2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

CNN-based multi-frame IMO detection from a monocular camera

Nolang Fanani[1], Matthias Ochs[1], Alina Stürck[1], Rudolf Mester[1,2]

***Abstract*— This paper presents a method for detecting independently moving objects (IMOs) from a monocular camera mounted on a moving car. A CNN-based classifier is employed to generate IMO candidate patches; independent motion is detected by geometric criteria on keypoint trajectories in these patches. Instead of looking only at two consecutive frames, we analyze keypoints inside the IMO candidate patches through multi-frame epipolar consistency checks. The obtained motion labels (IMO/static) are then propagated over time using the combination of motion cues and appearance-based information of the IMO candidate patches. We evaluate the performance of our method on the KITTI dataset, focusing on sub-sequences containing IMOs.**

I. Introduction

Classical visual odometry methods rely strongly on the rigidity of the depicted environment through which a robot (or car) moves. Deviations from this rigidity (e.g. other robots, cars, pedestrians, etc.) are traditionally excluded from egomotion analysis by using variants of RANSAC. Recently, methods have been proposed which are robust against other moving objects without RANSAC [5], [6]. However, these methods just 'blank out' all areas which are not conformant to the epipolar geometry induced by the ego-motion, but they do not analyze these areas further.

In the present paper, we present an approach which is kind of 'piggy-backed' on the mentioned recently developed propagation based tracking method (PbT) [6], and employs a CNN to produce candidate patches that correspond to single vehicle instances, thus potential *independently moving objects (IMOs)*. In the presented scheme, these patches are subsequently associated with each other over time using a dynamic motion model and simple appearance descriptors, and the conformity of these association with the epipolar geometry computed by PbT is checked. As a result, those IMOs that are moving differently from the motion of the ego-car can be detected.

There are two main contributions of our work. First, we propose a monocular IMO detection scheme which relies on multi-frame epipolar consistency checks. Hence, the case of epipolar-conformant IMOs is not in the scope of our work. Second, we propose a way to propagate and associate IMO candidates over time by combining the motion model with appearance information.

In this paper, after having presented related work, we present the detection and association principles used for this approach. Perspectives for additionally detecting and tracking 'epipolar-conformant' cars are provided in section VI-C.

[1]Visual Sensorics & Information Processing Lab, Goethe University Frankfurt am Main, Germany
[2]Computer Vision Laboratory, ISY, Linköping University, Sweden

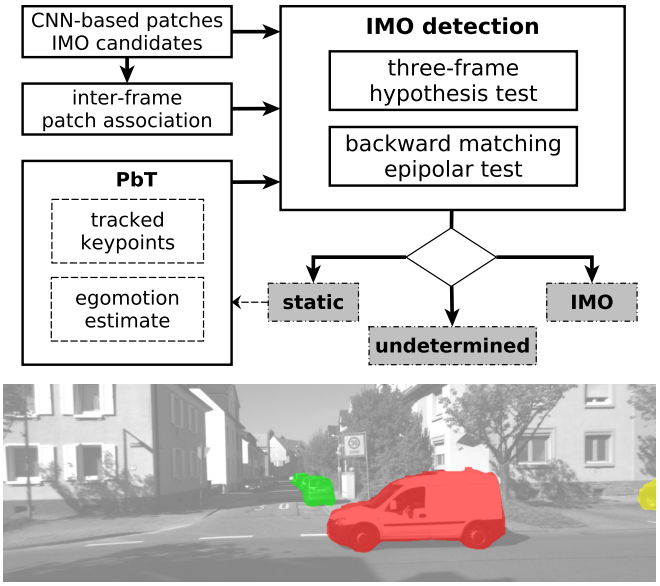

Fig. 1. **Top**: The scheme of the proposed IMO detection. **Bottom**: An example of car classifications into static (green), IMO (red) and undetermined (yellow).

We conclude with an evaluation of the method on the KITTI dataset.

II. Related work

The detection of independently moving objects (IMOs) from visual sensors is a vitally important part of many computer vision systems. A traditional application can be found in visual surveillance, where the camera is static, but recently, object detection from moving cameras is becoming more influential.

In this work we will focus on the detection of moving cars from a vehicle mounted camera. This scenario is very different to others such as handheld cameras (such as [12]) or general robot vision (such as [11]) in that motion is severely restricted.

In the development of other advanced driver assistance systems (ADAS) several approaches have been proposed. Many of those choose to work with additional information such as color images ([2], [17]) or a stereo system [24], [13]. In contrast to these we want to show that it is possible to reliably detect IMOs from a simple monocular camera.

Previously published monocular algorithms can be differentiated into two categories. Appearance-based approaches ([11], [15]) are often based on patch-matching or learning-techniques to determine the movement of any visible object.

Among such approaches are [27] who use features combined with a classification process for vehicle description or [28] who use an attention-inspired model to subtract less important image regions to obtain the moving foreground region.

On the other hand, motion-based approaches ([19], [26], [10]) aim to work with optical flow and other geometric constraints to determine varying motion patterns from IMOs.

We aim to provide an approach that combines cues from both the appearance of a car by employing a CNN-based detection as well as motion cues from optical flow based on the epipolar geometry to determine the presence of an independently moving object and track it. In this aspect our approach shares some similarities with [16] who use two separate CNNs to determine visual odometry and object localisation and fuse their results to obtain object localisations and also with [25] who use CNNs to obtain a rigidity score for each object and combine this with motion cues from optical flow. In contrast to their approach, we additionally use a series of hypothesis tests on keypoints lying on patches that have previously been identified as belonging to a vehicle in manner similar to [23] to track the regarded vehicle through consecutive frames. Bai et al [1] identifies IMOs by estimating the dense optical flows from each IMO candidate. In contrast, our approach utilizes sparse keypoints to identify static and IMOs.

III. Framework overview

The overall flow of the proposed method is presented in figure 1 (top image). It builds on a monocular visual odometry framework, the *propagation based tracking (PbT)* scheme [7]. An important principle of PbT is that the new relative pose for a new frame $n+1$ is *predicted* using the car ego-dynamics, and that a *refined* relative pose is computed only on the basis of keypoints that have been tracked at least twice, that is: keypoints which already passed a stringent test of being belonging to the static environment. All keypoints, including the new ones generated in sparsely covered areas of a new frame, are tracked in an epipolar-guided manner as discussed more detailed in section III-A.

In the present scheme, we detect IMO candidate patches for each new frame using the instance segmentation scheme of van den Brand et al. [22]. This results in M new IMO candidate patches for each new frame $n+1$. The generation of the CNN-based IMO candidate patches is discussed in section III-B.

All IMO candidate patches in image n are classified in one of the three states `static`, `IMO`, or `undetermined`. Keypoints that are located in IMO candidate patches are considered for pose estimation only if they have been classified as `static`.

When a new frame $n+1$ comes in, it is processed by the CNN in order to determine the IMO candidate patches. At the same moment, we have a set of old tracked keypoints in previous frame n. Some of those are already confirmed as `static` as they have been tracked at least twice and have shown 3D consistency (see section IV-A). Others have been tracked just once (from frame $n-1$ to n) and are only candidates for being considered as static.

Subsequently, the relative pose between frame n and $n+1$ is determined from all keypoints which are considered to be static in frame n in the standard manner used in propagation based tracking. With the resulting relative pose $n \rightarrow n+1$ being computed, individual keypoints can then be checked for being conformant with the epipolar geometry. By accumulating the checks from all keypoints on a patch, this leads to the classification of that patch into `IMO`, `static`, or `undetermined`. This IMO detection procedure is described in detail in section IV-A and section IV-B.

For each of the new IMO candidate patches in frame $n+1$, an association with the existing patches in frame n must be performed in order to propagate the motion state (`static`/`IMO`) over time. The association can be made *appearance-based*, that is: by comparing size and texture of the patches, or *tracking based*, that is: by checking matching residuals between keypoints inside of the patches. The inter-frame car patch association is discussed in section V.

A. PbT framework

As said, we build our scheme on the monocular visual odometry framework using propagation-based tracking (PbT) proposed in [5], [6]. Principles of keypoint generation and tracking are used in a similar way here, thus we give some details in the following. The egomotion of the ego-car is estimated using keypoints which have been confirmed to be static. These keypoints are the combination of keypoints outside CNN-based car masks and keypoints from car masks that have been classified as static cars. In addition, PbT with its epipolar constraint is able to propagate the static label of a car mask on subsequent frames as long as the keypoints inside that car mask are successfully tracked.

We find keypoints inside each IMO candidate patch by choosing both corner and edgel points, as proposed in [18]. This approach is an extension of the good-feature-to-track (GFTT) detector initially proposed by Shi and Tomasi [20]. As the matching and tracking processes used in the present paper are guided by the epipolar geometry, patches which have a local structure with only one dominant orientation (e.g. lines and straight edges) can be matched as long as the dominant orientation is sufficiently well inclined relative to the epipolar line under consideration.

Each keypoint is represented by a 15×15 patch centered on the keypoint and we use a 2D Gaussian filter with the same size of the patch as a masking weight $\mathcal{W}$ for each patch. In order to track the keypoint on subsequent frames, we employ an iterative matching which minimizes the photometric error between the patch correspondences.

When a new keypoint is tried to be matched for the first time, we initialize the matching with motion prior information as proposed in [4] followed by a Lucas-Kanade like optimization to find the final match. The matching results of a keypoint on consecutive frames form a keypoint trajectory. A keypoint is finally accepted and used for pose estimation

when it has been tracked on at least three consecutive frames which reflects its 3D-3D consistency.

B. CNN-based IMO candidate patches

IMO candidate patches are obtained from a CNN-based instance-level vehicle segmentation. We employ the 'deep contours' approach, proposed in [22]. The CNN has been trained to label cars using the Cityscapes dataset [3]. for training. The output of the CNN are 5 layers: one for the car labels and four layers for the left, right, top, and bottom edges of cars. Initially, the car label image marks every pixel belonging to a car; therefore overlapping cars will be part of the same patch. By removing all edges from the car label image, the overlapping car instances are separated from one another. To successfully separate the instances, the contour of a car created by the edges must be closed and the edges are dilated to prevent small gaps. The car instance image is then created by labeling each independent patch as car instance.

IV. IMO DETECTION

The method presented here is based on three main steps: identifying any IMO candidates in the field of view, determining whether an object is moving or static, and finally tracking the movement of each object. As mentioned earlier, we combine cues from the appearance of the object itself with cues about object movement obtained from visual odometry. In the current presentation, we constrain ourselves to the detection and tracking of vehicles, although of course many classes of IMOs such as pedestrians, cyclists, animals, etc. could be detected and processed similarly.

We employ an IMO detection scheme which combines a *three-frame hypothesis test* (see section IV-A) and a *backward epipolar matching test* (see section IV-B) that decide for each keypoint inside the IMO candidate patch whether it is compliant to a static point in the environment through which the ego-car (and thus the camera) is moving. The results from the backward epipolar matching test are obtained instantly at the most recent frame while the results from the three-frame hypothesis test are only available two frames later. Hence, the three-frame hypothesis test serves as a correction step whenever there is a difference between the results from the two tests.

A. Three-frame hypothesis test

In some situations, particularly if objects are far away or if they are only moving slowly with respect to the static environment, it is difficult to determine from the observation of only two frames that they move relative to the static scene. Thus, we perform an IMO test using three consecutive frames by checking the circular consistency as illustrated in fig. 2.

We begin the analysis of the found vehicle patches with the null hypothesis H_0 that supposes that a keypoint is static, and thus epipolar conformant. This means that the patch–patch correspondence is the same with and without applying the epipolar constraint coming from the relative pose determined earlier. The 'unconstrained' correspondence is

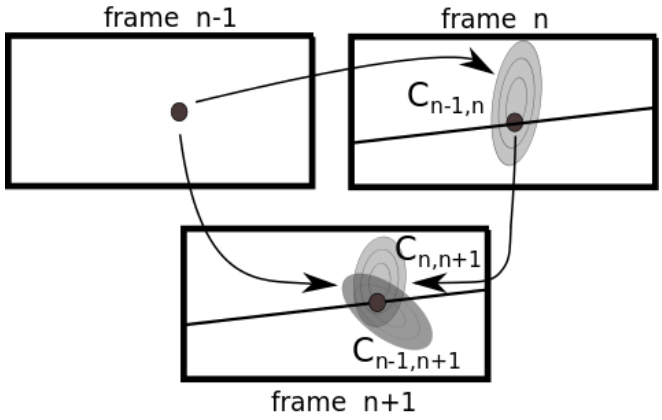

Fig. 2. Illustration of IMO detection using three-frame hypothesis test via circular consistency check. The epipolar constraints are shown by epipolar lines in frame n and $n+1$. The covariance matrix $\mathbf{C}$ is represented by an ellipse.

obtained from initializing a differential (Lucas-Kanade style) matching process with the best constrained match found on the epipolar line. Basically, the idea is that a truly epipolar match will essentially stay where it was initialized, whereas finding a *significantly* better match outside the individual area of confidence of that match indicates a non-epipolar-conformant match. We test this null hypothesis using a test statistic s which we compute from the observations, i.e. the Mahalanobis distance of the results of constrained and unconstrained matching, using the covariance matrix of the unconstrained matching. The spatial uncertainty of a differential (LK-style) match is depending on the distribution of the gradients in the regarded patch, and on the noise level of the images, assuming additive noise. The matching model predicts which distribution the test statistic s should have if the hypothesis H_0 is true.

1) Testing single trajectories for belonging to an IMO: Let us regard testing a single trajectory obtained from using the propagation-based tracking method [5]. We will consider a joint test for multiple trajectories associated with the same IMO candidate patch later in section IV-A.6.

We have to discriminate two cases:

1) the trajectory is epipolar consistent by construction, e.g. since it is the result of epipolar matching and tracking [5].
2) the trajectory is not constrained to the epipolar geometry induced by the sequence of poses associated with the trajectory.

The second variant is much harder to implement since we would need general long displacement matching to track general IMO motion. Hence, we focus on the hypothesis of static keypoints and consider a keypoint as belonging to an IMO whenever the hypothesis if rejected.

2) Finding the best matches with and without epipolar constraint: A basic task appearing over and over again in the presented method is to match a keypoint $\boldsymbol{m}_{n-1}$ from frame $n-1$ to frame n. This matching is performed by a

photometric comparison of the contents of the corresponding keypoint patches at times $n-1$ and n, considering the weight mask $\mathcal{W}$ (see section III-A). The patch at time $n-1$ is kept fixed with the center pixel at $\boldsymbol{m}_{n-1}$ (reference patch), whereas the position of the other one is varied in order to find an optimum match. This matching can be performed with or without consideration of the given epipolar relation induced by the relative pose between frame $n-1 \rightarrow n$.

The classical Lucas-Kanade (LK) matching optimization problem can be written as a sequence of local quadratic optimization steps. In each step, the weighted sum of squared differences (WSSD) Q between two patches (from two different frames), centred respectively at positions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, is minimized by small shifts $\boldsymbol{\Delta}_v$ from an initial position $\boldsymbol{x}_2 \mapsto \boldsymbol{x}_2 + \boldsymbol{\Delta}_v$.

$$Q(\boldsymbol{\Delta}_v) = \boldsymbol{\Delta}_v^T \cdot \mathbf{A} \cdot \boldsymbol{\Delta}_v + \boldsymbol{b}^T \cdot \boldsymbol{\Delta}_v + c \tag{1}$$

$\mathbf{A}$ is a symmetric 2×2 matrix built from the outer product of the gradient vectors around the moving patch, $\boldsymbol{b}$ is a 2×1 vector from the multiplication of gradient vectors and photometric error, and c is the sum of squared pixel-wise photometric errors inside the patch. The final estimate of the displacement vector $\boldsymbol{v}$ after N iterations is given by

$$\boldsymbol{v} = \boldsymbol{x}_2(0) - \boldsymbol{x}_1 + \sum_{i=1}^{N} \boldsymbol{\Delta}_v(i), \tag{2}$$

For tracking under an epipolar constraint, we add the epipolar constraint

$$\boldsymbol{x}_{2,h}^T \cdot \mathbf{F} \cdot \boldsymbol{x}_{1,h} = 0 \tag{3}$$

to the quadratic optimization function Eq. 1, where $\boldsymbol{x}_{i,h} = [\boldsymbol{x}_i^T, 1]^T$. This problem is solved with a Lagrange multiplier α, thus yielding Eq. 5 which leads to a closed form solution of $\boldsymbol{\Delta}_v$.

$$\begin{pmatrix} \mathbf{A} & \mathbf{F}' \cdot \boldsymbol{x}_{1,h} \\ (\mathbf{F}' \cdot \boldsymbol{x}_{1,h})^T & 0 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\Delta}_v \\ \alpha \end{pmatrix} \tag{4}$$

$$= \begin{pmatrix} -\boldsymbol{b} \\ -(\boldsymbol{x}_{1,h}^T \cdot \mathbf{F} \cdot \boldsymbol{x}_{1,h}) \end{pmatrix} \tag{5}$$

where $\mathbf{F}'$ corresponds to the first two rows of the $\mathbf{F}$ matrix.

Let us denote the best matches with and without epipolar constraint as $\boldsymbol{m}_n$ and $\tilde{\boldsymbol{m}}_n$, respectively. In our scheme, we first apply epipolar-constraint tracking to find $\boldsymbol{m}_n$. Afterwards, we start from $\boldsymbol{m}_n$ to search the best match $\hat{\boldsymbol{m}}_n$ without epipolar constraint.

3) Obtaining the spatial covariance matrix of a keypoint match: The nice thing about photometric matching is that the determination of a covariance matrix expressing the spatial uncertainty of the match in the final position of the sliding patch can be immediately obtained from the linear equation system (1).

Let $\mathbf{C}_{n-1,n}$ be the covariance matrix expressing the uncertainty level of choosing $\hat{\boldsymbol{m}}_n$ as the final match. The covariance of the final match $\mathbf{C}_{n-1,n}$ is given by matrix $\mathbf{A}$ from the last iteration of the LK matching optimization problem (see equation 1).

As a result, we have *two* entities expressing the (un)certainty of the given match:

- the final matching residual Q
- the curvature of the residual function, expressed by the covariance matrix

This information can be used in different ways:

- the final matching residual Q should be below a predetermined threshold
- the uncertainty of the position, as expressed by the covariance matrix, should be low enough

As we perform both epipolar constrained and unconstrained searches to find the best match, the two resulting positions are then compared:

- Both matching residuals (mean squared error) from constrained and unconstrained searches must be below threshold τ_{msd}. In addition, if the matching residual of the constrained position is significantly higher than the one of the pseudo-unconstrained[1] search, this indicates that the true match probably does not lie on the epipolar line.
- If the unconstrained match position is farther away from the constrained match position than suggested by the spatial covariance matrix, this is considered as an indicator that the true match is not on the epipolar line.

4) Computing the test statistic: Let us define a test statistic s which reflects how well the keypoint match in frame n follows the epipolar constraint. Let $\boldsymbol{w} = (w_x, w_y)$ be the distance vector between $\tilde{\boldsymbol{m}}_n$ and $\boldsymbol{m}_n$.

$$\boldsymbol{w}_n = \tilde{\boldsymbol{m}}_n - \boldsymbol{m}_n \tag{6}$$

The metric s we use for representing the difference between $\tilde{\boldsymbol{m}}_n$ and $\boldsymbol{m}_n$ is given by the squared Mahalanobis distance,

$$s = \boldsymbol{w}_n^T \cdot \mathbf{C}_{n-1,n}^{-1} \cdot \boldsymbol{w}_n \tag{7}$$

As the inverse covariance matrix in this 2nd order form decorrelates and 'whitens' the elements of the random vector, the test statistic s is known to be χ^2-distributed, and we can define a threshold for s above which the difference between $\tilde{\boldsymbol{m}}_n$ and $\boldsymbol{m}_n$ is considered to be too high. The value of the test statistic s can alternatively be transformed into a 'p-level' $p_{n-1,n}$ using the underlying χ^2 distribution.

5) Hypothesis check: We check the circular consistency between three consecutive frames: $n-1$, n, and $n+1$, as illustrated in figure 2. Hence, the steps to find the best match and to obtain the corresponding spatial covariance matrix are carried out not only between frames $\{n-1, n\}$, but also $\{n, n+1\}$ and $\{n-1, n+1\}$.

Since we assume all keypoints are IMO candidates in the beginning, a keypoint is only considered as a static object if the p-level exceeds a threshold τ_{ps} after three consecutive frames. With these p-levels, we can now decide whether the

[1]We denote this optimization as *pseudo-unconstrained* since it is not a true wide area search, but only a search in the vicinity of the constrained match position.

IMO candidate points are actually static, that is when they pass all of the following three checks:

$$p_{n-1,n} \underset{\text{IMO}}{\overset{\text{static}}{\gtrless}} \tau_{ps} \tag{8}$$

$$p_{n,n+1} \underset{\text{IMO}}{\overset{\text{static}}{\gtrless}} \tau_{ps} \tag{9}$$

$$p_{n-1,n+1} \underset{\text{IMO}}{\overset{\text{static}}{\gtrless}} \tau_{ps} \tag{10}$$

6) Testing sets of trajectories in the same IMO candidate patch: Let n_0 and n_1 be the number of keypoints on an IMO candidate patch which support the null hypothesis H_0 (= non-IMO) or the hypothesis $H_1 \rightarrow IMO$. The IMO candidate patch is classified as static if $n_0/(n_0+n_1) > \tau_h$ and as IMO if $n_1/(n_0+n_1) > \tau_h$.

B. Backward epipolar matching check

On top of the three frame hypothesis test from section IV-A, we also employ a backwards epipolar matching check to identify IMOs. There are several benefits in doing the epipolar matching check in backwards mode:

- We already have the IMO/static label of cars in the previous frame
- There is no delay in identifying IMOs since no future information is required.
- In a typical forward ego-motion, most keypoints in frame $n+1$ are also visible in frame n, but not the other way around due to objects leaving the camera view from frame n to $n+1$. Hence the chance of finding matches is higher when we match keypoints from frame $n+1$ to frame n (backward mode).

Let us assume that we have obtained all information up to frame n. When frame $n+1$ comes with its instance-level IMO segmentation, then we analyze each IMO candidate patch at this frame $n+1$.

First, we generate keypoints on the new found car patches in frame $n+1$, using the keypoint definition described in section III-A. Second, we try to find the matches of these newly generated keypoints in frame n under consideration of the epipolar constraint given by the relative pose computed by PbT, as illustrated by figure 3.

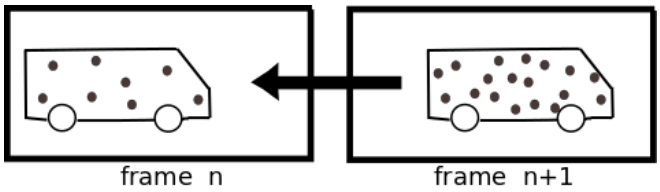

Fig. 3. Backward epipolar matching from frame $n+1$ to frame n. The ratio of successfully found matches in frame n versus the number of initial keypoints in frame $n+1$ is computed.

Let $k_{i,init}$ be the number of keypoints generated on an IMO candidate patch i in frame $n+1$ and let $k_{i,matched}$ be the number of successfully matched keypoints among the keypoints in patch i. We define r_i as the ratio of successful epipolar matching of patch i

$$r_i = \frac{k_{i,matched}}{k_{i,init}} \tag{11}$$

A high ratio of the epipolar matching ratio r_i reflects the epipolar conformity of the car patch i, and thus indicates a static object. On the other hand, a low value of r_i gives a strong hint that the IMO candidate patch does not follow the epipolar structure. The decision is made by considering a patch as backwards epipolar consistent (non-IMO) if $r_i > \tau_i$ with an empirically chosen threshold τ_i.

V. Propagation of label information

We determine the number of patches $P(n)$ in image n and the set of labels $\ell_i(n)$. Each car patch is represented by a feature vector $\boldsymbol{f}_i(n)$ consisting of its center of mass $\boldsymbol{c}_i(n)$, its size (pixel count) $s_i(n)$, its mean gray value $m_i(n)$ and its gray value standard deviation $\sigma_i(n)$. This information is collected for all patches by a single pass over the new label image generated by the CNN.

As the motion of those IMOs belonging to crossing traffic often exhibit very long 2D displacements in the image, they cannot be expected to be trackable by a differential LK-style motion tracker like the one which is applied in propagation based tracking (PbT). The association between 'old' patches in image n and new patches in image $n+1$ is performed by testing for each new patch the compatibility with each old patch.

The association is performed in a looping greedy manner (forward and backward) whenever car patches are observed in both image n and image $n+1$, subjecting each potential association between a patch j in image n represented by $\boldsymbol{f}_j(n)$, and a patch k in image $n+1$ represented by $\boldsymbol{f}_k(n+1)$. An association match between two car patches is accepted only when the pair of car patches reciprocally chooses each other as the best match.

We predict the position of the car patch j at frame $n+1$, represented by the predicted center of mass $\hat{\boldsymbol{c}}_j(n+1)$, using the information of past positions at up to three previous frames as presented in equation 13. We use the assumption of constant 2D acceleration to predict the pixel coordinate of the center of mass for each car patch.

$$\begin{aligned}&(\hat{\boldsymbol{c}}_j(n+1) - \boldsymbol{c}_j(n)) - (\boldsymbol{c}_j(n) - \boldsymbol{c}_j(n-1)) \\ &= (\boldsymbol{c}_j(n) - \boldsymbol{c}_j(n-1)) - (\boldsymbol{c}_j(n-1) - \boldsymbol{c}_j(n-2))\end{aligned} \tag{12}$$

which leads to

$$\hat{\boldsymbol{c}}_j(n+1) = 3\boldsymbol{c}_j(n) - 3\boldsymbol{c}_j(n-1) + \boldsymbol{c}_j(n-2) \tag{13}$$

A set of minimum requirements to accept patch matches between frame n and frame $n+1$ is defined using empirical thresholds τ_s, τ_m, τ_σ, and τ_c,

$$|s_j(n) - s_k(n+1)| < \tau_s \tag{14}$$

$$|m_j(n) - m_k(n+1)| < \tau_m \tag{15}$$

$$|\sigma_j(n) - \sigma_k(n+1)| < \tau_\sigma \tag{16}$$

$$\|\hat{\boldsymbol{c}}_j(n+1) - \boldsymbol{c}_k(n+1)\| < \tau_c \tag{17}$$

and subsequently confirming each 'surviving' association. Thus, there is no guarantee that this is the best association, but the procedure is fast. Obviously, some patches may be left unassociated, both in image n as well as in image $n+1$.

VI. Experiments

We tested our method on the KITTI dataset [9]. We use the following values used in the experiments: $\tau_{ps} = 0.95$, $\tau_{msd} = 0.20$, $\tau_h = 80\%$, $\tau_i = 25\%$, $\tau_s = 4000$, $\tau_m = 0.5$, $\tau_\sigma = 5.0$, $\tau_c = 200$.

First, we tested our method on the KITTI MoSeg dataset [21] which contains of 349 test images with annotated motion labels. Then we compare our method with MODNet [21] in terms of the precision of identifying static and moving objects.

Second, we tested our method on the KITTI odometry dataset, which is much bigger than the KITTI MoSeg dataset. Unfortunately, KITTI does not provide for this dataset neither instance labels for each vehicle nor patches, which classifies vehicles into static or moving objects. Thus, we have created our own dataset to evaluate our approach. This dataset will be available to the public after the work proposed in this paper has been accepted for publication. It is important to note that due to the nature of a monocular setup, IMOs moving parallel to the ego-car are out of the scope of our work (see section VI-C for further explanation.). Hence, we exclude these epipolar-conformant IMOs from our experiment.

A. KITTI odometry dataset with motion labels

For our new dataset, we used the 11 training sequences from the KITTI visual odometry dataset, which consists of 23201 images. The proposed CNN from van den Brand et al. [22] was utilized to generate candidate labels for the vehicle instances. In the current state of the dataset, we have limited the detected objects to vehicles only. This can be further extended to other objects, like pedestrians or bicycle in future work.

Given these segmented candidate labels, we manually assign to each candidate patch in all images one of the following class labels: 0 - background (non-vehicles), 1 - independently moving vehicle, 2 - static (non-moving) vehicle, 3 - far away vehicles (median distance greater than 50m) and 4 - undetermined. We labeled a candidate as undetermined, if the patch does not show a vehicle or if the patch is stretched over more than one vehicles, which do not fall into same category, like static or IMO. Some examples of this dataset are shown in fig 4. We will make this dataset public available after the reviewing process.

B. Accuracy analysis

Besides computing the accuracy of the results, we measure the decisiveness of the proposed method to give definite output (`IMO`/`static`) as compared to `undetermined`. We define the decisiveness level D as

$$D = \frac{n_{IMO} + n_{static}}{n_{IMO} + n_{static} + n_{undetermined}} \tag{18}$$

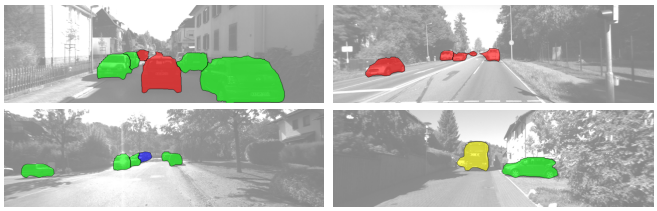

Fig. 4. Examples from our new IMO candidates dataset. The colored overlay encoding is as follows: red ↔ IMO (class 0), green ↔ static (class 1), blue ↔ too far away (class 2) and yellow ↔ undetermined (class 3).

TABLE I
Accuracy of IMO detection on the KITTI MoSeg dataset.

Method	P static	P moving	P average
MODNet [21]	0.65	0.67	0.66
Ours	**0.76**	**0.71**	**0.73**

where n_{IMO}, n_{static}, and $n_{undetermined}$ are respectively the number of outputs as `IMO`, `static`, and `undetermined`.

The accuracy of the IMO classification is expressed by precision P, recall R, and accuracy A. They are computed based on four metric occurrences: *true positive* (t_p), *false positive* (f_p), *true negative* (t_n) and *false negative* (f_n).

$$P = \frac{t_p}{t_p + f_p}\ ; \quad R = \frac{t_p}{t_p + f_n}\ ; \quad A = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{19}$$

*1) **Accuracy on the KITTI MoSeg dataset**:* Table I presents the precision of the IMO detection using our method and using MODNet [21]. The precision of our method is better on both identifying static cars and moving cars. The average precision of our method is 0.73 as compared to 0.66 of MODNet. Figure 5 shows the exemplary results of the IMO detection using our method and using MODNet.

Fig. 5. Exemplary results of the car classification into static and IMO labels on KITTI MoSeg dataset: using our method (**top**) and using MODNet (**bottom**). Red color represents IMO, green color represents static car, and yellow color represents undetermined. The comparison shows that our method correctly identifies a static parked car while MODNet wrongly classifies it as an IMO.

It is important to note that the KITTI MoSeg test dataset consists of not only the non-epipolar-conformant IMOs (e.g. crossing-IMOs) but also the epipolar-conformant IMOs (i.e.

TABLE II

ACCURACY OF IMO DETECTION ON THE KITTI ODOMETRY DATASET.

Metric	static	moving	average
Decisiveness (D)	0.90	0.90	0.90
Precision (P)	0.96	0.73	0.84
Recall (R)	0.84	0.92	0.88
Accuracy (A)		0.87	

parallel-moving IMOs). In order to compare fairly, we run the simulation fully on the test dataset thus also covering the case of parallel-moving IMOs, which are actually out of the scope of this paper. Nevertheless, the overall precision of our method is still better than MODNet.

*2) **Accuracy on the KITTI odometry dataset**:* The results of the proposed IMO detection on the KITTI odometry dataset are presented in table II. The decisiveness level has an average value of 90% which implies that `undetermined` outputs only happen at about 10% of the time. Most of the `undetermined` outputs occur when the cars just appear or are about to leave the image frame.

The average precision is 84%, with the precision for the static object detection is very high at 0.96 and the precision for the IMO detection is 0.73. The recall rate has an average of 88%. We obtain an overall accuracy of 87%.

Sequence 01 has one of the most challenging scenes as cars are gradually entering a highway, hence making a slow transition from a *non-epipolar-conformant* to a *fully-epipolar-conformant* state. This situation is illustrated in figure 6 (top image) where two cars are observed, one correctly classified as an IMO while another one wrongly perceived as a static car (in a very wrong distance) because its movement is almost parallel to the ego-car. This limitation is discussed further in section VI-C.

Figure 6 (middle and bottom images) shows the IMO detection for the KITTI odometry dataset sequence 03 and sequence 07. Both IMOs and static cars are correctly identified. The optical flow vectors used for pose estimation can also be observed on static cars but obviously not on IMOs.

C. Limitations of the current approach

Due to the monocular setup, our proposed method is unable to identify an IMO when the IMO moves in the same direction as the camera motion camera. One of the examples is shown in figure 7 where a moving car is correctly identified as an IMO at one frame, but wrongly classified as a static object when it moves in the same direction with the camera. This is because any IMO which moves in the same direction as the camera will automatically be conformant to the epipolar structure, hence an epipolar constraint check cannot detect the independent movement. This problem should be possible to solve when stereo setup is used instead of monocular one, or when coarse distance estimates are available for the regarded car patches, e.g. from [8], [14] The patches of cars which move in an epipolar conformant way imply a certain distance. A car in front of the ego-car moving with the same speed and direction is epipolar-conformant, but a triangulation would put it in infinite distance as its motion vectors are close to 0. The discrepancy between the appearance-based distance (knowledge of car size in 3D) and the motion-based distance estimate could probably be detected even in a monocular setup.

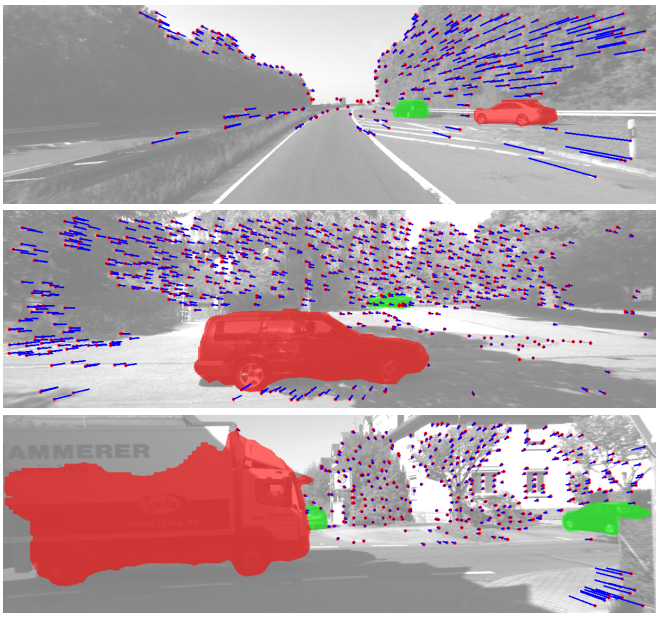

Fig. 6. Exemplary results of the car classification into static and IMO labels on the KITTI odometry dataset sequence 01 (top), sequence 03 (middle), and sequence 07 (bottom). Red color represents IMO while green color represents static car. Optical flow is also shown in blue lines.

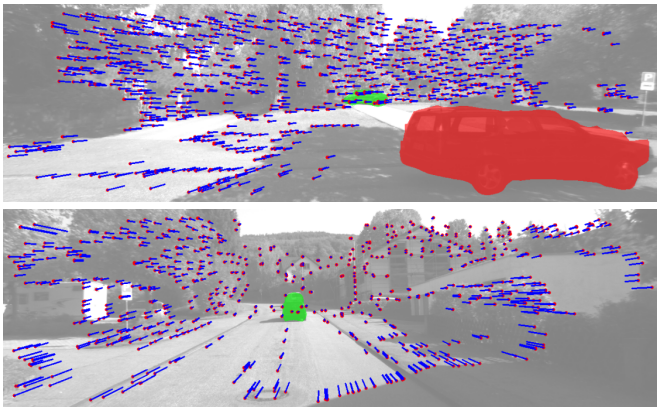

Fig. 7. A scene taken from the KITTI odometry dataset sequence 03. An IMO is correctly identified in the top image. However the same car is wrongly classified as a static object when it moves at the same direction with the camera, as shown in the bottom image.

VII. CONCLUSION

This paper has presented an IMO detection method using a monocular camera. The proposed method employs CNN-based classifiers to provide IMO candidates and subsequently checks them against the multi-frame epipolar consistency to identify them as IMOs. The motion label (IMO/static) is propagated over time by establishing patch label association

between two consecutive frames based on the cue combination of movement and appearance. Experiments on the KITTI dataset show the performance of the proposed method.

REFERENCES

[1] M. Bai, W. Luo, K. Kundu, and R. Urtasun, "Exploiting semantic information and deep matching for optical flow," in *European Conference on Computer Vision*. Springer, 2016, pp. 154–170.

[2] T. Brox, A. Bruhn, and J. Weickert, "Variational motion segmentation with level sets," in *Proc. ECCV 2006*. Springer, 2006, pp. 471–483.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[4] N. Fanani, M. Barnada, and R. Mester, "Motion priors estimation for robust matching initialization in automotive applications," in *International Symposium on Visual Computing (ISVC)*, 2015, pp. 115–126.

[5] N. Fanani, M. Ochs, H. Bradler, and R. Mester, "Keypoint trajectory estimation using propagation based tracking," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 933–939.

[6] N. Fanani, A. Stuerck, M. Barnada, and R. Mester, "Multimodal scale estimation for monocular visual odometry," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2017.

[7] N. Fanani, A. Stürck, M. Ochs, H. Bradler, and R. Mester, "Predictive monocular odometry (pmo): What is possible without ransac and multiframe bundle adjustment?" *Image and Vision Computing*, 2017.

[8] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, 2016, pp. 740–756.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[10] A. Jazayeri, H. Cai, J. Y. Zheng, and M. Tuceryan, "Vehicle detection and tracking in car video based on motion model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 583–595, 2011.

[11] B. Jung and G. S. Sukhatme, "Detecting moving objects using a single camera on a mobile robot in an outdoor environment," in *International Conference on Intelligent Autonomous Systems*, 2004, pp. 980–987.

[12] A. Kundu, C. V. Jawahar, and K. M. Krishna, "Realtime moving object detection from a freely moving monocular camera," in *2010 IEEE International Conference on Robotics and Biomimetics*, 2010, pp. 1635–1640.

[13] P. Lenz, J. Ziegler, A. Geiger, and M. Roser, "Sparse scene flow segmentation for moving object detection in urban environments," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 926–932.

[14] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1119–1127.

[15] F. J. López-Rubio and E. López-Rubio, "Foreground detection for moving cameras with stochastic approximation," *Pattern Recognition Letters*, vol. 68, pp. 161 – 168, 2015.

[16] G. L. Oliveira, N. Radwan, W. Burgard, and T. Brox, "Topometric localization with deep learning," 2017, arXiv preprint. [Online]. Available: https://arxiv.org/abs/1706.08775

[17] A. Ošep, W. Mehner, M. Mathias, and B. Leibe, "Combined image- and world-space tracking in traffic scenes," in *ICRA*, 2017.

[18] T. Piccini, M. Persson, K. Nordberg, M. Felsberg, and R. Mester, "Good edgels to track: Beating the aperture problem with epipolar geometry," in *ECCV Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving (CVRSUAD)*, Zurich, 2014, pp. 652–664.

[19] A. Ramirez, E. Ohn-Bar, and M. M. Trivedi, "Go with the flow: Improving multi-view vehicle detection with motion cues," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 4140–4145.

[20] J. Shi and C. Tomasi, "Good features to track," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[21] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "Modnet: Moving object detection network with motion and appearance for autonomous driving," *arXiv preprint arXiv:1709.04821*, 2017.

[22] J. van den Brand, M. Ochs, and R. Mester, "Instance-level segmentation of vehicles by deep contours," in *Asian Conference on Computer Vision 2016 – Workshops*. Springer, 2016, pp. 477–492.

[23] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *The International Journal of Robotics Research*, vol. 26, no. 9, pp. 889–916, 2007.

[24] A. Wedel, A. Meißner, C. Rabe, U. Franke, and D. Cremers, "Detection and segmentation of independently moving objects from dense scene flow," in *Energy minimization methods in computer vision and pattern recognition*. Springer, 2009, pp. 14–27.

[25] J. Wulff, L. Sevilla-Lara, and M. J. Black, "Optical flow in mostly rigid scenes," *arXiv preprint arXiv:1705.01352*, 2017.

[26] K. Yamaguchi, T. Kato, and Y. Ninomiya, "Vehicle ego-motion estimation and moving object detection using a monocular camera," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, 2006, pp. 610–613.

[27] Q. Yuan, A. Thangali, V. Ablavsky, and S. Sclaroff, "Learning a family of detectors via multiplicative kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 514–530, 2011.

[28] K. Yun, J. Lim, S. Yun, S. W. Kim, and J. Y. Choi, "Attention-inspired moving object detection in monocular dashcam videos," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2706–2711.