Detection of Moving Objects from a Moving Platform in Urban Scenes

Frank B. ter Haar, Richard J.M. den Hollander, Judith Dijk

Electro-Optical Systems, TNO Defence Security and Safety, Oude Waalsdorperweg 63, 2597 AK, The Hague, The Netherlands

ABSTRACT

Moving object detection in urban scenes is important for the guidance of autonomous vehicles, robot navigation, and monitoring. In this paper moving objects are automatically detected using three sequential frames and tracked over a longer period. To this extend we modify the plane+parallax, fundamental matrix, and trifocal tensor algorithms to operate on three sequential frames automatically, and test their ability to detect moving objects in challenging urban scenes. Frame-to-frame correspondences are established with the use of SIFT keys. The keys that are consistently matched over three frames are used by the algorithms to distinguish between static objects and moving objects. The tracking of keys for the detected moving objects increases their reliability over time, which is quantified by our results. To evaluate the three different algorithms, we manually segment the moving objects in real world data and report the fraction of true positives versus false positives. Results show that the plane+parallax method performs very well on our datasets and we prove that our modification to this method outperforms the original method. The proposed combination of the advanced plane+parallax method with the trifocal tensor method improves on the moving object detection and their tracking for one of the four video sequences.

Keywords: Moving object detection, Urban scenes, Fundamental matrix, Trifocal tensor, Plane+parallax

1. INTRODUCTION

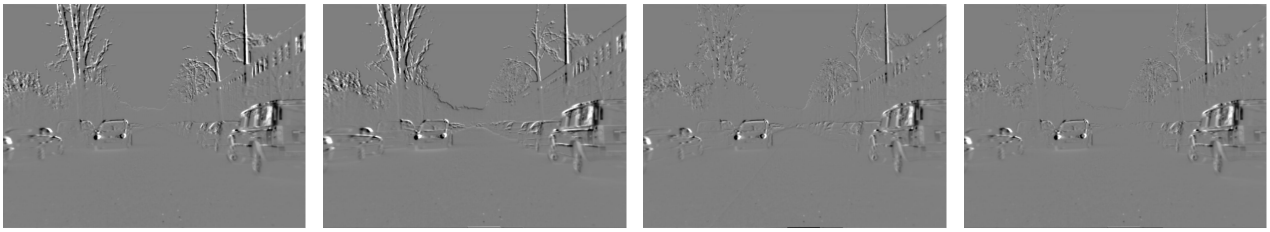
Moving object detection in video sequences has evolved over the past years from stationary video sequences, to dynamic video sequences with one or two 2D layers, and on to more general 3D scenes. This paper addresses the detection of moving objects in urban scenes from dynamic video sequences. The detection of moving objects, such as cars, bicycles, and pedestrians, increases the situational awareness in urban scenes. Knowing where moving objects are and where they are going helps to predict their position over time, which is useful for the guidance of autonomous vehicles, the monitoring of vehicles and people, and alerting drivers when cars are approaching. The challenge in driving through urban scenes is that all objects in the video sequence appear to move (including parked cars, light poles, and houses), and that static objects close to the camera may even appear to move faster (3D parallax) than the truly moving objects at a larger distance. Being able to distinguish between the static and the moving objects is important to reduce the number of false alarms in case of a surveillance system.

Moving object detection can be performed by estimating the motion model between two or more frames in order to compensate for the camera induced motion. Features that do not fit the motion model must have been subdue to individual motion. The required motion model may differ in complexity, and depends on the acquired video data. In case of a stationary video surveillance, all features with motion are independently moving objects. In case of a moving camera looking at a distant 3D scene (aerial surveillance), the static features can be considered planar which allows either a translation, or a affine or projective transformation to compensate for the camera motion and to directly detect the moving objects.[1] In case of a camera moving through a 3D scene these motion models do not suffice as illustrated in Fig. 1, and more general models or algorithms are required. One option is to estimate the epipolar geometry using the *fundamental matrix* (bifocal tensor) and to classify the model outliers as moving objects.[2] This two-frame geometry cannot detect objects moving along the epipolar lines, for which the three-frame geometry should be computed using the *trifocal tensor*[2] instead. These models impose either epipolar or trilinear constraints to separate parallax from moving objects.

An alternative is the *Plane+Parallax* approach in which the 2D projective transformation of a dominant plane is determined and additional geometric constraints are used to distinguish between the moving objects and the 3D parallax.[1,3,4] Many variants of the *Plane+Parallax* framework have been proposed. Irani et al.[1] detect the planar homography between

Contact email: Frank.terHaar@tno.nl

Visual Information Processing XIX, edited by Zia-ur Rahman, Stephen E. Reichenbach, Mark A. Neifeld,
Proc. of SPIE Vol. 7701, 770105 · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.848034



(a) none (b) translation (c) affine (d) projective

Figure 1. Subtraction of two images of the "frontal car seq." (a) directly, (b) after XY-translation, (c) after an affine transformation, (d) after a projective transformation. In all images the static car on the right appears to be moving faster than the moving car in the center. This is due to the parallax.

pairs of frames and apply a parallax rigidity constraint. In their method, the on-plane features are labeled as static features and for the off-plane features a rigidity constraint separates static features (showing parallax) from the moving features. To do so, the off-plane features are warped according to the 2D projective transformations (plane alignment) to the next frames imposing a non-rigid transformation for the moving features, in comparison to a static reference point. The advantage of their approach is that the epipole estimation is not required. However, they manually select the static reference point, which makes it difficult to apply in practical applications.

Sawhney et al.[3] also start with detecting the planar homography to compensate for the camera motion. If there is sufficient remaining motion, they apply the epipolar constraint between one pair of frames by estimating the *fundamental matrix* within the RANSAC framework.[5] Then the constancy of parallax is imposed for the second pair of frames (trifocal constraint), and the image regions that remain misaligned are the moving objects. The assumption in this shape constraint is that the planar homography is stable for three frames and the epipoles are correctly estimated.

Yuan et al.[4] present another *Plane+Parallax* approach. First they compute the planar homography to eliminate planar pixels as potential moving objects and the epipolar constraint is applied to detect some of the moving objects pixels (similar to[3]). To relax the assumption of a consistent planar homography for three frames, they apply a three-view geometric constraint that allows a different dominant plane in the sequential pairs of frames.

Yamaguchi et al.[6] proposed a feature based approach for moving object detection and tracking in urban scenes. They detect corresponding features between pairs of frames, estimate the *essential matrix* using 8-point RANSAC, detect moving objects using the epipolar constraint, and track the features.

The main difficulty is to estimate the true camera motion from two or three frames while the pixel or feature correspondences are inaccurate and can belong to moving objects. To deal with these outliers, a common approach is to select the model in a RANSAC scheme that for instance maximizes the number of model inliers[5,7] or the least median squares error.[4] The former would require a careful threshold selection to separate the model's outliers into parallax features and moving object features, while the latter assumes that less than half of the evaluated features are moving.

Contributions In this paper we perform moving object detection while the camera moves through challenging urban scenes. We modify the plane+parallax, the fundamental matrix, and trifocal tensor algorithms to operate on three sequential frames automatically, and test their ability to detect moving objects in two different video sequences. Our results quantify that (1) SIFT tracking improves on the reliability of the moving object detection, (2) the modified rigidity constraint of the plane+parallax method outperforms the original rigidity constraint, and (3) the combination of the modified plane+parallax method with the trifocal tensor method improves the moving object detection in one of the four scenes.

2. DATASETS

In this work we use frames from two different video sequences captured with different cameras mounted to a moving vehicle. The first set is our local dataset that consists of 2409 consecutive frames at 30fps and with a resolution of 640×480 pixels. From this set, we extracted three interesting sequences of 45, 80, and 80 frames at 30fps, the "frontal car", "side car", and the "left turn". In these sequences the camera moves (1) towards an approaching car, (2) towards a car driving from right to left, and (3) turns towards the left while passing a 'static' car, respectively. The latter is interesting for evaluating the number of false moving object detections. Each of these frames is corrected for the radial distortion imposed by the camera,

and cropped to 600×440 pixels to remove unknown pixels. The second set is the CamSeq01 Dataset[8] with multiple moving objects ("multiple objects"). It consists of 101 consecutive frames at 15fps and with a resolution of 960×720 pixels. To acquire frames with the same resolution as our local set, each frame is downsampled to 640×480 pixels and cropped to 600×440 pixels. Because moving objects captured at 30fps are only captured in a small number of frames, we do not reduce the frame rate of our local dataset.

3. FEATURE EXTRACTION

The adapted algorithms that we apply for moving object detection require only three consecutive frames (f_1-f_2-f_3) to detect the moving objects. From each of the frames, we extract SIFT keys[9,10] and search for the best match in its previous frame. The SIFT *matches* (p_1,p_2 and p_2,p_3) that exist for three consecutive frames are selected as SIFT *triplets* (p_1,p_2,p_3). The selected triplets are used for the detection and tracking of the moving objects in our video sequences. The advantage of SIFT keys over pixel classification, is the reduction of 600×440 pixels to approximately 1500 SIFT keys in our frames. This makes the estimation of the motion model much more efficient and pixel classification can still be done afterwards guided by the labeled features. To select reliable SIFT matches, a SIFT key in the current frame is matched against SIFT keys in the other frame on the same location within a 50 pixel window. The best match is only valid when it is significantly better (factor $\frac{5}{3}$) than the second best match. This is to avoid SIFT matches on less reliable lines and to focus on corners. These SIFT matches can be effectively used for tracking as we do in Sect. 5.

In order to compensate for the translation part of the camera motion, the frames are aligned using the average XY-translation of the SIFT matches. This translation is also applied to the features p_1 and p_2, such that their global position corresponds to that of p_3 in the current frame f_3. After this global 2D alignment, we can easily remove incorrect SIFT triplets using an angle rule and a relative displacement rule. We remove a triplet if,

- $d(p_1, p_2) + d(p_2, p_3) > 5$,
- $cos^{-1}(\frac{(p_1-p_2)}{||p_1||||p_2||} \cdot \frac{(p_3-p_2)}{||p_3||||p_2||}) < \frac{2}{3}\pi$,
- $max(\frac{d(p_1,p_2)}{d(p_2,p_3)}, \frac{d(p_2,p_3)}{d(p_1,p_2)}) > 3$,

where d is the Euclidean distance in pixels. The highly convex displacement of a SIFT key in three consecutive frame indicates an incorrect match and so does a sudden increase in the SIFT key displacement. Fig. 2 shows the SIFT triplets within the current frame by two lines, and some of these triplets are removed (as indicated with a circle) according to the angle and relative displacement rules.

4. MOVING OBJECT DETECTION

In this section we adjust three algorithms and apply them to the SIFT triplets for moving object detection (MOD), namely the plane+parallax decomposition, the fundamental matrix, and the trifocal tensor, which are described in detail below. Each of these algorithms estimates a model of the scene, for which we employ the RANSAC scheme to find the model with the maximum inlier support.

4.1 Plane + Parallax decomposition

Irani et al.[1] assume a dominant planar surface in frame pair f_1-f_2 that corresponds to the dominant planar surface in frame pair f_2-f_3. With this planar homography, previous frames f_1 and f_2 are warped to the current frame f_3. Features on the planar surface are assumed to be static, and the off-plane features are either static features or moving features depending on their rigidity. A static off-plane triplet is used as a reference to determine these rigidities. Each feature, including the reference, is warped to the current frame according to the estimated projective transformation. Because of the 3D parallax and independent motion, the features are not warped to the exact position in the current frame. The vectors μ between the current locations and these warped locations, are used for the rigidity constraint. This rigidity constraint is illustrated in Fig. 3. In this figure, p_i and p_j are two different off-plane triplets in the current frame, p_{wi} and p_{wj} are their warped locations from the *previous* frame to the *current* frame. Due to parallax or independent motion, the vectors μ_i and μ_j are not null vectors. C is the projection of p_i on the line through p_{wi} perpendicular to Δp_w, and B the projection of p_j on the same line. Despite the unreliable epipole estimation, the relative structure $\frac{AB}{AC}$ can be reliably computed.[1] A similar drawing can be made for the warp from the *pre-previous* frame to the *current* frame, obtaining the relative structure $\frac{AB'}{AC'}$. The *rigidity constraint* is as follows: if we assume p_i to be a static feature (with parallax), then p_j is static if and only if

(a) side car (b) side car

(c) left turn (d) left turn

Figure 2. SIFT keys in the current frame are consistently matched over the past two frames (lines in a and c), the global XY-translation compensates for the camera translation (b and d) and most incorrect SIFT triplets are removed (circles) using the angle and displacement rules. Notice the parallax on the static cars and trees.

$\frac{AB}{AC} = \frac{AB'}{AC'}$. In other words, the relative rigidity of two static features remains the same over consecutive frames, and the absolute difference can be used to detect moving objects as Irani did.

However, for the automatic detection of moving objects, there are two problems: (1) the reference feature has to be selected manually and (2) the relative structures can be less precise due to inaccuracies in the triplet extraction and the dominant plane selection, which both cause less accurate warp locations (especially those far from the selected planar surface). To adapt this method to be applicable in urban scenes, we assume that the majority of the off-plane features belong to static objects and relate the rigidity error to the residual error after the warp. The later is to compensate for the residual error that is usually higher for features with a larger frame-to-frame displacement.

In our implementation, we use the RANSAC scheme to estimate the planar homography between frames f_2 and f_3. In this scheme, four random triplets are selected to instantiate a plane and the inlier support is computed using the on-plane threshold t_H. The plane with the maximum support after m iterations is selected, and the supported on-plane features are used to find the same plane in frames f_1 and f_3 using the RANSAC scheme. This ensures the selection of the same plane in all three frames. All n off-plane SIFT triplets are warped according to the found planar homography from the frames f_1 and f_2 to the current frame f_3, and the parallax rigidity error (RE) is computed for each pair of off-plane triplets ($n \times n$)

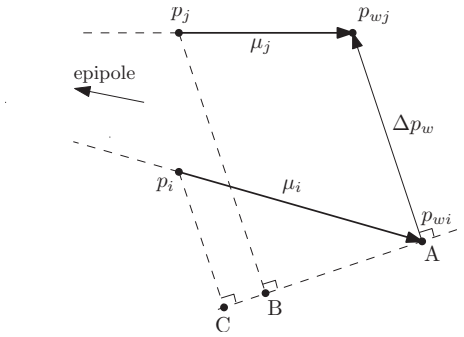


Figure 3. The relative structure $\frac{AB}{AC}$ for features p_i and p_j in two frames (f_2-f_3) provides one part of the rigidity constraint, when this structure corresponds to the relative structure $\frac{AB'}{AC'}$ of the other frame pair, then p_j is considered static if p_i is static too (image taken from[1]).

and divided by the averaged warp error $\frac{1}{2}||\mu_1^j||^2 + \frac{1}{2}||\mu_2^j||^2$:

$$RE_i = med_j \frac{\left|\frac{AB}{AC} - \frac{AB'}{AC'}\right|}{\frac{1}{2}||\mu_1^j||^2 + \frac{1}{2}||\mu_2^j||^2} \quad .$$

So, each feature i is assumed to be static once, and the rigidity of each other feature j is computed and divided by the distance to its warped feature. The median of row i in the rigidity matrix is determined and used as the final rigidity error of feature i. In case feature i is static and the majority of the other off-plane features are static too, then the median rigidity error of feature i is low. Otherwise if feature i moved, it is a bad reference point for all other static features, and the majority of rigidity errors are high and so is the final (median) rigidity error of feature i. A disadvantage of this Plane+Parallax decomposition method is that the rigidity errors are not related to the absolute pixel distances, which makes it hard to select a threshold. Instead, we use a dynamic rigidity threshold t_{Hmod} to select the moving object features.

4.2 Fundamental matrix (bifocal tensor)

The fundamental matrix relates the selected features between pairs of frames (f_1-f_2 and f_2-f_3).[2] We apply the RANSAC scheme using eight randomly selected triplets, instantiate the fundamental matrix from the feature pairs, compute the Sampson error, and determine the inlier support based on a threshold t_F. When a feature lies within the distance t_F (in pixels) from its expected epipolar line, it is considered to be an inlier of the model. The fundamental matrix with the highest inlier support is selected. Afterwards, we have selected two fundamental matrices for frames f_1-f_2 and f_2-f_3. We sum for each feature its two residual Sampson errors and label it as a moving feature when the error is above threshold $t_{Fmod} = 2.5 \times t_F$. This simple modification of the two-view fundamental matrix to three views increases its robustness. A known problem of the fundamental matrix for moving object detection is that movement along an expected epipolar line cannot be detected. Nevertheless, it is important to evaluate this method for our urban scenes, as it can be used as an additional MOD method for non-epipolar movement.

4.3 Trifocal tensor

The trifocal tensor is the generalization of the fundamental matrix to three views (f_1-f_2-f_3).[2] It describes the projective geometric relationships between these views using three 3×3 matrices T_1, T_2, and T_3. Whereas a point corresponds to a line in the bifocal case, the trifocal tensor recovers the point-to-point relation, which enables the detection of moving objects along the epipolar lines. We apply the RANSAC scheme using seven randomly selected triplets, instantiate the trifocal tensor, compute the trifocal transfer error, and determine the inlier support based on a threshold t_T. The trifocal tensor with the highest inlier support is selected. To compute the trifocal tensor error robustly, we determine the error of each triplet (p_1,p_2,p_3) by projecting p_1 via a line through p_3 towards p_2 using the trifocal tensor and compute the Euclidean distance between p_2 and the projected p_2. Because a randomly chosen line through p_3 that accidentally lies on the trifocal plane does not result in a correctly transferred point, we transfer each point via seven differently oriented lines through p_3 and keep the transferred point with the smallest Euclidean distance (as in Algorithm 1). In our experiments this modified trifocal transfer error performed better than the Sampson error. The outliers of the trifocal tensor (pixel error $> t_{Tmod}$, $t_{Tmod} = t_T$) are selected as moving objects.

Algorithm 1 trifocal transfer error(*tensor* T, *triplets*)

for each triplet (p_1,p_2,p_3) **do**
  $err_{triplet} \leftarrow \infty$
  for $\phi \in [0, \frac{1}{8}\pi, \frac{1}{4}\pi, \frac{3}{8}\pi, \frac{1}{2}\pi, \frac{5}{8}\pi, \frac{3}{4}\pi, \frac{7}{8}\pi]$ **do**
    define line l_3 through p_3 with orientation ϕ
    $H_{12} = [T_1 l'_3; T_2 l'_3; T_3 l'_3]$
    $p_{w2} = H_{12} p_1$
    $err_{triplet} = min(d(p_{w2}, p_2), err_{triplet})$
  if $err_{triplet} < t_T$ **then** triplet is an inlier of T
return all $err_{triplet}$

Figure 4. SIFT matches among pairs of frames results in longer tracks. The tracks over five frames (left) and over fifteen frames (right), both without the XY-translation.

5. MOVING OBJECT TRACKING

The SIFT keys are reliably selected and matched from one frame to another, resulting in SIFT triplets and longer SIFT tracks (Fig. 4). These tracks of triplets and their MOD results can be used to increase the confidence that a current model outlier is truly a moving object. In Sect. 3, we selected only the *reliable* SIFT matches and triplets for the model estimation and moving object detection. If an expected SIFT match is not established, then a SIFT track becomes disconnected and its MOD-history useless. To this extend, we select for each SIFT key in the current frame the best possible match in the previous frame and use those matches to bridge the gaps in the SIFT tracks. The additional points in history and the triplets that did not satisfy the angle and displacement rules are labeled as *undefined*, since they are not evaluated by the MOD algorithms. Otherwise each algorithm labels the triplet as either static or moving in its history. A triplet in the current frame is assigned the label *probably moving* (value '+1') iff its measured value is above $\frac{3}{4}$ times threshold t_{Hmod}, t_{Fmod} or t_{Tmod}, *moving* (value '+2') iff above these thresholds, or *static* (value '-1') otherwise. Based on a history of eight frames, including the current frame, we check the SIFT track of a current feature for the following three cases:

- If the sum of the last three frames is larger than four (e.g. {-1,1,1,2}), then the current feature is *moving*.
- If the sum of the last four frames is larger than four (e.g. {0,1,1,2}), then the current feature is *moving*.
- If the sum of the last eight frames is larger than five (e.g. {0,1,2,1,-1,1,1,0}), then the current feature is *moving* and if the current label is undefined or static it is set to *probably moving*.

The first rule ensures that a moving object can be detected after three evaluated triplets (i.e. five frames), the second rule is to overcome false-negative detections, and the latter ensures the tracking of reliably detected moving objects.

6. EXPERIMENTS AND RESULTS

In this section we evaluate the three different methods for their ability to detect and track the moving objects. Based on the manually segmented moving object(s) in each of the frames, each SIFT triplet can be easily labeled as truly moving or static.

For a fair comparison of the MOD algorithms, each method is allowed to use 60 iterations of the RANSAC scheme to estimate the motion model over three frames and to detect the moving objects. For the plane+parallax algorithm we use 50 iterations to find the dominant reference plane between frames f_1 and f_2, and use its on-plane features and another 10 iterations to find the same plane for frames f_2 and f_3. For the fundamental matrix algorithm we use 30 iterations to estimate the model for frames f_1 and f_2, and 30 for frames f_2 and f_3. The trifocal tensor uses 60 iterations to find the most plausible tensor. For each of the algorithms we have to set two thresholds, one to evaluate the estimated models (i.e. the planar homography, fundamental matrices, and trifocal tensor) and one to separate the parallax from the moving objects. The carefully selected thresholds t_H=0.4, t_F=0.5, t_T=1.5, are all related to the image resolution and feature accuracy and remain static during our experiments. For t_{Hmod} a different threshold is automatically selected in each of the frames, based on the median rigidities of the off-plane triplets. This threshold is set based on the rigidity errors (RE) of the off-plane features i (S) as:

$$t_{Hmod} = med_{i \in S}(RE_i) + 3stddev_{i \in S}(RE_i)$$

On each of the four video sequences, we apply each of the three algorithms to detect the moving features. Based on the ground truth data, the output of the MOD algorithms, and the SIFT track history, we determine the amount of *true positives* and *false negatives* both for the current frame data and the track history. To evaluate the performance of each algorithm, we use the fraction of summed *true positives* and summed *false negatives* over all frames in a sequence. Because each algorithm has a random component (RANSAC), we report the average result over three iterations. Note that taking the summed values over all frames in a sequence already compensates for most of the randomness. To show that the normalization of the rigidity error improves on the MOD results, we perform a run with ($H+$) and without (H) the denominator $\frac{1}{2}||\mu_1^j||^2 + \frac{1}{2}||\mu_2^j||^2$ in the rigidity constraint.

The results of the plane+parallax method (H), the normalized plane+parallax method ($H+$), the fundamental matrix (F), and the trifocal tensor (T) on the four sequences are shown in Tables 1, 2, 3, and 4. These results show the average MOD results over five runs and the corresponding standard deviation in brackets. For the 'left turn' sequence, we only report the false negatives because it has no positives. As an additional experiment, we have *combined* the MOD results of the modified plane+parallax method and the trifocal tensor method. To combine the two, we simply sum the positive MOD labels and divide it by two, or assign '-1' in case both labels are '-1'. The tracking rules are applied to acquire the final decision for this combined method (C). The seven possible cases (0,0), (-1,-1), (-1,1), (-1,2), (1,1), (1,2), and (2,2), result in values 0, -1, $\frac{1}{2}$, 1, 1, $1\frac{1}{2}$, and 2, respectively. Indeed different fusion techniques are possible.

From the results in Tables 1, 2, 3, and 4 we observe that: A significant increase of the *true positive* (TP) versus the *false positive* (FP) rate is achieved when the tracking history is used, especially the number of FPs is lowered. This means that many false alarms occur just once or twice at a certain location, whereas the truly moving features are consistently labeled as moving. As expected, the randomness of the model selection within the RANSAC framework has some influence on the number of TP and FP. Increasing the number of iterations increases the chance to select the model with most inliers, but makes the algorithm less efficient for practical use. For most of the video sequences our modification to the rigidity constraint improves on the TP/FP track rates (compare $H+$ to H). In fact, this method outperforms both the fundamental matrix and trifocal tensor methods for all sequences. The combination of the trifocal tensor and the normalized plane+parallax methods (C) performs best for the sequences with the frontal car and multiple objects, but for the other sequences the normalized plane+parallax method ($H+$) performs better. Another advantage of the combined method C is its high number of TPs, which makes it very well suited for additional techniques, such as MOD clustering and segmentation.

To elaborate on the number of frames in which a single car is detected, we show in Fig. 5 for each of the methods its number of tracked true positives per frame. In these sequences, the plane+parallax methods (H and H+) detect the car in more frames than the fundamental matrix and trifocal tensor methods do. The graphs in Fig. 5 clearly show that it is more challenging to detect a car moving in a parallel direction with respect to the moving camera, than a car moving in a perpendicular direction.

Some of the MOD results are shown in Fig. 6. In these figures, the yellow and red squares are the H and $H+$ MODs, the blue diamonds the F MODs, and the pink crosses the T MODs. The circles are the tracked MODs of the H (yellow), the $H+$ (red), the F (blue), and the T (pink) methods. The SIFT outliers are shown with black circles. Note that some of the false positives are cause by inaccurate SIFT features.

Table 1. MOD results of the frontal car video with 547 positives and 19744 negatives in 45 frames. The average (and std. dev.) number of true positives and false positives over three runs is shown with and without tracking.

	TP		FP		TP track		FP track		$\frac{TP}{FP}$	$\frac{TP}{FP}$ track
H	192	(20)	289	(25)	132	(15)	90	(31)	0,7	1,5
H+	196	(11)	245	(19)	141	(6)	82	(18)	0,8	1,7
F	100	(12)	310	(25)	51	(7)	40	(8)	0,3	1,3
T	152	(13)	393	(10)	98	(24)	48	(10)	0,4	2,0
C					144	(12)	47	(10)		3,0

Table 2. MOD results of the side car video with 313 positives. and 38677 negatives. in 80 frames. The average (and std. dev.) number of true positives and false positives over three runs is shown with and without tracking.

	TP		FP		TP track		FP track		$\frac{TP}{FP}$	$\frac{TP}{FP}$ track
H	130	(8)	146	(28)	97	(5)	28	(7)	0,9	3,5
H+	118	(8)	130	(9)	91	(10)	23	(3)	0,9	3,9
F	129	(6)	248	(17)	101	(4)	98	(5)	0,5	1,0
T	144	(5)	357	(15)	102	(4)	87	(5)	0,4	1,2
C					108	(6)	31	(9)		3,5

Table 3. MOD results of the left turn video with 0 positives. and 26677 negatives. in 80 frames. The average (and std. dev.) number of true positives and false positives over three runs is shown with and without tracking.

	FP		FP track	
H	272	(11)	65	(6)
H+	291	(9)	55	(12)
F	258	(8)	96	(4)
T	265	(3)	44	(3)
C			45	(6)

Table 4. MOD results of the multiple objects video with 4104 positives. and 44783 negatives. in 101 frames. The average (and std. dev.) number of true positives and false positives over three runs is shown with and without tracking.

	TP		FP		TP track		FP track		$\frac{TP}{FP}$	$\frac{TP}{FP}$ track
H	536	(35)	377	(41)	182	(25)	55	(22)	1,4	3,3
H+	485	(19)	381	(24)	155	(19)	50	(13)	1,3	3,1
F	679	(21)	471	(21)	350	(8)	149	(2)	1,4	2,4
T	757	(40)	656	(44)	331	(30)	126	(7)	1,2	2,6
C					257	(18)	66	(8)		3,9

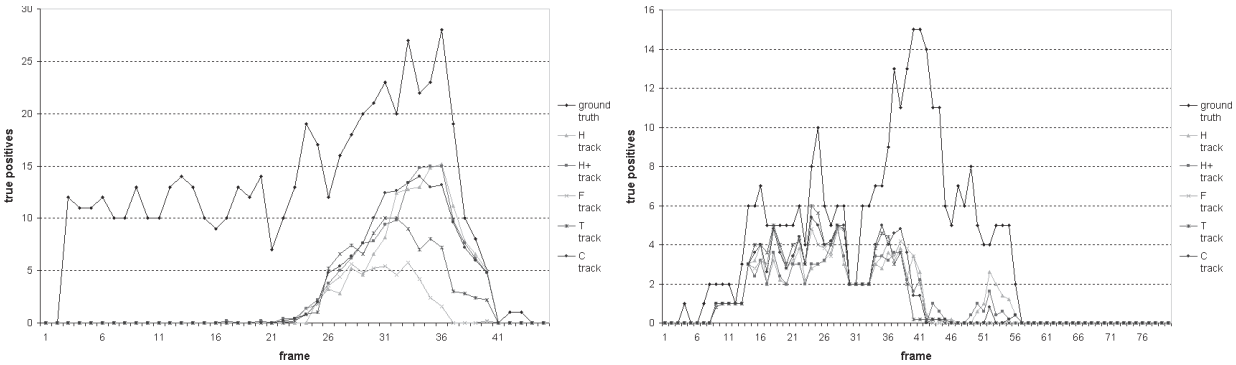


Figure 5. The number of true positive car detections in the frontal car (left) and side car (right) sequences averaged over five runs. The plane+parallax methods (H and H+) detect the car over a longer period.

7. CONCLUSION

In this paper we perform three-view moving object detection while the camera moves through challenging urban scenes. Because the camera moves through the scene, both static and moving objects appear to move and simple image subtraction fails in retrieving the moving objects. Because of the three-dimensionality of the scene, the assumption that the entire static scene can be transformed from one frame to another with a 2D planar homography fails as well.

Instead, we modified and applied the fundamental matrix, plane+parallax, and trifocal tensor methods to these urban scenes. We automated the plane+parallax algorithm and adjusted its rigidity constraint, we combined the results of two

fundamental matrices to improve this method for three views, and we modified the transfer error of the trifocal tensor. These modifications enable a more robust (and automatic) detection of moving objects in longer video sequences.

From the video sequences SIFT keys, SIFT matches, and SIFT triplets were extracted and used for the automatic detection and tracking of the moving objects. To quantitatively evaluate the performance of the three algorithms, we manually segmented the moving objects and report the fraction of true positives versus false positives detections. The plane+parallax method performs very well on our datasets and we prove that our modification outperforms the original method. Also the combination of the modified plane+parallax and the trifocal tensor method performs well. The fundamental matrix method, which was developed for two views and has some known flaws, has the lowest MOD performance.

ACKNOWLEDGMENTS

We thank Gijs Dubbelman and Julien Fauqueur for their video sequences.

REFERENCES

[1] Irani, M. and Anandan, P., "A Unified Approach to Moving Object Detection in 2D and 3D Scenes," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(6), 577–589 (1998).

[2] Hartley, R. and Zisserman, A., [*Multiple View Geometry in Computer Vision*], Cambridge University Press, Cambridge, UK, second edition (2004).

[3] Sawhney, H., Guo, Y., and Kumar, R., "Independent motion detection in 3D scenes," *PAMI* **22**(10), 1191–1199 (2000).

[4] Yuan, C., Medioni, G., Kang, J., and Cohen, I., "Detecting Motion Regions in the Presence of a Strong Parallax from a Moving Camera by Multiview Geometric Constraints," *PAMI* **29**(9), 1627–1641 (2007).

[5] Fischler, M. and Bolles, R., "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM* **24**(6), 381–395 (1981).

[6] Yamaguchi, K., Kato, T., and Ninomiya, Y., "Vehicle Ego-Motion Estimation and Moving Object Detection using a Monocular Camera," in [*ICPR*], 610–613 (2006).

[7] Torr, P., Zisserman, A., and Maybank, S., "Robust Detection of Degenerate Configurations for the Fundamental Matrix," *CVIU* **71**(3), 312–333 (1998).

[8] Fauqueur, J., Brostow, G., and Cipolla, R., "Assisted Video Object Labeling By Joint Tracking of Regions and Keypoints," in [*ICCV*], (2007).

[9] Lowe, D., "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV* **2**(60), 91–110 (2004).

[10] Mikolajczyk, K. and Schmid, C., "A Performance Evaluation of Local Descriptors," *PAMI* **27**(10), 1615–1630 (2005).

Figure 6. Moving object detection and tracking in the frontal car, side car, and multiple objects videos. Two examples are shown that are five frames apart.