Real-Time Mobile Object Detection Using Stereo

M. Derome, A. Plyer, M. Sanfourche, G. Le Besnerais

▶ To cite this version:

M. Derome, A. Plyer, M. Sanfourche, G. Le Besnerais. Real-Time Mobile Object Detection Using Stereo. ICARCV 2014, Dec 2014, MARINA BAY SANDS, Singapore. hal-01098035

HAL Id: hal-01098035

https://hal.science/hal-01098035

Submitted on 22 Dec 2014

Real-Time Mobile Object Detection Using Stereo

Maxime Derome, Aurélien Plyer, Martial Sanfourche and Guy Le Besnerais
ONERA - The French Aerospace Lab, Palaiseau, France
Emails: firstname.name@onera.fr

***Abstract*—This paper considers passive vision for robotics and focuses on devising a real-time process for moving object detection using a stereo rig. As several previous works, our method relies on the use of dense stereo and of optical flow. Observing that the main computational load of existing methods is related to the estimation of the optical flow, we propose to use a fast algorithm based on Lucas-Kanade's paradigm. We derive a new uncertainty model which explicitly takes into account all errors originating from each estimation step of the process. In contrast with most previous works, we describe a rigorous expansion of the error related to vision based ego-motion estimation. Finally, we present a comparative study of performance on the KITTI dataset, which demonstrates the effectiveness of the proposed approach.**

I. Introduction

A. Context and Problem Statement

Understanding complex environment in presence of dynamic objects is crucial for autonomous robotics. Such situation awareness could benefit to Advanced Driver Assistance Systems (ADAS) as well as Search And Rescue (SAR) missions. Vision sensors are particularly suited for this task as they are cheap, lightweight, and can provide, through dedicated fast algorithms, both scene perception and ego-motion estimation. Besides, using a stereo rig enables 3d reconstruction of the scene at each frames, which can be used for mobile objects detection. In the design of an embedded mobile object detection process, three main constraints have to be accounted for: real-time processing, high reactivity, and precise management of measurement and estimation errors to assess the reliability of the decisions. We address these three constraints in our work. We propose a new detection system which uses very fast algorithms for the low-level operations (stereo matching and optical flow (OF) estimation). The decision is based on the processing of two consecutive stereo images only. This features allows to maximize the reactivity of the system and also eases the modelling of error propagation. This last issue is rigourously addressed here thanks to a first order model based on the Implicit Function Theorem.

B. Related Works

Different approaches have been proposed to address the understanding of dynamic scenes from stereo-vision data. Algorithms based on sparse sets of feature points have been used in temporally integrated framework [1], or in graphical approaches to segment stereo-images according to 3D motion consistency [2]. However, because of their sparseness, these methods provide limited coverage of the scene.

A great deal of work has also been done using dense stereo-vision algorithms. Dense stereo provides the instantaneous 3d structure of the scene. It can be coupled with visual odometry that computes the camera rotation and translation [R,T] between two frames. From these informations, the scene geometry in a new camera frame can be predicted under the hypothesis of a static world. The discrepancies between the new observation and this prediction reveal the independent motions and are cues for the detection of moving objects. Detection then stems from thresholding some residual field.

Depending on the residual value which is used, or equivalently on the quantity which is predicted, two approaches can be distinguished. One can either synthesize a predicted image using previous image intensity (an approach which will be denoted by *image prediction methods* in the following) or directly predict geometrical quantities such as 3d points coordinates, optical flow (OF) and disparity (*direct methods*).

Direct methods have been applied with different residual values in the literature. For instance, [3],[4] and [5] consider the differences between observed and predicted 3d points — a vector field which is called *Scene Flow*. In [4] the authors reduce Scene Flow noise using a Kalman filter for each pixel, with a state vector made of 3d position and velocity. Unfortunately, such temporal filtering reduce the system reactivity, since multiple frames are required for the Kalman filters to converge. Furthermore, this model may encounter difficulties with non-uniformly moving objects. In [5] the authors include disparity changes estimation in a variational minimization framework that also computes OF. Variational minimization methods are well known in the field of OF estimation, as they provide smooth and accurate solutions. But they require several solver iterations to converge to a good solution. Hence, in practice they are not real-time for an embedded system using high resolution stereo images. Other direct methods consider residual values expressed in the image space: OF and disparity in [6], OF alone in [7].

Alternatively, image prediction methods have been investigated. Dense comparison between observed and predicted image can be done by computing OF [8], or by evaluation of some similarity index within a small neighbourhood of the current pixel: [9] uses Sum of Absolute Differences (SAD) while a Zero-mean version (ZSAD) appears in [10].

Except for [9] and [10], all previous approaches rely on the computation of some 2D or 3D residual field (which we denote by M in the following), and the thresholding of a pixelwise motion likelihood written as a weighted norm of M:

$$\xi(M) = \sqrt{M^T \Sigma_M^{-1} M}. \tag{1}$$

If the covariance matrix Σ_M models accurately the uncertainty about the residual field M, criterion (1) is called a Mahalanobis distance, and leads to optimal decision. The main issue is that the residual M depends on several variables (disparity fields, OFs, [R,T]) which stem from complex estimation processes. Estimating the resulting uncertainty on M is very difficult and requires some simplification. First attempts [6], [9] simply considered $\Sigma_M = Id$, which leads to poor results. A formulation of Σ_M depending on disparity and optical flow is proposed in [5], based on residual minimization energy. However, the authors disregard the rotation R that is assumed equal to the identity matrix Id_3, and model only camera translation uncertainty, which is a rather crude hypothesis, even in the context of urban navigation. A Bayesian formulation of OF error covariance Σ_{OF} is used in [7] to model Σ_M. The authors also consider [R,T] uncertainty, but assume independent rotational and translational errors without explicit mention of the ego-motion estimation process. In [8], the approach relies on first order expansion of the image displacement field with respect to the angular and translational velocity vectors $[\Omega, V]$. Both Σ_Ω and Σ_T are chosen as constant diagonal matrices which are evaluated a priori using a synthetic video. The first order expansion puts limits on the dynamic of the vehicle or on the framerate. Besides, the authors of [8] do not fully account for the $[\Omega, V]$ uncertainty but use 3σ bounds on the errors in the subsequent expressions. Finally, reference [3] suggests an heuristic to derive an approximate covariance matrix from the least-squares criterion (1) but does not explicitly include the error budget for disparity and OF estimation. As a conclusion, to our knowledge, there is no previous paper presenting a complete analysis of errors, especially regarding the uncertainty over ego-motion parameters [R,T].

C. Contribution and Outline of the Paper

Our contribution is threefold. We present a moving object detection system based on eFOLKI, a newly proposed fast OF method [11] which allows real-time processing of large images. We present a complete analytical formulation of the uncertainty model of both direct and image prediction methods. In particular, we account for the fact that [R,T] parameters derive from the optimization of an ego-motion criterion where image measurements (point matches) are also involved. This indirect relationship is rigourously handled thanks to the Implicit Function Theorem. Finally, we conduct a comparison of various methods and error models through an evaluation protocol based on challenging KITTI datasets [12]. This experimental study demonstrates the efficiency of the proposed image prediction method and the benefit of the presented error model.

The paper is organized as follows. Section II describes the detection process and discuss low level operations and choice of the residual value. The uncertainty model is detailed in Sec. III. The evaluation protocol and experimental results are presented in Sec. IV, then we conclude.

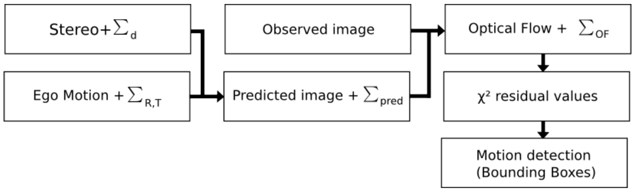
Stereo+Σ_d
Ego Motion + $\Sigma_{R,T}$
Observed image
Predicted image + Σ_{pred}
Optical Flow + Σ_{OF}
χ² residual values
Motion detection
(Bounding Boxes)

Fig. 1. Motion detection pipeline (see explanations in the text).

II. System Description

A. Overview

Fig. 1 presents a global overview of the moving object detection pipeline. Independently moving objects are detected by analysing two consecutive stereo images. Dense stereo is computed for each stereo acquisition time and dense optical flow is computed between successive times: these costly low-level operations are discussed in the following. We use the ”efficient Visual Odometry” (eVO) of [13] which can run at 20Hz on a single core of an embedded CPU: some details on this estimation process will be reviewed in Sec. III. With these informations we compute a residual field M that is null under rigid scene assumption. While the generic processing scheme of Fig. 1 pertains also for a direct approach, we adopt here a image prediction method based on the residual OF. Given the error covariance matrix Σ_M derived according to some model of uncertainty, see Sec. III, the Mahalanobis distance $\xi(M)$ of Eq. (1) is computed and thresholded. Bounding boxes are then fitted to the detected areas. In this section, we focus on low-level operations and choice of the residual value.

B. Low-level Operations

Several papers present algorithmic choices to reduce stereo computation time. A major breakthrough here is the publication of Semi Global Block Matching (SGBM) [14], a dense stereo algorithm that can be implemented on FPGA [15] and run at 25Hz on images of 740x480 pixels for a disparity range of 128. However, for larger images and wider disparity range, the real-time capability of SGBM can be questioned. Alternatively, one may consider a simple Block Matching (BM) algorithm that exhaustively searches stereo matches along the epipolar line. BM runs in real-time without needing a FPGA. The choice between SGBM and BM is discussed in [8], and their relative performances evaluated. The inconvenient of BM is that not only the disparity map is less accurate, but it is also often unavailable on large regions. This calls into question the benefit of dense methods, which is maximal coverage of the scene. Hence, we do not consider BM here.

Perhaps surprisingly, in previous works there are few discussions about the choice of the optical flow estimation algorithm. To our knowledge, all references use variational methods based on the framework originally presented by Horn and Schunck [16]. For instance, *Combined Local-Global Method* [17] is used in [7], while TV-L1 approaches close to the one presented in [18] are considered in [4] and [5]. However, these algorithms are not only computationally demanding, but also their robustness on real images is questionable. Indeed, TV-L1 approach requires expensive image pre-processing to deal with intensity changes in real world images, see for instance the computation of TV-L2 residual images discussed in [5].

One can conclude that the main point which precludes real-time operation of these methods is the use of such costly variational OF methods. Here we propose to use eFOLKI [11], a very fast OF algorithm based on Lucas-Kanade (LK) approach. Compared on the same GPU hardware, the runtime of eFOLKI is between one or two order of magnitude lower than variational methods such as TV-L1 [18] and Brox et al. [19]. Actually, looking at the OF benchmark of KITTI's website, eFOLKI appears among the very few methods able of real-time operation on 2 megapixels images. LK methods are generally considered as inaccurate in the computer vision community. However, Ref. [11] shows that it compares favourably with variational methods on the training dataset of KITTI, and that it provides useful solutions for various vision problems based on OF estimation. In the same line, we will show here that it lead to results of sufficient quality for our detection purpose.

C. Choice of the residual field

Here we present in more details several direct and image prediction methods, in order to introduce our original framework and compare various approaches in the experiments.

1) Direct methods: Direct methods have been applied either to Scene Flow, or to image quantities such as residual OF and disparity. These approaches differ essentially by the way they encode the depth information. We adopt the latter which eases the error modeling step. At each time instant t, we assume that disparity d_t and optical flow (u_t, v_t) between image I_{t-1} and I_t are available. In previous frame at $t-1$, each visible point of image coordinate (x, y) and disparity d_{t-1} is triangulated

$$X_{t-1}(x, y, d_{t-1}) = -\frac{b}{d_{t-1}} \begin{pmatrix} x - x_0 \\ y - y_0 \\ f \end{pmatrix} \qquad (2)$$

with b the stereorig baseline, and f the camera focal in pixel. Given the camera motion obtained from the visual odometry, the scene can be transferred into the coordinate frame at t:

$$X_t^{pred}(x, y, d_{t-1}) = RX_{t-1}(x, y, d_{t-1}) + T, \qquad (3)$$

under a rigid scene hypothesis. Then the predicted OF writes

$$\begin{pmatrix} u_{pred} \\ v_{pred} \end{pmatrix}(x, y) = \Pi\left(X_t^{pred}(x, y, d_{t-1})\right) - \begin{pmatrix} x \\ y \end{pmatrix}, \qquad (4)$$

and the predicted disparity:

$$d_{pred}(x, y) = \frac{-bf}{\begin{pmatrix} 0 & 0 & 1 \end{pmatrix} X_t^{pred}(x, y, d_{t-1})}, \qquad (5)$$

where Π is the projection operator.
The residual is then $M = \{u_t - u_{pred}, v_t - v_{pred}, d_t - d_{pred}\}$ — some authors [7] use the OF components only.

2) Image and disparity prediction method of [8]: The predicted image in [9] and [8] is computed from previous grayscale image intensity and from the predicted 3d structure of Eq. (3):

$$I_t^{pred}(x + u_{pred}, y + v_{pred}) = I_{t-1}(x, y) \qquad (6)$$

In Ref. [9], image correlation techniques are used to check the consistency of the predicted image with respect to the observed one. In [8], the residual optical flow $(\delta u, \delta v)$ is computed between the synthetized image I_t^{pred} and the observed one I_t. Note that pixel quantization, occlusions, etc., may lead to unallocated pixels in the predicted image: intensities taken from the observed image are used to fill these empty regions. Thanks to the robustness of OF codes, these problems affect the estimation only locally. Finally, the residual OF is also used to warp d_{pred} into

$$d_{pred}^w(x, y) = d_{pred}(x + \delta u, y + \delta v), \qquad (7)$$

and the resulting residual field is: $M = \{\delta u, \delta v, d_{pred}^w - d_t\}$

3) Proposed method: Our method is close to the one of Bak et al. [8], in the sense that we also compute a predicted image and then estimate a residual flow on it. However, unlike [8], we proceed backward by interpolating image intensities at $t-1$ from the reference frame coordinate at t. We first triangulate current observed points of image coordinates (x, y) and disparity d_t

$$X_t(x, y, d_t) = -\frac{b}{d_t} \begin{pmatrix} x - x_0 \\ y - y_0 \\ f \end{pmatrix} \qquad (8)$$

Compensating ego-motion and going back in the time, we predict the coordinates U_{t-1}^{pred} of these points, in previous frame, assuming a static scene:

$$U_{t-1}^{pred} = \Pi\left(R^{-1}\left(X_t - T\right)\right) \qquad (9)$$

The predicted image I_{t-1}^{pred} is obtained by interpolating image intensity I_{t-1} at the positions U_{t-1}^{pred}. We have to deal also with occlusions and we fill empty regions with current image intensity. Finally, we compute the residual OF $(\delta u, \delta v)$ between I_{t-1} and I_{t-1}^{pred}. The main benefit of this approach is to simplify image interpolation. Indeed, in our formulation we need to interpolate irregular data from data located on a regular grid, while the approach of Bak et al. requires the opposite, ie. to interpolate regular data from irregularly arranged ones, which is more computer demanding and may lead to local artifacts. Finally, for reasons which are justified below in the experimental study (see Fig. 2), we do not consider the disparity in our residual, which then writes $M = \{\delta u, \delta v\}$

D. Detection of mobile objects

Knowing the residual field and an estimation of its covariance, we can compute the Mahalanobis distance $\xi(M)$ of Eq. (1). As done in [7], we add a geometric constraint by only considering objects that are lower than $H_{max} = 2.5$m. Since we use KITTI datasets [12] in our experiments, we assume the camera horizontally oriented and positioned at $H_{cam} = 1.67$m from the floor. Under these assumptions, valid pixels satisfy:

$$H_{cam} + b\frac{y - y0}{d} < H_{max}. \tag{10}$$

We then extract the connected components so as to form detected blobs. For each blob, we compute a mean disparity that is used to calculate its depth attribute. To ease the merging process, blobs are considered as fronto-parallel planar regions, and merged if they are close enough (e.g. closer than 30cm) in 3D. When all neighboring blobs have been merged, small blobs aggregates are suppressed. To do so, we estimate the total surface of an aggregate by summing the surface associated to each pixel belonging the aggregate's blobs. We threshold this value (e.g. by 0.16m^2) and estimate bounding boxes for the remaining blobs aggregates, as well as their depth attributes.

Figure 2.a) shows an example of such estimated bounding boxes (in red) compared to ground truth bounding boxes (in blue), manually annotated using Vatic [20]. Let us recall that detections are made at each time independently.

III. Error Model

In this section, an error model for the residual field M, is studied. We focus on the image prediction method seen in II-C3, but the following reasoning can be adapted to other methods. We look for an expression of the error covariance matrix of M:

$$\Sigma_M = \Sigma_{OF} + \Sigma_{Pred}, \tag{11}$$

where Σ_{OF} and Σ_{Pred} are respectively the optical flow and the prediction error covariance matrices.

We will assume that Σ_{OF} can be modelled by $\sigma_{OF}^2 Id_2$ (e.g. $\sigma_{OF} = 0.5$ pixel). More sophisticated models, for instance based on estimates of the local textural content of the image, could be accounted for in the following analysis.

The chosen image prediction method of II-C3 is based on the following mapping:

$$(x, y, d) \mapsto U_{t-1}^{pred} = \Pi\left(R^{-1}\left(X_t(x, y, d) - T\right)\right). \tag{12}$$

We deduce from these dependencies, that U_{t-1}^{pred} estimation can be perturbed by an error in the estimation of (x, y, d) but also of ego-motion parameters [R,T]. The impact of (x, y, d) error has been considered in many articles. However, to the authors knowledge, only Alcantarilla et al. [3] have proposed an ego-motion uncertainty model directly related to the visual odometry without considering parameters learned a priori.

Assuming (x, y, d) and (R, T) error uncorrelated, one can study separately $\Sigma_{(x,y,d)}$ and $\Sigma_{R,T}$, as well as their respective contribution to Σ_M.

A. (x, y, d) Estimation Error

Because of pixel quantization, image coordinates (x, y) are prone to error. We model this by considering standard deviations σ_x and σ_y (e.g. equal to 0.1 pixel). Similarly, the error of the disparity obtained with a dense algorithm is represented by σ_d (e.g. equal to 1 pixel).

The contribution $\Sigma_{Pred(x,y,d)}$ of (x, y, d) error can be approximated using first order error propagation:

$$\Sigma_{Pred(x,y,d)} = J_{U_{t-1}^{pred}(x,y,d)} \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_d^2 \end{pmatrix} J_{U_{t-1}^{pred}(x,y,d)}^T \tag{13}$$

where $J_{U_{t-1}^{pred}(x,y,d)}$ is the Jacobian of $U_{t-1}^{pred}(x, y, d)$

B. (R, T) Estimation Error

To model $\Sigma_{R,T}$, we must know the energy minimized during the odometry. In our case, we choose the same odometry as the one used in [13], i.e. we minimize in a RANSAC procedure [21] the reprojection error

$$\mathcal{E}(R, T) = \frac{1}{N}\sum_{k=1}^{N} \|U_k^t - \Pi_t(RX_k^{t-1} + T)\|^2 \tag{14}$$

where $\{X_k^{t-1}\}_k$ is a set of triangulated feature points that have been extracted in I_{t-1}, and $\{U_k^t\}_k$ their location in I_t obtained by temporal matching.

This energy is minimized over $\Theta = (\theta_x, \theta_y, \theta_z, T_x, T_y, T_z)$, with the three first parameter being Euler's angles of R.

1) Heuristic formulation:
Alcantarilla et al. [3] estimate $\Sigma_{R,T}$ from the inverse of Hessian matrix of criterion (14):

$$H \approx J_{f(\Theta)}^T J_{f(\Theta)} \tag{15}$$

with:

$$\begin{cases} f(\Theta) & = \left(f_1(\Theta)^T, \cdots, f_N(\Theta)^T\right) \\ f_k(\Theta) & = U_k^t - \Pi_t(RX_k^{t-1} + T) \end{cases} \tag{16}$$

Unfortunately, with this approach, the estimation of uncertainty depends of any multiplicative factor applied to the energy $\mathcal{E}(\Theta)$: minimizing $\alpha\mathcal{E}(\Theta)$ leads to an Hessian matrix multiplied by α^2. Furthermore, such a model does not represent the error related to the estimation of $\{U_k^t, X_k^{t-1}\}_k$, nor their potentially correlated contributions to the error on Θ.

2) Analytical formulation:
The relation between Θ and input data $\{z_k^t\}_k = \{U_k^t, X_k^{t-1}\}_k$ is implicit but can be recovered by applying the well-known Implicit Function Theorem (cf. [22], chap 5). Considering the implicit function $\varphi : (\Theta, z) \mapsto \frac{\partial \mathcal{E}}{\partial \Theta}(\Theta, z)^T$, we then obtain the error covariance matrix below:

$$\Sigma_\Theta = H^{-1}\left(\frac{\partial \varphi}{\partial z}\right)\Sigma_z\left(\frac{\partial \varphi}{\partial z}\right)^T H^{-T} \tag{17}$$

where $H = \frac{\partial^2 \mathcal{E}}{\partial \Theta \partial \Theta}(\Theta, z) \in \mathbf{R}^{6\times 6}$, is supposed invertible. Assuming the error independent for each feature k, Eq. (17) becomes:

$$\Sigma_\Theta = \sum_k H^{-1}\left(\frac{\partial \varphi}{\partial z_k}\right)\Sigma_{z_k}\left(\frac{\partial \varphi}{\partial z_k}\right)^T H^{-T}. \tag{18}$$

As U_k^t and X_k^{t-1} are computed separately during the sparse temporal matching and the 3D reconstruction steps, we assume

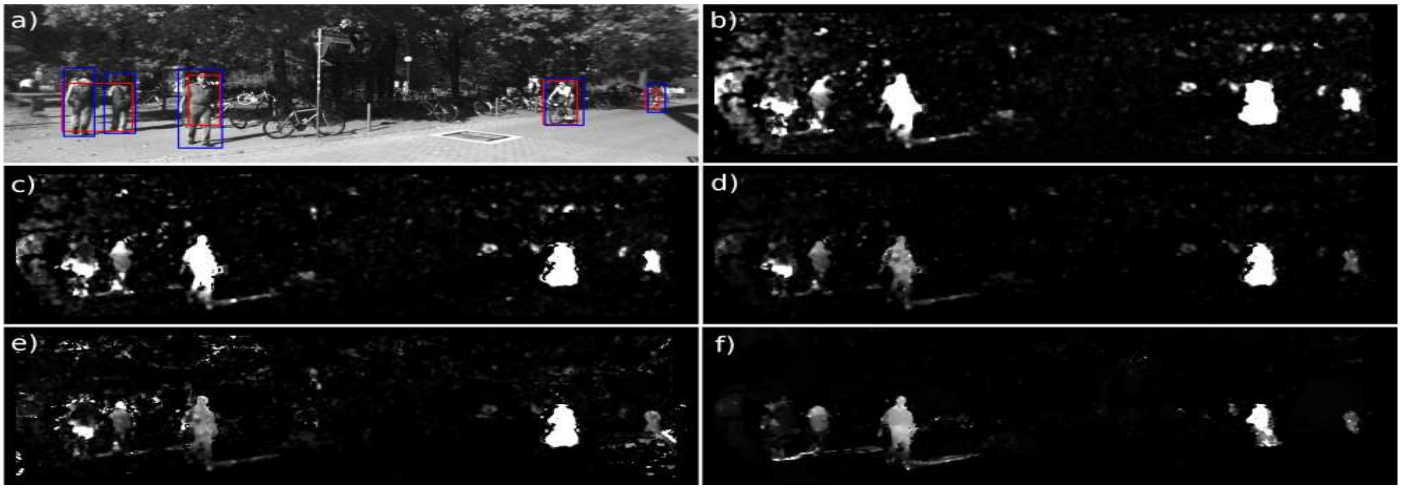

Fig. 2. The left camera image at $t = 55$ is displayed in (a) with annotated BB (blue) and estimated BB evaluated as correct (red) following protocol IV-B. In (b-d): Motion likelihoods $\xi^2(M)$ using eFOLKI and image prediction methods with Σ_{OF} (i.e. L2 norm) in (b), $\Sigma_{M(x,y,d)}$ in (c), $\Sigma_{M(x,y,d,\Theta)}$ in (d). In (e): eFOLKI + $\Sigma_{M(x,y,d,\Theta)}$ applied to direct method. In (f): Brox et al. OF [19] for image prediction method with $\Sigma_{M(x,y,d,\Theta)}$.

that they are not correlated, which leads to:

$$\Sigma_{z_k} = \left(\begin{array}{cc|c} \sigma_u^2 & 0 & \\ 0 & \sigma_v^2 & 0_{2\times 3} \\ \hline 0_{3\times 2} & & J_{X_k^{t-1}} \Sigma_{(x,y,d^*)} J_{X_k^{t-1}}^{T} \end{array}\right), \tag{19}$$

where the upper left diagonal matrix is the error model of the sparse temporal matching, and d^* the disparity of the feature point whose error is modelled by σ_{d^*} — assumed lower than σ_d (e.g. $\sigma_u = \sigma_v = \sigma_{d^*} = 0.5$ pixel).

C. Residual Field Error

The global prediction error model then writes:

$$\Sigma_{Pred} = \Sigma_{Pred(x,y,d)} + J_{U_{t-1}^{pred}(\Theta)} \Sigma_\Theta J^T_{U_{t-1}^{pred}(\Theta)}, \tag{20}$$

where $J_{U_{t-1}^{pred}(\Theta)}$ is the Jacobian matrix of U_{t-1}^{pred} regarded as a function of Θ. The uncertainty of the residual field Σ_M is estimated by Eq. (11) and used to obtain the motion likelihood by Eq. (1).

IV. Experimental Results

A. Residual Displacement Fields Comparison

We have evaluated different motion likelihoods ξ on two sequences (09/28-0037 and 09/29-0071) of the publicly available KITTI datasets [12]. Experiments have shown that $\xi(\delta u, \delta v, \delta d)$ is prone to more noise than $\xi(\delta u, \delta v)$, without improving the detection. Fig. 2 shows $\xi(\delta u, \delta v)$ obtained on sequence 09/28-0037 between $t = 54$ and $t = 55$: the camera undergoes translation and rotation, and there are both slow moving objects (pedestrians) and faster ones (cyclists).

Image prediction methods appear less subject to noise than direct ones. Indeed, the displacement between I_t^{pred} and I_t is smaller than the one between I_{t-1} and I_t, so the optical flow estimation is more accurate in the first case. Moreover, as shown in images (b-d), noise is reduced by a better uncertainty model, though at the cost of a lower SNR on moving objects.

We have also compared eFOLKI with the variational optical flow of Brox et al. [19] which is more accurate than TV-L1 on Kitti dataset, as shown in [11]. Parameters of eFOLKI are $J = 5$ resolution levels, $K = 5$ iterations, three window radii $\{12,8,4\}$ and rank order 4. Parameters of Brox et al. OF are $\alpha = 0.5$, $\gamma = 500$, 10/20/100 solver/inner/outer iterations. Comparing residual fields in Fig. 2 shows that the residual obtained with Brox et al. OF (f) is less noisy than eFOLKI's (d), but that slow moving objects or partially moving ones are more visible in the latter. Such behaviour is confirmed in Fig. 3 by the better recall of eFOLKI.

B. Evaluation Protocol

To evaluate the various tested approaches, we have manually annotated ground truth bounding boxes BB_{GT} using Vatic. Detection is based on a discrete overlap ratio:

$$\omega(BB, BB_{GT}) = \frac{\mathcal{A}_{BB\,\cap\,BB_{GT}}}{\mathcal{A}_{BB\,\cup\,BB_{GT}}}. \tag{21}$$

A detected BB is valid when there exists a BB_{GT} such that $\omega(BB, BB_{GT})$ is below some threshold (e.g. 25%). To avoid multiple instances of the same detection, we count one True Positive for each BB_{GT} whatever the number of valid BB it is associated to. Since some BB are too small to be considered as valid, we don't consider a BB as a False Positive if it belongs to a BB_{GT} detected by another valid BB. Other estimated BB are considered as False Positive, and BB_{GT} with no associated detections are False Negative. Several evaluations were done using different thresholds on $\xi^2(M)$ to construct the Precision/Recall curves displayed in Fig. 3. They demonstrates the gain of our method which returns a number more important of correct bounding boxes. A precision of approximately 80% and a recall of 50% is achieved on both sequences. This recall value may seem low, but only two frames are used in our process with no temporal or spatial filtering.

Let us remark that the benefit of a global error model including the pose uncertainty is more significant in video

TABLE I
PROCESSING TIME FOR EACH STAGE (EXCEPT DENSE STEREO)

Odometry	**Prediction**	**OF: eFOLKI/Brox**	$\xi^2(M)$	**CC + BB**
40ms	2ms	27ms/480ms	7ms	5-10ms

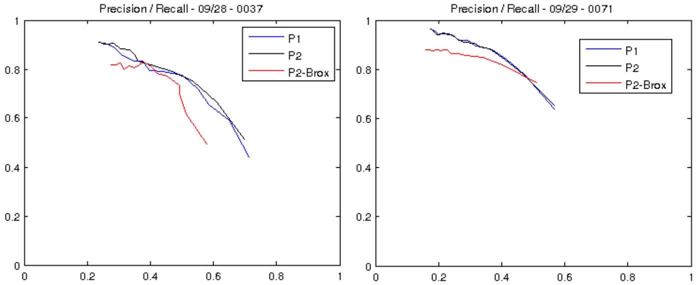

Fig. 3. Precision / Recall for two KITTI sequences, thresholding $\xi^2(M)$ from 1 to 15, using: eFOLKI + $\Sigma_{M(x,y,d)}$ (P1), eFOLKI + $\Sigma_{M(x,y,d,\Theta)}$ (P2), and Brox et al. optical flow [19] + $\Sigma_{M(x,y,d,\Theta)}$ (P2-BROX).

09/28-0037 than in 09/29-0071. Indeed, rotation is present only in video 09/28-0037 and rotation error affects more far objects which are better handled by Σ_Θ (Fig. 2 (c) and (d)).

C. Processing Time

Table I summarizes processing times with a CPU Intel Core i7 (12 cores) and a GPU GeForce GTX TITAN. Note that multi-threading the odometry would significantly decrease its runtime. We do not use a FPGA implementation of SGBM but the OpenCV CPU ones, which runs here in 425ms for 256 disparity levels.

The use of eFOLKI fast algorithm saves a considerable amount of time compared to variational OF and allows the whole process to achieve near video frame rate (ie 10Hz) leaving apart the calculation of disparity map. The latter operation remains the computational bottleneck of the process. SGBM on a FPGA could be a solution. One could also use geometrical informations returned by the system to speed-up stereo — e.g. the disparity range may be deduced from previous disparity map and camera pose.

V. CONCLUSION

We have presented a framework for mobile object detection from a moving stereo rig based on an image prediction strategy. It is compatible with real-time processing thanks to the fast dense OF method eFOLKI. A new complete error model has been derived, which allows to handle rigourously the uncertainty related to visual odometry. We have conducted an experimental study on several real videos from the KITTI website to compare our approach with various proposals of the literature. According to this study, image prediction strategy improves the SNR of the likelihood. It also shows that the fast OF algorithm eFOLKI is compatible with good detection rates. We now plan to add temporal filtering and to improve image prediction using a 3D scene model of higher level, such as, for instance, 3d mesh representation of [23].

ACKNOWLEDGMENT

This work was sponsored by the Direction Générale de l'Armement (DGA) of the French Ministry of Defense.

REFERENCES

[1] H. Badino and T. Kanade, "A headwearable shortbaseline stereo system for the simultaneous estimation of structure and motion." in *Machine vision and applications*, 2011, pp. 185–189.

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

[3] P. Alcantarilla, J. Yebes, J. Almazan, and L. Bergasa, "On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments," in *IEEE ICRA'12*, 2012, pp. 1290–1297.

[4] C. Rabe, T. Mller, A. Wedel, and U. Franke, "Dense, robust, and accurate motion field estimation from stereo image sequences in realtime," in *ECCV'10*, Springer, 2010, pp. 582–595.

[5] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers, "Stereoscopic scene flow computation for 3D motion understanding," *Int. J. of Computer Vision*, vol. 95, no. 1, pp. 29–51, Oct. 2011.

[6] A. Talukder and L. Matthies, "Realtime detection of moving objects from moving vehicles using dense stereo and optical flow," in *IROS'04*, vol. 4, 2004, pp. 3718–3725.

[7] V. RomeroCano and J. I. Nieto, "Stereobased motion detection and tracking from a moving platform," in *IEEE Intelligent Vehicles Symposium*, 2013, pp. 499–504.

[8] A. Bak, S. Bouchafa, and D. Aubert, "Dynamic objects detection through visual odometry and stereo-vision: a study of inaccuracy and improvement sources," *Machine vision and applications*, pp. 1–17, 2014.

[9] M. Agrawal, K. Konolige, and L. Iocchi, "Realtime detection of independent motion using stereo," in *Seventh IEEE Wksp on Appl. of Computer Vision*, vol. 2, 2005, pp. 207–214.

[10] A. Bak, S. Bouchafa, and D. Aubert, "Detection of independently moving objects through stereo vision and egomotion extraction," in *IEEE Intelligent Vehicles Symposium*, 2010, pp. 863–870.

[11] A. Plyer, G. Besnerais, and F. Champagnat, "Massively parallel Lucas-Kanade optical flow for real-time video processing applications," *J. of Real-Time Image Processing*, pp. 1–18, 2014.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE CVPR'12*, 2012, pp. 3354–3361.

[13] M. Sanfourche, V. Vittori, and G. Le Besnerais, "eVO: a realtime embedded stereo odometry for MAV applications," in *IEEE/RSJ IROS'13*, 2013, pp. 2107–2114.

[14] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[15] S. K. Gehrig, F. Eberli, and T. Meyer, "A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching," in *Proc. 7th Int. Conf. on Computer Vision Systems*, LNCS, M. Fritz, B. Schiele, and J. Piater, Eds., vol. 5815, Springer, 2009, pp. 134–143.

[16] B. K. Horn and B. G. Schunck, "Determining optical flow," in *Proc. SPIE*, vol. 0281, 1981, pp. 319–331.

[17] A. Bruhn, J. Weickert, and C. Schnrr, "Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods," *Int. J. of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[18] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tvl1 optical flow," in *Ann. Symp. German Assoc. Patt. Recogn*, 2007.

[19] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV'04*, Springer, 2004, pp. 25–36.

[20] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. of Computer Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[22] O. Faugeras, *Three dimensional computer vision: A geometric viewpoint*, the MIT Press, 1993.

[23] M. Lhuillier, "A generic error model and its application to automatic 3d modeling of scenes using a catadioptric camera," *Int. J. of Computer Vision*, vol. 91, no. 2, pp. 175–199, 2011.