3D-Connected Components Analysis for Traffic Monitoring in Image Sequences Acquired from a Helicopter

Matthieu Molinier, Tuomas Häme, Heikki Ahola

VTT Technical Research Center of Finland, Remote Sensing Group,
P.O.Box 1201, FIN-02044 VTT, Finland
{matthieu.molinier, tuomas.hame, heikki.ahola}@vtt.fi
http://www.vtt.fi/tte/research/tte1/tte14/

Abstract. The aim of the study was to develop methods for moving vehicle tracking in aerial image sequences taken over urban areas. The first image of the sequence was manually registered to a map. Corner points were extracted semi-automatically, then tracked along the sequence, to enable video stabilisation by homography estimation. Moving objects were detected by means of adaptive background subtraction. The vehicles were identified among many stabilisation artifacts and tracked, with a simple tracker based on spatiotemporal connected components analysis. While the techniques used were basic, the results turned out to be encouraging, and several improvements are under scrutiny.

1 Introduction

Traffic monitoring in urban areas has become a necessity for transportation and infrastructure planning. Ground-based video surveillance systems partially fulfil this need. However, this kind of systems are not easily transportable as they are often adapted to a given road configuration, and only cover a narrow area. On the contrary, aerial imagery acquired from a helicopter offers both a mobile system and a wide field of view, for monitoring several urban traffic spots in a same image sequence.

Hoogendoorn et. al [12] used a straightforward approach to monitor highway traffic from a helicopter, with camera motion compensation and residual motion detection by background subtraction. 90% of the vehicles were successfully detected by thresholding, then tracked by template matching. Yet this approach relied on greyscale imagery, used a same background image over the sequence regardless of illumination changes, and its efficiency in congested situations was altered. Han and Kanade [9] proposed a method to recover, given a monocular image sequence, the scene structure, the trajectories of the moving objects and the camera motion simultaneously. The technique is appealing since it does not require any camera motion compensation to detect and track vehicles. It has been successfully applied on an aerial video, but relies on the assumption that the objects are in linear motion. Fuse et al. [4], [5] developed a system for urban traffic monitoring with colour image sequences acquired from high-altitude

platforms. It includes video stabilisation, background subtraction then update by Kalman filtering, and vehicle tracking by an original spatiotemporal clustering of two features - the value after background subtraction and the optical flow value. Vehicle shadows were discarded following [3]. The overall approach turned out to be very effective, but is rather complicated and requires 8 thresholds. Moreover, little is said about how to handle artifacts due to an inaccurate video stabilisation followed by background subtraction. Still, an interesting conclusion of their study is that temporal resolution (frame rate) is more important than spatial resolution, if one wants to achieve good tracking performance.

Aiming at an operative system, we present a method for moving vehicle detection and tracking in high-resolution aerial colour image sequences. The goal of this study was to accurately estimate the speed of traffic objects like cars, buses or trams, and make the information available on GIS (Geographic Information System) for later analysis. The long-term purpose of this work is to collect traffic data over wide urban areas by means of aerial imagery, and assess overall traffic conditions for city planning. The method relied on simple techniques, yet achieved satisfying results.

2 Data and Pre-processing

2.1 Material

Acquisition campaigns were made over the city of Helsinki in June 2002, with a helicopter and a standard digital camera oriented near nadir direction. Monocular 25fps colour videos were acquired at altitudes ranging from 200m to 500m, over traffic spots like crossroads. The image sequence used in this study consists of 500 uncalibrated frames taken at 200m during 20s ; each frame has 768*560 pixels and a 17cm ground resolution (Fig. 1(a)). City of Helsinki provided us with digital map of the monitored urban areas (Fig. 1(b) covers an area of 134*110m).

2.2 Image Sequence Registration

Before detecting moving objects, the camera motion needed to be compensated. Following Censi et al. [1], the scene was assumed to be planar. Points of interest were manually extracted in the first frame, then tracked along the sequence. Other frames were automatically registered to the reference image, through the robust estimation of the plane-to-plane homography that links images of a planar scene taken at different viewpoints.

Homography Estimation between two Consecutive Frames. Because a fully automated selection of relevant points proved challenging in such a complex urban scene (with a lot of salient objects), points were first selected manually on the road surface. The Harris corner detector [10] extended to colour images [13] was then run to refine point extraction. These points were tracked in the following

frames by considering the local maximum response of the corner detector around their previous locations.

The homography can be estimated with 4 point correspondences. In practice, 20 point correspondences were established to solve an overconstrained linear system through Singular Value Decomposition [1], [7]. Points incorrectly tracked were declared as outliers by the X84 rule [6] ; noting $\widehat{H}$ the homography estimated between points p_j^k of frame I^k and p_j^{k+1} of frame I^{k+1}, outliers were found by examining statistics of the residuals r_j :

$$r_j = \|p_j^k - \widehat{H}p_j^{k+1}\| . \tag{1}$$

Only inliers were used for the final estimation of H. The expected positions of those points that generated an outlier residual were guessed by applying the robustly estimated homography [1]. The final homography estimate was applied to frames by inverse mapping, using nearest neighbour interpolation.

Image Sequence Geocoding and Homography Composition. Under the same assumption of scene planarity, a homography was estimated between the first image of the sequence and the GIS map, after selecting and tieing points manually, as shown in Fig. 1. The first frame was then aligned to the map, to produce a geocoded image - Fig. 2.

(a) Corner points in 1st frame   (b) Corresponding points in GIS
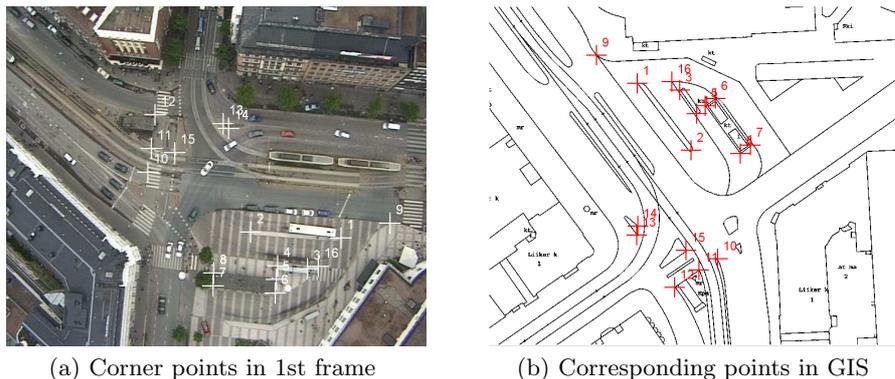
Fig. 1. Manual image-to-map registration

The actual point correspondences used for the homography estimation, in the stabilisation stage, were between a given frame and the geocoded first frame. By doing so, each frame of the sequence was only warped once to generate a geocoded, stabilised video sequence. The whole process of video stabilisation and geocoding is summed up on Fig. 3. Once the camera motion was compensated, the image sequence looked almost as if taken with a static camera, with the exception of buildings - mainly because of parallax effect. Being of no interest for vehicle tracking, the buildings were masked out before further processing.
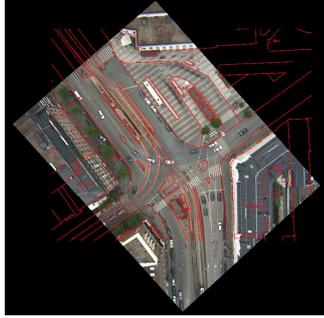
Fig. 2. First frame of the sequence registered to GIS
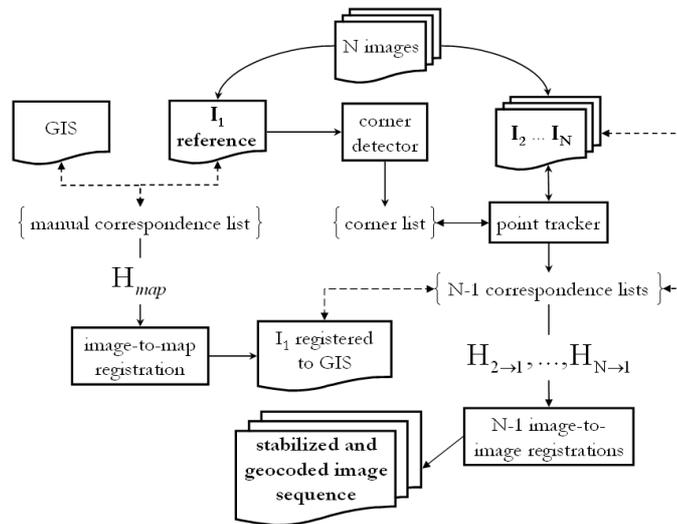
Fig. 3. Image sequence stabilisation and geocoding

3 Methods

3.1 Motion Detection by Adaptive Background Subtraction

Background subtraction is a widely used technique for motion detection in traffic monitoring applications [2]. In its adaptive form, a model of the background is generated and updated as the image sequence runs [8], to account for illumination changes. In our study, detection was carried out in HSV colour space (Hue Saturation Value), considering only the V component.

Background Initialisation. Because a vehicle-free image of a traffic scene does not generally exist, it has to be generated. The median was computed

pixel-wise over the 101 first frames of the sequence, forming an initial image of the background. The same approach for background initialisation was used in [4]. This requires that some frames of the sequence are known in advance ; in a real-time system, that could be done during a setup phase.

Because the video was not perfectly stabilised[1], computing the median over the sequence also introduced a spatial averaging : the background image appeared smoothed. Still, this resulted in a better initial background image than if we had taken the first image of the sequence. When initialising the background with the first frame of the sequence, all moving vehicles in the following frames already appear in the background, delaying their detection by the time the background updates.

Moving Object Detection. Once a background image was available, it was subtracted to the current image, forming the difference image (DI). Double thresholding was applied to DI in order to detect moving vehicles brighter (high threshold t_H) and darker (low threshold t_L) than the background. Both thresholds were set automatically by histogram analysis, considering a fixed percentage of the histogram maximum [8]. Values were empirically set to 5% for t_L and 3% for t_H, in other words t_L is the negative intensity level that has a number of elements representing 5% of the histogram maximum number of elements.

By thresholding DI, an object mask (OM) was obtained, that was cleaned by morphological operations to remove isolated pixels. Because a car windshield is a dark feature when seen from aerial imagery, its colour is close to that of the background (grey road surface). It often happened that the detected moving cars were split into two nearby blobs separated by their windshield, not detected as a moving region. A series of morphological operators was applied to the object mask OM to join the blobs and fill each object. A 8-connected components analysis was then run on OM to remove small objects.

Background Update. Once motion detection was done for the current image CI, the background image was updated as follows. First, the so-called instantaneous background was formed [8]. For all pixels in OM where a moving object was detected, the current background CB was sampled. For other pixels, where no motion was detected, the current image CI was sampled. The instantaneous background IB was thus computed as :

$$IB = OM. * CB + \overline{OM}. * CI .\tag{2}$$

where $.*$ denotes a pixel-wise multiplication and $\overline{OM}$ the complement to 1 of the binary mask. The background update BU was a weighted average of that instantaneous background obtained in Eq. 2 and the current background :

$$BU = \alpha IB + (1 - \alpha)CB .\tag{3}$$

---

[1] Strictly speaking, it can hardly be, even with a deformation model more complete than plane-to-plane homography.

α is the background learning rate : background updating has to be fast enough to adapt to illumination changes, while ignoring momentary changes [8]. A value of $\alpha = 0.3$ was found empirically to be satisfactory.

3.2 Vehicle Tracking by 3D-Connected Components Analysis

Motivation. The outcome of moving object detection is a sequence of masks containing pixels declared as in motion. Some of those pixels correspond to actual moving vehicles in the scene, but others can be gathered in regions not representing any real moving object, e.g. residual effects of illumination changes not modelled in the background update. Fig. 4 shows masks of pixels in motion, in two consecutive frames. While the actual vehicles appeared in both masks, regions appeared in Fig. 4(a) that were not present in the other mask. The region on top of Fig. 4(b) corresponds to a passenger crosswalk, detected as in motion because of misalignment of frames in the stabilisation process, and because its white colour over the dark road surface makes it a salient feature. These false detections had to be discarded before tracking vehicles.

A threshold on the size of the moving regions would not have been suitable for distinguishing stabilisation artifacts from vehicles, because some artifacts turned out to be bigger than cars. Based on the observation that artifacts flickered throughout the sequence (i.e. appeared then disappeared in consecutive frames), whereas actual vehicles appeared in all masks, we used temporal consistency as a criterion to identify vehicles, through 3D-connected components analysis.

Vehicle tracking was achieved at the same time, by labelling spatiotemporal regions. The video frame rate (25fps) guaranteed that there was an overlap between the same vehicle in two consecutive frames. This kind of approach may not have been used in traditional traffic monitoring applications, due to the many occlusions of vehicles by other vehicles when using a ground-based or aboveground camera. However, in aerial image sequences acquired near nadir direction, there is no such occlusion. Detected moving objects supposedly appear disjoint in each mask OM. Consequently, 3D-connected components analysis should allow vehicle tracking without risking that two vehicles are tracked as a single moving blob.

Tracker and Vehicle Speed Estimation. A 26-connected components analysis was run on the binary masks sequence after motion detection stage, defining 3D-regions in the spatiotemporal domain (Fig. 5(a)). Regions whose temporal extension did not exceed 25 frames (1s) were discarded as stabilisation artifacts, while vehicles left long trails in spatiotemporal domain (Fig. 5(b)).

Since the image sequence has been previously registered to a GIS, there exists a correspondence between dimensions in the image and real dimensions, allowing speed estimations. The centroids of the vehicles were extracted, and used to estimate their speed in an interval of 12 frames. The speed estimation would not be accurate if run between consecutive frames, because the inter-frame vehicle displacement is low at such frame rate. The speeds were then mapped

on the registered sequence (Fig. 6). For each vehicle, speed and position was recorded for later traffic data analysis.

4   Results and Discussion

Almost all moving vehicles were correctly detected in the sequence, but along with stabilisation artifacts (Fig. 4). Ways to address this issue of false detection could range from improving image stabilisation, to using a more elaborate background subtraction technique, e.g. some of those tested in [2].


(a) Frame 70   (b) Frame 71
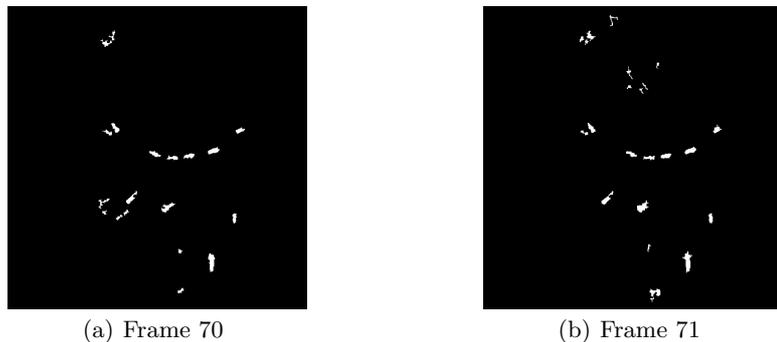
Fig. 4. Moving objects masks

Moving vehicle detection was problematic, especially with dark cars, which led us to empirically choose thresholds that would probably not be suited to another aerial image sequence with different lighting conditions. There were some frames in which a dark grey moving car was not detected at all. Besides, windshields tend to split moving cars into two moving blobs, which later altered the tracking quality. We used morphological operators in an attempt to join the two blobs of a same car. These operators included flood-fill to get full objects in the masks (OM). This was a straightforward attempt to obtain complete moving objects, that turned out to also accentuate false detections due to misalignment of frames. Instead of trying to join moving blobs in the masks, we could have used spatiotemporal clusters merging [4] as a post-processing after tracking.

The 3D-connected components analysis allowed to identify vehicles among stabilisation artifacts, and track them. It also provided a spatiotemporal visualisation of vehicles trajectories in the image sequence (Fig. 5).

Vehicles correctly detected were also successfully tracked. The dark car that was undetected during some frames in the middle of the sequence was not tracked during those frames, since tracking relied solely on detection. One way to circumvent the direct effect of misdetection on tracking efficiency, could be to add a predictor in the tracking process, such as the widely used Kalman filter. For

those vehicles correctly detected and tracked, the velocity vectors showed a consistent direction along the sequence (Fig. 6). Once the position of vehicles was known, speed estimation was rather straightforward.

(a) 3D-connected components   (b) After temporal consistency criterion
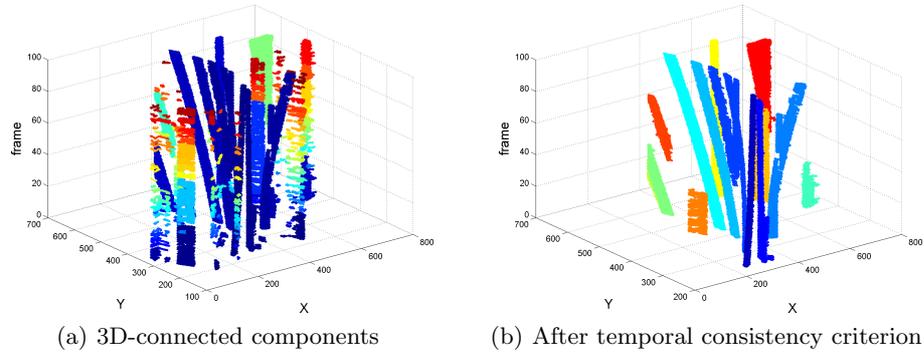
Fig. 5. Moving objects trajectories in spatiotemporal domain (2D+t view)
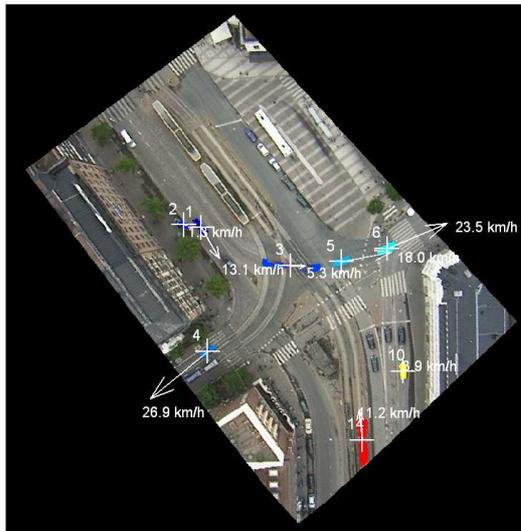
Fig. 6. Vehicles speeds

Spatiotemporal connected component analysis was a simple and appealing method to track vehicles, and at the same time to deal with false detections resulting from misalignment of frames, but it also has its drawbacks. With the

current implementation, it can only be applied to an image sequence treated as a whole, i.e. with all frames loaded in memory. The number of frames that could be processed was limited to about a hundred with Matlab and our computer, reducing the interest of the method for long image sequences. 3D-connected components analysis could still be applied in sliding spatiotemporal windows of some frames[2], for both tackling the memory limitation and allowing online tracking of objects, as the sequence runs. Another issue is that tracking efficiency was very dependent from the detection quality. There was no mechanism to prevent two cars moving closely together from being merged into a single spatiotemporal region, when their corresponding masks were 26-connected - which can occur when moving object detection is not accurate. Including colour, shape information or velocity continuity in the tracking procedure [4] might help increasing its robustness. Last, a connected components-based tracker is inherently limited as for the maximum speed of a vehicle that can be tracked. This depends on the video frame rate, image resolution, size and speeds of vehicles as well as the quality of moving object detection.

In the video stabilisation stage, point tracking relied on an unconventional technique, namely running a corner detector in each frame - the point correspondence problem was not rigorously solved. State-of-the-art point tracking algorithms, such as the Kanade-Lucas-Tomasi feature tracker [14], would certainly improve the point tracking quality, hence video stabilisation.

Last, the whole approach would need to be validated with ground truth data, ideally a vehicle of which we would know the speed or position at each moment during the sequence.

5 Conclusion

We proposed a simple processing chain for moving vehicle detection and tracking in image sequences acquired from a helicopter. After compensating the camera motion and registering the video to a GIS, residual motion was detected by adaptive background subtraction and vehicles tracked by spatiotemporal connected components analysis. The main loopholes of the system lie in the detection stage, and greatly affected the overall tracking performance. Cars correctly detected were successfully tracked hereafter, and the estimation of their speed seemed consistent - while this would need validation with ground truth data.

Lots of improvements can be made to catch up with state-of-the-art methods. The developed approach relied mainly on pixel-based methods, and some steps like video stabilisation still lack automation. Integration of higher-level image processing should also help making tracking more robust, especially when detection partly fails. One possibility would be to consider detailed vehicle models for tracking, such as Hinz [11] developed to detect cars in high-resolution images.

Yet, the results obtained were fairly good, and are encouraging for the development of an operative traffic monitoring system with aerial platforms.

---

[2] Enough frames so that a stabilisation artifact appears as a discontinuous region in spatiotemporal domain.

Acknowledgments

References

1. A. Censi, A. Fusiello, V. Roberto : Image stabilization by features tracking. Proceedings of the 9th International Conference on Image Analysis and Processing, Venice (1999).
2. S.-C. Cheung, C. Kamath : Robust techniques for background subtraction in urban traffic video. Proceedings of SPIE Electronic Imaging : Visual Communications and Image Processing, San Jose (2004).
3. R. Cucchiara, C. Grana, M. Piccardi, A. Prati : Detecting Moving Objects, Ghosts, and Shadows in Video Streams. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. **25** no. 10 (2003) 1337–1342.
4. T. Fuse, E. Shimizu, R. Maeda : Development of techniques for vehicle manoeuvres recognition with sequential images from high altitude platforms. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Corfu (2002), vol. **34** part. 5, 561–566.
5. T. Fuse, E. Shimizu, T. Shimizu, T. Honda : Sequential image analysis of vehicle manoeuvres : evaluation of vehicle position accuracy and its applications. Journal of the Eastern Asia Society for Transportation Studies, vol. **5** (2003) 1991–2002.
6. A. Fusiello, E. Trucco, T. Tommasini, V. Roberto : Improving Feature Tracking with Robust Statistics. Pattern Analysis and Applications, vol. **2** no. 4 (1999) 312–320.
7. R. Garcia Campos : A proposal to estimate the motion of an underwater vehicle through visual mosaicking. PhD Thesis, University of Girona (2001), chap. 5, 131–137.
8. S. Gupte, O. Masoud, R.F.K. Martin, N.P. Papanikolopoulos : Detection and classification of vehicles. IEEE Transactions on Intelligent Transportation Systems, vol. **3** no. 1 (2002) 37–47.
9. M. Han, T. Kanade : Reconstruction of a Scene with Multiple Linearly Moving Objects. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island (2000).
10. C. J. Harris, M. Stephens : A combined corner and edge detector. Proceedings of the 4th Alvey Vision Confeference, Manchester (1988), 147–151.
11. S. Hinz : Detection and counting of cars in aerial images. Proceedings of IEEE International Conference on Computer Vision, Barcelona (2003), vol. **III**, 997–1000.
12. S.P. Hoogendoorn, H.J. van Zuylen, M. Schreuder, B. Gorte, G. Vosselman : Microscopic traffic data collection by remote sensing. 82nd Annual Meeting of Transportation Research Board (TRB), Washington D.C. (2003).
13. P. Montesinos, V. Gouet, R. Deriche : Differential invariants for color images. Proceedings of IAPR International Conference on Pattern Recognition, Brisbane (1998), 838–840.
14. C. Tomasi, T. Kanade : Detection and tracking of feature points. Carnegie Mellon University Technical Report, CMU-CS-91-132, Pittsburgh, PA (1991).