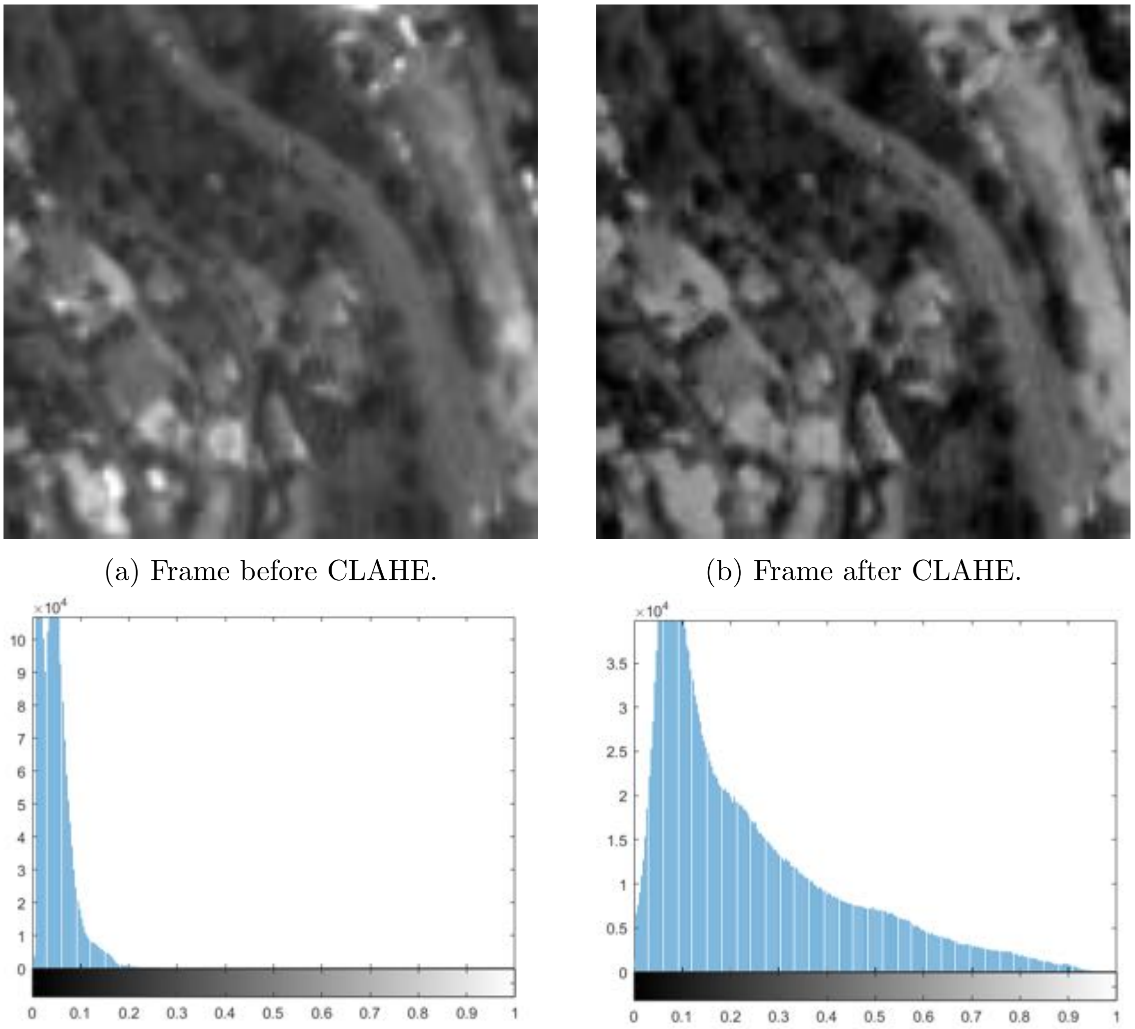
4.1 Pre-processing

The frames of the given satellite video have a high global contrast. Large reflective areas like roofs or large boats result in high intensity while, e.g., the ocean shows low intensity values. Local contrast, e.g., between cars and roads is low (see Figure 7). For detecting the foreground using background subtraction to be successful, local contrast has to be improved.

One of the simplest and most effective methods to enhance contrast is Histogram equalization (HE) [30]. HE works on the histogram of an image and spreads out the most frequent values [17]. HE only considers the global contrast, which leads to over-brightness and noise amplification because the frames of our dataset already have high global contrast (see Figure 8a). Therefore a variation of HE called Contrast Limited Adaptive Histogram Equalization [24] (CLAHE) is applied. CLAHE divides the given image into blocks (contextual regions) and applies HE to each tile individually [25]. Noise amplification is decreased by contrast limiting (see Figure 8b). Pixels contained in histogram bins that are above the specified contrast limit are clipped and redistributed. Noise reduction is performed using Gaussian filtering.

4.2 Global Motion

The frames of the dataset are not co-registered. Applying background subtraction without motion compensation would lead to a changing background and false positives. Optical flow estimation is used to compute the displacement field between two images [8]. The resulting 2D vector field is called the optical flow field. Dense optical flow describes correspondences between pixels while sparse optical flow describes correspondences between selected pixels, e.g. corners, caused by relative



(a) Frame before CLAHE.

(b) Frame after CLAHE.

(c) Histogram before CLAHE.

(d) Histogram after CLAHE.

Figure 7: Frame containing cars before and after CLAHE. The contrast between road and cars is considerable low in the top left image and got improved using CLAHE in the top right image.

motion between the scene and an observer [23]. Optical flow estimation is used for video stabilization, structure from motion and object tracking among others.

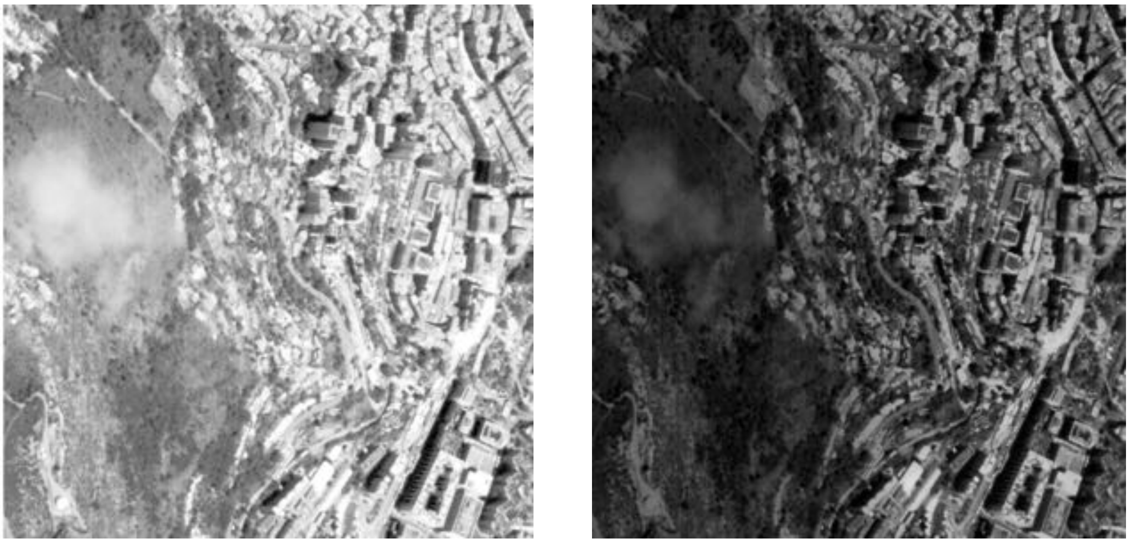
Optical flow works under the assumption of brightness constancy i.e., the intensity value of a pixel does not change between immediate frames [29].

This means the pixel intensities are translated from frame to frame [13],

$$I(\vec{x}, t) = I(\vec{x} + \vec{v}, t + 1). \tag{1}$$

$I(\vec{x}, t)$ denotes the intensity of the pixel at spatial position $\vec{x} = (x, y)$ and temporal position t, and $\vec{v} = (u_1, u_2)$ is the velocity vector. Assuming that equation 1 can be approximated by a Taylor series stopping after the first derivative:

$$I(\vec{x} + \vec{v}, t + 1) \approx I(\vec{x}, t) + \vec{v} \cdot \nabla I(\vec{x}, t) + I_t(\vec{x}, t), \tag{2}$$



(a) Image after common HE.

(b) Image after CLAHE.

Figure 8: Common HE leads to possible information loss due to over-brightness and noise amplification while CLAHE preserves information and decreases noise amplification.

where I_t denotes the temporal partial derivative of image I. Substituting equation 2 into equation 1 gives the optical flow constraint equation

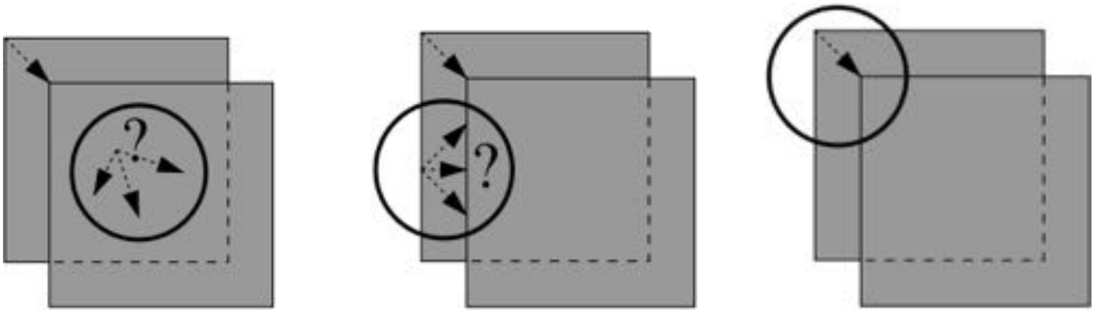
$$\nabla I(\vec{x}, t) \cdot \vec{v} + I_t(\vec{x}, t) = 0. \tag{3}$$

The image gradient ∇I and the gradient along time I_t are known but the velocity vector $\vec{v}$ is unknown which leads to an infinite number of solutions because we only have a single equation [31] which is not sufficient to solve both components of $\vec{v}$. Methods used to solve this problem consider additional assumptions considering properties of the optical flow field.

The aperture problem is a result of the indefinite optical flow constraint equation. [5]. As a consequence of the optical flow constraint, the component of the flow in the direction of the gradient is determined, but the component of the flow along an edge cannot be determined. If an untextured object is viewed through an aperture, it is not possible to extract the motion within the objects area. At edges without additional information the motion can be estimated only along the gradient in one dimension. In order to extract two dimensional motion more information is needed like corners or texture.

The optical flow algorithm applied in the proposed framework is the Lucas-Kanade method. This method assumes that the flow in a local neighborhood of the candidate pixel is more or less constant. It solves the optical flow constraint equation by calculating the optical flow for each pixel contained in the local neighborhood. Using a $3x3$ neighborhood this leads to 9 equations and is therefore overdetermined and can be solved using the least squares criterion. The Lucas-Kanade method is more robust to noise than point-wise methods which is important given the noisy dataset [4]. Because Lucas-Kanade computes a sparse optical flow, it needs feature points to track.

For each frame pair Shi-Tomasi features [26] are extracted (11a). Shi-Tomasi



(a) No texture is present, therefore the motion can not be extracted.

(b) Only one edge is present, the motion can only be estimated along one dimension.

(c) A corner is present and the motion can be estimated.

Figure 9: Visualization of the aperture problem[5].

features are chosen with tracking in mind. The principle is the same as with the Harris Corner Detector [15] but they use a different scoring function. The Harris Corner Detector and Shi-Tomasi features use a sliding window. The goal is determining patches in the given image I where the sliding window generates large variations when moved around. The difference between the original and moved window is given by

$$E = \sum_{x,y} w(x,y)[I(x+u, y+v) - I(x,y)]^2 \tag{4}$$

where $w(x,y)$ is the sliding window at position (x,y), u the offset of the window in x direction and v the offset of the window in y direction. Using a first order Taylor approximation E can be expressed as

$$E \approx \begin{bmatrix} u \\ v \end{bmatrix} M \begin{bmatrix} u & v \end{bmatrix} \tag{5}$$

where M is the second moment matrix

$$M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{6}$$

from which the eigenvalues λ_1 and λ_2 are extracted. The Shi-Tomasi scoring function is given by:

$$R = min(\lambda_1, \lambda_2). \tag{7}$$

By effectively applying a threshold to the smaller eigenvalue the authors determine that λ_1 and λ_2 are sufficiently large to correspond with a large variation and thus a reliable trackable pattern like corners or salt-and-pepper textures. Trackable in this context means the extracted feature points contain enough information to estimate motion in all image planes. This problem is known as the above mentioned aperture problem (see Figure 9).

After calculation of features in one frame, they are tracked to the subsequent frame. As error metric for the tracking, the Forward-Backward error [16] has been implemented. It is based on the assumption that video tracking is symmetric regarding the image sequence. The ground truth of a tracked feature is not known, but the path has to be the same in both tracking directions:

$$Track(I_1, I_2, I_3, \ldots) \equiv Track(\ldots, I_3, I_2, I_1) \tag{8}$$

To calculate the Forward-Backward error feature point X_t is tracked to X_{t+1} and then to X_{t+k}. X_{t+k} is then tracked backwards to $\widehat{X_{t+1}}$ and then to $\hat{X}_t$. If the error between Xt and $\hat{X}_t$ is greater than a certain threshold, the feature point is discarded (see Figure 10). All remaining feature points are used to estimate the homography between each frame pair. These homographies are then used to register the frame pairs to each other, which minimizes global motion (see Figure 11b).

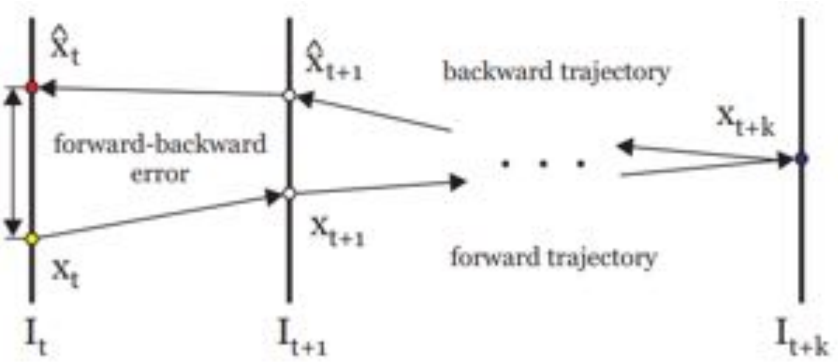
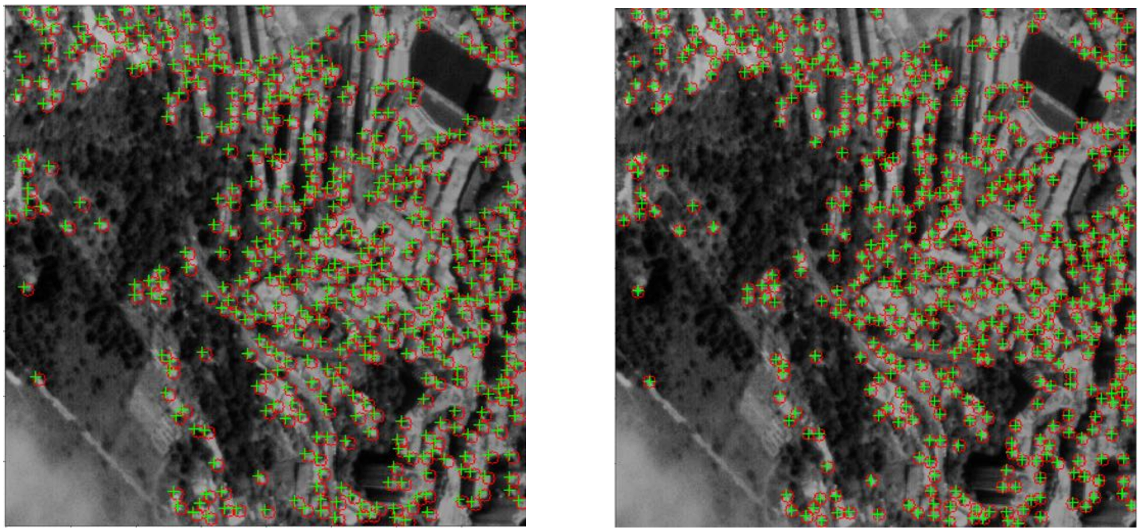


Figure 10: Forward-Backward error calculation is used to determine faulty registration between frames.



(a) Corresponding feature points before motion compensation.

(b) Corresponding feature points after motion compensation.

Figure 11: Illustration of motion compensation. Red circles correspond to the reference frame and green crosses to the frame to be aligned.

4.3 Local Motion

After compensation for the global motion local motion is estimated using Background Subtraction. The background model is represented by a mixture of Gaussians.

Instead of modeling each pixel with a single Gaussian a multitude of adaptive Gaussians is used for each pixel. This allows not only to account for lighting changes over time but multiple surfaces in the area containing the given pixel. The different values of a pixel over time are called its history. The history of a pixel x_0, y_0 at time t is denoted by

$$\{X_1, \ldots, X_t\} = I(x_0, y_0, i) : 1 \leq i \leq t, \tag{9}$$

I representing the image sequence. The probability of a pixel value X_t modeled by a mixture of K Gaussian distributions is given by

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}). \tag{10}$$

$\omega_{i,t}$ is the estimated weight of the i^{th} Gaussian distribution at time t, $\mu_{i,t}$ is the mean of the i^{th} Gaussian distribution at time t and $\Sigma_{i,t}$ is the covariance matrix of the i^{th} Gaussian distribution at time t. η is a Gaussian probability function. Every time there is a new pixel value it is matched to one of the Gaussian distributions using a K-means approach. Stauffer et. al define a match as having a pixel value within 2.5 standard deviations of a distribution [28]. If no match can be found the distribution with the lowest probability is replaced with a new Gaussian having the pixel value as its mean. The background model is ultimately composed of those Gaussian distributions having the most supporting evidence and the least variance. The result of background subtraction is a binary image for every frame depicting the detected foreground components.

Background subtraction using Gaussian mixture models performs well with light changes, cluttered-regions and slow moving objects. Most importantly it is able to learn repetitive variations which makes it more robust to the motion caused by not moving objects in the used satellite video which is mostly repetitive. It has been developed with outdoor scenes in mind [28].

4.4 Post-processing of binary image

The binary image obtained from background subtraction contains not only motion from vehicles and other independent objects, but also noise and artifacts from not moving objects like high buildings and areas suffering from geometric distortions (see Figure 15a). In order to avoid false positives a few simple heuristics are employed. Connected components containing less than 5 pixels or more than 25 pixels are discarded. Despite cars being about 4 pixels large in the original video resulting components from the background subtraction step are larger in diameter.

Connected components with a bounding box ratio larger than 1:2 are also removed. They usually represent borders of large buildings, a result from motion artifacts caused by the elevation angle of the image sensor (see Figure 15b).

4.5 Detection of motion not caused by moving vehicles

High structures, mainly buildings, appear to move because satellite images are captured at an off-nadir angle as illustrated in Figure 4. This effect is characterized by constant speed and direction for all corresponding components. In order to detect motion caused by stationary structures, the space-time trajectories of the centroids of all remaining components are analyzed. In a first step, a point cloud is generated from the centroids of all connected components over time (see Figure 12).

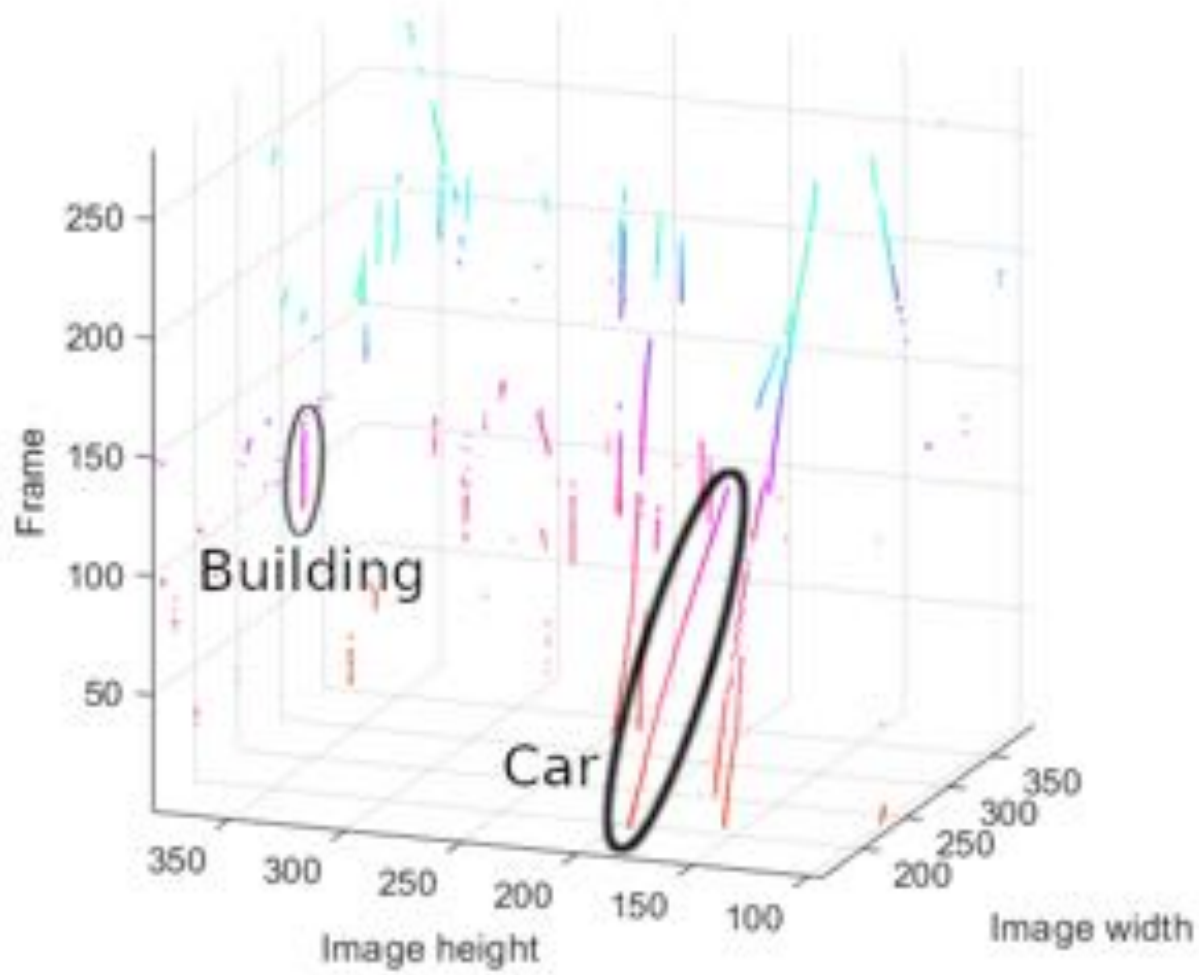


Figure 12: Point cloud generated from centroids of connected components over time. Sloped trajectories depict faster moving objects than trajectories parallel to the frame axis.

Principal component analysis (PCA) is used to analyze and dimensionally compress multivariate data [10].

Given a set of possibly correlated values, PCA enables us to compute a set of values of linearly uncorrelated variables which are called principal components. Using an orthogonal transformation, PCA is used to re-express the given data-set with another basis, which is a linear combination of the original basis [27].

In order to re-express given data in the most meaningful way PCA maximizes the signal to noise ratio (SNR). The SNR can be interpreted as ration of variances:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{11}$$

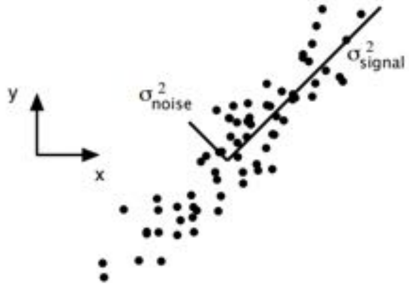


Figure 13: 2D data-set illustrating signal and noise variances. The direction of the largest variances are depicted by two lines [27].

In data with a high SNR structures of interest are on the axes with maximal variance (see Figure 13). Figure 14 depicts different levels of redundancy in 2D datasets with two separate measurements $r1$ and $r2$. In the right panel one can examine two highly correlated variables. In order to reduce dimensionality, it would be more efficient to calculate, e.g., $r1$ from $r2$. The direction of the largest variance corresponds with the largest eigenvector of the covariance matrix of the data. Finding the principal components is therefore done using eigenvector decomposition.

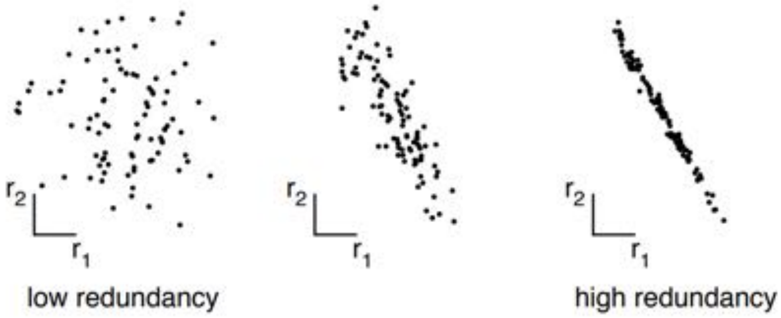
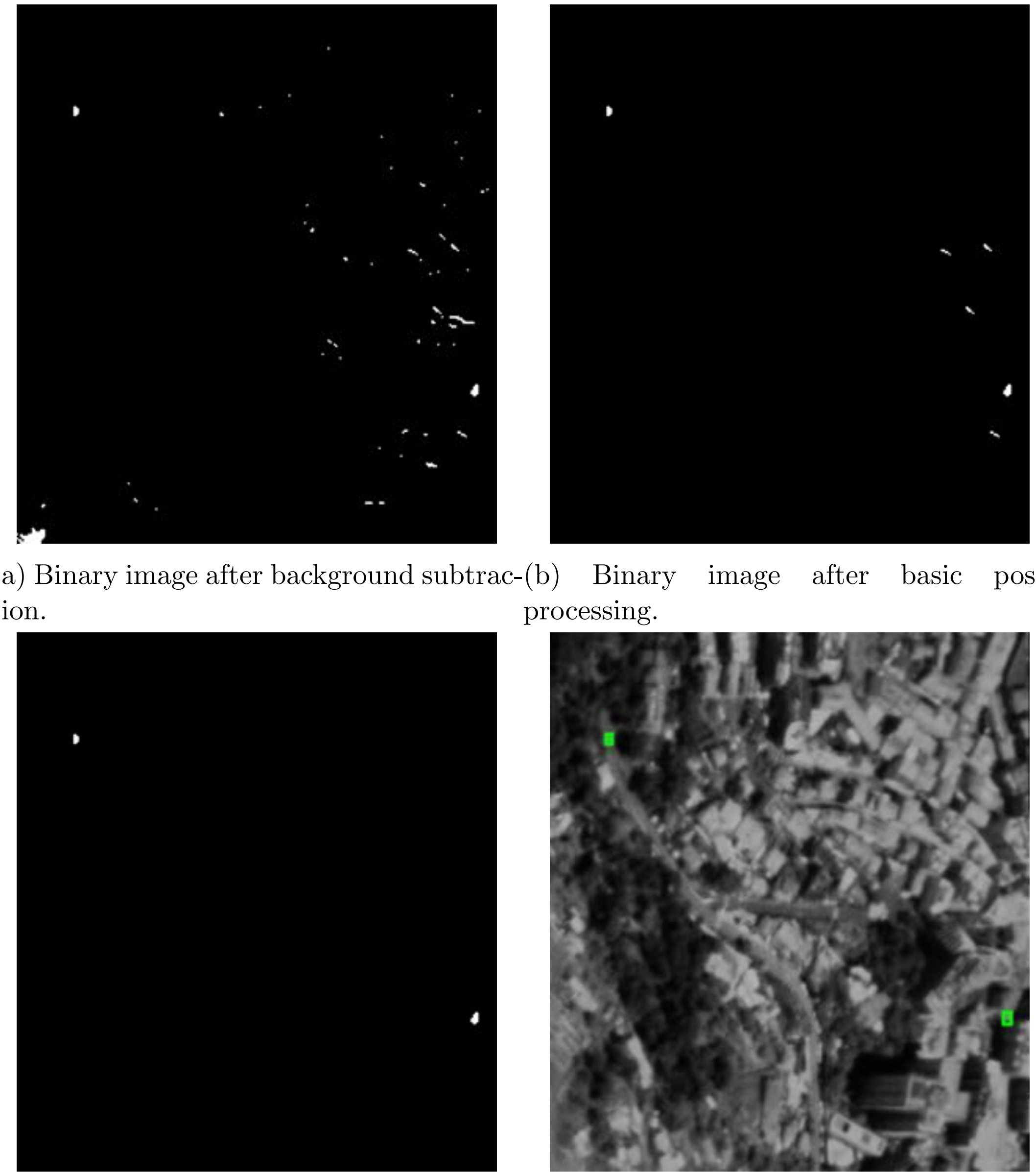


Figure 14: Different degrees of redundancy in a 2D data-set [27].

Transferring the motion data into a point cloud produces trajectories for both moving objects and motion caused by stationary objects. The motion data over time is correlated within a local neighborhood. Assuming that trajectories caused by cars and motion effects are locally linear we can extract them by analyzing the point cloud holding the centroids of connected components over time. Therefore motion caused by stationary objects is detected by a local variant of PCA in the next step. Local PCA (LPCA) [32] applies the PCA algorithm only to a subset of data points [6]. By moving the analysis window along the first principal component linear trajectory segments are obtained. The obtained trajectory segments are examined to discriminate real moving objects from artifacts caused by other structures. Because artifacts caused by stationary structures are results of the position and alignment of the satellite sensor, trajectory segments caused by them exhibit similar orientation and length. Therefore similar trajectories are clustered using K-means clustering. For each cluster the mean angle between trajectory and the time-axis is determined. If it falls below a certain threshold, the corresponding connected components to the trajectories in the current cluster are classified as motion not caused by moving objects.



(a) Binary image after background subtraction.

(b) Binary image after basic post-processing.

(c) Binary image after analysis of space-time trajectories.

(d) Frame with overlaid detections.

Figure 15: After background subtraction basic heuristics are used to eliminate components that unlikely correspond to vehicles. Elements resulting from motion artifacts are removed by analyzing their space-time trajectories.

5 Results

The data set consists of a 30 seconds video provided by the company Planet Labs Inc.[1]. The video is composed of 900 frames, depicting the view of Monaco as can be seen in Figure 1. The provided video is not stabilized, i.e., the scene and the camera are constantly in motion. The motion imagery has been captured by a Skysat satellite in the panchromatic spectrum. Ground sample distance is between $0.72m$ and $0.86m$. No radiometric correction for atmosphere or other geometric distortions has been applied[2]. The frames have a resolution of 2560 x 1080 pixels.

The proposed method has been evaluated on 3 different sections of the provided satellite video each containing 250 frames. The sequences have been labeled manually beforehand using bounding boxes. Pre-processing and motion compensation has been implemented using the OpenCV library [7], while the remaining steps have been implemented using Matlab. The parameters for each step have been chosen empirically. They consist of the number of Gaussian distributions used for the background model, the learning rate which specifies how fast the model parameters are updated, the angle between a trajectory and the frame axis at which an object in considered as local motion and the size of the LPCA filter kernel in pixels. The chosen parameters can be seen in Table 1.

Number of Gaussians	Learning rate	Angular threshold	Kernel size
5	0.01	5 degrees	5 pixels

Table 1: Empirically determined parameters.

The ROIs have been chosen in order to represent different conditions like density of cars and roads or the presence of large buildings. **ROI 1** (see Figure 16a) features a narrow network of streets with many occlusions and a tower prone to the leaning-building effect.

ROI 2 (see Figure 16b) shows a sparsely used road without high buildings but big clouds that are occluding the targets.

ROI 3 (see Figure 16c) depicts high buildings as well as large differences in terrain height. The distribution of cars is dense.

The results for all three regions of interest (ROI) with omitting the non vehicular motion detection are listed in Table 2. The evaluations using non vehicular motion detection can be seen in Table 3.

In ROI 1 and especially ROI 2 clouds entering the frame are leading to an increased number of false positives. Activating detection of local motion not caused by moving objects effectively decreases the number of false positives in both ROIs while preserving true positives and false negatives, but slightly decreasing true negatives. Motion artifacts from stationary objects are strong in ROI 3 and lead to an increased number of false positives in combination with moving clouds. The recall rate without non vehicular motion detection leads to a slightly higher F1 score.

Background subtraction using Gaussian mixtures did not detect cars with low contrast in the presence of noise which leads to an overall poor recall rate. Due

[1] https://www.planet.com/

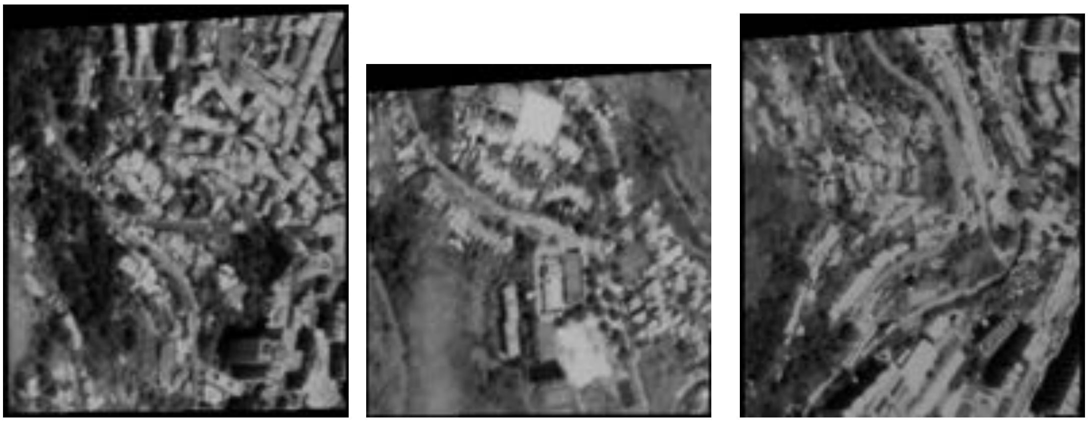
[2] https://www.planet.com/products/satellite-imagery/files/Planet_Combined_Imagery_Product_Specs_December2017.pdf

	ROI 1	ROI 2	ROI 3
TP	751	230	2083
FP	628	3034	1018
TN	7552	13225	6097
FN	532	280	4122
Precision	0.54	0.07	0.67
Recall	0.58	0.45	0.34
F1 score	0.56	0.12	0.45

Table 2: Results of the proposed method without detection of motion caused by stationary objects.

	ROI 1	ROI 2	ROI 3
TP	670	227	1832
FP	243	1592	596
TN	7937	14667	6519
FN	613	283	4373
Precision	0.73	0.12	0.75
Recall	0.52	0.44	0.3
F1 score	0.61	0.19	0.42

Table 3: Results of the proposed method including detection of motion caused by stationary objects.



(a) ROI 1. (b) ROI 2. (c) ROI 3.

Figure 16: The three scenes used for evaluation.

to the resolution of the satellite sensor, vehicles often vanish for a couple of frames or are reduced to a single pixel. The vanishing of vehicles could be improved by using some kind of prediction model, e.g. based on Kalman filtering. Choosing the sampling rate parameter of the background subtraction model is very sensitive as can be seen in Figure . A low learning rate leads to false negatives while a high learning rate increases noise and detection of motion not caused by moving objects. The sampling rate has been chosen to not lead to cluttered scenes because clutter leads to wrong trajectories in the LPCA step.

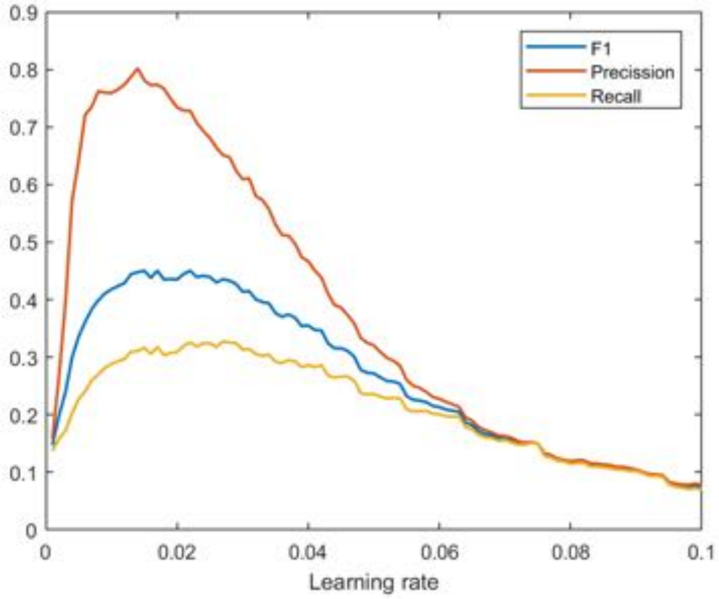


Figure 17: Influence of the learning rate parameter for background subtraction on the result of object detection.

The proposed detection of motion caused by stationary objects falsely classifies slowly moving cars traveling in a similar direction as fake local motion. This leads to an increase of false negatives. Still, it reduced the detection of false positives for about 48%.

The proposed algorithm does not deal well with clouds because they often result in multiple small connected components which cannot be distinguished from moving vehicles. Also they tend to occlude other moving objects.

6 Conclusion and Future Work

Object tracking in satellite videos bears multiple challenges. Due to motion effects from stationary objects, objects near the sampling resolution, noise, clouds and low contrast robust tracking remains a challenge.

A method has been proposed that uses temporal information instead of spatial features to identify moving objects. The problem of motion not caused by moving objects has been formulated and an approach for reducing it has been shown. In order to use background subtraction the video is stabilized using Shi-Tomasi features and optical flow. Using CLAHE the local contrast between vehicles and roads has been improved.

Experiments have been conducted on 3 different sections of our dataset video. The video has been manually annotated in order to gain meaningful evaluations. for each ROI the calculation of recall, precision, and F1 score is used to rate the tracking performance.

Especially the parameters of the background subtraction set are very sensitive to the underlying video material. Noise and artifacts resulting from background subtraction are dealt with using heuristics on the properties of the connected components. Noise to an extent is removed successfully but also small under-sampled cars are removed. The proposed method for detection of motion not caused by moving objects works well on not cluttered scenes with distinct trajectories. Therefore the parameters for the background model have to be chosen accordingly as not to amplify effects caused by stationary objects.

As the underlying background subtraction algorithm has proved itself to be rather sensitive to the given data, the focus for future work lays in finding more robust methods to extract local motion. Viable options would involve, e.g., Large Displacement Optical Flow[9]. The use of supervised learning methods is restricted due to the lack of training data. The algorithm for detecting motion caused by stationary objects could be improved by not only analyzing each trajectory segments, but clusters of segments. Incorporating Structure from motion could be used to detect large objects.

References

[1] Columbus large image format dataset. https://www.sdms.afrl.af.mil/index.php?collection=clif2007, 2007. [Online; accessed 20-November-2018].

[2] Aigong Xu, Jiaqi Wu, G. Z. S. P. T. W. Y. J., and Shen, X. Motion detection in satellite video. *Journal of Remote Sensing & GIS 6*, 2 (2017), 1–9.

[3] Barnich, O., and Van Droogenbroeck, M. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing 20*, 6 (2011), 1709–1724.

[4] Bauer, N., Pathirana, P., and Hodgson, P. Robust optical flow with combined lucas-kanade/horn-schunck and automatic neighborhood selection. In *2006 International Conference on Information and Automation* (Dec 2006), pp. 378–383.

[5] Beauchemin, S. S., and Barron, J. L. The computation of optical flow. *ACM Comput. Surv. 27*, 3 (Sept. 1995), 433–466.

[6] Beleznai, C., Fruhstuck, B., and Bischof, H. Multiple object tracking using local pca. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (2006), vol. 3, IEEE, pp. 79–82.

[7] Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[8] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. High accuracy optical flow estimation based on a theory for warping. In *Computer Vision - ECCV 2004* (Berlin, Heidelberg, 2004), T. Pajdla and J. Matas, Eds., Springer Berlin Heidelberg, pp. 25–36.

[9] Brox, T., and Malik, J. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 3 (2011), 500–513.

[10] Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM) 58*, 3 (2011), 11.

[11] Chen, X., Xiang, S., Liu, C., and Pan, C. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters 11*, 10 (Oct 2014), 1797–1801.

[12] Du, B., Cai, S., Wu, C., Zhang, L., and Tao, D. Object tracking in satellite videos based on a multi-frame optical flow tracker. *CoRR abs/1804.09323* (2018).

[13] Fleet, D., and Weiss, Y. *Optical Flow Estimation.* Springer US, Boston, MA, 2006, pp. 237–257.

[14] Han, J., Zhang, D., Cheng, G., Guo, L., and Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing 53*, 6 (June 2015), 3325–3337.

[15] Harris, C., and Stephens, M. A combined corner and edge detector. In *Alvey vision conference* (1988), vol. 15, Citeseer, pp. 10–5244.

[16] Kalal, Z., Mikolajczyk, K., and Matas, J. Forward-backward error: Automatic detection of tracking failures. In *2010 20th International Conference on Pattern Recognition* (Aug 2010), pp. 2756–2759.

[17] Kim, Y.-T. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE Transactions on Consumer Electronics 43*, 1 (Feb 1997), 1–8.

[18] Kopsiaftis, G., and Karantzalos, K. Vehicle detection and traffic density monitoring from very high resolution satellite video data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (July 2015), pp. 1881–1884.

[19] LaLonde, R., Zhang, D., and Shah, M. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In *Computer Vision and Pattern Recognition* (2018).

[20] Liang, P., Teodoro, G., Ling, H., Blasch, E., Chen, G., and Bai, L. Multiple kernel learning for vehicle detection in wide area motion imagery. In *2012 15th International Conference on Information Fusion* (July 2012), pp. 1629–1636.

[21] Lucas, B. D., and Kanade, T. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2* (San Francisco, CA, USA, 1981), IJCAI'81, Morgan Kaufmann Publishers Inc., pp. 674–679.

[22] Meng, L., and Kerekes, J. P. Object tracking using high resolution satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 5*, 1 (Feb 2012), 146–152.

[23] Patel, D., and Upadhyay, S. Optical flow measurement using lucas kanade method. In *International Journal of Computer Applications* (01 2013), vol. 61, pp. 6–10.

[24] Pizer, S. M. Intensity mappings to linearize display devices. *Computer Graphics and Image Processing 17*, 3 (1981), 262 – 268.

[25] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing 39*, 3 (1987), 355–368.

[26] Shi, J., and Tomasi, C. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Jun 1994), pp. 593–600.

[27] Shlens, J. A tutorial on principal component analysis. *CoRR abs/1404.1100* (2014).

[28] Stauffer, C., and Grimson, W. E. L. Adaptive background mixture models for real-time tracking. In *cvpr* (1999), IEEE, p. 2246.

[29] Sun, D., Roth, S., Lewis, J. P., and Black, M. J. Learning optical flow. In *Proceedings of the 10th European Conference on Computer Vision: Part III* (Berlin, Heidelberg, 2008), ECCV '08, Springer-Verlag, pp. 83–97.

[30] Vasile V. Buzuloiu, Mihai Ciuc, R. M. R. C. V. Adaptive-neighborhood histogram equalization of color images. *Journal of Electronic Imaging 10* (2001), 10 – 10 – 15.

[31] Wedel, A., and Cremers, D. *Optical Flow Estimation.* Springer London, London, 2011, pp. 5–34.

[32] Xu, L. Multisets modeling learning: an unified theory for supervised and unsupervised learning. In *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on* (June 1994), vol. 1, pp. 315–320 vol.1.

[33] Yang, T., Wang, X., Yao, B., Li, J., Zhang, Y., He, Z., and Duan, W. Small moving vehicle detection in a satellite video of an urban area. *Sensors 16*, 9 (2016), 1528.