Multispectral Stereo Odometry

Tarek Mouats, *Student Member, IEEE*, Nabil Aouf, *Member, IEEE*,
Angel Domingo Sappa, *Senior Member, IEEE*, Cristhian Aguilera, and Ricardo Toledo

***Abstract*—In this paper, we investigate the problem of visual odometry for ground vehicles based on the simultaneous utilization of multispectral cameras. It encompasses a stereo rig composed of an optical (visible) and thermal sensors. The novelty resides in the localization of the cameras as a stereo setup rather than two monocular cameras of different spectrums. To the best of our knowledge, this is the first time such task is attempted. Log-Gabor wavelets at different orientations and scales are used to extract interest points from both images. These are then described using a combination of frequency and spatial information within the local neighborhood. Matches between the pairs of multimodal images are computed using the cosine similarity function based on the descriptors. Pyramidal Lucas–Kanade tracker is also introduced to tackle temporal feature matching within challenging sequences of the data sets. The vehicle egomotion is computed from the triangulated 3-D points corresponding to the matched features. A windowed version of bundle adjustment incorporating Gauss–Newton optimization is utilized for motion estimation. An outlier removal scheme is also included within the framework to deal with outliers. Multispectral data sets were generated and used as test bed. They correspond to real outdoor scenarios captured using our multimodal setup. Finally, detailed results validating the proposed strategy are illustrated.**

***Index Terms*—Egomotion estimation, feature matching, multispectral odometry (MO), optical flow, stereo odometry, thermal imagery.**

I. Introduction

IN RECENT years, the field of intelligent transportation systems (ITS) has noticed a remarkable shift in researchers' interests particularly toward advanced driver-assistance systems (ADAS). These systems play a significant role in the development of intelligent vehicles. Localization of vehicles, in particular, is a key component of such systems. Ordinarily, localization information is provided by the Global Positioning System (GPS). However, this system suffers from a number of shortcomings where solutions have been investigated in the literature to temporarily filling signal gaps or completely replacing GPS information. In this context, cameras present an interesting sensing alternative. Its main advantages reside in cost effectiveness, low power consumption and the meaningfulness of its content. During the last decade, the automotive industry witnessed the introduction of a variety of cameras to enhance vehicles' safety (e.g., thermal and parking cameras). Developing localization techniques on these grounds represents an interesting research path for the coming years. Studies on using visual information for self-localization have been conducted over the last decades. Visual odometry (VO) along with visual simultaneous localization and mapping represent the main vision driven localization solutions. VO involves the estimation of the egomotion of an agent using only visual information from one or multiple cameras. It has been widely investigated in computer vision and robotics. Early attempts to recover motion from vision were made as far as three decades ago [1]. VO was coined as so for the first time in [2]. Its applications span a variety of domains such as robotics, automotive, and space missions. In the context of driving assistance and autonomous systems, self-localization represents a fundamental issue. The vehicle's own movement (egomotion) is a prerequisite for higher level tasks (e.g., scene perception). In general, this task is performed using wheel odometry, Inertial Measurement Units (IMUs), or GPS devices. Another way to accomplish that task is through VO, which takes advantage of cameras. These can overcome negative aspects of wheel odometry, particularly in slippery terrain. In addition cameras can mitigate the drawbacks of IMUs by providing less drifty estimates of the motion. Furthermore, GPS devices, although very costly, can suffer shortages or inaccuracies. In this case also, VO comes as a cheaper and reliable alternative. Up to now, and to the best of the authors' knowledge, all the attempts to achieve such task have been conducted using cameras working on the same spectral band namely visible. These include monocular, stereo and omnidirectional cameras. In this paper, the feasibility of egomotion estimation from cameras working in different spectral bands is investigated. The aim is to extend the concepts of VO to multispectral odometry (MO). In the field of driving assistance systems, these types of cameras are already deployed to tackle a variety of problems. Infrared (IR) cameras are used to improve night-time driving experience as they are able to capture scene elements in the dark. Pedestrian detection and collision avoidance mechanisms based on day-time cameras were extended to night-time using IR technology. Our aim is to take advantage of equipment already in place to get more functionality. The vehicle motion is estimated incrementally on a frame-to-frame basis using only the acquired stereo image pairs with no prior knowledge of the environment. The system is capable of estimating its 6 degrees of freedom (DOF) without use of filtering techniques. These are generally used with SLAM algorithms, where the choice of the filter influences the accuracy of the motion estimates [3].

Manuscript received December 26, 2013; revised May 13, 2014 and August 8, 2014; accepted August 28, 2014. Date of publication September 26, 2014; date of current version May 29, 2015. The Associate Editor for this paper was P. Grisleri.

T. Mouats and N. Aouf are with the Centre for Electronic Warfare, Cranfield University, Defence Academy of the United Kingdom, Shrivenham SN6 8LA, U.K. (e-mail: t.mouats@cranfield.ac.uk; n.aouf@cranfield.ac.uk).

A. D. Sappa, C. Aguilera, and R. Toledo are with the Computer Vision Center, Universitat Autónoma de Barcelona, 08193 Barcelona, Spain (e-mail: asappa@cvc.uab.es; caguilera@cvc.uab.es; ricardo.toledo@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2014.2354731

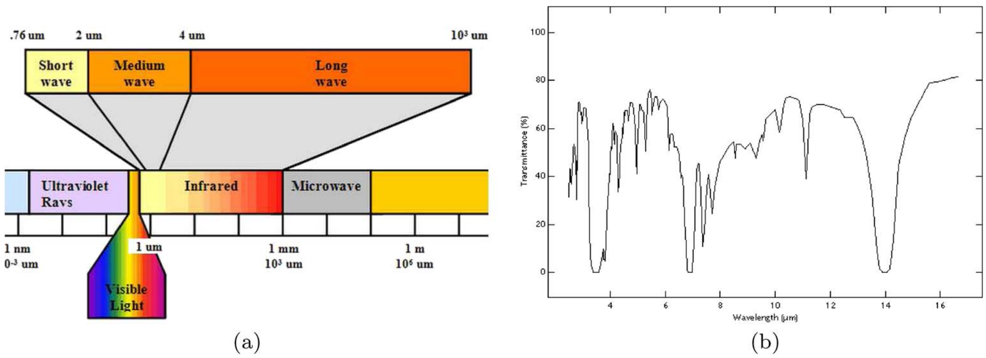


Fig. 1. (a) IR spectrum and (b) IR transmittance in the atmosphere.

A. Related Work

Many implementations of VO can be found in the literature that can be classified according to the sequential correspondence search strategy. There are those relying on temporal feature matching [4]–[6] and those based on tracking algorithms [7], [8]. VO techniques could also be classified according to the number of used cameras. Typically, there are monocular approaches [9]–[11] and those using stereo rigs [4], [12]–[15]. In general, approaches based on multiple cameras provide better performance than monocular algorithms. Many investigations were conducted to enhance the performance of VO. One aspect relates to the utilization of digital maps [16] and extended digital maps [17] to accurately localize vehicles with respect to their surrounding environment. However, the approach adopted in [17] is dedicated to intersection scenarios. Other works involve the localization of other sensors to aid the VO as in [18], where IMU measurements are fused with VO outputs using a Kalman filter. Agrawal *et al.* [19], [20] fused GPS and wheel encoders information within the VO framework. These multisensory fusion schemes were claimed to tackle long-term drifts. Keeping drift small can also be done through local optimization over the last N camera poses. Sliding window (windowed) bundle adjustment (WBA) represents one of the approaches used in several works. It was demonstrated by Konolige *et al.* [21] to track the motion over distances of 10 km with a small error. A similar approach is adopted in the motion estimation part of this paper, where the current and previous pairs of images are used. Other estimation algorithms have been used in works tackling VO. Milella and Siegwart [13] used the iterative closest point technique, whereas singular value decomposition was adopted in [18]. An essential prerequisite for most VO approaches lies in the so-called low-level image processing task of feature extraction. In the literature, many interest point extraction and matching techniques have been utilized. For instance, Shi and Tomasi's Good Features to Track [22] was used in [5], [18] to detect features in the images. SURF [23] descriptors were used to match the detected features in [5] based on the Euclidean distance. In [16], the well-established Scale-invariant Feature Transform (SIFT) [24] feature detector was used. Although the matching process was not described, it is usually performed using the Euclidean distance. The Harris Corner Detector [25] was used in [4] and matching was achieved using normalized correlation. Works mentioned so far rely mainly on intensity and gradient information for matching features. Unfortunately, these algorithms provide very poor results (or fail completely) in the multimodal stereo configuration tackled in our work. The most obvious reason would be the fact that in a cross-spectral image pair, the relationship between pixel intensities of the thermal and visible images is nonlinear. That is, pixels that appear bright in a thermal image might appear dark in its visible stereo pair and vice versa. In this sense, having information on pixels in one modality (i.e., thermal or visible) does not provide any information on the corresponding pixel of the other modality. It is also to be noted that all of the methods mentioned above rely on images acquired in the visible part of the spectrum. These differ greatly from thermal imagery, which is characterized by low resolution and lack of textureness. Jung *et al.* [26] proposed an algorithm for egomotion estimation from a monocular infrared (IR) camera. Focus of expansion was used as a basis for feature matching and egomotion was computed using reprojection errors. However, the shown results did not include egomotion trajectories. In [27], authors proposed a similar monocular approach but using two cameras: thermal and visible. A handover mechanism was introduced, but once again, each camera was separately used for monocular SLAM. The cross-spectral setup was not used in a stereo fashion, and hence, no matching was required between images.

B. Background

In contrary to what is generally believed, any object that has a temperature greater than the absolute zero is a source of thermal radiation (even cold objects) as it emits heat in the IR spectrum. Within the electromagnetic spectrum, IR light lies between the visible and microwave bands [see Fig. 1(a)]. It is traditionally divided into three subbands: near-, mid-, and far-IR. From these, only portions are of interest as most of the radiation is absorbed by water and carbon dioxide molecules present in the atmosphere [see Fig. 1(b)]. These are the short-wavelength IR (SWIR/NIR) [0.7–1.4 μm]; the medium-wavelength IR (MWIR/MIR) [3–5 μm]; the long-wavelength IR (LWIR/FIR) [8–14 μm]. The subband of interest in our work is the far-IR or thermal. FIR cameras are widely used for night-time vision-based driver assistance applications (e.g., pedestrian detection [28]). In the far-IR band, heat reflectance of observed objects hardly contributes to the captured image. Instead, it is composed mainly of the objects' thermal radiation. Intensity information contained within these images does not vary with lighting conditions but rather with changes in temperature. This makes it possible for thermal cameras to see in the dark (i.e., nighttime) and through smoke or fog. Nonetheless, IR imagery exhibits a number of challenges compared with visible imagery [29] namely: 1) high noise and low spatial resolution, where the former invalidates the smoothness model, and the latter means losing high-frequency data; 2) history effects caused by the fact that the brightness of a pixel in a thermal image depends on objects' self-emissions, which are function of their temperature as well as the environment's temperature. In contrary to light variations, temperature variations take time in general. This means that the information captured by thermal sensors does not depend only on the instantaneous states of the objects being imaged but also on the effects of the history of changes. Therefore, pixel information captured by the camera may relate to an object that is no longer present in the scene (i.e., ghost objects); and 3) image saturation due to the nature of thermal imagery, and more precisely, thermal self-emissions. Objects being imaged emit radiation, of which strength is proportional to the fourth degree of the objects temperature. Therefore, very dark and very bright objects are expected to be seen every time. In such case, local texture information is lost as either the bright objects are overexposed or the dark objects are underexposed. These drawbacks increase the difficulties encountered within the different subtasks of VO. For instance, due to low resolution, the number of extracted interest points is lower in thermal images than in visible images. Furthermore, because of the history effects, which invalidate the basic assumption of the brightness constancy constraint, optical flow-based trackers cannot be used to track features in sequential thermal images. This paper is organized as follows: Section II describes the proposed feature extraction and matching approach. Motion estimation scheme is explained in Section III and the experimental results are detailed in Section IV. Conclusions are drawn in Section V, where some insights into our future work are highlighted.

C. Motivations

Multispectral vision systems are commonly used on military ground and air vehicles. This is dictated by the 24-h all-weather operation capability required for these military assets in terms of target/threat detection and identification. Similarly, in recent years, cars manufacturers have been equipping new vehicles with networks of sensors, from visible to thermal cameras. One can take advantage of such multispectral setup to get more functionality out of it and provide added value. This forms the main motivation of this paper, where the aim is to develop strategies that take advantage of what is already available in terms of sensors and accomplish more tasks than what they were designed for. One concrete instance of utilization is onboard military vehicles, where the former setups would be used for egomotion estimation in addition to the inherited "Detection, Recognition, and Identification" military-oriented functions. This can be particularly beneficial when the military assets are facing GPS signal loss or jamming, and alternatives must be used to allow self-localization with respect to the surrounding environment. Therefore, the prime aim of this paper is to demonstrate the feasibility of egomotion estimation from a multispectral setup. The same task has been widely demonstrated in the literature using a pair of (or a single) visible band cameras. Indeed, the idea of this proposed solution is to exploit the existence/availability of different modality sensors such as thermal cameras that are usually dedicated to the detection and tracking for either day (cluttered) time or nighttime. The goal is to improve and complement a monocular visible band camera based motion estimation setup into a more efficient stereo multispectral motion estimation setup without the need of installing two visible cameras. Tackling night-time VO is not considered in the scope of this paper.

II. Feature Extraction and Matching

Feature extraction is a low-level image processing task that represents a prerequisite for most computer vision applications. This is particularly true in the case of autonomous navigation applications, where essential information contained within an image needs to be extracted. Our interest point extraction and matching strategy is detailed in our previous work on multimodal stereo matching [30]. Nevertheless, in order to help the reader and make this paper self-contained, a summary of the approach is given here.

A. Feature Extraction

The goal is to represent a given image by a set of distinctive interest points or features. These should be stable enough to be repeatedly detected, invariant to geometric transformations and robust to noise. Phase congruency (PC), the adopted feature detector, is derived from the work done by Morrone and Owens [31] based on the local energy model (LEM). This model was shown to successfully explain a number of psychophysical effects in human feature perception [32]. The LEM assumes that image features are located in the frequency domain, where their Fourier components are maximally in phase. Traditionally, intensity-based extractors assume them to be at points of maximal intensity gradients. These classical operators exhibit a common behavior. The corner response varies considerably with image contrast and changes in lighting conditions making the setting of appropriate thresholds a difficult task. In [33], Kovesi represented the PC at a position x as follows:

$$\mathrm{PC}_2(x) = \frac{\sum_n W(x) \lfloor A_n(x)\Delta\Phi(x) - T \rfloor}{\sum_n A_n(x) + \epsilon} \quad (1)$$

where

$$\Delta\Phi(x) = \cos\left(\Phi_n(x) - \bar{\Phi}(x)\right) - \left|\sin\left(\Phi_n(x) - \bar{\Phi}(x)\right)\right|. \quad (2)$$

In (1) and (2), $A_n(x)$ and $\Phi(x)$ represent, respectively, the amplitude and phase of the nth component at position x; $W(x)$ is a factor that weights for frequency spread; $\Delta\Phi(x)$ is the phase deviation; T is the estimated noise influence; and ϵ is a small constant added mainly to avoid division by zero. The symbols $\lfloor \; \rfloor$ denote that the enclosed quantity is equal to itself

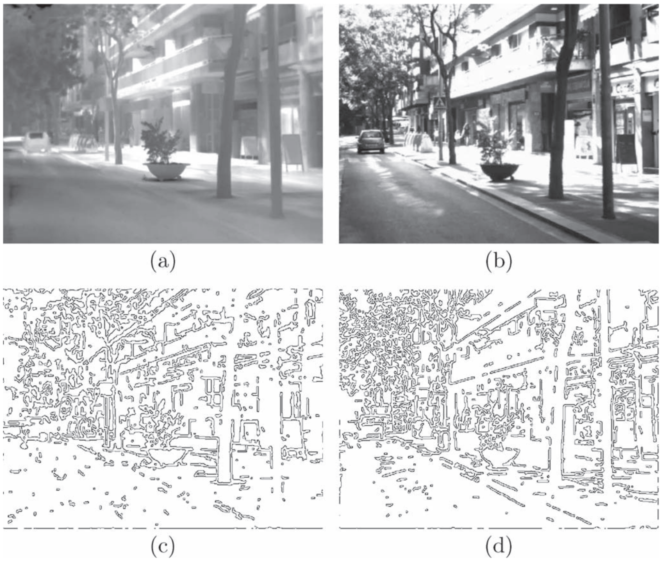


Fig. 2. IR and visible stereo pair with corresponding edge maps. (a) and (b) IR and visible images. (c) and (d) Corresponding edge maps. (Images from our data set).

when its value is positive and zero otherwise. This means that only energy values that exceed the noise level T are taken into account in the result. In (2), $\bar{\Phi}(x)$ represents the weighted mean phase angle. In practice, the PC is computed using banks of Log-Gabor filters at different frequencies and orientations. Our implementation comprises a set of 24 Log-Gabor filters corresponding to six orientations at four frequencies. They are used to obtain the PC map of the images used to extract edges and corners by calculating the maximum (M) and minimum (m) moments

$$M = \frac{1}{2}\left(c + a + \sqrt{b^2 + (a-c)^2}\right) \tag{3}$$

$$m = \frac{1}{2}\left(c + a - \sqrt{b^2 + (a-c)^2}\right) \tag{4}$$

where

$$a = \sum (\mathrm{PC}(\theta)\cos\theta)^2 \tag{5}$$

$$b = \sum (\mathrm{PC}(\theta)\cos\theta)\,(\mathrm{PC}(\theta)\sin\theta) \tag{6}$$

$$c = \sum (\mathrm{PC}(\theta)\sin\theta)^2 \tag{7}$$

where $\mathrm{PC}(\theta)$ represents the PC value determined at orientation θ and the sum operation is performed for the set of the used orientations (06). At this stage, a given pixel is labeled *edge* if its maximum moment is large. It is labeled *corner* if, at the same time, its minimum moment is also large. Fig. 2 shows the resulting edge map for a multispectral image pair. In order to improve our detection and matching cross-spectral approach [30], nonmaxima suppression, and feature spreading were introduced.

1) Nonmaxima Suppression: Once corners are extracted, a common observation is that these might be clustered. This can possibly add ambiguity when matching those features. One solution to tackle this problem is the use of nonmaxima suppression. It is used in computer vision applications and more specifically in feature extraction algorithms [24]. It mainly consists in keeping only corners larger than all their neighbors. Fig. 3 illustrates the obtained corners before and after applying the nonmaxima suppression using a three-pixel neighborhood.

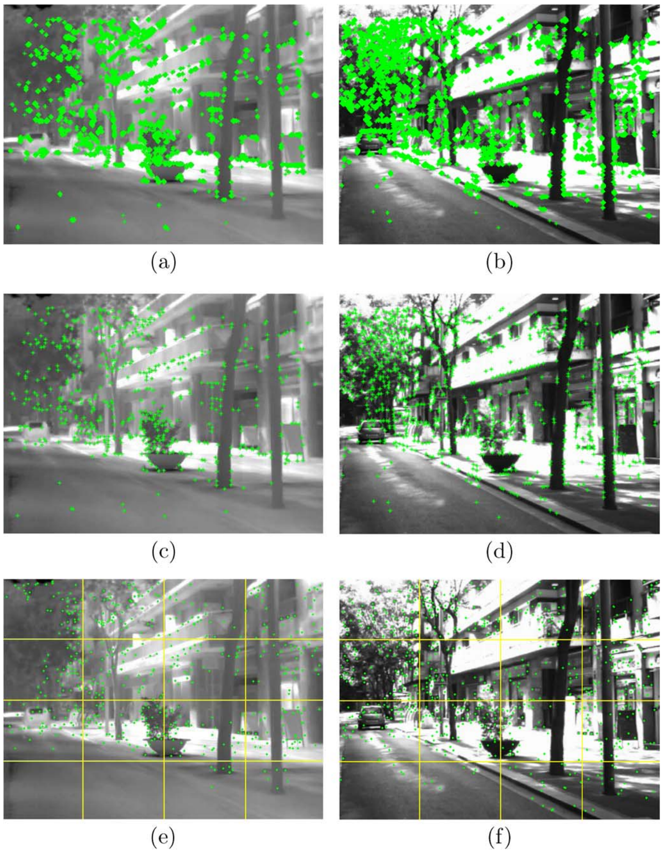


Fig. 3. Extracted features in a stereo pair (left:IR, right:visible). (a) and (b) Raw. (c) and (d) Using nonmaxima suppression. (e) and (f) Using subimage extraction.

2) Spreading Features Across the Image: A common problem with feature detectors is that some areas of the images are overloaded with interest points, whereas other regions are left featureless (i.e., nearly empty). This is due to the fact that the detection process is carried out at a small scale where only a restricted area around a given pixel is considered. Fortunately, there are alternatives for this limitation. In our work, the considered image is subdivided into subimages, where the detection takes place. A maximum number of features are allowed per subimage to guarantee the spread of interest points to all image regions if there is enough texture. Fig. 3 illustrates an example contrasted to the original detection scheme.

B. Feature Description

The next step is matching the extracted keypoints. For this aim, descriptors are computed based on the edge histogram descriptor (EHD) [30] and combined with the Log-Gabor coefficients (24 elements) computed in the previous step. This yields a larger descriptor incorporating frequency and shape information. Such descriptor choice is primarily dictated by the nonlinear relationship intensity-wise between multispectral images, which make the use of gradient-based descriptors useless. Furthermore, as concluded in [34] and shown in Fig. 2, FIR images tend to preserve the same boundaries as visible images. This implies that a descriptor based on the shape information around the keypoints should provide better matching performance. There are many edge detection algorithms in the literature (e.g., Canny [35], Laplacian of Gaussian [36]) that can be used to compute the first part of the descriptor (EHD). However, we opted for the edge map obtained from the PC as it is computed in the previous step meaning that no extra computation is required. In addition, experiments with classical algorithms were conducted where no significant improvements were noticed. The spatial component of the descriptor is obtained as follows:

- Select a region of $P \times P$ pixels centered at the keypoint of interest from the edge map.
- Divide the region into 16 (4 × 4) subregions.
- Compute local edge histograms for each subregion where five bins are used to categorize an edge: horizontal, vertical, 45° diagonal, 135° diagonal, and isotropic (no orientation).

The resulting histogram vector formed of 80 bins (4 × 4 × 5) is then normalized. Combining the two parts creates the sought descriptor consisting of 104 elements.

C. Matching

There are mainly two types of matching tackled within the scope of this paper. This is driven by the fact that at any time t, the algorithm is fed with four input images: left and right at times $t-1$ and t. Therefore, in addition to the stereo matching that takes place every time a stereo image pair is acquired; there is a temporal (sequential) matching that needs to be addressed. For this dual objective, the cosine similarity function is used to compare features descriptors. Let D_L be the descriptor of the feature f_L at position (x_L, y_L) in the left image. Similarly, let D_R be the descriptor of a potential match f_R at position (x_R, y_R) in the right image within a search window $disp_x \times disp_y$ centered at (x_L, y_L). $disp_x$ and $disp_y$ account for the maximum expected horizontal and vertical disparities, respectively. The similarity function is given by

$$S(D_L, D_R) = \frac{\sum_j d_{Lj} d_{Rj}}{\sqrt{\sum_j {d_{Lj}}^2 \sum_j {d_{Rj}}^2}} \tag{8}$$

where (D_L, D_R) are the descriptors of the compared features; d_{L_j}, d_{R_j} are, respectively, the jth coefficients of (D_L, D_R). The feature in the right image that maximizes the similarity function for a given feature in the left image is selected as a potential match. A threshold is then applied to keep only strong matches. As stated above, the algorithm is fed with four images: previous left $(im_{L_{t-1}})$, previous right $(im_{R_{t-1}})$, current left (im_{L_t}), and current right (im_{R_t}). The matching is carried out in a loop fashion [14] to keep only features that find their correspondences across all four images. Fig. 4 illustrates the different steps. We first start by finding stereo matches between $(im_{L_{t-1}})$ and $(im_{R_{t-1}})$ (I). Then, sequential matches are found between $(im_{R_{t-1}})$ and (im_{R_t}) (II). Another stereo matching is performed between (im_{L_t}) and (im_{R_t}) (III). Finally, a last sequential matching is performed between $(im_{L_{t-1}})$ and (im_{L_t}) (IV). At this stage, if the starting and ending feature points are identical, then the match is accepted. Otherwise, it is simply rejected. This process is carried out for all the features extracted in the first image $(im_{L_{t-1}})$.

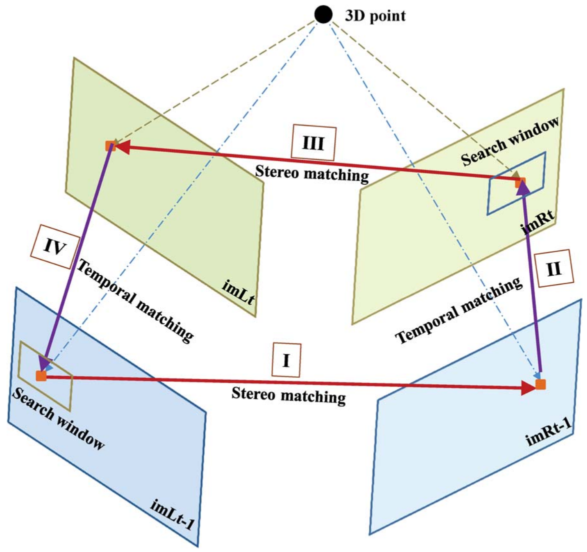


Fig. 4. Illustration of the loop matching steps.

As the multispectral image pairs are to be rectified, the search window (2-D) reduces to a search line (1-D) in the stereo matching process. Correspondences are expected to be found on the same line (i.e., epipolar constraint) of the left and right images. However, this is not the case with the sequential matching where a 2-D search would be still required.

III. Motion Estimation

The proposed algorithm for egomotion estimation is based on a reduced version of the wide variety of bundle adjustment algorithms surveyed in [37]. This version is called WBA as it analyzes only a portion of the image set to derive the motion estimates. In our case, only the previous and current image pairs of the sequence are used at each time step. First, features are extracted and matched in all four images as described in Section II. Egomotion estimation is achieved using these matches by minimizing reprojection errors using Gauss–Newton optimization within the WBA framework. An outlier rejection scheme based on random sample consensus (RANSAC) [38] is included prior to the final motion optimization step. Outliers that occur due to false matches or matches detected on independently moving objects are dealt with. Each of the aforementioned steps is detailed here along with a reminder of the camera model.

A. Camera Model

In the current work, a multispectral stereo vision setup is considered. The intrinsic and extrinsic calibration parameters of the camera are assumed to be known. Let K be the calibration parameters matrix. In what follows, the left camera is considered as the reference camera. The relationship between the homogeneous image coordinates $\hat{x} = (u, v, 1)$ and the camera coordinates $\mathbf{X_C} = (X_C, Y_C, Z_C)$ is given by

$$\hat{x} = K.\mathbf{X_C}. \tag{9}$$

It is worth mentioning that the parameters matrix K is identical for both cameras after rectification of the images. Considering the projections on the left and right images, this yields

$$K = K_L = K_R = \begin{pmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{10}$$

where α_u, α_v correspond to the focal length u_0, v_0 the principal point coordinates. Therefore, the projections $\hat{x}_L = (u_L, v_L, 1)$ and $\hat{x}_R = (u_R, v_R, 1)$ on the left and right cameras, respectively, are given by

$$\hat{x}_L = K.\mathbf{X_C} \tag{11}$$

$$\hat{x}_R = K.\left(\mathbf{X_C} - (b_L, 0, 0)^T\right) \tag{12}$$

where b_L denotes the stereo baseline. Note that v_L and v_R are identical. It is then convenient to define a vector $y = (u_L, v_L, u_R)$ of the projected coordinates on the stereo images obtained by applying the projection function π to a 3-D point $\mathbf{X}$ (with respect to the left camera)

$$y = f(\mathbf{X}) = \begin{pmatrix} u_L \\ v_L \\ u_R \end{pmatrix} = \begin{pmatrix} \alpha_u \left(\frac{X}{Z}\right) - u_0 \\ \alpha_v \left(\frac{Y}{Z}\right) - v_0 \\ \alpha_u \left(\frac{(X - b_L)}{Z}\right) - u_0 \end{pmatrix}. \tag{13}$$

We assume that the camera parameters do not change with time allowing the bundle adjustment to not recompute them again.

B. Motion Parameters

The camera/vehicle motion can be regarded as a combination of rotations and translations embodied in a motion parameters vector $m = (\phi, \theta, \psi, t_x, t_y, t_z)$. The first three parameters correspond to the Euler rotations and form the rotation matrix $R = (\phi, \theta, \psi)$, whereas the last parameters form the translation vector $t = (t_x, t_y, t_z)$. Writing the transformation matrix $M_p(m)$ derived from the motion parameters gives

$$M_p(m) = T_{xyz}(t).R_x(\phi).R_y(\theta).R_z(\psi). \tag{14}$$

This transformation matrix, in homogeneous coordinates, represents the evolution of the motion of a given vector according to the 6 DOF parameters m.

In order to retrieve the motion parameters m, the following bundle adjustment formulation of the reprojection error function is minimized:

$$S(m) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{q} r_j \left(m, X^{(i)}\right)^2 \tag{15}$$

where r_j represent the residuals that are functions of the motion vector m. $X^{(i)}$ correspond to the observations. By observations, it meant the 3-D coordinates obtained from the triangulation of matched features across a stereo image pair. According to [37], Gauss–Newton optimization postulates that the optimal solution m to (15) can be computed in an iterative manner by calculating an increment δm at each iteration using the Jacobian matrix $J \equiv dr/dm$ of the residuals vector with respect to the motion parameters m as

$$(J^T.J).\delta m = -J^T.r \tag{16}$$

where $r \in \mathbb{R}^n$ is the residual vector and $(J^T.J)$ represents an approximation of the Hessian matrix [37]. There are typically two reprojection strategies for motion estimation where either points from the previous pair are reprojected into the current frame or the other way round. However, as stated in [5], combining both reprojections yields better estimates of the motion. Following the same strategy, the residuals are defined as $r_d \in \mathbb{R}^6$

$$r_d = \left(r_f^T, r_b^T\right)^T \tag{17}$$

where

$$r_f = y_k - \hat{y}_k = y_k - f\left(M_k(\hat{m}).X_{k+1}\right) \tag{18}$$

$$r_b = y_{k+1} - \hat{y}_{k+1} = y_{k+1} - f\left(M_k^{-1}(\hat{m}).X_k\right). \tag{19}$$

In (18), $\hat{y}_k$ correspond to the estimated coordinates of the feature on the previous camera frame. Similarly, in (19), $\hat{y}_{k+1}$ are the estimated coordinates of the feature on the current frame.

C. Outlier Rejection

In order to improve the accuracy of the motion estimation, the algorithm has to get rid of outliers. These are generally caused by matched features belonging to nonstationary objects or simply undetected false matches from the matching process. One way to deal with outliers is constraining the reprojection error residuals relative to a feature to be bound by a user-defined threshold ϵ. This constraint is expressed by

$$\left(\sum_{j=1}^{q} r_j \left(m, X^{(i)}\right)^2\right) < \epsilon. \tag{20}$$

To this end, the bundle adjustment estimation is wrapped in a RANSAC scheme. At each iteration, three matched points are randomly selected to estimate the motion parameters. The rest of the points are tested and classified as inliers or outliers according to (20). The winning solution with the largest number of inliers is then used to refine the motion parameters m.

D. Additional Constraints

The randomly selected three points for motion estimation need to be spread across the image. Therefore, we incorporated an additional constraint into the algorithm. Let $a(x_a, y_a)$, $b(x_b, y_b)$, and $c(x_c, y_c)$ be the three candidate points in the image. If the area covered by the triangle abc is larger than a user-defined portion of the image area then the points are accepted. Combining this constraint with the feature extraction constraint ensures that motion estimation is carried out on features that are spread across the image. Thus, near and far objects are considered to obtain more precise egomotion estimation.

E. Large Motion Challenges

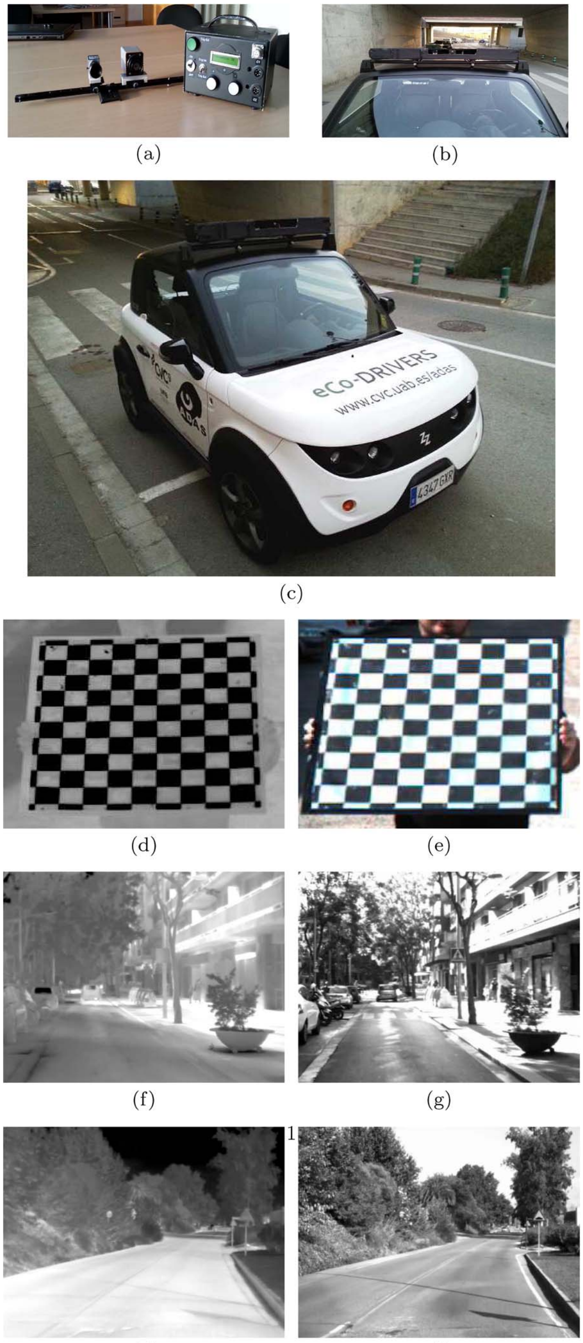
As it will be shown in the experimental section, the proposed approach provides promising results. However, during testing, it came to our attention that the temporal matching might occasionally fail to find correspondences in some specific and challenging conditions. This is particularly true when the vehicle goes into speed bumps at a relatively high speed. Losing sequential matches means that egomotion cannot be estimated at those frames and therefore motion information is lost. We investigated a solution to tackle this problem. Instead of using descriptors based matching with search windows, we opted for an optical flow-based method, namely, the pyramidal implementation of the Lucas– Kanade (LK) tracker [39] for feature tracking. The choice of the pyramidal version was motivated by the fact that it can deal with large motions by using different image scales in the tracking process. The original LK tracker [40] handles only small pixel displacements. However, optical flow cannot be used on IR images due to the reasons mentioned in Section I-B. Therefore, tracking was instead performed on visible images. Matching features is carried out following the same strategy explained in Section II-C. However, temporal matches are obtained using optical flow within the visible images. The loop cannot be closed as before. Instead, the descriptor of the starting feature and the one computed from stereo matching between (im_{L_t}) and (im_{R_t}) are compared. If the distance between them exceeds a user-defined threshold, then the loop is closed, and the match is accepted. Otherwise, the match is rejected.

IV. Experimental Results

A. Setup Overview

This section details the multispectral stereo head used in our experiments together with the calibration and rectification steps. Fig. 5 shows an illustration of the whole platform consisting of the stereo head [see Fig. 5(a)] and the electric car [see Fig. 5(c)] used to generate the data sets.

The stereo head consists of a pair of cameras separated by a baseline of about 12 cm and a nonverged geometry. One camera works in the IR spectrum, more precisely long-wavelength IR and is referred to as FIR. It detects radiations in the range of 8–14 μm. The other camera, which is referred to as visible (VS), responds to the visible spectrum. Images captured by the multispectral stereo head are calibrated and rectified using [41]; a process similar to the one presented in [42] is followed. It consists of a reflective metal plate with an overlain chessboard pattern. This chessboard can be visualized in both spectrums making possible the cameras' calibration and image rectification. Fig. 5 shows a pair of calibration images.



(a) (b)

(c)

(d) (e)

(f) (g)

(h) (i)

Fig. 5. Dataset acquisition system, calibration images, and typical set of images from semiurban and rural scenarios. (a) Multispectral stereo rig. (b) Stereo rig mounted on the car. (c) Electric vehicle used as mobile platform to generate the data sets. (d) IR image of the checkerboard pattern. (e) Visible image of the checkerboard. (f) and (g) Typical pairs of stereo images from urban sequences. (h) and (i) Typical pairs of stereo images from rural sequences.

The FIR camera (Gobi-640-GigE from Xenics) captures images up to 50 fps with a resolution of 640 × 480 pixels. The VS camera is an ACE acA645-100gc from Basler that provides up to 100 fps with a resolution of 658 × 492 pixels. Both cameras are synchronized using an external trigger [see Fig. 5(a)]. The focal lengths of the cameras were set so that pixels in both images contain a similar amount of information of the observed scene. The whole platform is placed on the roof of a vehicle for driving assistance applications [see Fig. 5(c)].

TABLE I
MULTISPECTRAL VIDEO SEQUENCES USED FOR EXPERIMENTS

Video #	Scenario Type	Travelled Distance (m)
Vid00	Urban	240
Vid01	Urban	470
Vid02	Urban	450
Vid03	Rural	350
Vid04	Rural	260
Total		**1770**

TABLE II
ESTIMATED MO VARIATION WITH NUMBER OF DETECTED FEATURES

# Detected Features	MO (m)	Distance Error $(\%)$	RMSE (m)
500	173	29.88	5.90
1000	231	5.71	4.21
1500	236	3.67	3.22
2000	216	11.84	3.65
2500	224	8.5	3.53

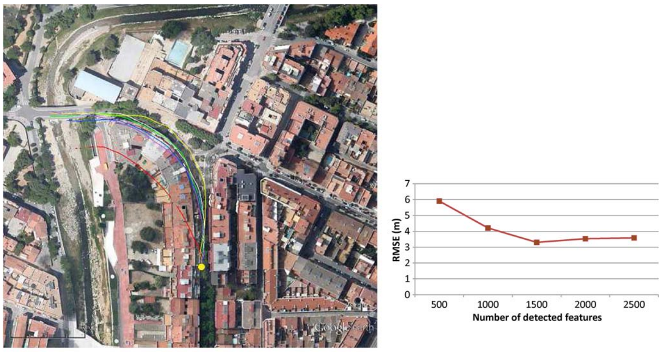
Once the FIR and VS cameras have been calibrated, their intrinsic and extrinsic parameters are estimated. Additionally, ground truth geopositional information is obtained from a low-cost GPS connected to the acquisition system. Hence, every frame from the camera is enriched with the latest latitude-longitude information from the GPS. However, the GPS data is updated almost 2 times per second, which is considerably slower than the camera frame rate.

B. Results

Our MO technique was tested against a series of real outdoor sequences captured from our experimental vehicle. These scenarios are split into semiurban and rural scenarios (see Fig. 5) and detailed in Table I. The former are richer in terms of extractable features than the latter. However, at the same time, they present more probabilities of containing nonstationary objects (i.e., vehicles, pedestrians…). All these sequences represent real traffic conditions with strong illumination variations and lack of texture. The texture issue applies more to thermal images and can be explained by the fact that FIR pixel brightness depends on heat variations. Most of the lower part of images is composed of ground, where heat does not vary a lot. This means that this part of the image would be textureless and therefore not used in the matching process. Unfortunately, this limitation impacts the motion estimation in a sense that closer objects within the scene cannot be included in the computation process.

1) Feature Extraction and Matching Analysis:

Number of Detected Features: Here, we show the influence of the number of detected features on the estimated MO. With this approach, thresholds used to extract interest points are implicitly adapted (i.e., adaptive thresholding) in order to get the same feature numbers in thermal and visible images. Experiments were conducted using different amounts of features $n = \{500, 1000, 1500, 2000, 2500\}$. Table II shows the distance errors $(\%)$ as well as the RMSE (in $meters$) for different values of extracted features and Fig. 6(a) illustrates the estimated trajectory for each value on an initial data set. It corresponds to Vid00 in Table I, which consists in an urban sequence containing moving objects and a left bend. The estimated MO accuracy increases with the number of detected features. However, in our experiments, it reaches a maximum and stabilizes at a certain number of features (1500), as shown in Fig. 6(b). This could relate to the added number of false matches not detected in the matching process.



(a) (b)

Fig. 6. Influence of the number of detected features on the motion estimation. (a) MO trajectories (yellow line: GPS; red line: 500; magenta line: 1000; green line: 1500; blue line: 2000; white line: 2500). (b) Graph showing the variation of the RMSE with respect to the number of detected features.

Feature Correspondence: Here, we present some results regarding the performance of feature correspondence. We compare our correspondence approach to state-of-the-art intensity-based algorithms, namely, SIFT and SURF, as well as binary techniques such as BRISK [43] and ORB [44]. Note that we do not dispose of the disparity ground truth and the comparisons in terms of correct matches are qualitative. A visual inspection was carried out to determine good matches when produced by the algorithms. All the algorithms were implemented using the OpenCV library [45] with heuristically tuned parameters. The tuning process was independently carried out for the visible and thermal images. This comes from the observation that setting the same thresholds for both modalities led to disparities in terms of the number of extracted features. For instance, setting a similar contrast threshold (e.g., 0.04) for the SIFT detector yields the extraction of 503 features for the thermal image in contrast to 2781 for the visible image. In order to have a fair comparison, the parameters of all the tested detectors/descriptors were tuned to have an equal number of features in both modalities (2000). In addition, the epipolar constraint is imposed on all detectors when computing matches to keep only those on the same scan line (rectified images). Indeed, when this constraint is not implemented, the intensity-based detectors provide almost no match [see Fig. 7(c)]. This observation is valid for the other algorithms as well. Fig. 7 shows a sample of the results obtained when using intensity-based algorithms. Results obtained using our approach on the same image pair are shown in Fig. 8(f) (to avoid duplication). Table III summarizes the results obtained by the tested algorithms on a sample of 100 images from Vid01. In general, all tested techniques provide very poor correspondences. BRISK and SURF struggle to extract features from thermal images despite tuning of their parameters. However, SIFT extracts the required number of

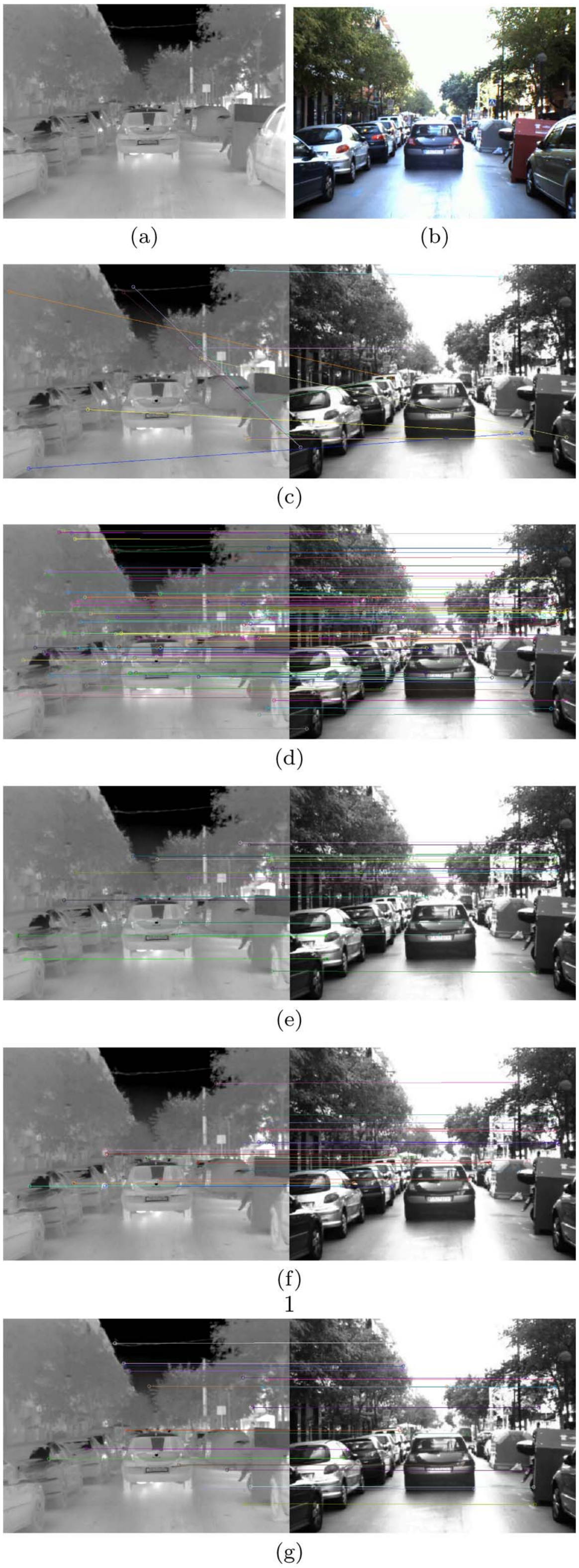


Fig. 7. Feature matching results using intensity-based detectors/descriptors. (a) and (b) Original thermal and visible images. (c) Correspondence using SURF without enforcing epipolar constraint. (d) Matching using SURF with epipolar constraint. (e) SIFT matching results. (f) ORB matching results. (g) BRISK matching results.

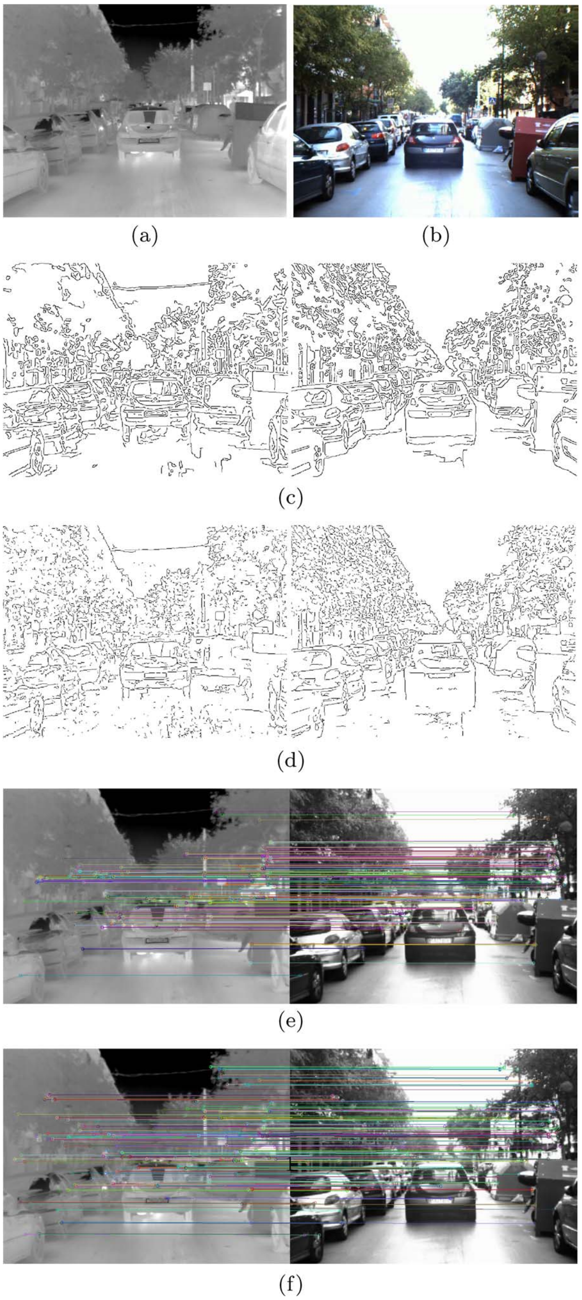


Fig. 8. PC edge map versus Canny edge map. (a) and (b) Original thermal and visible images. (c) Corresponding edge maps using Canny. (d) Corresponding edge maps using PC. (e) Matching results using Canny. (f) Matching results using PC.

TABLE III
Feature Correspondence Results on Sample Images

Method	Detected (IR)	Matched	Ratio
SIFT	1983	18	0.956
SURF	1548	135	8.77
ORB	1739	9	0.53
BRISK	1323	25	1.90
Our method	2000	620	31

features once tuned. The ratio matched/detected highlighted in Table III indicate that intensity-based algorithms provided poor correspondence in the multispectral scenario. When inspecting the matches visually, it seems that on average these algorithms find 3∼six correspondences successfully. In contrast, our approach provides around 250∼300 matches in average.

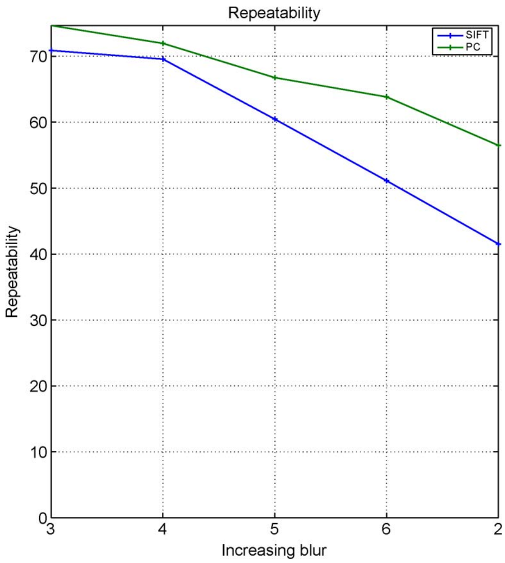
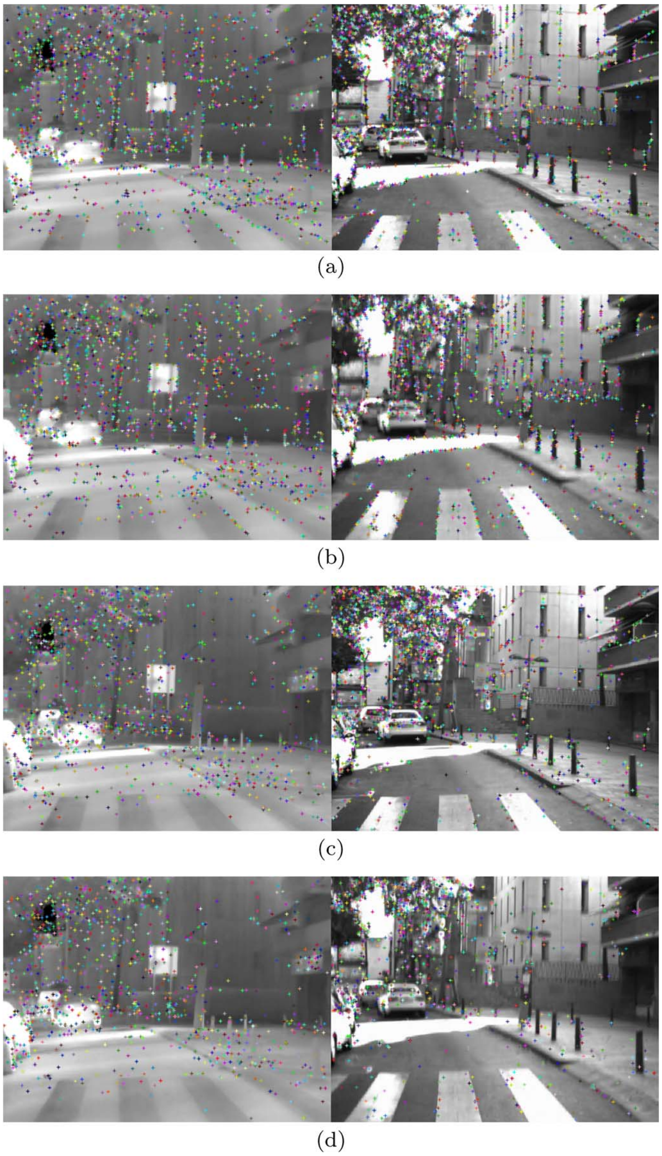


Fig. 9. Repeatability results on the bikes dataset.

Edge Filter: We also show the influence of the edge detector used to compute features' descriptors on the matching performance. For that, we compare the PC edge map (currently implemented in the descriptor) with the well-known Canny edge detector. Once again, different thresholds had to be independently tuned for the visible and thermal images for the Canny detector, whereas the same values are used with PC. The tests show that no significant gain is obtained when replacing the PC edge map by Canny's [see Fig. 8(c) and (d)]. In addition, it would induce more computational burden on the algorithm (when using Canny). This is due to the fact that when using PC, both corners and edges are computed in one run. Another aspect related to the Canny operator is that lines are detected twice unlike PC, which provides a single response. The same observation was also made in [46]. In general, the matching algorithm gives roughly the same numbers of features using PC and Canny [see Fig. 8(e) and (f)].

Motion Blur: Image blur induced by the motion of the sensing system introduces more challenges on most feature extraction algorithms. In order to show the robustness of our algorithm to blurring effects, we used the *bikes* data set [47] to evaluate the performance of the feature extraction compared with SIFT. The data set contains images captured with varying image blur. The VLBenchmark [48] was used to compute the repeatability of both extractors. The repeatability can be defined as the percentage of detected features that survive some transformation or disturbance (blur) between two images. The higher it is, the better is the detection algorithm. The results are shown in Fig. 9, where it can be clearly seen that PC outperforms SIFT. In addition, we noticed that motion blur occurred in some parts of our data sets. For this reason, we picked sample images from Vid01 (Fig. 10) to compare the feature extraction on blurred and normal images representing the same scene (acquired at 0.1-s interval). Once again, the frequency-based extractor performs better than SIFT. Indeed, in Table IV, we can see that even in the presence of blurred images, PC can extract the required 2000 features (without modifying the extraction parameters), whereas SIFT barely extracted half the required number of features.


(a)

(b)

(c)

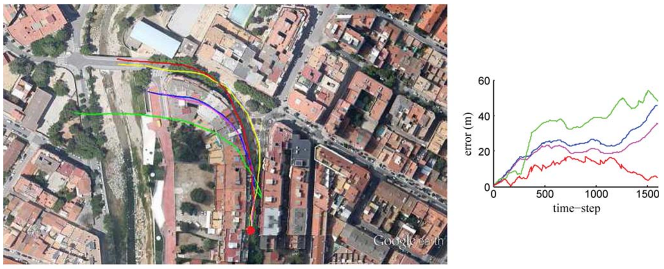
(d)

Fig. 10. Feature extraction on a sample pair of images from Vid01. (a) and (b) PC feature extraction on normal and blurred images, respectively. (c) and (d) SIFT feature extraction on normal and blurred images, respectively.

TABLE IV
EXTRACTED FEATURES WITH AND WITHOUT MOTION BLUR

image	PC		SIFT	
	IR	VS	IR	VS
Non-blurred	2000	2000	1513	2000
Blurred	1934	2000	1227	982

2) Outlier Removal: Experiments were also conducted to test the motion estimation subprocess. To this end, different outlier rejection threshold values were used to determine their influence on the MO outputs for Vid00. Setting a threshold value underlines a tradeoff because high values allow more wrong matches, whereas low values might discard valid matches. Fig. 11(a) illustrates the obtained results on the same sequence (Vid00). The resulting travelled errors are shown in Fig. 11(b) for the different values $\epsilon = \{1.5, 3.5, 5.5\}$. Fig. 11 also shows the effect of not using the outlier rejection scheme within



(a) (b)

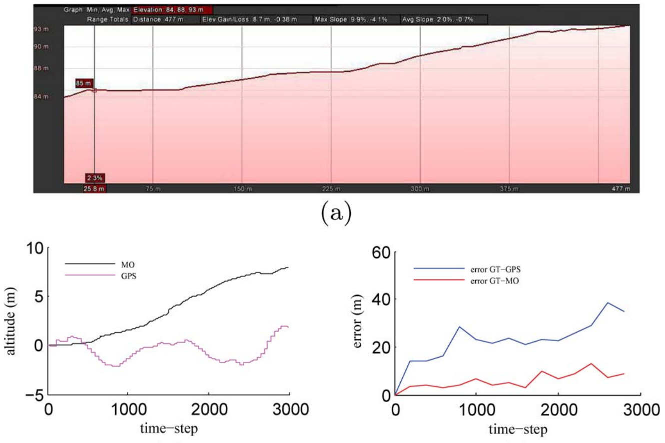
Fig. 11. Influence of the outlier rejection threshold on the motion estimation. (a) MO trajectory (yellow line: GPS; red line: $\epsilon = 1.5$; blue line: $\epsilon = 3.5$; magenta line: $\epsilon = 5.5$; green line: without outlier rejection) (b) corresponding travelled errors (same color legend applies).

the estimation of the trajectory. Note the divergence of the outputted trajectory with respect to the other thresholds. As expected, the accuracy of the motion estimation decreases with the increase of the outlier rejection threshold. This is due to the fact that more outliers are used in the process when the threshold is increased.

3) Multimodal Odometry: The following results are based on both types of sequences namely semiurban (Vid01 & Vid02) and rural (Vid03 & Vid04). Note that the term *semiurban* is used instead of *urban* as a good portion of the images contain vegetation [see Fig. 5(g)]. Vid01 represents the simplest scenario, where the vehicle is travelling along a straight road. Vid02 is a more challenging data set within the same environment, which contains speed bumps and bends. Both sequences have proven to be challenging as many nonstationary objects and significant illumination variations were experienced. Vid03 corresponds to a straight road followed by a left bend in a rural environment, where moving vehicles were overtaking our car and where severe lighting conditions were encountered at and after the bend. Vid04 represents a U-turn at a roundabout, where MO suffered from blurred images at the level of the roundabout. The major challenge in this type of scenario is that images lack of nearby features. It is mainly due to the nature of thermal imagery, where thermal response of the road varies less than in the visible spectrum. This causes the corresponding image region to be textureless. The direct impact is an underestimation of the motion as noted in [49]. It is believed to be due to the lack of close-by features combined to the short baseline of the stereo rig. This issue will be addressed in future work.

Based on the previous results (Section IV-B1 and B2), a fixed number of features (1500) is used. This guarantees a reasonable amount of matches for odometry. In addition, an outlier rejection threshold $\epsilon = 1.5$ is selected.

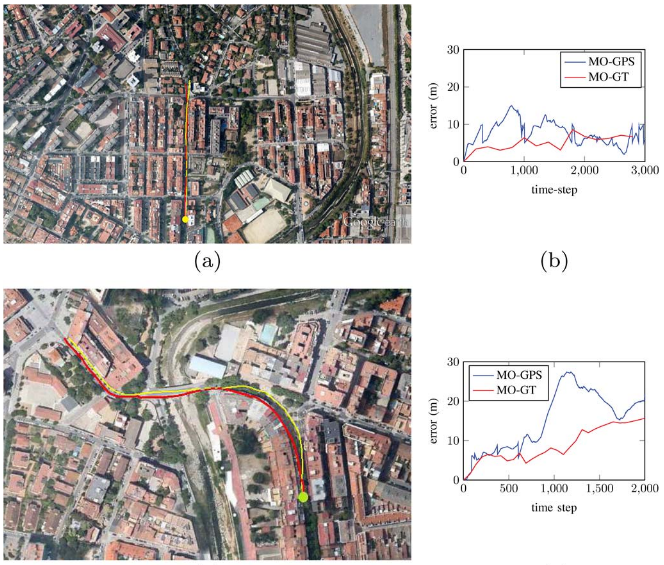
As stated in Section IV-A, the geopositional information is provided by a low-cost GPS considered as 'drifty' ground truth. For this reason, a more precise Google Earth-based ground truth (GT) was manually generated. It was created by introducing control points (based on the images) every 100 frames in Google Earth. This allowed us to obtain a more precise ground truth for comparison with MO estimated trajectories. Fig. 12(b) illustrates the altitudes estimated by our MO compared with the GPS readings for Vid01. Fig. 12(a) represents the elevation profile extracted from Google Earth corresponding to the same sequence. It shows that our MO estimates are far more accurate than the GPS measurements. The same observation applies for the estimated trajectory (of the same sequence) as it can be noted in Fig. 12(c). It illustrates errors between Google Earth-based GT and GPS measurements as well as between GT and MO. Note that errors between GT-GPS are larger than between GT-MO. Following these findings, the same strategy was adopted for all the sequences, where two error graphs are always plotted along with the estimated trajectory. The first error is computed for every frame between MO and the GPS readings that are linearly interpolated due to their low update rate. The second error is the one computed every 100 frames between GT and MO. The errors that are adopted to evaluate the MO performance are based on the GT-MO graphs. Therefore, hereinafter, errors are always expressed from the GT-MO graphs.



(a)

(b) (c)

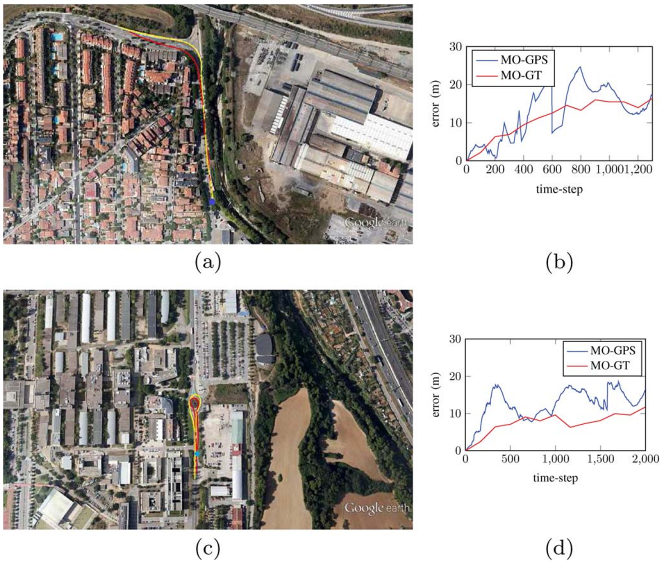
Fig. 12. Comparison of the MO estimate of the altitude against GPS measurements for Vid01. (a) Google Earth elevation profile of the trajectory. (b) MO estimation of the altitude against GPS measurements. (c) Errors between GT-GPS and GT-MO.



(a) (b)

(c) (d)

Fig. 13. MO trajectories and travelled errors for semiurban sequences. (a) and (c) MO trajectories for Vid01 and Vid02, respectively (yellow line: GPS; red line: MO; yellow circle: starting point). (b) and (d) Corresponding travelled errors.



(a) (b)

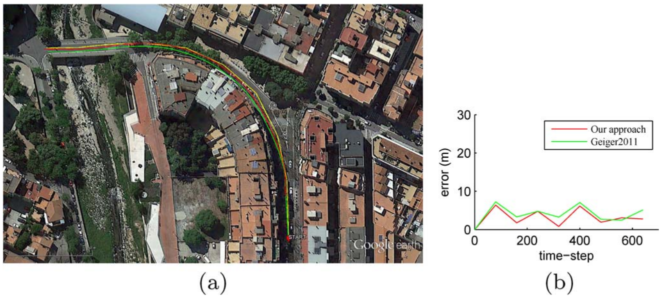
(c) (d)

Fig. 14. MO trajectories and travelled errors for rural sequences. (a) and (c) MO trajectories for Vid03 and Vid04, respectively (yellow line: GPS; red line: MO; blue circle: starting point). (b) and (d) Corresponding travelled errors.

Fig. 13 illustrates the trajectories computed by the GPS and MO for the semiurban sequences (Vid01, Vid02). The peaks in errors (MO-GPS) are due to the imprecision of the GPS and are given as indication only. From Fig. 13(b) and (d), it can be seen that the achieved results are successful reaching errors as low as 2% and 3% for Vid01 and Vid02, respectively, and defined as

$$\text{error}(\%) = \frac{100.\text{mean}(\text{errors})}{\text{travelled distance}}. \tag{21}$$

These errors do not correspond to the ones commonly provided in the literature and defined as the ratio of the last offset (endpoint) to the travelled distance. The latter errors do not provide information on the behavior of the system along the whole trajectory in contrary to the errors provided here. The estimated trajectories from the rural sequences are shown in Fig. 14. Errors obtained in this scenario (rural environment) are slightly higher than in urban sequences for the aforementioned reasons. These errors are shown in Fig. 14(b) and (d). They correspond to 5% and 4% for Vid03 and Vid04, respectively. In general, the system is able to temporally track 40–50 features due to the textureless nature of thermal imagery. However, in the case of rural scenarios, most of them correspond to far features therefore increasing the errors in the estimation process. In severe lighting conditions, the number of tracked features falls considerably (6–10), making the motion estimation even noisier. Restrictions imposed by the RANSAC-based outlier rejection deal with the wrong matches and allow more robust estimations. Features on nonstationary objects are also discarded by the same restrictions.



(a) (b)

Fig. 15. Comparison of the estimated trajectories on the visible band data set. (a) MO trajectories (yellow line: GPS; red line: our approach; green line: Geiger_2011). (b) Corresponding travelled errors.

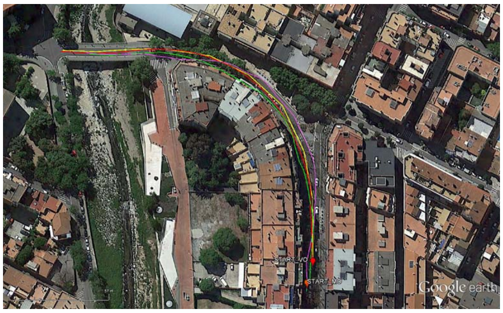
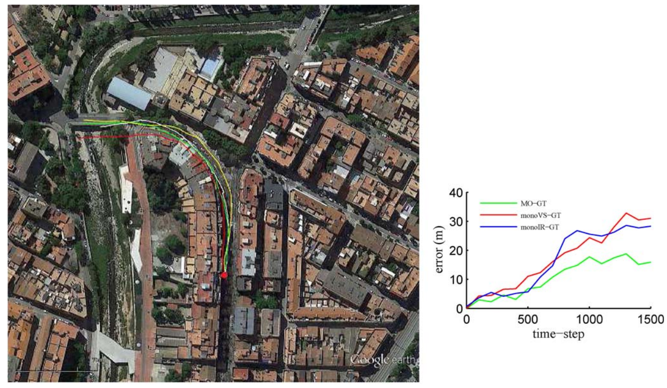


Fig. 16. Illustration of MO and VO on similar data sets.

4) Visible Band Odometry: In order to be more complete in the assessment of the proposed approach, we also compare our approach to the algorithm introduced in [14] (Geiger_2011) using a pure visible data set captured in the same conditions (same route). This is mainly added to demonstrate the validity of our approach in the pure visible band-based VO. For this particular experiment, we used a stereo rig that consists in two IDS ueye cameras (resolution: 1280 × 1024) capturing images at 10 fps. We recorded the same trajectory as in Vid00 using this setup. For the Geiger_2011 algorithm, we used the parameters reported in their paper. As shown in Fig. 15, our approach yields a better estimate of the trajectory. In average, our approach computes 400 matches with 50% inliers, whereas Geiger_2011 obtains 300 matches with 35% inliers. Inliers are defined as the correct matches whose reprojection errors are below a user-defined threshold (20). Note that the same threshold ($\epsilon = 1.5$) was used for both algorithms. In terms of travelled errors (Fig. 15(b)), the proposed framework outperformed Geiger_2011 achieving errors of 1.28% compared with 1.66% for the latter.

5) Multispectral Versus Visible Odometry: The visible band data set generated for the experiment described in Section IV-B4 was also used to compare the behavior of our approach on multispectral data sets with respect to visible band data sets. We illustrate the estimated trajectories obtained from both data sets on a similar sequence (see Fig. 16). In other words, we captured the same route using both types of setups each with its own ground truth. As shown in Fig. 16, one can see that quite similar results are obtained when using multispectral or visible band data sets. A quantitative comparison is not possible at this stage and will be addressed in future work.

6) MO Versus Monocular VO: Here we show a comparison between trajectories estimated using the proposed MO scheme and a classical monocular VO algorithm. In order to obtain a fair comparison, images from both the visible spectrum and thermal cameras were used to estimate the trajectory. Moreover, we used the implementation of [50] with finely tuned parameters to compare the results. Their monocular VO approach uses $L\infty$ norm with convex optimization for the motion estimation part. SIFT is used to extract features, and matching is based

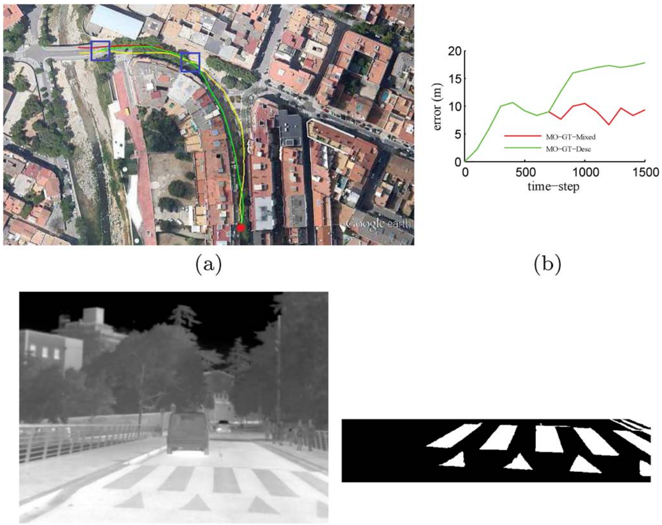


(a) (b)

Fig. 17. Comparison of MO and monocular VO. (a) MO and monocular VO trajectories (yellow line: GPS; green line: MO; red line: monocular VO using visible band images; white line: monocular VO using IR images). (b) Corresponding travelled errors.

on the Euclidean distance between interest points. To deal with the scale ambiguity relative to the projective effects in monocular systems, they use an $H\infty$ filter for the frame-to-frame scale estimation. As shown from Fig. 17(a), the proposed MO algorithm outperforms the monocular scheme in terms of the estimated trajectory and the travelled errors [see Fig. 17(b)]. In addition, Fig. 17 shows that monocular VO provides less accurate estimates than MO for both types of imagery (i.e., visible and thermal). However, although the travelled errors from monocular VO are approximately the same, each modality coped differently with the images of the travelled path. The end-points of the two monocular algorithms in Fig. 17(a) as well as the complete trajectories illustrate it clearly. This enforces our motivation to take advantage of the available multimodal setup to estimate the motion of the vehicle rather than using each camera separately in a monocular fashion.

7) Optical Flow-Based Strategy for Challenging Sequences: Fig. 18 illustrates the improvement obtained by switching the temporal feature matching technique for sequences containing speed bumps. In their presence, optical flow is used for a number of frames to track interest points within the visible images. The descriptor-based matching fails in these images to obtain good sequential correspondences. However, in normal operating conditions, the descriptor-based matching provides the egomotion estimation subtask with less outliers than the optical flow tracking. For this reason, it is utilized throughout the sequences except when the vehicle encounters speed bumps and large motions may occur. In this case, the MO output is improved by approximately 15%, as shown in Fig. 18(b). Note that speed bumps are detected using a simple yet effective approach. Speed bumps are usually marked on the ground using rectangular and triangular patterns in white paint. Therefore, the lower part of the image is analyzed every N frames to look for a similar pattern as in Fig. 18(d). Histogram thresholding and blobs detection are used on FIR images to extract the sought shapes. FIR images are used due to the large contrast between the painted and nonpainted regions of the road induced by the change in their thermal reflection (due to the paint). A basic template matching scheme is then applied to test if the patterns are present within the bottom part of the image. When detected, the matching scheme automatically switches to the optical flow tracking for a limited number of frames (i.e., time to get past the bump).



(a) (b)

(c) (d)

Fig. 18. Improvement of multimodal odometry in sequences containing speed bumps. (a) MO trajectories (yellow line: GPS; red line: MO with mixed matching; green line: MO with descriptor matching only; red dot: starting point; blue rectangles: positions of the speed bumps on the road). (b) Travelled errors for both strategies. (c) Speed bumps painted pattern. (d) Processed bottom part of the image showing the sought pattern.

V. Conclusion and Future Work

A multispectral stereo odometry solution has been introduced. To the best of our knowledge, it represents the first attempt in the literature. Features are extracted using a frequency-based detector, namely, PC, and described using a combination of spatial and frequency information. Motion is retrieved using a sliding WBA incorporating Gauss–Newton optimization and RANSAC for outlier removal. The experimental part involved the setting-up of a multimodal stereo rig on a vehicle and the capture of our own data sets. Tests were performed under real traffic conditions. Shown results validate our approach and more importantly demonstrate the possibility to achieve MO. As part of our future work, we intend to improve the temporal matching scheme using available on-board sensors such as the IMU in order to guide the search region for feature matching. In addition, a pure night-time VO work is currently under investigation using either a monocular approach, where only one thermal camera is used or a stereo framework by using two IR cameras (if available). In addition, we are investigating other motion estimation techniques, where the aim is to further improve the accuracy obtained from multimodal odometry. Finally, generating data sets with varying distances, weather conditions, and times of the day while incorporating different types of stereo cameras is also under work.

Acknowledgment

The authors would like to thank TIN2011-25606 project, German Ros and the ADAS group of CVC. C. A. Aguilera-Carrasco was supported by "Universitat Autonoma de Barcelona."

REFERENCES

[1] H. P. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 1980.

[2] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, vol. 1, pp. 652–659.

[3] A. Nemra and N. Aouf, "Robust airborne 3D visual simultaneous localization and mapping with observability and consistency analysis," *J. Intell. Robot. Syst.*, vol. 55, no. 4/5, pp. 345–376, Jan. 2009.

[4] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *J. Field Robot.*, vol. 23, no. 1, pp. 3–20, Jan. 2006.

[5] D. Rodriguez and N. Aouf, "Robust egomotion for large-scale trajectories," in *Proc. IEEE Int. Conf. MFI*, Sep. 2012, pp. 156–161.

[6] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2010, pp. 486–492.

[7] J. Hedborg, P. Forssén, and M. Felsberg, "Fast and accurate structure and motion estimation," in *Proc. 5th Int. Symp. Adv. Vis. Comput., Part I*, Las Vegas, NV, USA, 2009, pp. 211–222.

[8] V. Grabe, H. H. Bulthoff, and P. Robuffo Giordano, "Robust optical-flow based self-motion estimation for a quadrotor UAV," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 2153–2159.

[9] J.-P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 2531–2538.

[10] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 4293–4299.

[11] M. Kharbat and N. Aouf, "Recursive estimation of three-dimensional unmanned aerial vehicle motion and structure based on the L-norm," *Proc. Inst. Mech. Eng. Part G, J. Aerosp. Eng.*, vol. 226, no. 7, pp. 751–762, Nov. 2011.

[12] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *J. Field Robot.*, vol. 24, no. 3, pp. 169–186, Mar. 2007.

[13] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel tracking and iterative closest point," in *Proc. 4th IEEE ICVS*, 2006, p. 21.

[14] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3d reconstruction in real-time," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2011, pp. 963–968.

[15] O. Pink, F. Moosmann, and A. Bachmann, "Visual features for vehicle localization and ego-motion estimation," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2009, pp. 254–260.

[16] I. Parra Alonso *et al.*, "Accurate global localization using visual odometry and digital maps on urban environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1535–1545, Dec. 2012.

[17] S. Nedevschi, V. Popescu, R. Danescu, T. Marita, and F. Oniga, "Accurate ego-vehicle global localization at intersections through alignment of visual data with digital map," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 673–687, Jun. 2013.

[18] L. R. García Carrillo, A. E. Dzul López, R. Lozano, and C. Pégard, "Combining stereo vision and inertial navigation system for a quadrotor UAV," *J. Intell. Robot. Syst.*, vol. 65, no. 1–4, pp. 373–387, Aug. 2011.

[19] M. Agrawal, K. Konolige, and R. Bolles, "Localization and mapping for autonomous navigation in outdoor terrains: A stereo vision approach," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Feb. 2007, pp. 7–7.

[20] M. Agrawal and K. Konolige, "Real-time localization in outdoor environments using stereo vision and inexpensive GPS," in *Proc. 18th ICPR*, 2006, pp. 1063–1068.

[21] K. Konolige, M. Agrawal, and J. Sola, "Large-scale visual odometry for rough terrain," in *Proc. Int. Symp. Robot. Res.*, 2007, pp. 201–212.

[22] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. CVPR*, 1994, pp. 593–600.

[23] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vis. Conf.*, 1988, pp. 147–152.

[26] S.-H. Jung, J. Eledath, S. Johansson, and V. Mathevon, "Egomotion estimation in monocular infra-red image sequence for night vision applications," in *Proc. IEEE WAC V*, Feb. 2007, p. 8.

[27] M. Magnabosco and T. P. Breckon, "Cross-spectral visual simultaneous localization and mapping (SLAM) with sensor handover," *Robot. Auton. Syst.*, vol. 61, no. 2, pp. 195–208, Oct. 2012.

[28] J. H. Lim, O. Tsimhoni, and Y. Liu, "Investigation of driver performance with night vision and pedestrian detection systems—Part I: Empirical study on visual clutter and glance behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 670–677, Sep. 2010.

[29] S. Lin, "Review: Extending visible band computer vision techniques to infrared band images," Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep. MS-CIS-01-04, pp. 1–23, 2001.

[30] T. Mouats and N. Aouf, "Multimodal stereo correspondence based on phase congruency and edge histogram descriptor," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2013, pp. 1981–1987.

[31] M. C. Morrone and R. Owens, "Feature detection from local energy," *Pattern Recog. Lett.*, vol. 6, no. 5, pp. 303–313, Dec. 1987.

[32] M. C. Morrone and D. C. Burr, "Feature detection in human vision: A phase-dependent energy model," *Proc. R. Soc. London B Biol. Sci.*, vol. 235, no. 1280, pp. 221–245, Dec. 1988.

[33] P. Kovesi, "Phase congruency detects corners and edges," in *Proc. Australian Pattern Recog. Soc. Conf., DICTA*, 2003, pp. 309–318.

[34] N. J. W. Morris, S. Avidan, W. Matusik, and H. Pfister, "Statistics of infrared images," in *2007 IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.

[35] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Jun. 1986.

[36] D. Marr and E. Hildreth, "Theory of edge detection," *Proc. R. Soc. London Series B, Containing papers Biol. Character R. Soc. (Great Britain)*, vol. 207, no. 1167, pp. 187–217, Feb. 1980.

[37] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—A modern synthesis," in *Proc. ICCV*, London, U.K., 2000, vol. LNCS 1883, pp. 298–372.

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

[39] J.-Y. Bouguet, *Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm*. Santa Clara, CA, USA: Microprocessor Res. Labs, Intel Corp., 2001.

[40] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. IJCAI*, 1981, pp. 674–679.

[41] J. Y. Bouguet, *Camera Calibration Toolbox for Matlab*. Natick, MA, USA: MathWorks, 2008.

[42] F. Barrera Campo, F. Lumbreras Ruiz, and A. D. Sappa, "Multimodal stereo vision system: 3D data extraction and algorithm evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 437–446, Sep. 2012.

[43] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2548–2555.

[44] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[45] G. Bradski, "The openCV library," *Dr. Dobb's J. Softw. Tools*, vol. 27, no. 7, pp. 122–125, Jul. 2000.

[46] P. Kovesi, "Phase congruency: A low-level image invariant," *Psychological Res.*, vol. 64, no. 2, pp. 136–148, Jan. 2000.

[47] K. Mikolajczyk and C. Schmid, "Performance evaluation of local descriptors.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[48] K. Lenc, V. Gulshan, and A. Vedaldi, Vlbenchmkars, 2011. [Online]. Available: http://www.vlfeat.org/benchmarks/

[49] J.-P. Tardif, M. George, M. Laverne, A. Kelly, and A. Stentz, "A new approach to vision-aided inertial navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, no. 2008, pp. 4161–4168.

[50] M. Boulekchour and N. Aouf, "L_Inf norm based solution for visual odometry," in *Proc. CAIP*, 2013, pp. II:185–II:192.

Tarek Mouats (S'13) received the computer science engineering degree from the National Polytechnic School, Algiers, Algeria, in 2005 and the M.Sc. degree in defense sensors and data fusion from Cranfield University, Shrivenham Campus, Shrivenham, U.K., in 2008, where he is currently working toward the Ph.D. degree in the Department of Informatics and Systems Engineering.

His research focuses on multimodal image processing, intelligent transportation systems, localization techniques, and more specifically visual odometry.

Nabil Aouf (M'13) is a Reader with the Centre of Electronic Warfare, Cranfield University, Shrivenham, U.K. He has more than 100 publications of high calibre in his domains of interest. His research interests are aerospace and defense systems, information fusion and vision systems, guidance and navigation, tracking and control, and autonomy of systems.

Dr. Aouf is an Associate Editor of the INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE IN CONTROL.

Angel Domingo Sappa (S'94–M'00–SM'12) received the electromechanical engineering degree from National University of La Pampa, General Pico, Argentina, in 1995 and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1999.

In 2003, after holding research positions in France, the U.K., and Greece, he joined the Computer Vision Center, where he is currently a Senior Researcher. He is a member of the Advanced Driver Assistance Systems Research Group. His research interests span a broad spectrum within the 2-D and 3-D image processing. His current research focuses on stereo image processing and analysis, 3-D modeling, cross-spectral imaging, dense optical flow estimation, and visual SLAM for driving assistance.

Cristhian Aguilera received the B.S. degree in automation engineer from the Universidad del Bío-Bío, Concepción, Chile, in 2008. He is currently studying the M.S. degree in computer vision with the Universitat Autónoma de Barcelona, Barcelona, Spain.

His current research focuses on cross-spectral imaging.

Ricardo Toledo received the degree in electronic engineering from the Universidad Nacional de Rosario, Rosario, Argentina, in 1986, the M.Sc. degree in image processing and artificial intelligence from the Universitat Autónoma de Barcelona (UAB), Barcelona, Spain, in 1992, and the Ph.D. degree in 2001.

Since 1989, he has been giving lectures with the Department of Computer Science, UAB, and participating in R+D projects. Currently he is a full time Associated Professor. In 1996, he participated in the foundation of the Computer Vision Centre, UAB. He has participated in national and international R+D projects being the leader of some of them and is the coauthor of more than 40 articles, all these in the field of computer vision, robotics, and medical imaging.