Superpixel based road user tracker

Ardö, Håkan; Nilsson, Mikael; Laureshyn, Aliaksei

2014

Superpixel based road user tracker.

Håkan Ardö Mikael Nilsson Aliaksei Laureshyn

November 18, 2013

Abstract

A superpixel based tracker is tested on the two Tracking sequences from PDTV[7], Minsk and Sherbrooke. It detects all vehicles from the Minsk dataset although a few of them are splitted. The pedestrians are too small and thus all missed. The results for the Sherbrooke are not as good, especially in the areas far way from the camera where the intersection s viewed at a low angle. Also the sign in the foreground causes misses.

Results on the positioning precision are also presented and how these can be improved by fitting 3D models to the detections.

1 Introduction

Comparing video analytics algorithms on the same datasets is important to assess their relative performance. To facilitate this we here present the result of a superpixel based tarcker on two Tracking sequences from PDTV [7], Minsk and Sherbrooke, publicly available online. The system extract the moving road users from the video data and produces trajectories on the ground consisting of both positions and orientations.

Advanced traffic applications like detection of serious breakdowns in interactions (traffic conflicts) or extraction of large amounts of microscopic data for calibration of traffic models require quite high level of accuracy in detection and tracking of the road users in video. In the trade-off between the processing time and complexity of calculation for higher output accuracy, the priority is given to the latter. The problem, however, is that the more advanced video processing techniques are used, the more parameters have to be set up, and the performance might vary a lot depending on the conditions at which the input video was taken (camera angle, resolution, distance to the objects). To test the universality of the developed technique, it needs to be tested on videos taken in different places and in different conditions.

2 Tracking

The moving road users are extracted from the video using a sequence of video analytics operations. The result after each of those steps are shown for one example frame in Figure 1.

The first step performs a background foreground segmentation. The algorithm used[1] calculates image gradients and builds a background model as the temporal mean over the observed gradient directions. This background model is compared with the gradient directions in the current frame. By using the gradient magnitudes as weights it asserts the reliability of those matches and produces a probability of foreground for each pixel. This probability becomes close to 0.5 in uniform areas of the image where the gradient directions are unreliable. In structured areas on the other hand it is close to 1.0 for foreground pixels and close to 0.0 for background pixels. The second image in Figure 1 shows an example.

The next step turns this probabilistic segmentation into a binary segmentation by performing a Markov random field segmentation [2, 4]. The single pixel probabilities are used as unary terms and constant probabilities are used as binary times. The probability of the pixels belonging to the same class, i.e. both being foreground or both being background is set to

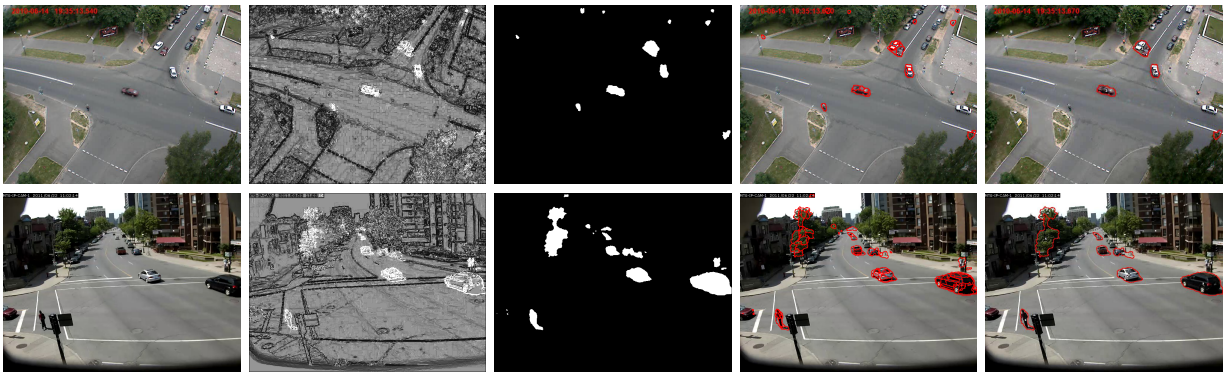

Figure 1: Results after each processing step for a single frame. Top row from the Minsk sequence and bottom row from the Sherbrooke seqeunce. Columns from left to right: input frame, probabilistic segmentation, binary MRF segmentation, superpixel segmentations, final tracks.

0.9, and a prior is introduced saying that on average 30 % of the pixels in the image are foreground. The result of this step is that the uncertain pixels are filled in by looking at their neighbours, see the third image in Figure 1. The parameters are chosen in such a way as to never split a single road user into several connected components. This means that quite often two or more road users are merged into a single connected component.

The connected segments are further segmented into super pixels. We use the energy function suggested by Conrad et al. [3] to define good segmentations, but optimize it using hierarchical clustering. That is initially each normal pixel is assigned to it's own super pixel. Then, iteratively, two super pixels are selected and fused into a single super pixel. The selection is made by locating the fuse that will reduce the energy function the most. This process is continued until all super pixels are larger than some threshold and any additional fusing would increase the energy function significantly. An example of the resulting segmentations is shown in the forth image of Figure 1.

Finally, the super pixels are matched form one frame to the next. This is achieved by minimizing the distance to the closest super pixel in (r, g, b, x, y)-space over translations. Here (r, g, b) represents the color of the pixel and (x, y) its position within the image. The optimization is regularized by also penalizing variations in distances between neighbouring super pixels. This connects the segments produced by the background foreground segmentation between the frames into a potentially complicated graph of segments that are splitting and merging as different object becomes close to each other. This graph is then splitted into tracks under the assumption that all super pixels forming a single object belongs to the same connected foreground segment in each frame. This means that in theory it is enough for an object to become a foreground segment of it's own for a single frame during it's trajectory to allow it to be segmented out during it's full trajectory.

3 Metric Promotion

Camera calibration is performed by manually selecting points in the image and their 3D positions. They can be measured with a GPS (such as *Leica GX1230 GG*) or using areal images from for example google earth. These points are manually positioned in an image from a static mounted camera. The calibration procedure is performed using well chosen points, following guidelines described in [5]. Once the world and image corresponding points are measured and positioned, the camera is calibrated [8].

Given the calibration, a defined search space in the ground plane and a 3D model a 3D search with

context is conducted to find metric position and orientation in the ground-plane [6]. The pipeline involves foreground/background segmentation and non-maximum suppression. Two models were used, one van-sized box model and one sedan car using 60 triangles, see Fig. 2.

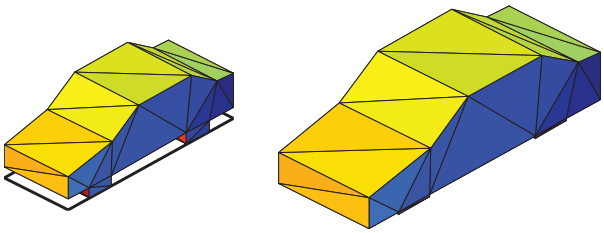

Figure 2: 3D model of a sedan car using 60 triangles and its rectangle footprint (left) and the enlarged 3D model (right) used to capture context.

	Bus	Car	Van	Ped	?
True tracks	1	33	3	4	1
Detected tracks	1	33	3	0	1
Missed tracks	0	0	0	4	0
Extra tracks	0	8	0	0	0
True states	28	888	83	261	66
Detected states	26	834	79	0	53
Missed states	2	54	4	261	13
Extra states	3	221	0	0	7

Table 1: Pixel level detection results from the Minsk sequence.

4 Results

The system was tested on the two Tracking sequences from PDTV [7], Minsk and Sherbrooke. The first step yields pixel level bounding boxes in the image. Those image tracks were compared to the ground truth as suggested by PDTV [7]. Both the tracking result and the ground truth was restricted to the region of interest before comparing. The result is presented in Table 1 and Figure 3 for the Minsk sequence and in Table 2 and Figure 4 for the Sherbrooke sequence.

There are a lot of missed tracks in the in Sherbrooke sequence. The vehicles are however tracked fine in the center of the intersection. The mistakes are made on the remote side of the intersection where the viewing angle is low and the vehicles are small. Also the sign in foreground causes vehicles going behind it to be missed.

The Minsk sequence comes with a camera calibration that allows the pixel level result to be promoted to meters. The 3D model fitting approach described above was used with two different 3D models that were fitted to all tracks. Figure 5 shows the results of fitting a car 3D model and compares the result with

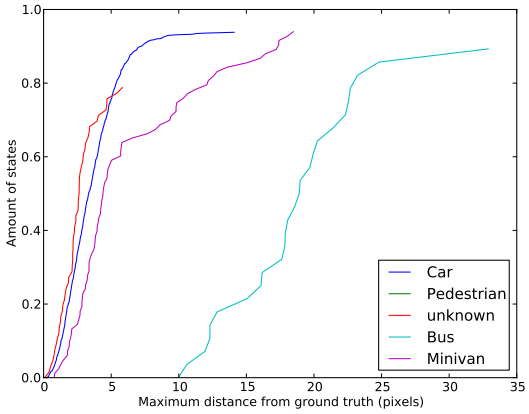

Figure 3: Pixel level tracking precision for the Minsk sequence.

	car	pedestrians	unknown
True tracks	15	4	2
Detected tracks	7	1	1
Missed tracks	8	3	1
Extra tracks	15	0	0
True states	2965	2045	568
Detected states	684	162	49
Missed states	2281	1883	519
Extra states	793	0	0

Table 2: Pixel level detection results from the Sherbrooke sequence.

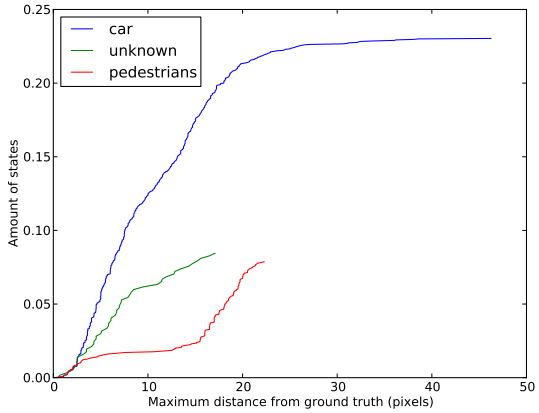

Figure 4: Pixel level tracking precision for the Sherbrooke sequence.

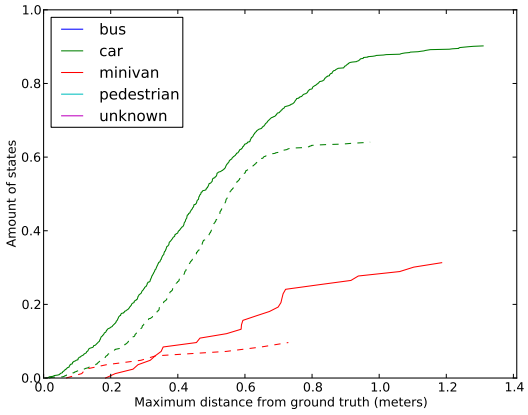

Figure 5: Metric tracking precision for the Minsk sequence using a car 3D model. Dotted lines are the pixel bounding box centers projected onto the ground plane for comparison.

the position achieved by projecting the center of the pixel bounding box onto the ground plane. Figure 6 uses a van sized 3D box instead. Note how the position is improved when a suitable 3D model is used. That is the position estimate of the cars are improved when the car model is used and the position estimate of the vans are improved when the box model is used.

5 Conclusion

We've shown how super pixels can be used to extract road user trajectories form video sequences and that the estimated position of those road users can be improved by fitting 3D models to the detections.

References

[1] H. Ardo and L. Svärd. "Bayesian Formulation of Gradient Orientation Matching". In: *Submitted to CVPR 2014* () (cit. on p. 1).

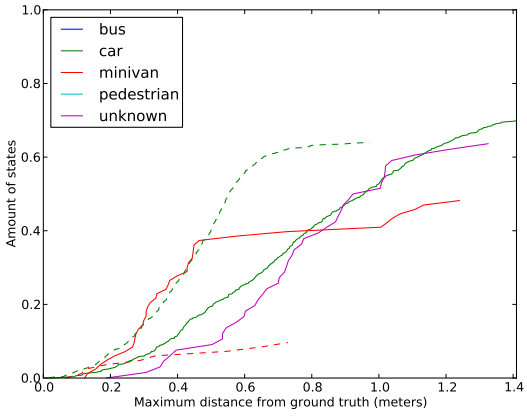

Figure 6: Metric tracking precision for the Minsk sequence using a van sized box as 3D model. Dotted lines are the pixel bounding box centers projected onto the ground plane for comparison.

[2] Y. Boykov and V. Kolmogorov. "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 26.9 (2004), pp. 1124–1137 (cit. on p. 1).

[3] C. Conrad, M. Mertz, and R. Mester. "Contour-Relaxed Superpixels". In: *Energy Minimization Methods in Computer Vision and Pattern Recognition* (2013) (cit. on p. 2).

[4] P. Kohli and P. H. S. Torr. "Dynamic Graph Cuts for Efficient Inference in Markov Random Fields". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.12 (Dec. 2007), pp. 2079–2088. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1128. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=04359296 (cit. on p. 1).

[5] A. Laureshyn, M. Nilsson, H. Ardö, Å. Svensson, and A. Persson. "How accurate can we measure from video?" In: *The 26th ICTCT workshop*. Maribor, Slovenia, 2013 (cit. on p. 2).

[6] M. Nilsson and H. Ardö. "In Search of a Car - Utilizing a 3D Model with Context for Object Detection". In: *Accepted for presentation in International Conference on Computer Vision Theory and Applications (VISAPP)*. 2014 (cit. on p. 3).

[7] N. Saunier, H. Ardö, J.-P. Jodoin, A. Laureshyn, M. Nilsson, Å. Svensson, L. Miranda-Moreno, G.-A. Bilodeau, and K. Åström. "A Public Video Dataset for Road Transportation Applications". In: *Transportation Research Board (TRB) 93rd Annual Meeting* (2014) (cit. on pp. 1, 3).

[8] R. Tsai. "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off- the-shelf TV cameras and lenses". In: *Robotics and Automation, IEEE Journal of* 3.4 (1987), pp. 323–344. ISSN: 0882-4967. DOI: 10.1109/JRA.1987.1087109 (cit. on p. 2).