Robotics and Autonomous Systems 85 (2016) 26–36

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Monocular visual odometry: A cross-spectral image fusion based approach

Angel D. Sappa [a,b,*], Cristhian A. Aguilera [a,c], Juan A. Carvajal Ayala [b], Miguel Oliveira [d,e], Dennis Romero [b], Boris X. Vintimilla [b], Ricardo Toledo [a,c]

[a] Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain

[b] Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ingeniería en Electricidad y Computación, CIDIS, Campus Gustavo Galindo Km 30.5, Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador

[c] Computer Science Dept., Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain

[d] INESC TEC - INESC Technology and Science, R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[e] IEETA - Institute of Electronics and Informatics Engineering of Aveiro, Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

HIGHLIGHTS

- Monocular visual odometry based on a fused image approach.
- DWT image fusion parameters selected according to a quantitative evaluation metric.
- Experimental results with two public data sets illustrate its validity.
- Comparisons with other approaches are provided.

ARTICLE INFO

Article history:
Available online 26 August 2016

Keywords:
Monocular visual odometry
LWIR-RGB cross-spectral imaging
Image fusion

ABSTRACT

This manuscript evaluates the usage of fused cross-spectral images in a monocular visual odometry approach. Fused images are obtained through a Discrete Wavelet Transform (DWT) scheme, where the best setup is empirically obtained by means of a mutual information based evaluation metric. The objective is to have a flexible scheme where fusion parameters are adapted according to the characteristics of the given images. Visual odometry is computed from the fused monocular images using an off the shelf approach. Experimental results using data sets obtained with two different platforms are presented. Additionally, comparison with a previous approach as well as with monocular-visible/infrared spectra are also provided showing the advantages of the proposed scheme.

1. Introduction

The usage of cross-spectral imaging has been increasing due to the drop in price of cameras working at different spectral bands. That increase is motivated by the possibility of developing robust solutions that cannot be obtained if a single band were used. These robust solutions can be found in domains such as thermal inspection [1], video surveillance [2], face detection [3], driving assistance [4] and visual odometry [5], which is the focus of the current work. Before tackling one of the problems mentioned above, the information provided by the cameras working at different spectral bands needs to be fused into a single and compact representation for further processing, assuming an early fusion scheme is followed.

Visual Odometry (VO) is the process of estimating the egomotion of an agent (e.g., vehicle, human or a robot) using only the input of a single or multiple cameras attached to it. This term has been proposed by Nister [6] in 2004, which has been chosen for its similarity to wheel odometry. In wheel odometry, the motion of a vehicle is obtained by integrating the number of turns of its wheels over time. Similarly, VO operates by incrementally estimating the pose of the vehicle by analyzing the changes induced by the motion in the images of the onboard vision system.

* Corresponding author at: Computer Vision Center, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain.
E-mail addresses: asappa@cvc.uab.es (A.D. Sappa), caguilera@cvc.uab.es (C.A. Aguilera), jacarv@espol.edu.ec (J.A. Carvajal Ayala), m.riem.oliveira@gmail.com (M. Oliveira), dgromero@espol.edu.ec (D. Romero), boris.vintimilla@espol.edu.ec (B.X. Vintimilla), ricard@cvc.uab.es (R. Toledo).

http://dx.doi.org/10.1016/j.robot.2016.08.005
0921-8890/

State of the art VO approaches are based on monocular or stereo vision systems; most of them working with cameras in the visible spectrum (e.g., [7,8]). The approaches proposed in the literature can be coarsely classified into *feature based* methods, *image based* methods and *hybrid* methods. The feature based methods rely on visual features extracted from the given images (e.g., corners, edges) that are matched between consecutive frames to estimate the egomotion. On the contrary to feature based methods, the image based approaches directly estimate the motion by minimizing the intensity error between consecutive images. Finally, hybrid methods are based on a combination of the approaches mentioned before to reach a more robust solution. All the VO approaches based on visible spectrum imaging, in addition to their own intrinsic limitations, have additional ones related with the nature of the images (i.e., photometry). Having in mind these limitations (i.e., noise, sensitivity to lighting changes, etc.) monocular and stereo vision based VO approaches, using cameras in the infrared spectrum, have been proposed (e.g., [9,10]) and more recently cross-spectral stereo based approaches have been also introduced (e.g., [11,5]). The current work proposes a step further by tackling the monocular vision odometry problem with an image resulting from the fusion of a cross-spectral imaging device. The goal behind such an approach is to take advantage of the strengths of each band according to the characteristics of the scenario (e.g., daytime, nighttime, poor lighting conditions, etc.). A difference to a previous approach published in [5] is that in the current work fusion parameters are adapted to the characteristics of the given images.

Image fusion is the process of combining information from two or more images of a given scene into a single representation. This process is intended for encoding information from source images into a single and more informative one, which could be suitable for further processing or visual perception. There are two different cases where image fusion takes place; firstly, the case of images obtained from different sensors (multisensory), which could also work at different spectral band (multispectral). Secondly, the case of images of the same scene but acquired at different times (multitemporal). The current work is focused on the first case, more specifically, fusing pair of images from visible and infrared spectra obtained at the same time by different sensors. It is assumed that the images to be fused are correctly registered [12]; otherwise a process of cross-spectral feature detection and description should be followed in order to find the correspondences between the images (e.g., [13,14]).

During the last decades, the image fusion problem has been largely studied, mainly for remote sensing applications (e.g., [15,16]). Most of these methods have been proposed to produce a high-resolution multispectral representation from a low-resolution multispectral image fused with high-resolution panchromatic one. The difference in image resolution is generally tackled by means of multi-scale image decomposition schemes that preserve spectral characteristics but represented at a high spatial resolution. Among the different proposals, wavelet based approaches have shown some of the best performance by producing better results than standard methods such as intensity–hue–saturation (IHS) transform technique or principal component analysis (PCA) [17]. Wavelet based image fusion consists of two stages. Firstly, the given images are decomposed into two components (more details are given in Section 2.1.1); secondly, the components from the given images are fused in order to generate the final representation. Hence, the main challenge with wavelet based fusion schemes lies on finding the best setup for both the image decomposition approach (i.e., number of levels, wavelet family and its configurations) and the fusion strategy to merge the information from decomposed images into a single representation (e.g., min, max, mean, rand, etc., from the two approximations and details obtained from the given images at element-wise by taking, respectively, the minimum, the maximum, the mean value, or a random element). The selection of the right setup for fusing the given images will depend on the way the performance is evaluated. Hence a special care should be paid to the quantitative metric used to evaluate the obtained result, avoiding psychophysical experiments that will result in qualitative values [18].

The current paper addresses the problem of cross-spectral fused image visual odometry by using the algorithm proposed by Geiger et al. in [19], which is referred to as LibVISO2. The main novelty of the current approach is to take advantage of information obtained at different spectral bands when visual odometry is estimated. In this way, robust solutions are obtained independently of the scenario's characteristics (e.g., daytime). Fused images are obtained by a Wavelet based scheme. Different fusion schemes are quantitatively evaluated looking for the best one, evaluations are performed by means of a quality metric based on Mutual Information. Once the best configuration is found, the fused image based visual odometry is computed and compared with a previous cross-spectral based approach [5] and classical visible/infrared based approaches.

The manuscript is organized as follows. Section 2 presents the proposed approach detailing the discrete wavelet transform based image fusion and its setups together with the off the shelf monocular visual odometry algorithm used to compute the vehicle odometry. Experimental results and comparisons are presented in Section 3. Finally, conclusions are given in Section 4.

2. Proposed approach

This section presents the Discrete Wavelet Transform image fusion scheme, the evaluation metric used to find the best setup and the monocular visual odometry approach used in the current work.

2.1. Wavelet based image fusion

Wavelet theory has been largely studied in digital signal processing and applied to several subjects (from noise reduction [20] to texture classification [21], jut to mention a couple). At this section, the basic concepts and elements of Discrete Wavelet Transform (DWT) in the context of image fusion are introduced. Let I_{VS} and I_{IR} be the original images, of $m \times n$ pixels, in the visible (*VS*) and Long Wavelength Infrared (*LWIR*) spectra, respectively. We assume the given pair of images are already registered. Let I_F be the image, also of $m \times n$ pixels, resulting from their fusion. In the wavelet based image fusion, the given images are decomposed at their corresponding approximation (*A*) and detail (*D*) components, which correspond to the low pass and high pass filtering for each decomposition level. These decompositions can be represented through sub-images. The detail representations correspond to the vertical details (*VD*), horizontal details (*HD*) and diagonal details (*DD*), respectively (see Fig. 3). Fig. 1(*right*) depicts illustrations of one level DWT decompositions obtained from the original images Fig. 1(*left*) (different approaches used to decompose the given images are introduced in Section 2.1.1).

Once the coefficients (approximations and details) from each decomposition level are obtained, a fusion scheme is applied to *catch* the most relevant information from each representation. The most widely used fusion schemes proposed in the literature to merge the information are reviewed in Section 2.1.2. Finally, the inverse DWT is applied to the result in order to obtain the sought fused image (I_F), which is used in the current work as a monocular image to compute the visual odometry. Fig. 2 presents a classical DWT based image fusion pipeline. In order to cope with misalignments, extensions to this basic pipeline have been also proposed in the literature, such as for instance the dual-tree complex wavelet transform [22]. In the current work just DWT is considered since images to be fused are correctly registered.

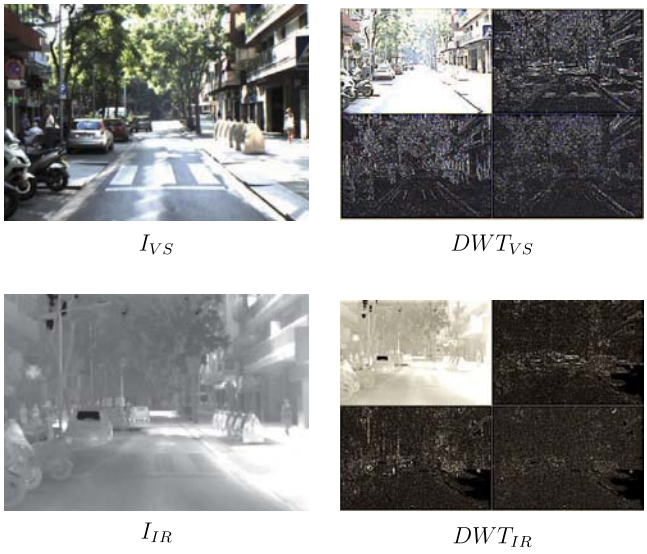

Fig. 1. (*left*) Pair of images (VS-LWIR) to be fused. (*right*) DWT decompositions (one level) of the input images.

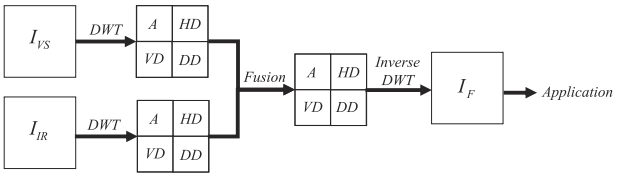

Fig. 2. Illustration of a DWT based image fusion scheme.

2.1.1. Discrete Wavelet Transform (DWT)

At this section, basic concepts of discrete wavelet transform are introduced. The DWT can be represented as a bank of filters, where at each level of decomposition the given signal is split up into high frequency and low frequency components. The low frequency components can be further decomposed until the desired resolution is reached. If multiple levels of decomposition are applied, it is referred to as multiresolution decomposition. Although there is no rule, in general, in the image fusion problem just one level of decomposition is considered. In the current work, the optimum number for the level of decomposition is found by evaluating different configurations.

Several wavelet *families* have been proposed in the literature, each family has a wavelet function and a scaling function. These two functions can be represented by means of a high pass filter (the wavelet function) and a low pass filter (the scaling function). A wavelet family is normally represented by only its wavelet function [23]. Within each of these families some subclasses exist that depend on the number of vanishing moments in the wavelet function. This is just a mathematical property that can directly relate to the number of coefficients. Each of these wavelet functions and their subclasses represent a different way of decomposing a signal. In the current work, looking for the best visual odometry result under different scenarios (daytime), different wavelet families have been evaluated (see Table 1). Details about the evaluation metric approach are presented in Section 2.2.

2.1.2. Fusion strategies

Once the given images are split up into the corresponding approximation images and detail images (i.e., horizontal details, vertical details and diagonal details) the fused image (I_F) is obtained by using a merging scheme that takes into account the approximation and detail information from both images—a correct registration is assumed. Some of the most used merging schemes are summarized below [24]:

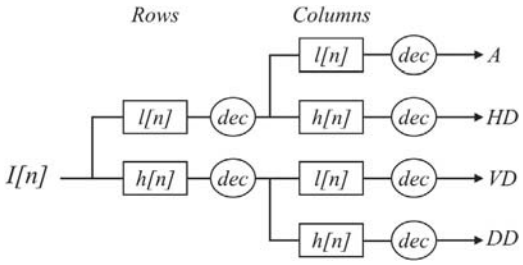

Fig. 3. Two dimensional wavelet decomposition scheme (*l*: low pass filter; *h*: high pass filter; *dec*: decimation).

Substitutive wavelet fusion: in this scheme, the information from one image is completely replaced with information from the other image. In other words, the approximation from one image is merged with the detail of the other image. Once the information is merged the inverse transform is computed to obtain I_F.

Additive wavelet fusion: as indicated by the name, at this scheme the approximations from one image are added to the other one. The same happens for the detail information. If multiple decompositions were applied, the details at each resolution level are added. Finally, after merging the information the inverse transform is performed resulting in the sough I_F. In our implementation, this scheme is implemented by considering the mean value, instead of just the result from the addition.

Weighted models: at this scheme a user tuned merging strategy is applied. Depending on the application and the kind of input images approximations and details are combined according to some statistic values (μ, σ) or according to some other relevant criteria. This scheme is not considered in the current work because input images are of the same resolution, and the performance of fusion based on DWT of infrared and visible images in a general way is evaluated.

Other schemes have been proposed in the literature, which somehow can be considered as combinations of the ones presented above; for instance in this work a strategy that considers the minimum value from each image (approximation or detail images), the maximum value or a random selection was also considered.

From the three approaches presented above, in the current work four different options have been considered as fusion strategies: mean (mean value between approximation coefficients and mean value between detail coefficients); max (the coefficients with maximum value are selected, in both cases approximation and details); min (the coefficients with minimum values are selected); rand (coefficients of approximation and details are randomly selected).

2.2. Performance evaluation

The main challenge with wavelet based fusion schemes lies on finding the best setup for both, the image decomposition approach (Section 2.1.1) (i.e., number of levels, wavelet family and its configurations) and the fusion strategy (Section 2.1.2) used to merge the information from decomposed images into a single representation. Trying all the possible combinations presented above (combinations from Table 1 plus fusion strategies from 2.1.2), 2600 different configurations can be obtained. Since there is not a clear indication in the literature as to which should be the best configuration, all the possibilities are quantitatively evaluated looking for the best one. In the current work, before computing the visual odometry, a set of pairs of images are first selected and evaluated to find the best DWT based fusion configuration. It is expected that a different setup would be required at each daytime

Table 1
Wavelet families evaluated in the current work.

Wavelet name	Comments	Setups
Haar (haar)	Orthogonal wavelet linear phase.	haar
Daubechies (dbN)	Daubechies' external phase wavelets. *N* refers to the number of vanishing moments.	db1, db2, …, db8.
Symlets (symN)	Daubechies' least asymmetric wavelets. *N* refers to the number of vanishing moments.	sym2, sym3,…, sym8.
Coiflets (coifN)	In this family, *N* is the number of vanishing moments for both the wavelet and scaling function.	coif1, coif2, …, coif5.
Biorthogonal (biorNr.Nd)	Biorthogonal wavelets with linear phase. Feature pair of scaling functions (with associated wavelet filters), one for decompositions and one for reconstruction, which can have different number of vanishing moments. Nr and Nd represent the number of vanishing moments.	bior1.1, bior1.3, bior1.5, bior2.2, bior2.4, bior2.6, bior2.8, bior3.1, bior3.3, bior3.5, bior3.7, bior3.9, bior3.5, bior3.7, bior3.9, bior4.4, bior5.5, bior6.8
Reverse Biorthogonal (rbioNr.Nd)	Reverse of the Biorthogonal wavelet explained above.	rbio1.1, rbio1.3, rbio1.5, rbio2.2, rbio2.4, rbio2.6, rbio2.8, rbio3.1, rbio3.3, rbio3.5, rbio3.7, rbio3.9, rbio4.4, rbio5.5, rbio6.8
Discrete meyer approximation (dmey)	Approximation of meyer wavelets leading to FIR filters that can be used in DWT.	dmey

or weather condition. Once this set up is obtained, all the images from the data set are fused and the visual odometry estimated.

Quantitative evaluation of fused images has been an active research topic in recent years (e.g., [25,26]). Proposed approaches can be classified into two categories depending on the existence or not of a reference image [25]. In the case a reference image is available, it can be used as a ground truth to evaluate results by means of quality metrics such as Root Mean Square Error (RMSE), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Mutual Information (MI) among others (e.g., [27,26]). On the other hand, when there is no reference image, the quality of the results is indirectly measured through some metrics such as Entropy (a high entropy value indicates the fused image as rich in information content), Standard Deviation (high values indicate high contrast) and Fusion Mutual Information (the larger the value the better quality fused images) (e.g., [28,29]).

Although there is no reference image, in the current work an adaptation of a quality metric based on the Mutual Information (MI) approach recently presented in [28] is considered. This implementation is a fast version of the original one [29]. It evaluates the performance of the fusion algorithm by measuring the amount of information carried from the source images to the fused image by means of mutual information. Our adaptation consists in computing this metric twice, once assuming the visible image as a reference and once with the infrared image as a reference. Then, the average value is considered:

$$\mathrm{FMI}_{VS-IR} = (\mathrm{MI}_{F,VS}/(H_F + H_{VS}) + \mathrm{MI}_{F,IR}/(H_F + H_{IR}))/2, \quad (1)$$

where MI is the mutual information value and H_k, with $k = \{VS, IR, F\}$, are the histogram based entropies of the visible, infrared and fused images, respectively, as presented in [29].

2.3. Monocular visual odometry

The fused images obtained with the best configuration as mentioned above are used in the monocular version of the well-known algorithm proposed by Geiger et al. in [19], which is referred to as LibVISO2. The algorithm is briefly presented below, for more details see [19].

Generally, results from monocular systems are up to a scale factor; in other words they lack of a real 3D measure. This problem affects most of monocular odometry approaches. In order to overcome this limitation, LibVISO2 assumes a fixed transformation from the ground plane to the camera (parameters given by the camera height and the camera pitch). These values are updated at each iteration by estimating the ground plane. Hence, features on the ground as well as features above the ground plane are needed for a good odometry estimation. Roughly speaking, the algorithm consists of the following steps:

- Compute the fundamental matrix (**F**) from point correspondences using the 8-point algorithm.
- Compute the essential matrix (**E**) using the camera calibration parameters.
- Estimate the 3D coordinates and [**R**|**t**].
- Estimate the ground plane from the 3D points.
- Scale the [**R**|**t**] using the values of camera height and pitch obtained in previous step.

3. Experimental results

This section presents experimental results and comparisons with classical approaches based on visible spectrum or infrared images. Additionally, comparisons with the results presented in [5] are provided showing the improvements when a better setup is considered for the fusion algorithm. In all the cases GPS information is used as ground truth data to evaluate the performance of evaluated approaches. Below, the acquisition platforms are introduced and then experimental results are depicted.

3.1. Acquisition platforms

The proposed approach has been evaluated using images obtained from two different platforms. Fig. 4(*left*) shows the electric car with the cross-spectral stereo head used in [5]. The stereo head consists of a pair of cameras arranged in a non verged geometry. One of the camera works in the infrared spectrum, more precisely Long Wavelength Infrared (LWIR), detecting radiations in the range of 8–14 μm. The other camera, which is referred to as Visible Spectrum (VS) responds to wavelengths from about 390 to 750 nm (visible spectrum). Both cameras are synchronized using an external hardware trigger. The images provided by the cross-spectral stereo head are calibrated and rectified using [12]; a process similar to the one presented in [30] 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. With this acquisition platform three video sequences have been obtained (see [5] for more details about them) and used in the current work. They will be referred to as *CVC − VidNN*[1]

Fig. 4(*right*) depicts an illustration of the acquisition cross-spectral system from [29]. In this case, also LWIR and VS images are obtained, but the cameras are arranged differently. In this case

[1] Data set available at: https://ngunsu.github.io/cvc_vod/.

Fig. 4. Acquisition systems (cross-spectral imaging on the top): (*left*) Electric vehicle from the Computer Vision Center (Barcelona, Spain) [5] (CVC video sequences); (*right*) Car from the Korea Advanced Institute of Science and Technology (Seoul, Korea) [29] (KAIST video sequences).

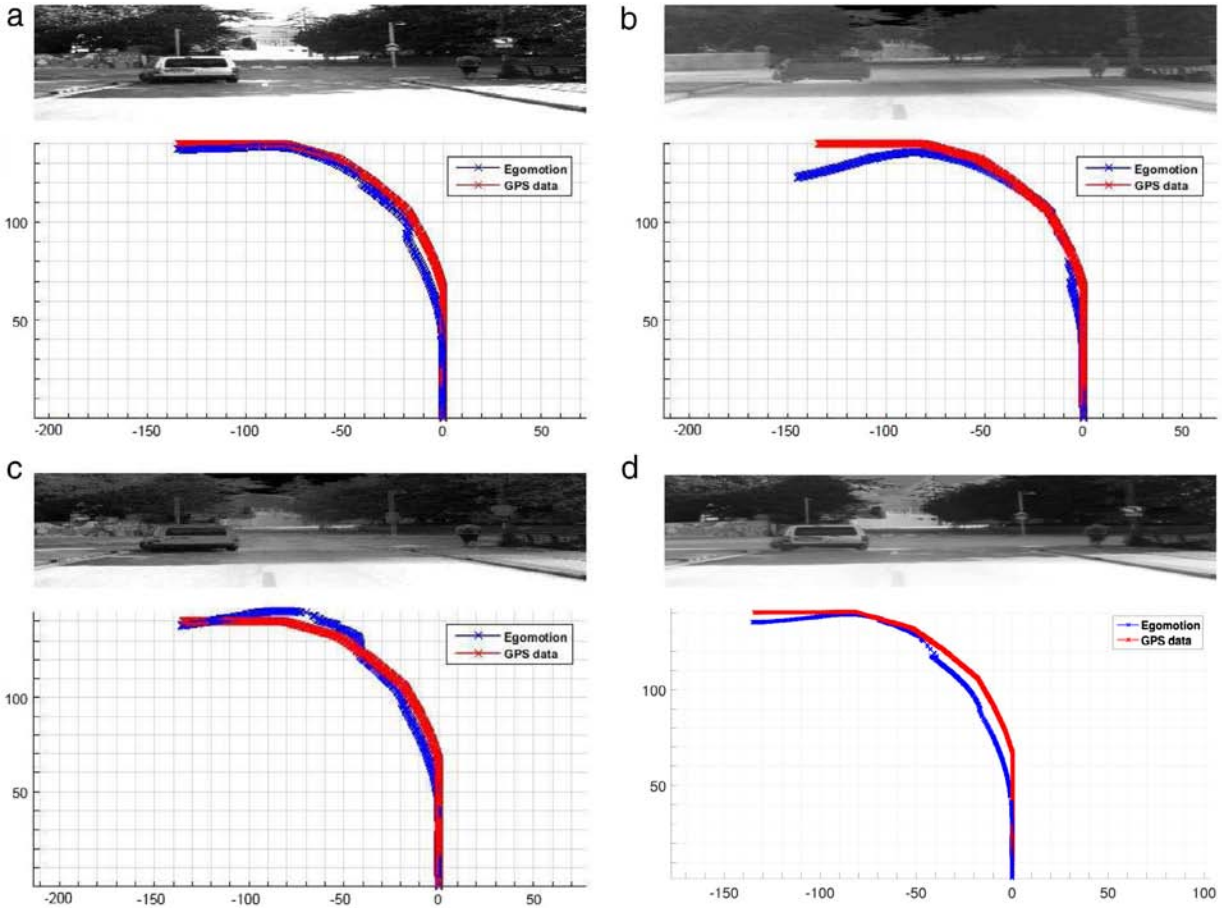

Fig. 5. Estimated trajectories for the CVC-Vid00 video sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT fused images from [5]; and (d) DWT fused images, proposed approach.

a beam splitter (made of Zinc coated Silicon wafer) is used, so that both cameras capture images from the same point of view. Like in the previous case, both cameras are synchronized using an external hardware trigger. A calibration process similar to the one presented above is applied (in this case using squares milled onto a thin cooper board—a printed circuit board). Images are also calibrated and rectified using the toolkit from [12]. With this platform, a data set containing video sequences obtained at six different times of the day has been generated. In the current work, the following three videosequences have been used: (i) 5 AM; (ii) 12 AM; (iii) 10 PM just to evaluate results from scenes that contain different light (from early morning till dark night). These video sequences will be referred to as *KAIST − VidNNN*.

The CVC video sequences were all obtained at day light time (about midday), hence it is difficult to appreciate the advantages of the proposed cross-spectral visual odometry approach. In other words, in the CVC video sequences just a visible spectrum based approach would be enough to compute visual odometry. In order to appreciate the advantages of using the proposed cross-spectral based approach three video sequences from KAIST [29] have been evaluated. These video sequences correspond to different times of day (from early morning till late evening). More precisely, they

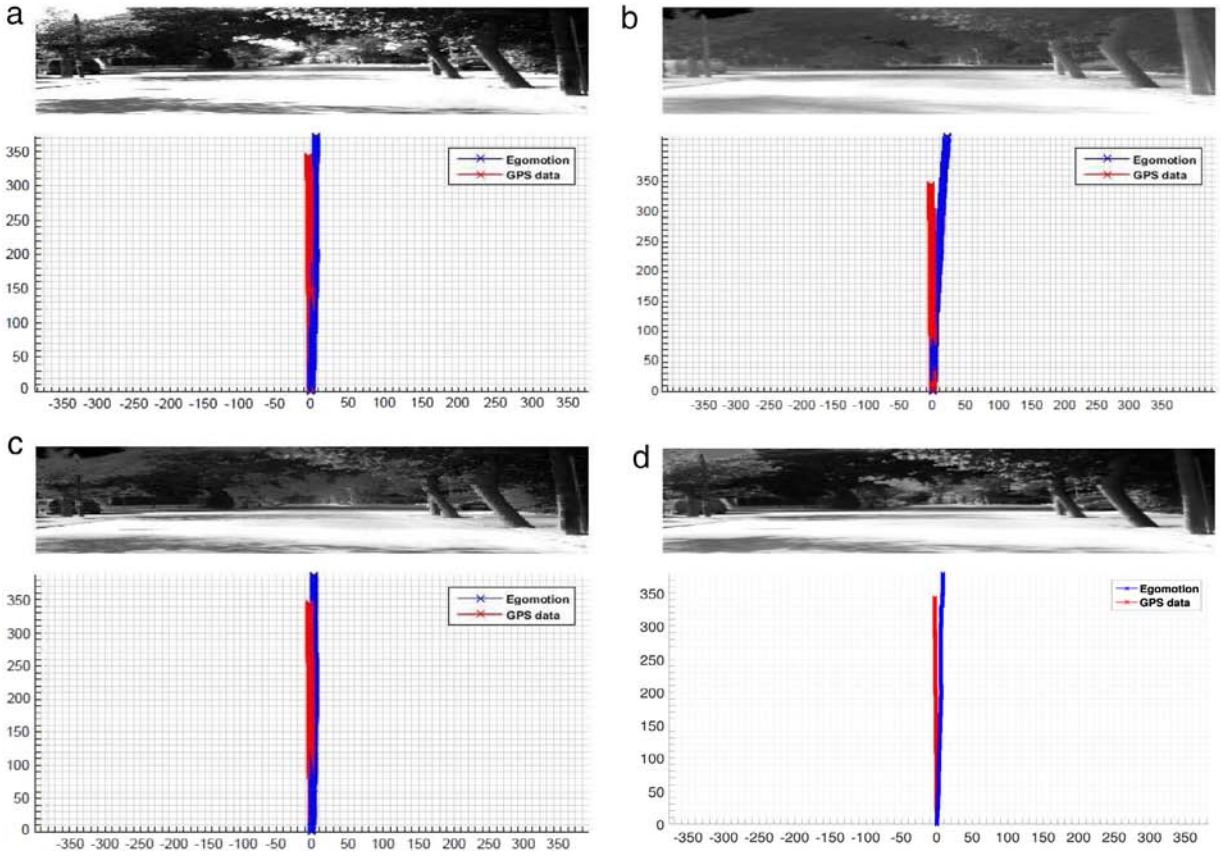

Fig. 6. Estimated trajectories for CVC-Vid01 sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT based fused image, result from [5]; and (d) DWT based fused image, proposed approach.

were obtained at 5 AM, 10 AM and 10 PM. In the three sequences, the car has traveled about 300 m at an average speed of 35 km/h. Results with these two data sets are presented below.

3.2. Visual odometry results

In this section, experimental results and comparisons, with the video sequences from the two platforms introduced above, are presented. In order to have a fair comparison, the user defined parameters for the VO algorithm (LibVISO2) have been tuned accordingly to the image nature (visible, infrared, fused) and characteristics of the video sequence. These parameters were empirically obtained looking for the best performance in every image domain. In all the cases, ground truth data from GPS are used for comparisons. Additionally, the average number of matches and percentage of inliers per video sequence evaluated are provided as complementary information; there is no correlation between these values and the final position error.

3.2.1. CVC-Vid00 video sequence

It consists of a large curve in a urban scenario. The car travels more than 200 m at an average speed of about 17 km/h. The VO algorithm (LibVISO2) has been tuned as presented in [5]. Fig. 5 depicts the plots corresponding to the four different cases (visible, infrared, fused [5] and fused (proposed approach)) when they are compared with ground truth data (GPS information). Quantitative results corresponding to these four trajectories are presented in Table 2. VO computed with the visible spectrum video sequence gets the best result since the sequence has been obtained at day light; it can be appreciated that the DWT tuned with the proposed approach (selecting the best configuration using the FMI metric) gets better results than the one presented in [5]. The visual odometry computed with the infrared spectrum video sequence gets the worst results; this is mainly due to the lack of texture in the images.

Table 2
VO results in the **CVC-Vid00 video sequence** using images from different spectrum and fusion approaches (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fusion using Discrete Wavelet Transform; DWT [Prop. App]: fusion using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 235 m)	234.88	241.27	245	240.3
Final position error (m)	2.9	18	5.4	5.1
Average number of matches	2053	3588	4513	2123
Percentage of inliers	71.5	61.94	60	65.1

3.2.2. CVC-Vid01 video sequence

It is a simple straight line trajectory on a urban scenario consisting of about 350 m; the car travels at an average speed of about 25 km/h. The (LibVISO2) algorithm has been tuned as presented in [5]. Fig. 6 depicts the plots of the visual odometry computed over each of the four representations (VS, LWIR, DWT based fused images [5] and DWT based fused images with the best setup) together with the corresponding GPS data. Like in the previous case, the visual odometry computed with the infrared video sequence gets the worst result, as can be easily appreciated in Fig. 6 and confirmed by the final position error value presented in Table 3. The results obtained with the other three representations (visible spectrum, DWT based image fusion [5] and proposed approach) are similar both qualitatively and quantitatively. Once

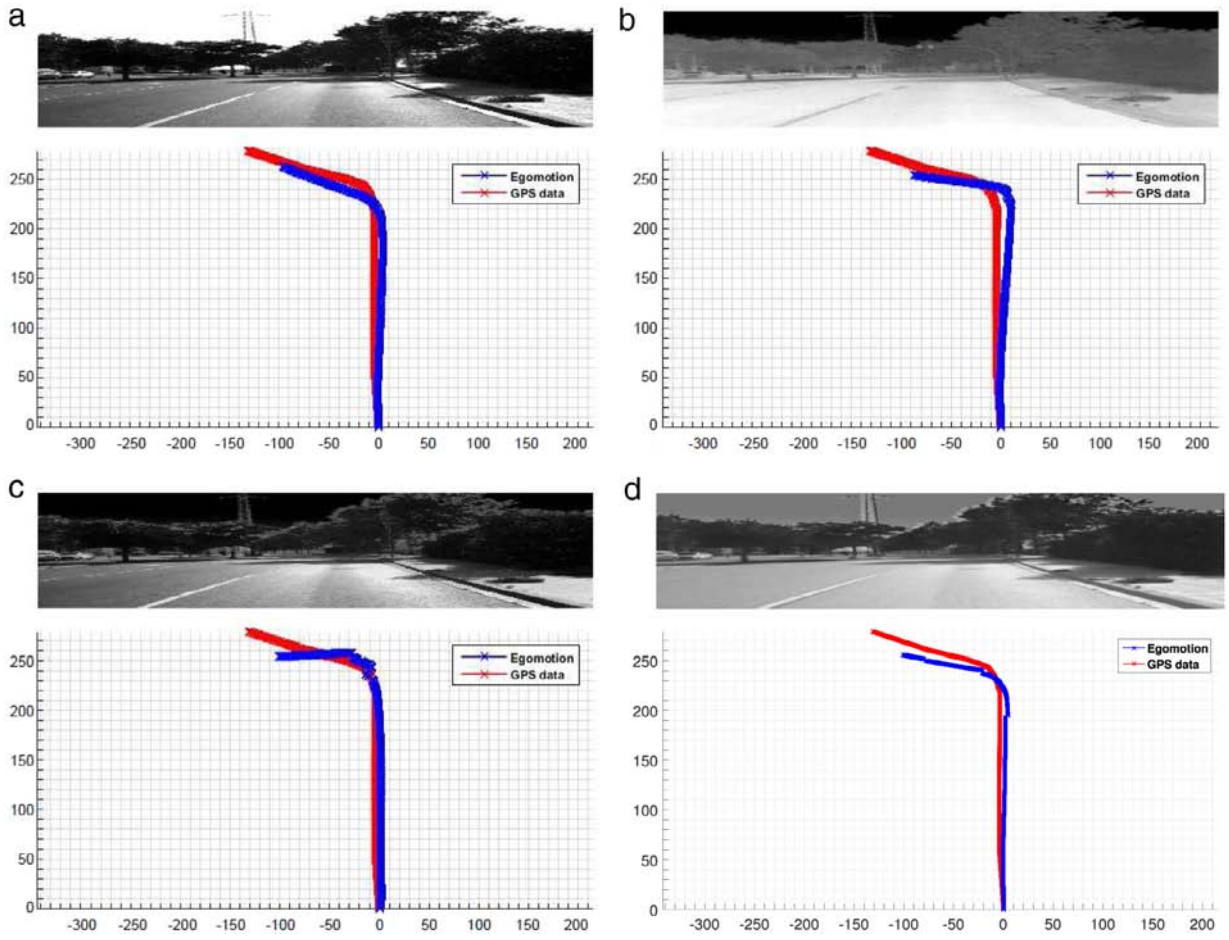

Fig. 7. Estimated trajectories for CVC-Vid02 sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT based fused image, result from [5]; and (d) DWT based fused image, proposed approach.

Table 3
VO results in the **CVC-Vid01 video sequence** using images from different spectrum and fusion approaches (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fusion using Discrete Wavelet Transform; DWT [Prop. App]: fusion using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 365 m)	371.8	424	386	379
Final position error (m)	32.6	84.7	44	38.2
Average number of matches	1965	1974	2137	2071
Percentage of inliers	72.6	67.8	61.5	66.3

Table 4
VO results in the **CVC-Vid02 video sequence** using images from different spectrum and fusion approaches (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fusion using Discrete Wavelet Transform; DWT [Prop. App]: fusion using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 370 m)	325.6	336.9	354.4	334.7
Final position error (m)	37.7	48.7	37.2	38.0
Average number of matches	1890	1028	1952	1719
Percentage of inliers	70	65.8	61	64.1

again, like in the previous case, results from the proposed approach are considerably better than the ones obtained from [5], which suggests the need for selecting the best configuration of the fusion's parameters.

3.2.3. CVC-Vid02 video sequence

It is a "L" like shape trajectory on a sub-urban scenario. It is the longest trajectory (370 m) and the car has traveled faster than in the previous cases (about 30 km/h). The (LibVISO2) algorithm has been tuned as presented in [5]. In this particular video sequence, the visible spectrum and both fused based approaches get similar results (see Fig. 7 and Table 4). Although the DWT based approach from [5] gets the smallest final position error, the difference with respect to the results obtained in the visible spectrum and the proposed approach is smaller than one meter. On the contrary, it can be appreciated that the traveled distance in [5] is considerably higher, in other words, although [5] gets the smallest final position error, both, the visible spectrum and the proposed approach result in trajectories quite similar to the one obtained with the GPS.

As a conclusion from the CVC-video sequences, it can be appreciated that, although they correspond to day light sequences, the usage of fused images results in quite stable solutions. The best setup of the DWT based image fusion corresponds to Reverse of the Biorthogonal wavelet family with vanishing moments (Nr $=$ 2 and Nd $=$ 8) for the DWT (rbio2.8 in Table 1); and (mean, max) for the fusion strategy (see Section 2.1.2). The same setup for the fusion configuration has been used in the three video sequences. This setup has been empirically obtained by evaluating quantitatively a set of frames of CVC video sequence using the FMI metric presented in Section 2.2.

3.2.4. KAIST-5AM video sequence

This is the first video sequence where the advantages of cross-spectral based approaches can be easily appreciated. At this

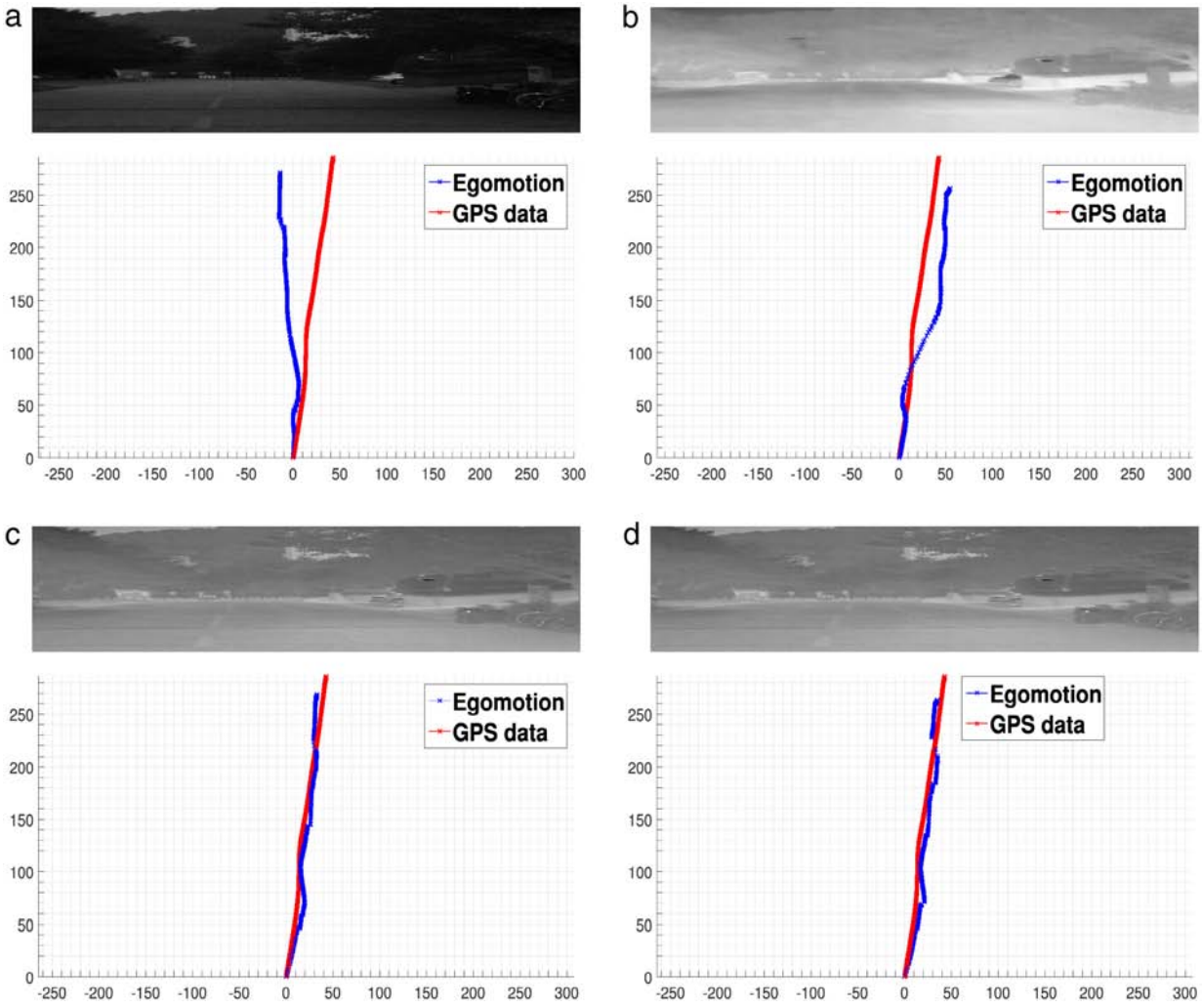

Fig. 8. Estimated trajectories for KAIST-5AM video sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT based fused image, result using [5]; and (d) Fused with the proposed DWT based approach, by selecting the best setup.

daytime, although there is already a little bit of light, results from visible spectrum are the worst. On the contrary, the visual odometry computed with the infrared video sequence gets better results. Finally, the best VO results are obtained with the images fused with the proposed approach (selecting the best configuration of fusion's parameters). It should be noticed that the results (final position error) obtained with the proposed approach are two times better than the one from visible spectrum and almost 50% better than infrared spectrum (see quantitative values in Table 5). The estimated trajectories can be found in Fig. 8; it can be appreciated how the results from the proposed approach keep almost attached to the GPS trajectory. Regarding the fusion strategy, the best configuration corresponds to Biorthogonal wavelet family with vanishing moments (Nr $=$ 3 and Nd $=$ 5) for the DWT (bio3.5 in Table 1); and (max, min) for the fusion strategy (see Section 2.1.2).

Table 5
VO results for the 5AM video sequence using different images (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fused images using Discrete Wavelet Transform; DWT [Prop. App]: fused images using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 287.9 m)	278.7	279.5	278.9	274.6
Final position error (m)	58.4	32.1	25.1	24.1
Average number of matches	3197	2659	2555	2584
Percentage of inliers	51.7	64.1	53.6	54.1

3.2.5. KAIST-10AM video sequence

On the contrary to the previous case, here the results are similar to the ones obtained with the CVC video sequences. It makes sense since the sequence has been obtained at day light time (10 AM). In other words, visible spectrum and cross-spectral based approaches reach to similar results. This can be quantitatively appreciated looking at Fig. 9. Quantitative values are provided in Table 6. As expected, infrared based visual odometry gets the worst results. Like in the CVC video sequence, in this case the best setup of the DWT based image fusion corresponds to Reverse of the Biorthogonal wavelet family with vanishing moments (Nr $=$ 2 and Nd $=$ 8) for the DWT (rbio2.8 in Table 1); and (mean, min) for the fusion strategy (see Section 2.1.2).

3.2.6. KAIST-10PM video sequence

Finally, this last sequence corresponds to a quite dark night, which is a challenging scenario for visible spectrum video sequence. Like in the KAIST-5AM video sequence, the worst result is obtained using visible spectrum images. This worst result is due to the lack of light in the scenario, just car's light is present. Visual odometry computed with infrared video sequence gets quite acceptable results. The best VO result was obtained with images fused with the proposed approach (by selecting the best

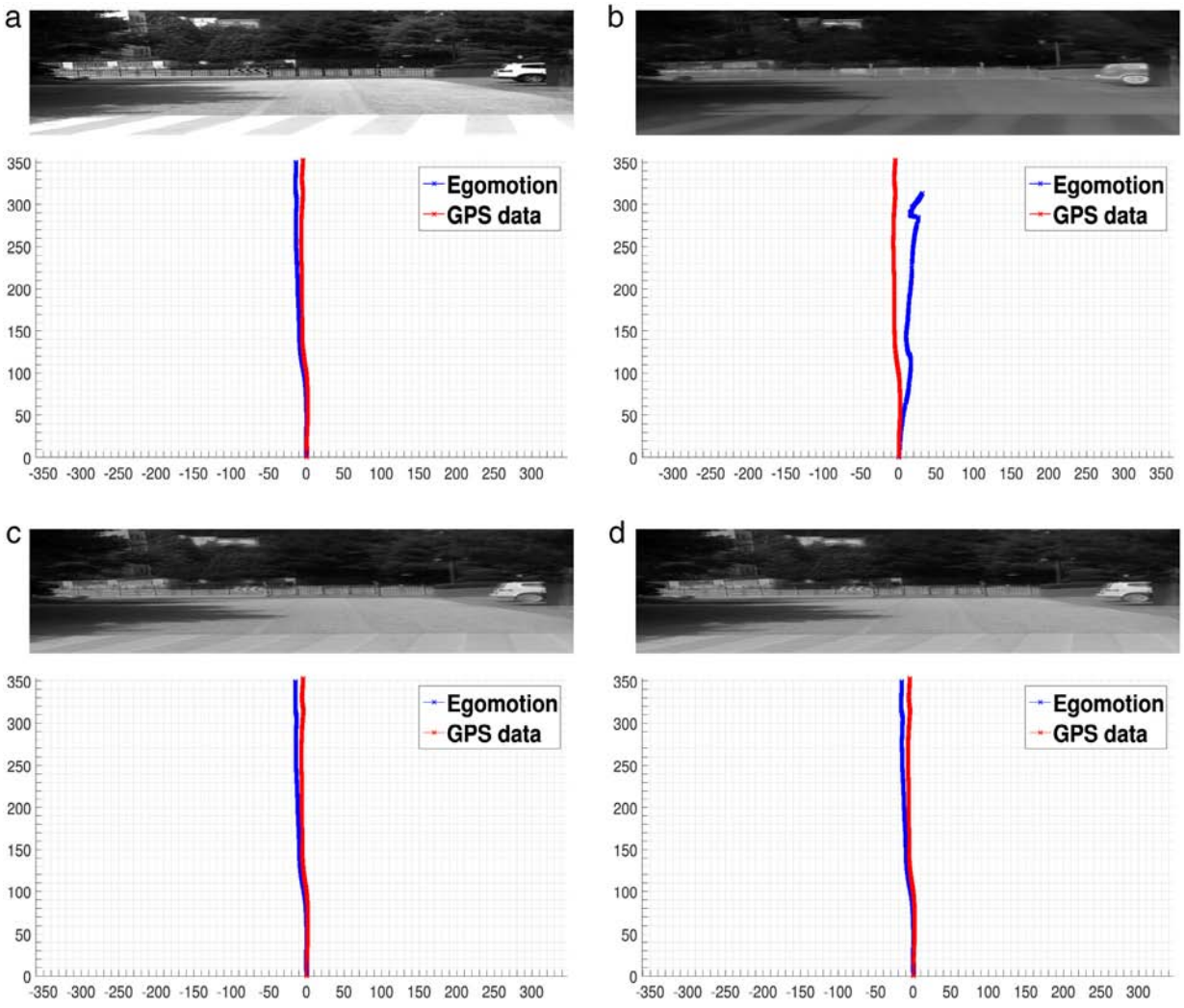

Fig. 9. Estimated trajectories for KAIST-10AM video sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT based fused image, result using [5]; and (d) Fused with the proposed DWT based approach, by selecting the best setup.

Table 6
VO results for the 10AM video sequence using different images (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fused images using Discrete Wavelet Transform; DWT [Prop. App]: fused images using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 351.6 m)	348.2	330.2	347.8	347.3
Final position error (m)	10.1	53.9	12.6	11.8
Average number of matches	12 696	4278	11 078	11 119
Percentage of inliers	93.8	69.7	90.6	89.9

Table 7
VO results for the 10PM video sequence using different images (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum; DWT [5]: fused images using Discrete Wavelet Transform; DWT [Prop. App]: fused images using Discrete Wavelet Transform selecting the best setup).

Results	VS	LWIR	DWT [5]	DWT [**Prop. App.**]
Total traveled distance (m) (GPS traveled dist.: 222.3 m)	281.90	218.07	226.15	222.79
Final position error (m)	42.20	24.99	15.2	12.62
Average number of matches	271	2308	1065	2301
Percentage of inliers	44.89	61.71	52.29	61.73

configuration for the fusion's parameters); in this case, VO from fused images, the final position error (see Table 7) is two times smaller than the obtained with infrared images and more than three times when compared with the result from visible spectrum video sequence. The resulting trajectories can be appreciated in Fig. 10.

4. Conclusion

The manuscript evaluates the performance of a classical monocular visual odometry when cross-spectral fused images are used. The best fusion strategy is selected by using a novel mutual information based metric. The obtained visual odometry results are compared with a previous approach as well as with classical ones (based on visible and infrared spectrum, respectively). While at day light time the performance of classical visible spectrum based approach is quite similar to the one obtained with proposed approach, results show that the proposed approach is the best option to tackle challenging scenarios, in particular those with dark or poor lighting conditions. As a future work, other challenging scenarios, including different weather conditions (fog and rain), will be evaluated.

Acknowledgments

This work has been partially supported by the Spanish Government under Project TIN2014-56919-C3-2-R; the PROMETEO

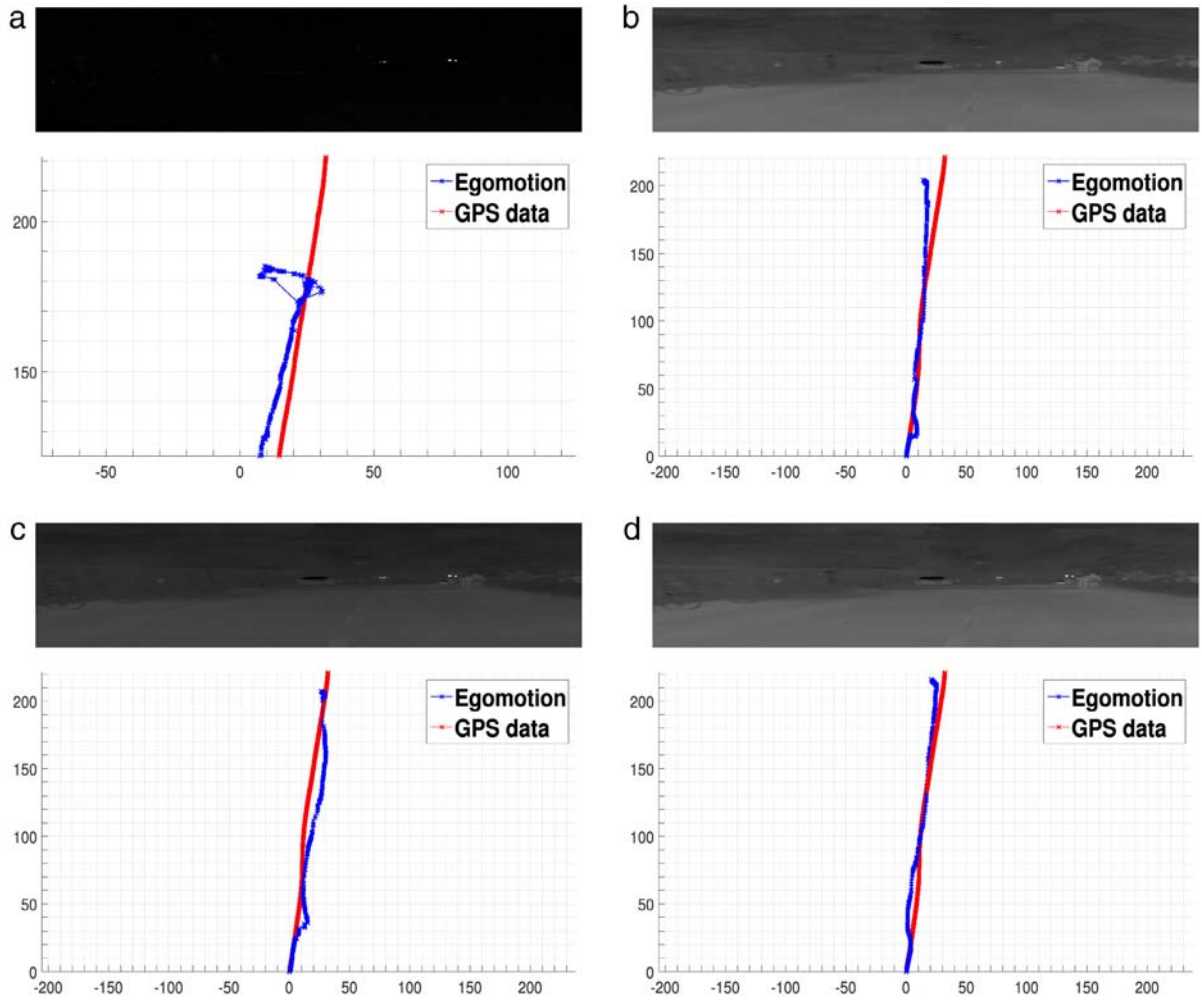

Fig. 10. Estimated trajectories for KAIST-10PM video sequence: (a) Visible spectrum; (b) Infrared spectrum; (c) DWT based fused image, result using [5]; and (d) Fused with the proposed DWT based approach, by selecting the best setup.

Project of the "Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador"; Ecuador; the ESPOL project Pattern recognition: case study on agriculture and aquaculture (M1-D1-2015); and the "Secretaria d' Universitats i Recerca del Departament d' Economia i Coneixement de la Generalitat de Catalunya" (2014-SGR-1506). C. Aguilera has been supported by the Universitat Autònoma de Barcelona (09-2013,09-2017). M. Oliveira has been supported by the Portuguese Foundation for Science and Technology "Fundação para a Ciência e Tecnologia (FTC)", under grant agreement SFRH/BPD/109651/2015 and projects POCI-01-0145-FEDER-006961 and UID/CEC/00127/2013. This work was also financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020.

References

[1] D. Borrmann, A. Nüchter, M. akulović, I. Maurović, I. Petrović, D. Osmanković, J. Velagić, A mobile robot based system for fully automated thermal 3D mapping, Adv. Eng. Inf. 28 (4) (2014) 425–440.
[2] P. Shah, B.C.S. Reddy, S.N. Merchant, U.B. Desai, Context enhancement to reveal a camouflaged target and to assist target localization by fusion of multispectral surveillance videos, Signal Image Video Process. 7 (3) (2013) 537–552.
[3] T. Bourlai, N. Kalka, A. Ross, B. Cukic, L. Hornak, Cross-spectral face verification in the short wave infrared (swir) band, in: 2010 20th International Conference on Pattern Recognition, (ICPR), IEEE, 2010, pp. 1343–1347.
[4] Y. Choi, et al. All-day visual place recognition: Benchmark dataset and baseline, in: IEEE International Conference on Computer Vision and Pattern Recognition Workshops, CVPRWVPRICE, 2015.
[5] J. Poujol, C. Aguilera, E. Danos, B. Vintimilla, R. Toledo, A.D. Sappa, Visible-thermal fusion based monocular visual odometry, in: ROBOT'2015: Second Iberian Robotics Conference, in: Advances in Intelligent Systems and Computing, vol. 417, Springer Verlag, Lisbon, Portugal, 2015, pp. 517–528.
[6] D. Nistér, O. Naroditsky, J. Bergen, Visual odometry, in: IEEE Intgernational Conference on Computer Vision and Pattern Recognition, vol. 1, 2004, pp. I–652.
[7] D. Scaramuzza, F. Fraundorfer, R. Siegwart, Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC, in: IEEE International Conference on Robotics and Automation, 2009, pp. 4293–4299.
[8] D. Scaramuzza, F. Fraundorfer, Visual odometry [tutorial], IEEE Robot. Autom. Mag. 18 (4) (2011) 80–92.
[9] A. Chilian, H. Hirschmüller, Stereo camera based navigation of mobile robots on rough terrain, in: IEEE International Conference on Intelligent Robots and Systems, IROS, IEEE, 2009, pp. 4571–4576.
[10] E. Nilsson, C. Lundquist, T. Schön, D. Forslund, J. Roll, Vehicle motion estimation using an infrared camera, in: 18th IFAC World Congress, Milano, Italy, 28 August-2 September, 2011, Elsevier, 2011, pp. 12952–12957.
[11] T. Mouats, N. Aouf, A.D. Sappa, C.A. Aguilera-Carrasco, R. Toledo, Multispectral stereo odometry, IEEE Trans. Intell. Transp. Syst. 16 (3) (2015) 1210–1224.
[12] J.-Y. Bouguet, Camera calibration toolbox for matlab, July 2010.
[13] P. Ricaurte, C. Chilán, C.A. Aguilera-Carrasco, B.X. Vintimilla, A.D. Sappa, Feature point descriptors: Infrared and visible spectra, Sensors 14 (2) (2014) 3690–3701.
[14] C. Aguilera, F. Barrera, F. Lumbreras, A.D. Sappa, R. Toledo, Multispectral image feature points, Sensors 12 (9) (2012) 12661–12672.
[15] L. Dong, Q. Yang, H. Wu, H. Xiao, M. Xu, High quality multi-spectral and panchromatic image fusion technologies based on curvelet transform, Neurocomputing 159 (2015) 268–274.
[16] Z. Wang, D. Ziou, C. Armenakis, D. Li, Q. Li, A comparative analysis of image fusion methods, IEEE Trans. Geosci. Remote Sens. 43 (6) (2005) 1391–1402.
[17] R. Gharbia, A.T. Azar, A.H.E. Baz, A.E. Hassanien, Image fusion techniques in remote sensing, CoRR abs/1403.5473.
[18] A.E. Hayes, G.D. Finlayson, R. Montagna, Rgb-nir color image fusion: metric and psychophysical experiments, in: Proc. SPIE, vol. 9396, 2015.
[19] A. Geiger, J. Ziegler, C. Stiller, Stereoscan: Dense 3D reconstruction in real-time, in: Intelligent Vehicles Symposium (IV), 2011.

[20] M. Lang, H. Guo, J.E. Odegard, C.S. Burrus, R. Wells Jr., Noise reduction using an undecimated discrete wavelet transform, IEEE Signal Process. Lett. 3 (1) (1996) 10–12.
[21] T. Chang, C.J. Kuo, Texture analysis and classification with tree-structured wavelet transform, IEEE Trans. Image Process. 2 (4) (1993) 429–441.
[22] A. Barri, A. Dooms, P. Schelkens, The near shift-invariance of the dual-tree complex wavelet transform revisited, J. Math. Anal. Appl. 389 (2) (2012) 1303–1314.
[23] K. Amolins, Y. Zhang, P. Dare, Wavelet based image fusion techniques — An introduction, review and comparison, ISPRS J. Photogramm. Remote Sens. 62 (2007) 249–263.
[24] I. Mehra, N.K. Nishchal, Wavelet-based image fusion for securing multiple images through asymmetric keys, Opt. Commun. 335 (2015) 153–160.
[25] P. Jagalingam, A.V. Hegde, A review of quality metrics for fused image, Aquat. Proc. 4 (2015) 133–142. International Conference on Water Resources, Coastal and Ocean Engineering (ICWRCOE'15).
[26] W. Zhou, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[27] C. Yang, J.-Q. Zhang, X.-R. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2) (2008) 156–160.
[28] M. Haghighat, M. Razian, Fast-fmi: Non-reference image fusion metric, in: IEEE 8th International Conference on Application of Information and Communication Technologies, 2014, pp. 1–3.
[29] M.B.A. Haghighat, A. Aghagolzadeh, H. Seyedarabi, A non-reference image fusion metric based on mutual information of image features, Comput. Electr. Eng. 37 (5) (2011) 744–756. Special Issue on Image Processing.
[30] F. Barrera, F. Lumbreras, A.D. Sappa, Multimodal stereo vision system: 3D data extraction and algorithm evaluation, IEEE J. Sel. Top. Sign. Proces. 6 (5) (2012) 437–446.

Angel Domingo Sappa 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 UK, and Greece, he joined the Computer Vision Center, Barcelona, Spain, where he is currently a Senior Researcher, member of the Advanced Driver Assistance Systems Group. Since 2016 he is also with the Electrical and Computer Science Engineering school of the ESPOL, Guayaquil, Ecuador, as an invited Full Professor. His research interests span a broad spectrum within the 2D and 3D image processing. His current research focuses on stereo image processing and analysis, 3D modeling, dense optical flow estimation and multispectral imaging.

Cristhian Aguilera received the B.S. degree in automation engineer from the Universidad del Bío-Bío Concepción, Chile, in 2008 and the M.Sc. degree in computer vision from the Autonomous University of Barcelona, Barcelona, Spain, in 2014. He is currently working towards the Ph.D. degree in computer science from the Autonomous University of Barcelona. Since 2015, he is an editor assistant of the Electronic Letter on Computer Vision and Image Analysis journal. His current research focuses in cross-spectral image similarity, stereo vision and deep convolutional networks.

Juan Carvajal Ayala received the Bachelor's degree in Electronic Communications Engineering from the University of Navarra School of Engineering, San Sebastian, Spain, in 2014. He did a research internship at Fraunhofer IIS in Erlangen, Germany, and is currently a research assistant at Center for Research, Development and Innovation of Computer Systems (CIDIS), ESPOL. His research interests are image fusion, pattern recognition, and deep learning.

Miguel Oliveira received the Mechanical Engineering and M.Sc. in Mechanical Engineering degrees from the University of Aveiro, Portugal, in 2004 and 2007, respectively. Later in 2013 he obtained the Ph.D. in Mechanical Engineering specialization in Robotics, on the topic of Autonomous Driving and Drivers Assistance Systems. Currently he is a researcher at both the Institute for Systems and Computer Engineering,Technology and Science in Porto, Portugal, as well as the Institute of Electronics and Telematics Engineering of Aveiro, Portugal. In addition, he is an assistant professor at the Department of Mechanical Engineering, University of Aveiro, Portugal, where he teaches computer vision courses. His research interests include visual object recognition in open-ended domains, scene reconstruction from multi-modal sensor data, image and 3D data processing, computer vision and robotics.

Dennis G. Romero received the Computer Engineering degree from Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador, in 2007, and the Ph.D. degree in Electrical Engineering from Universidade Federal do Espírito Santo, UFES, Vitória, Brazil, in 2014. In 2014, he joined the Center for Research, Development and Innovation of Computer Systems (CIDIS). He is a member of the Pattern recognition research group at ESPOL. His research interests center on improving the data understanding from sensor fusion, mainly through the application of machine learning and statistics for pattern recognition. His current research focuses on Pattern Recognition from microscope images of shrimps for identification of diseases.

Boris X. Vintimilla received his degree in mechanical engineering in 1995 at the Escuela Superior Politécnica del Litora—ESPOL, Guayaquil, Ecuador, and his Ph.D. degree in industrial engineering in 2001 at the Polytechnic University of Catalonia, Barcelona, Spain. In May 2001, he joined the Department of Electrical and Computer Science Engineering of the ESPOL as associated professor and in 2008 became a full professor. Dr. Vintimilla has been the director of the Center of Vision and Robotics from 2005 to 2008. He did his post-doctorate research in the Digital Imaging Research Center at Kingston University (London, UK) from 2008 to 2009. Currently, he is director of the Center for Research, Development and Innovation of Computer Systems (CIDIS) at ESPOL. His research areas include topics related with image processing and analysis, and vision applied to mobile robotics. Dr. Vintimilla has been involved in several projects supported by international and national organizations, as result of these researches he has published more than 40 scientific articles and book chapters.

Ricardo Toledo received the degree in Electronic Engineering from the Universidad Nacional de Rosario (Argentina) in 1986, the M.Sc. degree in image processing and artificial intelligence from the Universitat Autònoma de Barcelona (UAB) in 1992 and the Ph.D. in 2001. Since 1989 he was lecturer in the Computer Science Dept. of the UAB, and was involved in R + D projects. In 1996, he participated in the foundation of the Computer Vision Center (CVC) at the UAB. Currently, he is a full time associated professor at the Computer Science Dept., Cordinator of the Master in Informatics at the Escola d'Enginyeria (UAB) and member of the Computer Vision Centre. Ricardo has participated in several national and international/EU R + D projects, being the leader of some of them, is author/co-author of more than 40 papers, all these in the field of computer vision, robotics and medical imaging and has supervised several Master and Ph.D. thesis.