Mono-vision based moving object detection in complex traffic scenes

Vincent Frémont, Sergio Alberto Rodriguez Florez, Bihao Wang

Mono-Vision based Moving Object Detection in Complex Traffic Scenes

Vincent Frémont[1] , Sergio Alberto Rodríguez Florez[2] and Bihao Wang[1]

***Abstract*—Vision-based dynamic objects motion segmentation can significantly help to understand the context around vehicles, and furthermore improve road traffic safety and autonomous navigation. Therefore, moving object detection in complex traffic scene becomes an inevitable issue for ADAS and autonomous vehicles. In this paper, we propose an approach that combines different multiple views geometry constraints to achieve moving objects detection using only a monocular camera. Self-assigned weights are estimated online moderating the contribution of each constraint. Such a combination enhances the detection performance in degenerated situations. According to the experimental results, the proposed approach provides accurate moving objects detections in dynamic traffic scenarios with large camera motions.**

***Index Terms*—Moving object detection, Monocular vision, Multiple views geometric constraints, Dynamic scene analysis**

I. Introduction

Traffic scene understanding [1], [2] has been a popular topic for the past few years, especially in the field of autonomous vehicles. Among these methods, vision based moving object detection from a moving vehicle is still one of the most challenging subjects, because of the complexity of motion models, changing illumination conditions and limited embedded processing capabilities.

In this area, existing approaches can be mainly structured into three main categories: Motion clustering methods [3], [4], foreground and background segmentation [5], [6] and geometric constraints based detection [7], [8], [9], [10]. Each category meets different requirements for specific applications. For example, clustering methods usually incorporate subspace constraints to segment the different motions. These methods can provide precise results. But most of them, like [3], rely on prior assumptions and are restricted to short video sequences. The advanced background segmentation methods in [5], [6], [11] can handle both spatial and temporal information at the same time. However, it is difficult to avoid the background model from being contaminated by foreground pixels in cases of complex environment with strong illumination changes or similar texture mixed together. Geometric constraints on the other hand, are more effective for moving object detection related to 3D scene reconstruction, such as multi-body Structure-from-Motion (SfM) [12], [7], [13] or Simultaneous Localization, Mapping and Moving Object Tracking (SLAMMOT) [14], [15].

In [7], the authors propose an incremental approach to detect moving objects at different speed by accumulating information from two views through the epipolar constraint. However, using this two-views tensor alone shows limitation when facing degenerated motion cases, e.g. surrounding vehicles moving in the parallel direction with ego-vehicle. To cope with this problem, plane+parallax methods have been broadly discussed, and new constraints have been proposed. For example, Flow Vector Bound constraint [8] is proposed by finding the reasonable bound of parallax range for static points. Any point with a parallax value falling out of the range will be given a high probability of being mobile. This constraint is also combined with graph-based clustering to segment motions recursively in a later work [14]. On the other hand, the authors in [9] proposed an algebraic three-view geometric constraint: The structure consistency constraint that encapsulates the plane+parallax information. In [10], the authors improved this approach by replacing the epipole with a reliable tracked feature point set as reference for projective depth calculation. This modification avoids noisy information introduced by epipole estimation. The advantage of these two approaches is that no reconstruction and no constant reference plane are needed. However, existing approaches rely on manually tuned parameters for constraint combination and they have been evaluated only on datasets with small baseline camera motion.

In this paper, we propose an enhanced geometric constraint-based approach for moving objects detection. Both two-views and three-views geometric constraints are applied in our approach: The epipolar, the trifocal tensor re-projection and the structure consistency constraints. All constraints contributions are combined through a flexible weight assignment procedure so as to infer the likelihood of a point being mobile. The resulting residual motion image is then refined using both road segmentation and connected components labeler in order to retrieve on-road moving objects entities. An evaluation of our proposed approach was conducted on the KITTI dataset[16] and the experimental results indicate that our approach can handle challenging dynamic traffic scenes with large camera motions, while providing an accurate detection of the moving objects using only a monocular video sequence.

The authors are with [1]*Sorbonne Universites, universite de technologie de Compiegne, CNRS, UMR 7253 Heudiasyc-CS 60 319, 60 203 Compiegne Cedex, France,* [2]*SATIE, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 94235 Cachan, France.* *This work has been carried out in the SIVALab joint laboratory between UTC, CNRS and Renault.*

The paper is organized as follows. Section II comes back on multiple views geometric constraints and the definitions of residuals errors for each of them. Then Section III, presents the new constraint combination approach with an algorithm evaluation on pixel level. An application orientated system and the corresponding experimental results are presented in Section IV, and the result is evaluated on object level in real traffic scenarios. Finally, Section V concludes the paper and presents some future works.

II. Multiple View Constraints

A. 2-views Tensor

Let a 3D point $\boldsymbol{P}$ in the world coordinate be observed from two views (see Fig. 1). The perspective projections coordinates of $\boldsymbol{P}$ in the two images are denoted by points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in homogeneous coordinates through a fully calibrated pinhole camera model. The fundamental matrix $\mathbf{F}_{21}$ (2-views tensor) defines a linear mapping of the point $\boldsymbol{x}_1$ to its corresponding epipolar line $\boldsymbol{l}_2$ in the first view as follows [17]:

$$\boldsymbol{l}_2 \sim \mathbf{F}_{21}\boldsymbol{x}_1 \tag{1}$$

Where $\sim$ mean an equality up to an unknown scale factor. If $\boldsymbol{P}$ is static in the observed scene, the corresponding point $\boldsymbol{x}_2$ of $\boldsymbol{x}_1$ belong to the epipolar line, i.e. $\boldsymbol{x}_2^T\boldsymbol{l}_2 \sim 0$. Considering a set of matched/tracked points between the two images, the fundamental matrix can be robustly estimated from the algorithms presented in [17]. Then it is possible to calculate the residual errors (see Eq. 2) for matched/tracked points and to detect potential moving points:

$$r_F = \sum_{i=1}^{2} d(\boldsymbol{l}_i, \boldsymbol{x}_i) \tag{2}$$

Where $d(\boldsymbol{l}_i, \boldsymbol{x}_i)$ represent the point-to-line distance in the image.

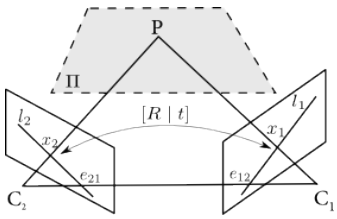


Figure 1: Epipolar Geometry between two views.

B. 3-views Tensor

The trifocal tensor is a closed-form representation of the geometry relations between three different camera viewpoints. In its matrix representation is composed of a set of three 3 × 3 matrices $\{\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3\}$.

Let assume that the camera matrices of three views are represented by canonical projection matrices: $\mathbf{P}_1 = [\mathbf{I} \mid \mathbf{0}]$ and $\mathbf{P}_2 = [\mathbf{A} \mid \mathbf{a}_4]$, $\mathbf{P}_3 = [\mathbf{B} \mid \mathbf{b}_4]$, where $\mathbf{P}_2$ and $\mathbf{P}_3$ are defined with respect to the first camera frame. $\mathbf{A}$ and $\mathbf{B}$ are 3×3 matrices representing the infinite homographies from the first view to the second and to the third views respectively. The 3×1 vectors $\mathbf{a}_4$ and $\mathbf{b}_4$ are the epipoles in second view and the third view, arising from the first camera. The trifocal tensor is then formalized as follows [17]:

$$\mathcal{T} = [\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3] \tag{3}$$

With

$$\mathbf{T}_i = \mathbf{a}_i\mathbf{b}_4^T - \mathbf{a}_4\mathbf{b}_i^T \tag{4}$$

Where, the vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ are the i^{th} columns of the camera matrix $\mathbf{P}_2$ and $\mathbf{P}_3$ for $i = 1, ..., 3$. As described in [17], the trifocal tensor transfer a point $\boldsymbol{x}_1$ from the first image to a point $\boldsymbol{x}_3'$ in the third image. Ideally, the projected point $\boldsymbol{x}_3'$ satisfies $\| \boldsymbol{x}_3 - \boldsymbol{x}_3' \| = 0$ where $\boldsymbol{x}_3$ represents the image coordinates of the matched/tracked point $\boldsymbol{x}_1$ through the 3 views. Thus, as for the 2-views tensor, the residues of the trifocal tensor constraint can be defined as:

$$r_T = \| \boldsymbol{x}_3 - \boldsymbol{x}_3' \| \tag{5}$$

It is worth noting that the use of the trifocal tensor based point transfer avoids degenerated motion configurations that cannot be handled with the fundamental matrix.

C. Structure Consistency

Based on the concept of induced plane homography [17], a 3D scene can be represented by a dominant 3D plane and the off-plane points located through a residual parallax noise with respect to the homography plane. Assuming that such a reference plane Π in 3D space is estimated between two views (see Fig. 1), it introduces the homography matrix $\mathbf{H}_{12}$ that transfers all the in-plane points from the second image to the first image. As stated previously, for a general point $\boldsymbol{P}$ in 3D space, its projections in the two images are denoted by points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Let a point $\boldsymbol{P}'$ be the intersection of the projection ray $\boldsymbol{C}_2\boldsymbol{P}$ with plane Π. The projection of point $\boldsymbol{P}'$ in the first image can be obtained by the homography transform induced by the plane Π as follows:

$$\boldsymbol{x}_1' \sim \mathbf{H}_{12}\boldsymbol{x}_2$$

where the second camera center $\boldsymbol{C}_2$, the in-plane point $\boldsymbol{P}'$ and the off-plane point $\boldsymbol{P}$ are collinear. According to the invariant properties of a projective transform, projections in the first view, i.e. epipole $\boldsymbol{e}_{12}$, point $\boldsymbol{x}_1'$ and point $\boldsymbol{x}_1$ must remain collinear. Hence, the point $\boldsymbol{x}_1$ can be represented as:

$$\boldsymbol{x}_1 \sim \mathbf{H}_{12}\boldsymbol{x}_2 + \kappa_{12}\boldsymbol{e}_{12} \tag{6}$$

Where the scalar κ_{12} corresponds to the projective depth relative to the reference plane Π [17]. As proposed in [18], κ_{12} can be estimated by:

$$\kappa_{12} = \frac{(\mathbf{H}_{12}\boldsymbol{x}_2 \times \boldsymbol{x}_1)^T(\boldsymbol{x}_1 \times \boldsymbol{e}_{12})}{\| \boldsymbol{x}_1 \times \boldsymbol{e}_{12} \|^2} \tag{7}$$

Eq. 7 is derived from Eq. 6 by cross-multiplying both sides of the equation with $\boldsymbol{x}_1$. It is then clear that point $\boldsymbol{P}$ is on the

plane Π, if $\kappa_{12} = 0$. Otherwise, the sign of κ_{12} indicates on which side the point $\boldsymbol{P}$ stands with respect to the reference plane Π.

Knowing the projective depth, the 3D points $\boldsymbol{P}$ can be represented by a projective structure constructed from the 2 views:

$$\widetilde{\boldsymbol{P}}_{12} = (\boldsymbol{x}_1; \kappa_{12}) = [u_1, v_1, 1, \kappa_{12}]^T \tag{8}$$

Then, considering a third view, $\boldsymbol{P}$ also can be represented by the projective structure between the second view and the third one: $\widetilde{\boldsymbol{P}}_{23} = (\boldsymbol{x}_2; \kappa_{23}) = [u_2, v_2, 1, \kappa_{23}]^T$, where, κ_{23} is the projective depth to a new reference plane connecting the second view and the third view. There exists a relationship that links the two projective structures from a static point in quadratic form. This relationship is denoted structure consistency constraint and is formalized in Eq. 9:

$$r_G = \left\| \widetilde{\boldsymbol{P}}_{23}^T \mathbf{G} \widetilde{\boldsymbol{P}}_{12} \right\| \sim 0 \tag{9}$$

where, $\mathbf{G}$ is a 4×4 matrix representing a bilinear constraint for 3D projective structures of the same point [9]. Therefore, the residues r_G of the structure consistency constraint can be used to detect moving from static points given pixel matching/tracking in three views. To notice that, the matrix $\mathbf{G}$ encapsulates the normal vectors of two reference planes, the camera's relative orientation, and two unknown scale factors κ_{12} and κ_{23}. It directly relates the pair of projective structures from views $1 \leftrightarrow 2$ and views $2 \leftrightarrow 3$ without knowing the camera configuration and the plane position. Employing the estimated projective structures, the matrix $\mathbf{G}$ is computed by solving Eq. 9, enforcing $\| \mathbf{G} \| = 1$ using both a linear solution and a non-linear optimization on the residual errors using the Levenberg-Marquardt algorithm as proposed in [9].

D. Modified Structure Consistency

To build the projective structures $\widetilde{\boldsymbol{P}}_{12}$ and $\widetilde{\boldsymbol{P}}_{23}$ for the estimation of the matrix $\mathbf{G}$ through three views, the projective depth κ_{12} and κ_{23} are required. Eq. 7 is usually used for the parallax based approaches such as structure from motion [19]. However, this equation can not be evaluated for all image points because of singularities.

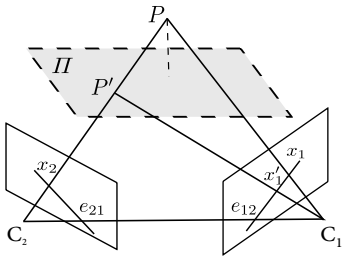
As shown in Fig. 2b, the cross products in Eq. 7 can be represented by two parallelograms (red one and green one respectively). For collinear vectors, this product is zero regardless of their scale. Thus, the Eq. 7 is undefined for images points lying on the line defined by the origin of the image coordinate $\boldsymbol{O}$ and the epipole $\boldsymbol{e}_{12}$. Fig. 3 shows some results of moving point detection using Eq. 7. Pixels passing through the image origin and the epipole are wrongly detected as moving pixels because of the invalid projective depth calculation.

To cope with this situation, we propose in this paper, a better conditioned calculation of the projective depth calculation:

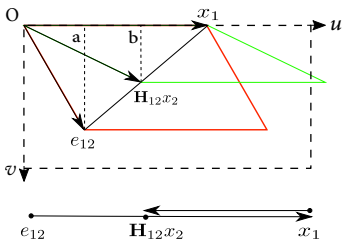
$$\kappa_{12} = \cos\theta \cdot \frac{\| \mathbf{H}_{12}\boldsymbol{x}_2 - \boldsymbol{x}_1 \|}{\| \boldsymbol{x}_1 - \boldsymbol{e}_{12} \|}, \ (\boldsymbol{x}_1 \neq \boldsymbol{e}_{12}) \tag{10}$$

with

$$\cos\theta = \frac{(\mathbf{H}_{12}\boldsymbol{x}_2 - \boldsymbol{x}_1) \cdot (\boldsymbol{x}_1 - \boldsymbol{e}_{12})}{\| \mathbf{H}_{12}\boldsymbol{x}_2 - \boldsymbol{x}_1 \| \| \boldsymbol{x}_1 - \boldsymbol{e}_{12} \|} \tag{11}$$



(a) Off-plane point P is observed in two views, the relationship between its projections composed of a planar part and parallax part.



(b) The plane plus parallax composition figured in the first view

Figure 2: Plane+Parallax geometric configuration

According to the property of the cross product, for any $\boldsymbol{x}_1 \neq \boldsymbol{e}_{12}$, the parameter κ_{12} can be considered as the signed area proportion of the parallelograms sided by $(\overrightarrow{\boldsymbol{O}\boldsymbol{e}_{12}}, \overrightarrow{\boldsymbol{O}\boldsymbol{x}_1})$ and $(\overrightarrow{\boldsymbol{O}\mathbf{H}_{12}\boldsymbol{x}_2}, \overrightarrow{\boldsymbol{O}\boldsymbol{x}_1})$. Besides, the area of the parallelogram can also be computed by the product of its base and height:

$$A = d \cdot h \tag{12}$$

In Fig. 2b, the height h of parallelogram sided by $(\overrightarrow{\boldsymbol{O}\boldsymbol{e}_{12}}, \overrightarrow{\boldsymbol{O}\boldsymbol{x}_1})$ is denoted by a; the height of parallelogram sided by $(\overrightarrow{\boldsymbol{O}\mathbf{H}_{12}\boldsymbol{x}_2}, \overrightarrow{\boldsymbol{O}\boldsymbol{x}_1})$ is denoted by b. While the basis of the two parallelograms are the same, so $d = \| \overrightarrow{\boldsymbol{O}\boldsymbol{x}_1} \|$. Hence, the scale of κ_{12} can be simplified as the proportion of parallelogram's heights. Using similar triangles rule, we obtain:

$$| \kappa_{12} | = \frac{b}{a} = \frac{\| \mathbf{H}_{12}\boldsymbol{x}_2 - \boldsymbol{x}_1 \|}{\| \boldsymbol{x}_1 - \boldsymbol{e}_{12} \|}$$

The sign of κ_{12} indicates the direction of the point $\boldsymbol{P}$ to the plane Π. Projected into the second view, the sign is defined by the direction of point $\boldsymbol{x}_1$ to point $\mathbf{H}_{12}\boldsymbol{x}_2$. As shown in Fig. 2b, the points $\boldsymbol{e}_{12}$, $\boldsymbol{x}_1$, $\mathbf{H}_{12}\boldsymbol{x}_2$ are collinear, and the direction of the vector $(\mathbf{H}_{12}\boldsymbol{x}_2 - \boldsymbol{x}_1)$ can be represented by its intersection angle θ with vector $(\boldsymbol{x}_1 - \boldsymbol{e}_{12})$. If the two vectors are in the same direction, $\theta = 0$ and $\cos\theta = 1$,

therefore, κ_{12} is positive. On the contrary, if the two vectors are in opposite directions, κ_{12} is negative.

Comparing to the original method presented in [18] to calculate κ_{12}, our proposed formula can be used for most of the points in image plane except for the epipole.

Figure 3: Example of unstable detection result caused by unmodified projective depth calculation. Top: Original image. Bottom: Moving pixels detected by Eq.7

III. Pixels Motion Segmentation using Multiple View Constraints

A. Residues distribution models

If the epipolar constraint in Eq. 2 is established, the points should lie on their corresponding epipolar lines. Ideally, if a point is static, r_F should be equal to 0, however, because of the image noise, it is usually a positive value close to 0. Assuming that the noise of points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ follows a normal distribution, for the static points, their re-projection distances from two views should follow a Chi-squared χ^2_m distribution with m degrees of freedom (DOF). According to [17], the residues from inlier points can be modeled by χ^2_1. Fig. 4 shows an example of the residual values distribution followed by the inlier points obtained from the epipolar constraint estimation on real images selected in the KITTI dataset.

Regarding the structure consistency constraint, if the residues r_G converge to 0, then the two projective structures are consistent with the $\mathbf{G}$ matrix and the corresponding 3D point is static. After normalization, we can assume the noise of each element in projective structures have the same normal distribution deviation. The residues of structure consistency constraint follow a χ^2_3 distribution as for the trifocal tensor based point transfer residue r_T [17].

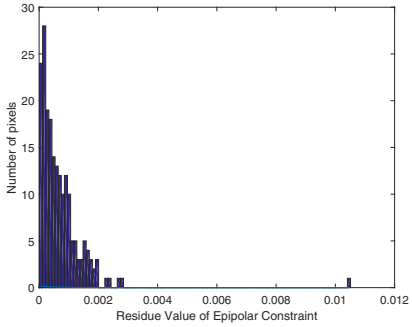


Figure 4: The residues distribution of r_F on the inliers.

B. Constraints Likelihood Definition

With knowing the inlier residues distribution, an interval threshold τ with confidence level of $1-\alpha = 0.95$ of χ^2_m distribution can be defined. For example, $\tau_F = 3.84\sigma^2$, where σ^2 represent the data scale and it can be obtained by means of a Maximum-likelihood estimator (MLE).

For points whose residual value is out of the 0.95 confidence interval, the larger the residual value is, the more likely the point is mobile. Based on this analysis, moving points likelihood functions can be built, for each geometric constraint, as:

$$\mathcal{L}_i(\boldsymbol{x}) = \begin{cases} 1 - e^{-\frac{(r_i(\boldsymbol{x}) - \tau_i)}{\tau_i}} & r_i(\boldsymbol{x}) > \tau_i \\ 0 & r_i(\boldsymbol{x}) \leqslant \tau_i \end{cases} \quad (i = F,\ G,\ T) \tag{13}$$

where x corresponds to a pixel tracked over multiple views and $F,\ G,\ T$ stand for respectively epipolar, structure consistency and trifocal tensor point transfer constraints. Compared to the likelihood definition of [9], all the thresholds are dynamically estimated from the monocular video sequence.

C. Likelihoods combination

The combined motion likelihood of an image point $\boldsymbol{x}$ is defined by a weighted sum of the three geometric constraints:

$$\mathcal{L}(\boldsymbol{x}) = \sum w_i \cdot \mathcal{L}_i(\boldsymbol{x}),\ (i = F,\ G,\ T) \tag{14}$$

The variable w_i defines the weight accorded to each constraint in the current frame, satisfying $\sum w_i = 1$. The assigned weight is adaptive and is defined by analyzing the residues distribution.

Because of the image noise and the accumulated errors like imprecise inliers set estimation, the distribution of residues from inliers could have some deviation. MLE provides the instantaneous parameter of the distribution for each frame, and it usually do not follow exact χ^2_m distribution. Hence, likelihoods proposed in Eq. 13 might not be accurate. This accuracy is measured by the difference between estimated DOF and the corresponding DOF of χ^2_m distribution. We also consider the skewness of the distribution, since high skew value indicates inliers that are more likely to be distinguished from outliers. For that purpose, a coefficient of variation $cv = \frac{\sigma}{\mu}$ is introduced so as to measure the skewness of each distribution. Because of the dynamics, the weights distributed to each constraint are assigned considering both the current DOF differences and the coefficient of variation cv for each frame. Smaller values contribute to bigger weights for constraints combination:

$$w_F : w_G : w_T = \frac{1}{\Delta_F + cv_F} : \frac{1}{\Delta_G + cv_G} : \frac{1}{\Delta_T + cv_T}, \quad \text{with } \sum w_i = 1,\ (i = F,\ G,\ T)$$

D. Moving points segmentation

The purpose of our algorithm is to identify all the moving objects in the scene. Every potential moving points which might leads to a dangerous situation should be labeled out. From this consideration, we set a threshold for the moving points segmentation: For all points which has a combined likelihood of being mobile , those whose likelihood is bigger than 65% is considered as mobile points in the scene.

$$M(\boldsymbol{x}) = \begin{cases} 1, & \mathcal{L}(\boldsymbol{x}) \geq 0.65 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$M(\boldsymbol{x})$ is the state of a pixel $\boldsymbol{x}$ being mobile or not, state 1 means that the pixel belongs to a moving object, 0 means that the pixel is static. In the end, the multiple geometric constraints based moving points detection algorithm can be summarized as:

Algorithm 1 Moving points segmentation algorithm

Input: Corresponding points in three different views $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$

Output: Segment moving points from static ones

1: ▶ Estimate geometric constraints $\mathbf{F}, \mathbf{G}, \mathcal{T}$
2: ▶ Estimate static points residual distribution model from the inliers residuals using MLE.
3: ▶ Compute the motion likelihood for each point using constraints expressed in Eq. 13
4: ▶ Constraints combination (Eq. 14)
5: ▶ Moving points classification (Eq. 15)

IV. Application to Moving Objects Detection

A. Implementation Details

In the proposed moving object detection system, a background subtraction approach based on homography registration is first applied to preserve potential moving pixels in a residual image [20]. If the camera is static, the moving objects in the scene are exactly the result of background subtraction. On contrary, if the camera is moving, geometric constraints as fundamental matrix $\mathbf{F}$ and the trilinear structure consistency matrix $\mathbf{G}$ need to be estimated for potential moving pixels classification. The trifocal tensor is added to enhance the moving pixels classification. Before applying the geometric constraints on the residual image, the road detection results detailed in [21] are used to define a driving space area in order to reduce the computation time and the number of false alarms. To notice that other road detection approaches evaluated within the KITTI Road benchmark can also be used. The corresponding points of potential moving pixels in the three views are obtained through dense optical flow estimation [22]. Then a likelihood is assigned to each constraint (Eq. 13) based on the detection results. Finally, the likelihood combination function (Eq. 14) is applied for combining information from the different constraints to segment the moving pixels. After removing the moving pixels outside the driving space area, on-road moving pixels are then clustered using connected components labeler [23].

Algorithm	Average time (s)
Dense optical flow [22]	19.2
Feature detection and tracking [25]	0.07
Fundamental matrix estimation (500 points) [17]	0.3
Trifocal tensor estimation (500 points) [17]	0.55
Road detection[21]	1.02
Moving pixels segmentation (14,15)	3.54
Connected components labeler[23]	3.22
Global average computation time (for 3 frames)	27.9

Table I: Average computation time for 3 frames

B. Experimental Results

The proposed moving object detection approach has been tested on the KITTI dataset [24]. Two different video sequences have been selected:

- Dataset 1: KITTI raw data, 2011_09_26_drive_0005 with a minivan and a cyclist continuously appearing in the sequence.
- Dataset 2: KITTI raw data, 2011_09_29_drive_0071 with narrow street passing through a commercial center, with many pedestrians and other traffic participants moving in different directions.

The algorithm implementation and the experiments were conducted on a standard PC with Windows 7 Enterprise OS, Intel CPU of 2.66 GHz and Matlab R2015a. The geometric constraints have been estimated using features detection and tracking from [25]. The dense optical flow estimation is performed with the code of [22]. The average error of the optical flow estimation using this method is about 3 to 5 pixels for the KITTI dataset. This result is obtained by evaluating the flow estimation results using the KITTI-flow benchmark. The average computation time of our approach, under Matlab without c++ mex function with GPU acceleration, for three consecutive images is about $27.9s$, knowing that it is possible to perform a sliding buffer strategy in order to use previous estimations and save computation time. It is also important to notice that many parts of the approach, especially the optical flow computation, can be processed using GPU implementation for example on the DRIVE PX2 Nvidia embedded platform.

From the results of Dataset 1, a false alarm appears after the constraints combination stage. It is located on the left side of the image, and is due to trees and parallax. This is a common example, where most of the false alarms were induced by trees or the occluded parallax. This is because in such a kind of regions, the dense optical flow cannot be correctly estimated. Fortunately, such regions mostly appears along or outside the road. Hence, applying the road space constraint can greatly reduce such false alarms. The monovision-based road detection provides a driving space area, but there still exist many false positives/negatives that will impact negatively on the moving object detection. In Dataset 2 (see Fig. 5b), on the left side of the road there are two pedestrians walking away. They are very well detected by the geometric constraints, but according to the ROI, they are not in the traffic area, thus they are eliminated from the final detection result. This is because the left part of

the road surface are completely covered by the pedestrians. This situation is hard to avoid in cluttered environments. One solution is to introduce tracking strategies or object recognition to predict the presence of moving object in that situation.

As we can see from Tab. II, the detection rate of Dataset 1 is less than Dataset 2. The reason is that, when the objects are moving in the same direction as the host vehicle, a degenerate configuration may appear. Indeed, in this situation, the geometric constraints cannot segment the moving pixels from static background. The false alarm rate is higher in Dataset 1 since the scene is cluttered and there are more parallax than in the structured urban road. Meanwhile, the redundant detections that are caused by the default of background subtraction are more frequent in Dataset 2.

	Detection rate	Mis-detection	False alarms	Redundant detection
Dataset 1	50.80%	49.20%	29.84%	3.66%
Dataset 2	74.64%	25.36%	6.69%	28.03%

Table II: General evaluation of the moving object detection by monovision

V. Conclusion and further discussions

In this paper, we have presented a complete system for on-road moving objects detection based on monocular vision. It integrates multiple geometric constraints to detect the moving pixels in an estimated driving space. All the components together improved the efficiency and flexibility of the system: Efficiency because it is concentrated on detecting the traffic participants, and flexibility because the system can change its detection strategy according to the motion state of the camera/vehicle. We also analyzed the strength and limitations of each constraint. Especially, for the structure consistency constraint, we correct the formula for calculating projective depth. This simple correction may help to improve the reliability of the approach. For each constraint, we defined a likelihood function. Furthermore, we introduced the coefficients of variation as criteria to infer the importance of each constraints in the process of fusion of likelihoods.

Future works will be devoted on adding object tracking to improve the stability of the results and also on the use of geometric constraints in Deep Learning architectures for detecting moving objects by combining dense optical flow and pixel-wise semantic labels.

References

[1] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun. 3d traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):1012–1025, May 2014.

[2] Philip Lenz, Julius Ziegler, Andreas Geiger, and Martin Roser. Sparse scene flow segmentation for moving object detection in urban environments. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 926–932. IEEE, 2011.

[3] Shankar R Rao, Allen Y Yang, S Shankar Sastry, and Yi Ma. Robust algebraic segmentation of mixed rigid-body and planar motions from two views. *International journal of computer vision*, 88(3):425–446, 2010.

[4] Peter Ochs and Thomas Brox. Higher order motion models and spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 614–621. IEEE, 2012.

[5] Dong-Sun Kim and Jinsan Kwon. Moving object detection on a vehicle mounted back-up camera. *Sensors*, 16(1):23, 2015.

[6] Kwang Moo Yi, Kimin Yun, Soo Wan Kim, Hyung Jin Chang, and Jin Young Choi. Detection of moving objects with non-stationary cameras in 5.8 ms: Bringing motion detection to your mobile device. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 27–34, 2013.

[7] Soumyabrata Dey, Vladimir Reilly, Imran Saleemi, and Mubarak Shah. Detection of independently moving objects in non-planar scenes via multi-frame monocular epipolar constraint. In *Computer Vision–ECCV 2012*, pages 860–873. Springer, 2012.

[8] Abhijit Kundu, K Madhava Krishna, and Jayanthi Sivaswamy. Moving object detection by multi-view geometric techniques from a single camera mounted robot. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 4306–4312. IEEE, 2009.

[9] Chang Yuan, Gerard Medioni, Jinman Kang, and Isaac Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1627–1641, 2007.

[10] Fuyuan Xu, Guohua Gu, Kan Ren, and Weixian Qian. Motion segmentation by new three-view constraint from a moving camera. *Mathematical Problems in Engineering*, 2015, 2015.

[11] Daniya Zamalieva and Alper Yilmaz. Background subtraction for the moving camera: A geometric approach. *Computer Vision and Image Understanding*, 127:73 – 85, 2014.

[12] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(6):1134–1141, 2010.

[13] R. Sabzevari and D. Scaramuzza. Multi-body motion estimation from monocular vehicle-mounted cameras. *IEEE Transactions on Robotics*, 32(3):638–651, June 2016.

[14] Rahul Kumar Namdev, Abhijit Kundu, K Madhava Krishna, and CV Jawahar. Motion segmentation of multiple objects from a freely moving monocular camera. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4092–4099. IEEE, 2012.

[15] Abhijit Kundu, K Madhava Krishna, and CV Jawahar. Realtime motion segmentation based multibody visual slam. In *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*, pages 251–258. ACM, 2010.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[18] Andrea Fusiello, Stefano Caldrer, Sara Ceglie, Nikolaus Mattern, and Vittorio Murino. View synthesis from uncalibrated images using parallax. In *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on*, pages 146–151. IEEE, 2003.

[19] Ting Li, Vinutha Kallem, Dheeraj Singaraju, and René Vidal. Projective factorization of multiple rigid-body motions. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[20] Andrews Sobral and Antoine Vacavant. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122:4 – 21, 2014.

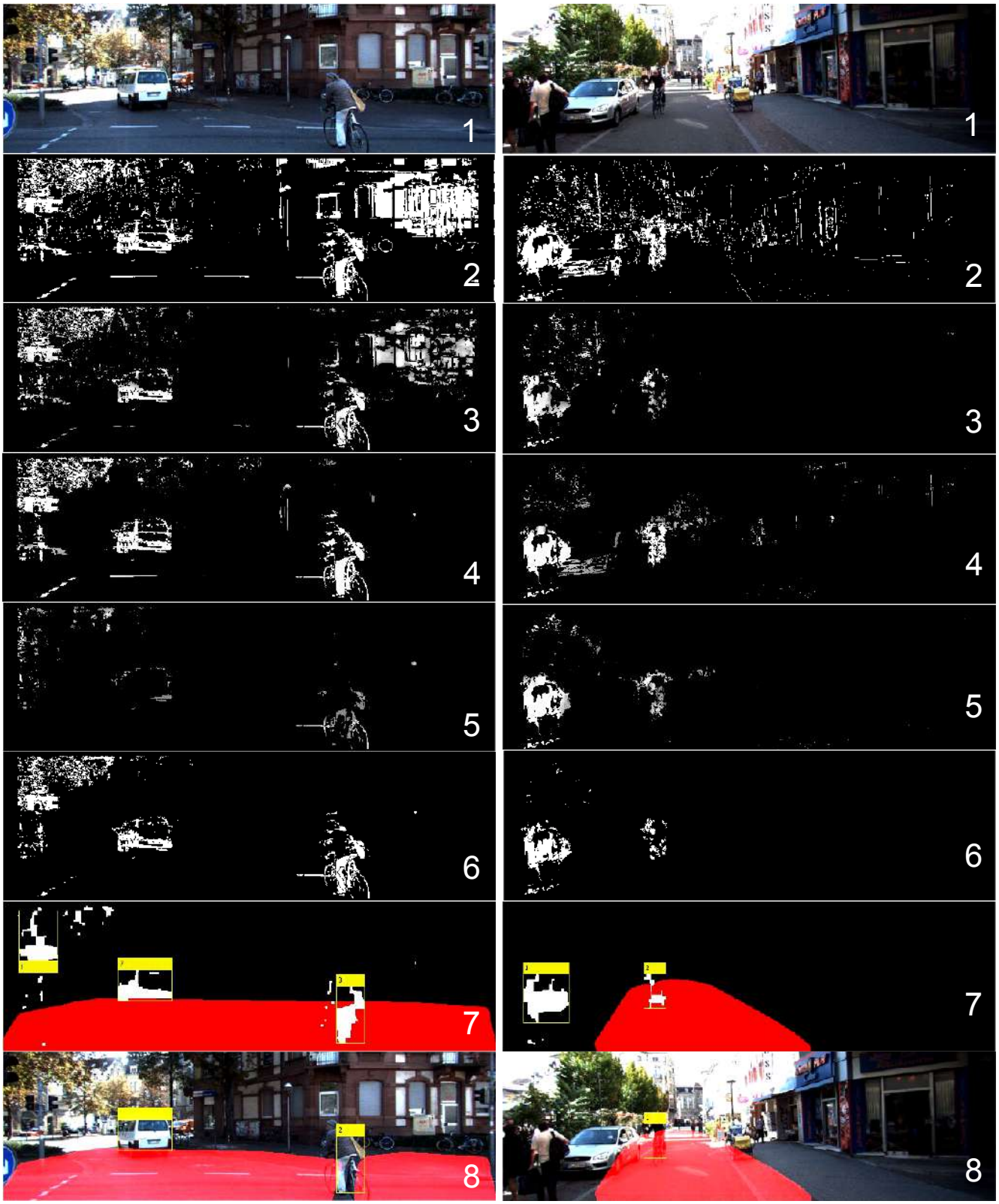
[21] B. Wang, V. Fremont, and S. A. Rodriguez. Color-based road detection and its evaluation on the kitti road benchmark. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 31–36, June 2014.

[22] Ce Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, MIT, 2009.

[23] Costantino Grana, Federico Bolelli, Lorenzo Baraldi, and Roberto Vezzani. YACCLAB - Yet Another Connected Components Labeling Benchmark. In *23rd International Conference on Pattern Recognition*. ICPR, 2016.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[25] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV)*, 2011.



(a) Example of moving detection in Dataset 1

(b) Example of moving detection in Dataset 2

Figure 5: On-road moving object detection in two datasets: First row to the end: 1- original image; 2- residual image after background subtraction; 3- confidence map of the epipolar constraint; 4- confidence map of structure consistency constraint; 5- confidence map of trifocal tensor constraint; 6- combined likelihood based detection result; 7- traffic area construction and blob analysis; 8- final detection result of on-road moving object detection.