A Visible-Thermal Fusion based Monocular Visual Odometry

Julien Poujol[1], Cristhian A. Aguilera[1,2], Etienne Danos,[1],
Boris X. Vintimilla[3], Ricardo Toledo[1,2] and Angel D. Sappa[1,3]

[1]Computer Vision Center, Edifici O, Campus UAB
08193, Bellaterra, Barcelona, Spain
[2]Computer Science Department,
Universitat Autònoma de Barcelona, Campus UAB,
Bellaterra, Spain
[3]Escuela Superior Politécnica del Litoral, ESPOL,
Facultad de Ingeniería en Electricidad y Computación,
Campus Gustavo Galindo Km 30.5 Vía Perimetral, P.O. Box 09-01-5863,
Guayaquil, Ecuador

Abstract. The manuscript evaluates the performance of a monocular visual odometry approach when images from different spectra are considered, both independently and fused. The objective behind this evaluation is to analyze if classical approaches can be improved when the given images, which are from different spectra, are fused and represented in new domains. The images in these new domains should have some of the following properties: *i*) more robust to noisy data; *ii*) less sensitive to changes (e.g., lighting); *iii*) more rich in descriptive information, among other. In particular in the current work two different image fusion strategies are considered. Firstly, images from the visible and thermal spectrum are fused using a Discrete Wavelet Transform (DWT) approach. Secondly, a monochrome threshold strategy is considered. The obtained representations are evaluated under a visual odometry framework, highlighting their advantages and disadvantages, using different urban and semi-urban scenarios. Comparisons with both monocular-visible spectrum and monocular-infrared spectrum, are also provided showing the validity of the proposed approach.

Keywords: Monocular Visual Odometry; LWIR-RGB cross-spectral Imaging; Image Fusion.

1 Introduction

Recent advances in imaging sensors allow the usage of cameras at different spectral bands to tackle classical computer vision problems. As an example of such emerging field we can mention the pedestrian detection systems for driving assistance. Although classically they have relied only in the visible spectrum [1], recently some multispectral approaches have been proposed in the literature [2]

showing advantages. The same trend can be appreciated in other computer vision applications such as 3D modeling (e.g., [3], [4]), video-surveillance (e.g., [5], [6]) or visual odometry, which is the focus of the current work.

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 [7] in 2004; it has been chosen for its similarity to wheel odometry, which incrementally estimates the motion of a vehicle 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 to the images of the onboard vision system.

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., [8], [9], [10], [11]). 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. Generalizations to the 3D domain has been also proposed in the literature [12]. 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 sepctrum imaging, in addition to their own limitation, have those 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., [13], [14]) and more recently cross-spectral stereo based approaches have been also introduced [15]. The current work proposes a step further by tackling the monocular vision odometry with an image resulting from the fusion of a cross-spectral imaging device. In this way the strengths of each band are considered and the objective is to evaluate whether classical approaches can be improved by using images from this new domain.

The manuscript is organized as follow. Section 2 introduces the image fusion techniques evaluated in the current work together with the monocular visual odometry algorithm used as a refernce. Experimental results and comparisons are provided in Section 3. Finally, conclusions are given in Section 4.

2 Proposed Approach

This section presents the image fusion algorithms evaluated in the monocular visual odometry context. Let I_v be a visible spectrum (VS) image and I_{ir} the corresponding one from the Long Wavelength Infrared (LWIR) spectrum. In the current work we assume the given pair of images are already registered. The image resulting from the fusion will be referred to as F.

2.1 Discrete Wavelet Transform based Image Fusion

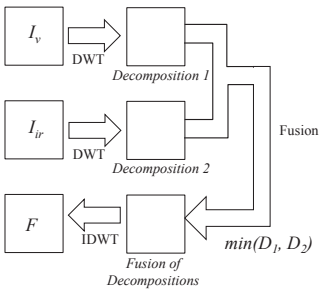

Fig. 1. Scheme of the Discrete Wavelet Transform fusion process.

The image fusion based on discrete wavelet transform (DWT) consists on merging the wavelet decompositions of the given images (I_v, I_{ir}) using fusion methods applied to approximations coefficients and details coefficients. A scheme of the DWT fusion process is presented in Fig. 1. Initially, the process starts by decomposing the given images into frequency bands. They are analyzed by a fusion rule to determine which component $(D_i = \{d_1, ..., d_n\})$ is removed and which one is preserved. Finally, the inverse transform is applied to get the fused image into the spacial domain. There are different fusion rules (e.g., [16], [17]) to decide which coefficient should be fused into the final result. In the current work high order bands are preserved, while low frequency regions (i.e., smooth regions) are neglected. Figure 2 presents a couple of fused images obtained with the DWT process. Figure 2$(left)$ depicts the visible spectum images (I_v) and the corresponding LWIR images (I_{ir}) are presented in Fig. 2$(middle)$. The resulting fused images (F) are shown in Fig. 2$(right)$.

2.2 Monochrome Threshold based Image Fusion

The monochrome threshold image fusion technique [18] just highlights in the visible image hot objects found in the infrared image. It works as follows. Firstly, an overlay image $O(x, y)$ is created using the thermal image $I_{ir}(x, y)$ and an user defined temperature threshold value τ (see Eq. 1). For each pixel value greater than the threshold value τ a new customized HSV value is obtained, using a predefined H value and the raw thermal intensity for the S and V channels. In the current work the H value is set to 300—this value should be tuned according with the scenario in order to easily identify the objects associated with the target temperature:

$$O(x,y) = \begin{cases} HSV(H, I_{ir}(x,y), I_{ir}(x,y)) & \text{if } I_{ir}(x,y) > \tau \\ HSV(0,0,0) & \text{otherwise} \end{cases} \quad (1)$$

Fig. 2. Illustrations of DWT based image fusion. (*left*) VS image. (*middle*) LWIR image. (*right*) Fused image.

Secondly, after the overlay has been computed, the fused image $F(x, y)$ is computed using the visible image $I_v(x, y)$ and the overlay image $O(x, y)$ (see Eq. 2). The α value is an user defined opacity value that determines how much we want to preserve of the visible image in the fused image:

$$F(x,y) = \begin{cases} I_v(x,y)(1-\alpha) + O(x,y)\alpha & \text{if } I_{ir}(x,y) > \tau \\ I_v(x,y) & \text{otherwise} \end{cases} \tag{2}$$

As a result we obtain an image that is similar to the visible image but with thermal clues. Figure 3 presents a couple of illustrations of the monochrome threshold image fusion process. Figure 3(*left*) depicts the visible spectrum images (I_v); the infrared images (I_{ir}) of the same scenarios are shown in Fig. 3(*middle*) and the resulting fused images (F) are presented in Fig. 3(*right*).To obtain these results the alpha was tuned to 0.3. That leads, if IR pixel intensity is higher than the temperature threshold, to a resulting pixel intensity blend by 30 percent from infrared and 70 percent from visible image.

2.3 Monocular Visual Odometry

The fused images computed above are evaluated using the monocular version of the well-known algorithm proposed by Geiger et al. in [19], which is referred to as LibVISO2.

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

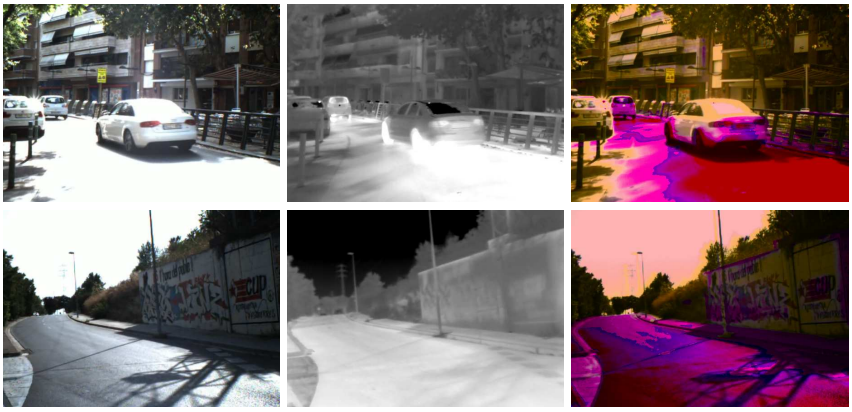

Fig. 3. Illustration of monochrome threshold based image fusion. (*left*) VS image. (*middle*) LWIR image. (*right*) Fused image.

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 obtained with different cross-spectral video sequences. In all the cases GPS information is used as ground truth data to evaluate the performance of evaluated approaches. The GPS ground truth must be considered as a weak ground truth, since it was acquired using a low-cost GPS receiver. Initially, the system setup is introduced and then the experimental result are detailed.

3.1 System Setup

This section details the cross-spectral stereo head used in the experiments together with the calibration and rectification steps. Figure 4 shows an illustration of the whole platform (from the stereo head to the electric car used for obtaining the images).

The stereo head used in the current work consists of a pair of cameras set up in a non verged geometry. One of the camera works in the infrared spectrum,

Fig. 4. Acquisition system (cross-spectral stereo rig on the top left) and electric vehicle used as mobile platform.

more precisely Long Wavelength Infrared (LWIR), detecting radiations in the range of $8-14\ \mu m$. The other camera, which is referred to as (VS) responds to wavelengths from about 390 to 750 nm (visible spectrum). The images provided by the cross-spectral stereo head are calibrated and rectified using [20]; a process similar to the one presented in [3] 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.

The LWIR camera (Gobi-640-GigE from Xenics) provides images up to 50 *fps* with a resolution of 640×480 pixels. The visible spectrum camera is an ACE from Basler with a resolution of 658×492 pixels. Both cameras are synchronized using an external trigger. Camera focal lengths were set so that pixels in both images contain similar amount of information from the given scene. The whole platform is placed on the roof of a vehicle for driving assistance applications.

Once the LWIR and VS cameras have been calibrated, their intrinsic and extrinsic parameters are known, being possible the image rectification. With the above system setup different video sequences have been obtained in urban and semi-urban scenarios. Figure 5 shows the map trajectories of three video sequences. Additional information is provided in Table 1.

Fig. 5. Trajectories used during the evaluations: (*left*) Vid00 path; (*middle*) Vid01 path; (*right*) Vid02 path.

Table 1. Detailed characteristics of the three datasets used for the evaluation.

Name	Type	Duration (sec)	Road length (m)	Average speed (km/h)
Vid00	Urban	49.9	235	17.03
Vid01	Urban	53.6	365	24.51
Vid02	Semi-urban	44.3	370	30.06

3.2 Visual Odometry Results

In this section experimental results and comparisons, with the three video sequences introduced above (see Fig. 5 and Table 1), 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.

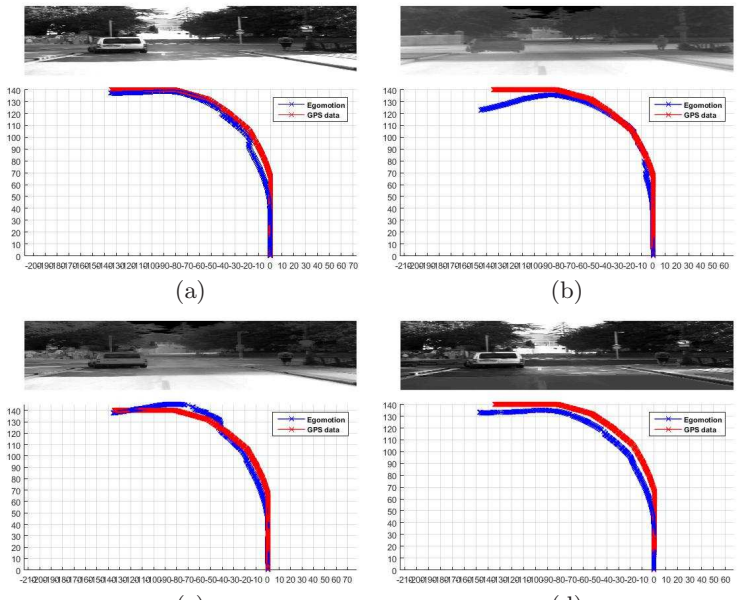

Fig. 6. Estimated trajectories for the Vid00 sequence: (*a*) Visible spectrum; (*b*) Infrared spectrum; (*c*) DWT fused images; and (*d*) Monochrome threshold fused images.

Table 2. VO results in the **Vid00 video sequence** using images from: visible spectrum (VS); Long Wavelength Infrared spectrum (LWIR); fusion using Discrete Wavelet Transform (DWT); and fusion using Monochrome Threshold (MT).

Results	VS	LWIR	DWT	MT
Total traveled distance (m)	234.88	241.27	245	240.3
Final position error (m)	2.9	18	5.4	14.4
Average number of matches	2053	3588	4513	4210
Percentage of inliers	71.5	61.94	60	67.9

Vid00 Video sequence: it consists of a large curve in a urban scenario. The car travels more than 200 meters at an average speed of about 17 Km/h. The VO algorithm (LibVISO2) has been tuned as follow for the different video sequences (see [19] for details on the parameters meaning). In the **visible** spectrum case the bucket size has been set to 16×16 and the maximum number of features per bucket has been set to 4. The τ and match radius parameters were tuned to 50 and 200 respectively. In the **infrared** video sequence the bucket size has been also set to 16×16 but the maximum number of features per bucket has been increased to 6. Regarding τ and match radius parameters, they were set to 25 and 200 respectively. Regarding the VO with fused images the parameters were set as follow. In the video sequences obtained by the **DWT fusion based approach** the bucket size was set to 16×16 and the maximum number of features per bucket to 6; τ and the match radius parameters were set to 25 and 200 respectively. Finally, in the **Monochrome Threshold fusion based approach** the bucket size has been also set to 16×16 but the maximum number of features has been increased to 6. The τ and match radius parameters were tuned to 50 and 100 respectively. The refining at half resolution is disabled, since the image resolution of the cameras is small. Figure 6 depicts the plots corresponding to the different cases (visible, infrared and fused images) when they are compared with ground truth data (GPS information). Quantitative results corresponding to these trajectories are presented in Table 2. In this particular sequence, the VO computed with the visible spectrum video sequence get the best result just followed by the one obtained with the DWT video sequence. Quantitatively, both have a similar final error, on average the DWT relay on more matched points, which somehow would result in a more robust solution. The visual odometry computed with the infrared spectrum video sequence get the worst results; this is mainly due to the lack of texture in the images.

Vid01 Video sequence: it is a simple straight line trajectory on a urban scenario consisting of about 350 meters; the car travels at an average speed of about 25 Km/h. The (LibVISO2) algorithm has been tuned as follow. In the **visible** spectrum case the bucket size was set to 16×16 and the maximum number of features per bucket has been set to 4. The variables τ and match radius parameters are respectively tuned to 25 and 200. The user defined parameters in

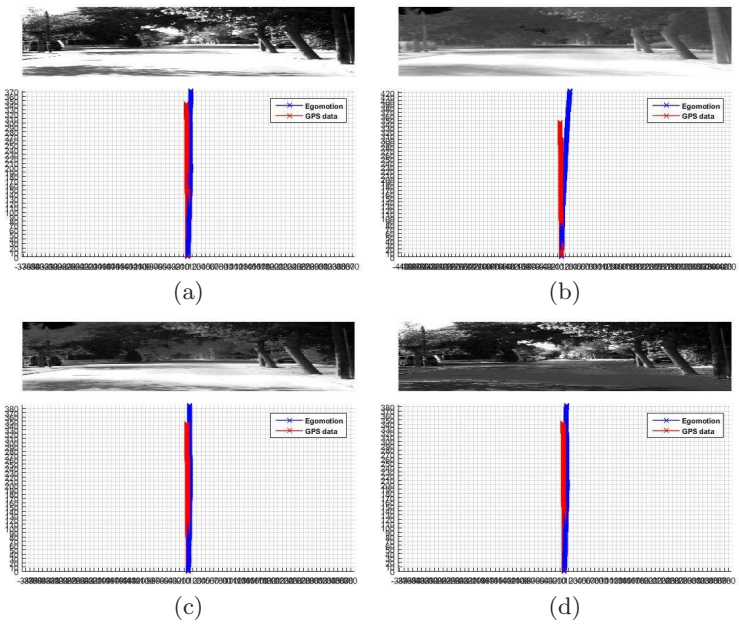

Fig. 7. Estimated trajectories for Vid01 sequence: (*a*) Visible spectrum; (*b*) Infrared spectrum; (*c*) DWT based fused image; and (*d*) Monochrome threshold based fused image.

the **Infrared** case have been set as follow. The bucket size was defined as 16×16 and the maximum number of features per bucket has been set to 50 and 200 respectively. The half resolution was set to zero. The LibVIS02 algorithm has been tuned as follow when the fused images were considered. In the **DWT fusion based approach** the bucket size was set to 16×16 and the maximum number of features per bucket set to 4. The τ and match radius parameters are respectively tuned to 25 and 200. Finally, in the **Monochrome Threshold fusion based approach** the bucket size was set to 16×16 and the maximum number of features per bucket was set to 4. The τ and match radius parameters are respectively tuned to 25 and 200. Figure 7 depicts the plots of the visual odometry computed over each of the four representations (VS, LWIR, DWT fused and Monochrome threshold fused) together with the corresponding GPS data. The visual odometry computed with the infrared video sequence gets the worst result, as can be easily appreciated in Fig. 7 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 and Monochrome Threshold based image fusion) are similar both qualitatively and quantitatively.

Vid02 Video sequence: it is a "L" shape trajectory on a sub-urban scenario. It is the longest trajectory (370 meters) and the car has traveled faster than

Table 3. VO results in the **Vid01 video sequence** using images from: visible spectrum (VS); Long Wavelength Infrared spectrum (LWIR); fusion using Discrete Wavelet Transform (DWT); and fusion using Monochrome Threshold (MT).

Results	VS	LWIR	DWT	MT
Total traveled distance (m)	371.8	424	386	384
Final position error (m)	32.6	84.7	44	42.7
Average number of matches	1965	1974	2137	2060
Percentage of inliers	72.6	67.8	61.5	65.4

in the previous cases (about 30 Km/h). The (LibVISO2) algorithm has been tuned as follow. In the **visible** spectrum case the bucket size was set to 16×16 and the maximum number of features per bucket set to 4. Regarding τ and match radius parameters, they were tuned as 25 and 200 respectively. In the **infrared** case the bucket size has been set to 16×16 and the maximum number of features per bucket set to 4. τ and match radius parameters were respectively tuned to 50 and 100. In the fused image scenario the LibVIS02 algorithm has been tuned as follows. First, in the **DWT fusion based approach** the bucket size has been set to 16×16 and the maximum number of features per bucket set to 4. Like in the visible case, the τ and match radius parameters were tuned to 25 and 200 respectively. Finally, in the **Monochrome Threshold fusion based approach** the bucket size has been defined as 16×16 and the maximum number of features per bucket set to 4. The τ and match radius parameters were respectively tuned to 50 and 200. In this challenging video sequence the fused based approaches get the best results (see Fig. 8). It should be highlighted that in the Monochrome Threshold fusion the error is less than half the one obtained in the visible spectrum (see values in Table 4).

Table 4. VO results in the **Vid02 video sequence** using images from different spectrum and fusion approaches (VS: visible spectrum; LWIR: Long Wavelength Infrared spectrum, DWT: fusion using Discrete Wavelet Transform, MT: fusion using Monochrome Threshold).

Results	VS	LWIR	DWT	MT
Total traveled distance (m)	325.6	336.9	354.4	371.5
Final position error (m)	37.7	48.7	37.2	14.3
Average number of matches	1890	1028	1952	1374
Percentage of inliers	70	65.8	61	66

In the general, the usage of fused images results in quite stable solutions; supporting somehow the initial idea that classical approaches can be improved when the given cross-spectral images are fused and represented in new domains.

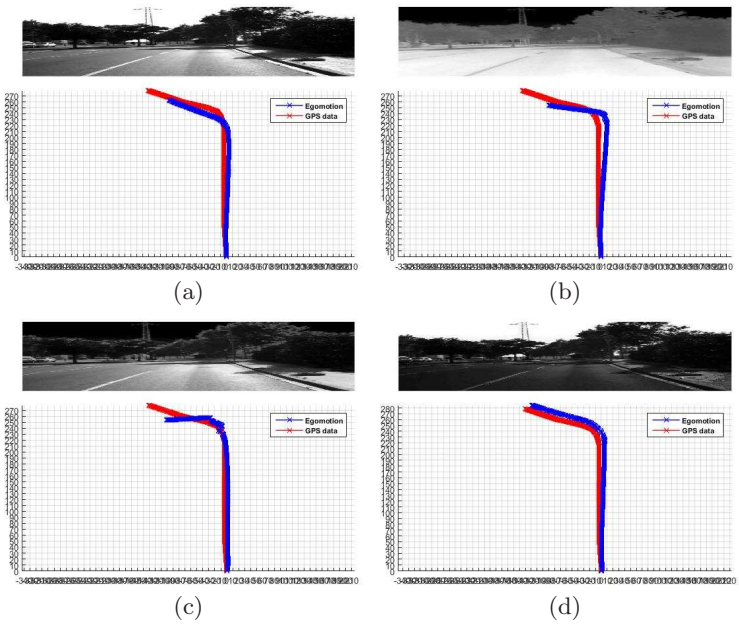

Fig. 8. Estimated trajectories for Vid02 sequence: (*a*) Visible spectrum; (*b*) Infrared spectrum; (*c*) DWT fused image; and (*d*) Monochrome threshold based fused image.

4 Conclusion

The manuscript evaluates the performance of a classical monocular visual odometry by using images from different spectra represented in different domains. The obtained results show that the usage of fused images could help to obtain more robust solutions. This evaluation study is just a first step to validate the pipeline in the emerging field of image fusion. As future work other fusion strategies will be evaluated and a more rigorous framework set up.

Acknowledgments. This work has been supported by: the Spanish Government under Project TIN2014-56919-C3-2-R; the PROMETEO Project of the "Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador"; and the "Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya" (2014-SGR-1506). C. Aguilera was supported by Universitat Autònoma de Barcelona.

References

1. Geronimo, D., Lopez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(7) (2010) 1239–1258

2. Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: IEEE International Conference on Computer Vision and Pattern Recognition. (2015)
3. Barrera, F., Lumbreras, F., Sappa, A.D.: Multimodal stereo vision system: 3D data extraction and algorithm evaluation. IEEE Journal of Selected Topics in Signal Processing **6**(5) (2012) 437–446
4. Barrera, F., Lumbreras, F., Sappa, A.D.: Multispectral piecewise planar stereo using manhattan-world assumption. Pat. Recognition Letters **34**(1) (2013) 52–61
5. Conaire, C.O., O'Connor, N.E., Cooke, E., Smeaton, A.: Multispectral object segmentation and retrieval in surveillance video. In: IEEE International Conference on Image Processing. (2006) 2381–2384
6. Denman, S., Lamb, T., Fookes, C., Chandran, V., Sridharan, S.: Multi-spectral fusion for surveillance systems. Comp. & Electrical Engineering **36**(4) (2010) 643–663
7. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: IEEE Intgernational Conference on Computer Vision and Pattern Recognition. Volume 1. (2004) I–652
8. Scaramuzza, D., Fraundorfer, F., Siegwart, R.: Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC. In: IEEE International Conference on Robotics and Automation. (2009) 4293–4299
9. Tardif, J.P., Pavlidis, Y., Daniilidis, K.: Monocular visual odometry in urban environments using an omnidirectional camera. In: IEEE International Conference on Intelligent Robots and Systems IROS. (2008) 2531–2538
10. Howard, A.: Real-time stereo visual odometry for autonomous ground vehicles. In: International Conference on Intelligent Robots and Systems. (2008) 3946–3952
11. Scaramuzza, D., Fraundorfer, F.: Visual odometry [tutorial]. IEEE Robotics & Automation Magazine **18**(4) (2011) 80–92
12. Comport, A.I., Malis, E., Rives, P.: Accurate quadrifocal tracking for robust 3D visual odometry. In: IEEE International Conference on Robotics and Automation, ICRA, 10-14 April 2007, Roma, Italy. (2007) 40–45
13. Chilian, A., Hirschmüller, H.: Stereo camera based navigation of mobile robots on rough terrain. In: IEEE International Conference on Intelligent Robots and Systems IROS, IEEE (2009) 4571–4576
14. Nilsson, E., Lundquist, C., Schön, T., Forslund, D., Roll, J.: Vehicle motion estimation using an infrared camera. In: 18th IFAC World Congress, Milano, Italy, 28 August-2 September, 2011, Elsevier (2011) 12952–12957
15. Mouats, T., Aouf, N., Sappa, A.D., Aguilera-Carrasco, C.A., Toledo, R.: Multispectral stereo odometry. IEEE Transactions on Intelligent Transportation Systems **16**(3) (2015) 1210–1224
16. Amolins, K., Zhang, Y., Dare, P.: Wavelet based image fusion techniques — an introduction, review and comparison. ISPRS Journal of Photogrammetry and Remote Sensing **62**(4) (2007) 249–263
17. Suraj, A., Francis, M., Kavya, T., Nirmal, T.: Discrete wavelet transform based image fusion and de-nosing in FPGA. Journal of Electriacal Systems and Information Technology **1** (2014) 72–81
18. Rasmussen, N.D., Morse, B.S., Goodrich, M., Eggett, D., et al.: Fused visible and infrared video for use in wilderness search and rescue. In: Workshop on Applications of Computer Vision (WACV), IEEE (December 2009) 1–8
19. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3D reconstruction in real-time. In: Intelligent Vehicles Symposium (IV). (2011)
20. Bouguet, J.Y.: Camera calibration toolbox for matlab (July 2010)