Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation

Article *in* IEEE Journal of Selected Topics in Signal Processing · August 2012

DOI: 10.1109/JSTSP.2012.2204036

Fernando Barrera
University of Strasbourg
13 PUBLICATIONS **209** CITATIONS
SEE PROFILE

Felipe Lumbreras
Autonomous University of Barcelona
57 PUBLICATIONS **1,025** CITATIONS
SEE PROFILE

Angel Domingo Sappa
Autonomous University of Barcelona
188 PUBLICATIONS **2,638** CITATIONS
SEE PROFILE

Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation

Fernando Barrera Campo, Felipe Lumbreras Ruiz, and Angel Domingo Sappa, *Member, IEEE*

***Abstract*—This paper proposes an imaging system for computing sparse depth maps from multispectral images. A special stereo head consisting of an infrared and a color camera defines the proposed multimodal acquisition system. The cameras are rigidly attached so that their image planes are parallel. Details about the calibration and image rectification procedure are provided. Sparse disparity maps are obtained by the combined use of mutual information enriched with gradient information. The proposed approach is evaluated using a Receiver Operating Characteristics curve. Furthermore, a multispectral dataset, color and infrared images, together with their corresponding ground truth disparity maps, is generated and used as a test bed. Experimental results in real outdoor scenarios are provided showing its viability and that the proposed approach is not restricted to a specific domain.**

***Index Terms*—Color and infrared images, multimodal stereo rig, sparse 3D maps.**

I. Introduction

The coexistence of visible (VS) and infrared (IR) cameras has opened new perspectives for the development of multimodal systems. In general, visible and infrared cameras are used as *complementary* sensors in applications such as video surveillance (e.g. [1], [2]) and driver assistance systems (e.g. [3]). Visible cameras provide information at diurnal scenarios while infrared cameras are used as night vision sensors. More recently, Near-InfraRed (NIR) and visible images are used under a common framework to improve the accuracy of the registration (e.g., [4]). It works by capturing near-infrared and visible images at the same time using a single sensor. In other words, every pixel of the 2D image contains information about the red, green, blue and NIR channels.

All the approaches mentioned above involve registration and fusion stages, resulting in an image that even though contains several channels of information lies in the 2D space. The current work goes beyond classical registration and fusion schemes by formulating the following question: "*is it possible to obtain 3D information from a multispectral stereo rig?*". It is clear that if the objective is to obtain depth maps close to state-of-the-art, classical binocular stereo systems (VS/VS) are more appropriated. Therefore, the motivation of current work is to show that the generation of 3D information from images belonging to different spectral bands is possible. The proposed multispectral stereo rig is built with two cameras, which are rigidly mounted and oriented in the same direction. These cameras work at different spectral bands, while one measure radiation in the visible band the other one registers infrared radiation. From now on, this system will be referred to as *multimodal stereo head*, which is able to provide a couple of multispectral images.

The role of cameras in the proposed multimodal stereo system is not only restricted to work in a *complementary* way (as it is traditionally) but also in a *cooperative* fashion, being able to extract 3D information. This challenge represents a step forward in the state-of-the-art of 3D multispectral community, and results obtained from this research by sure can benefit applications in the driver assistance or video surveillance domains, where the detection of an object of interest can be enriched with an estimation of its aspect or distance from the cameras.

The performance of a stereo vision algorithm is directly related to its capacity to find good correspondences (*matching*) between pairs of images, this task relies on the similarity function used to match features. In the multispectral case similarity functions such as: SAD (sum of absolute differences), NCC (normalized cross correlation), SSD (sum of squared differences) or Census transform cannot be used since a linear correlation between the data cannot be assumed [5]. In the current work a non linear similarity function, that establish the relationship between multispectral images is presented. In other words, it is able to associate information content between IR and VS images. Through this manuscript VS and color images will refer to images obtained by classical color cameras; these terms are used interchangeably herein.

Multispectral matching has been widely studied in registration and fusion problems, specially in medical imaging (e.g., [6]–[8]). However, there are few research related with the correspondence problem when infrared and color images are considered. Hence, it is not clear how to exploit visible and infrared imaging in a combined framework to obtain 3D information.

Most of the stereo heads presented in the literature, and other commercially available, are built from cameras that have the same specifications (i.e., sensor and focal length). This choice constrains the problem and facilitates the reuse of software and

Manuscript received July 31, 2011; revised December 27, 2011 and April 12, 2012; accepted June 02, 2012. Date of publication June 18, 2012; date of current version August 10, 2012. This work was supported in part by the projects TIN2011-29494-C03-02 and TIN2011-25606, and in part by the research program Consolider-Ingenio 2010: MIPRCV (CSD2007-00018). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chunrong Yuan.

F. Barrera Campo and A. D. Sappa are with the Computer Vision Center, 08193 Bellaterra, Barcelona, Spain (e-mail: jfbarrera@cvc.uab.es; asappa@cvc.uab.es).

F. Lumbreras Ruiz is with the Computer Vision Center, 08193 Bellaterra, Barcelona, Spain, and also with the Computer Science Department, Universitat Autònoma de Barcelona, 08290 Barcelona, Spain (e-mail: felipe@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/JSTSP.2012.2204036

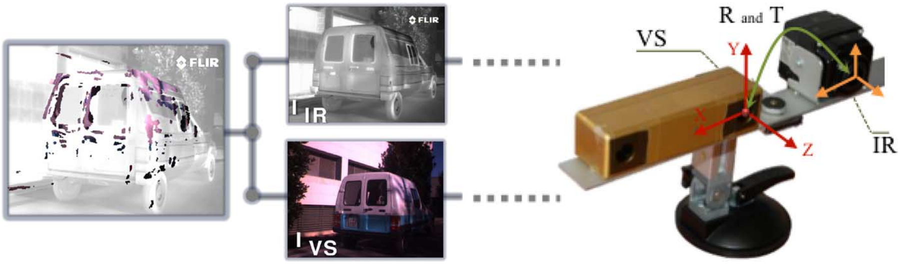


Fig. 1. Proposed framework: (*left*) sparse 3D map, over the corresponding IR image, obtained with the proposed approach; (*middle*) a couple of images from the IR and color (left camera of Bumblebee) cameras; (*right*) multispectral stereo head.

published methods. However, the case tackled in the current work is far more complex since heterogeneous sensors are used, besides the intrinsic problems due to multimodality. So, the alignment of two views coming from cameras with different sensors and intrinsic parameters should be taken into account, which is more difficult than a classical VS/VS stereo heads.

The use of multimodal stereo heads (IR/VS) has attracted interest of researchers in different computer vision fields, for examples: human detection [9], video surveillance [10], and 3D mapping of surface temperature [11], [12]. Recently, [13] presents a comparison of two stereo systems, one working in the visible spectrum (composed of two color cameras) and the other in the infrared spectrum (using two IR cameras). Since the study was devoted to pedestrian detection, the authors conclude that both, color and infrared based stereo, have a similar performance for such a kind of applications. However, in order to have a more compact system they propose a multimodal trifocal framework defined by two color cameras and an IR camera. In this framework, infrared information is not used for stereoscopy but just for mapping IR information over the 3D points computed from the VS/VS stereo head. This allows to develop robust approaches for video surveillance applications (e.g., [10]).

On the contrary to the previous approaches, a multimodal stereo head constructed with just two cameras: an infrared and a color one is presented in [14]. This minimal configuration is adopted in the current work since it is the most compact architecture in terms of hardware and software. Critical issues such as camera synchronization, control signaling, bandwidth, image processing, among other have a minimal impact in the overall performance, and can be easily treated by an acquisition system such as the one presented in [15]. In Krotosky *et al.* [14] this compact multimodal stereo head (IR/VS) is used for matching regions that contain human body silhouettes. Since their contribution is aimed at person tracking some assumptions are applied, for example a foreground segmentation for disclosing possible human shapes, which are corresponded by maximizing mutual information [16]. Although, these assumptions are valid, they restrict the scope of applications.

A more general solution should be envisaged, allowing such a kind of multimodal stereo head to be used in different applications. In other words, the matching should not be constrained to regions containing human body silhouettes. The current paper has three main contributions. Firstly, a robust approach that allows to compute sparse depth maps from a multimodal stereo head is proposed. Since it is not restricted to a specific application it can be used in any scenario. The second contribution is the adaptation to the multispectral case of a recently presented methodology for comparing and evaluating stereo matching algorithms. This evaluation method has been proposed for classical stereo heads where both cameras work in the same spectral band [17]. It is based on Receiver Operating Characteristics (ROC) curves that capture both error and sparsity. Finally, a dataset with VS and IR images, together with their corresponding disparity maps and 3D models, is publicly available for evaluating different approaches. Up to our knowledge there is not such a kind of dataset in the research community to be used as a test bed.

Although the proposed approach is motivated for recovering 3D information, optionally it could help to solve other multimodal problems. Due to the fact that most existing multimodal systems are affected by the same problem. That is, statistical independence between the different modalities, which makes difficult its correlation. Our approach offers a non-heuristic based solution, which is a novel feature with respect to state of the art. The current work presents an approach that reveals the information shared by the modalities, and from these correspondences find the match between blobs or image regions. The latter is relevant for multimodal applications such as moving target detection, medical imaging, video fusion, among other.

The paper is organized as follows. Section II presents the multimodal stereo head and the proposed approach for computing sparse depth maps. Section III introduces the adaptation of the evaluation methodology to tackle the multispectral stereo case. Additionally, it presents in details the multispectral dataset. Experimental results with different scenarios are presented in Section IV, together with the technique used for setting the parameters of the algorithm. Conclusions and final remarks are detailed in Section V.

II. Multimodal Stereo

This section presents in detail the multimodal stereo head together with the proposed algorithm for computing sparse 3D maps. Fig. 1 shows an illustration of the multimodal platform and a couple of images used to compute a sparse 3D representation. The different challenges of the tackled problem can be appreciated in this illustration, from the image acquisition and depth map estimation to the evaluation of the performance of the algorithm. The different stages of the proposed multispectral stereo are presented in detail below.

A. Multimodal Stereo Head

In the current work, a multimodal stereo head with an IR camera (PathFindIR from Flir)[1] and a color camera is built. The color camera, by convenience, corresponds to the left camera of a commercial stereo vision system (Bumblebee, from Point Grey).[2] The Bumblebee stereo head is used for validating the results and consists of two cameras Sony ICX084 with Bayer pattern CCD sensors, and 6 mm focal length lenses. It is a pre-calibrated system that does not require in-field calibration. In summary, two stereo systems coexist (see Fig. 1(*right*)). The left camera coordinate system of Bumblebee is used as a reference system for both stereo heads. In this way, a kind of ground truth for the depth of each pair of images (infrared and color) is obtained from the Bumblebee stereo head.

The IR camera, which will be referred just as IR, detects radiations in the range 8–14 $\mu\mathrm{m}$ (long-wavelength infrared), whereas the color camera, referred to as VS, responds to wavelengths from about 390 to 750 nm (visible spectrum).

B. Calibration and Rectification

The multimodal stereo head has been calibrated using Bouguet's toolbox [18]. The main challenge in this stage is to make visible the calibration pattern in both cameras. In order to do this, a special metallic checkerboard has been made using a thin aluminium metallized paper. Black squares over this surface are generated by means of a laser printer, being able to detect them from both VS and IR cameras. Fig. 2(*left*) shows a pair of calibration images (IR and color). Despite of using a metallic calibration pattern, the junctions of black and white squares are not correctly detected due to thermal diffusion. Hence, calibration points are extracted using a saddle point detector, instead of a classical corner detector. In our particular case the use of saddle points results in a more stable detection; it is due to the fact that thermal variation between black and white squares are not enough to generate step edges, and the structure of junctions looks more like saddle points than corners [19]. Fig. 3(*top*) shows three illustrations of junctions obtained with the saddle point detector; note that even though the contrast of these infrared images is different the junctions are correctly detected. Fig. 3(*bottom*) depicts local structure indicated by the red windows in Fig. 3(*top*); the green points are saddle points while red ones are corners; straight lines show diagonal directions where their intersection corresponds to the most likely position of junctions. As can be seen in these plots, the green points are nearer to the intersections than the corresponding red ones.

Three independent calibration processes under different temperature were performed to study the robustness of intrinsic parameters of IR camera when the saddle point detector is used; as a result, the obtained intrinsic parameters were stables beside the changes in temperature. Notice that the IR images in Fig. 3(*top*) correspond to one image of those calibration sequences.

Once the IR and VS cameras have been calibrated, their intrinsic and extrinsic parameters are known, being possible, not only the image rectification, but also to calculate the disparity map of the scene. The image rectification was done, using the method proposed in [20], with an accuracy improvement due to the inclusion of the radial and tangential distortion coefficients into their camera model. An example of rectified images is shown in Fig. 2(*right*).

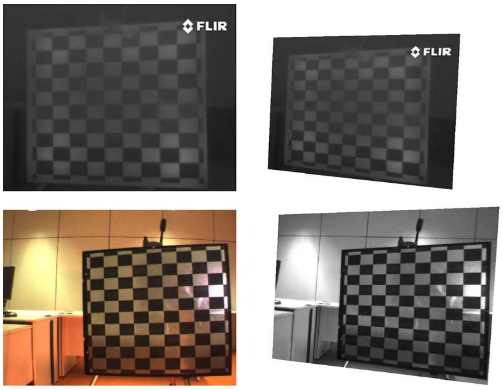


Fig. 2. (*top-left*) Infrared image of the checkerboard pattern. (*top-right*) Infrared rectified image. (*bottom-left*) Original color image. (*bottom-right*) Rectified image.

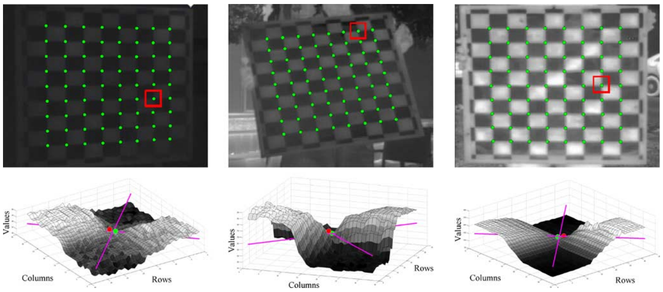


Fig. 3. Saddle points extracted from infrared images of the checkerboard pattern at different temperatures.

C. Matching Cost Computation

The definition of a cost function able to find the good matching between information provided by the VS and IR cameras is a challenging task, due to their poor correlation [21], [22]. In spite of that, recent works on computational stereo [23] have shown that mutual information is a nonparametric cost function able to address nonlinear correlated signals. However, we have found undesirable behaviors when it is used as a cost function in the multispectral stereo problem. Mainly, due to its low capability to distinguish the correct match from a given set of patterns.

In the current work the problem mentioned above is tackled by enriching mutual information (I) with gradient information (G), although it has been presented in [24] for a template matching problem, we have used it as a similarity function for multispectral signals. Additionally, in the current work its accuracy is improved through a Parzen window estimator with a Gaussian distribution, as will be described later.

Correspondence search is done as follows. A window $w_l(x,y)$ with size wz × wz and centered at a point (x,y) is selected from the left image. Then, it is compared to a set of windows (w_r) extracted from the right image. Thus, a cost value is obtained for each pair $w_l(x,y)$ and $w_r(x,y+d)$, where d is the disparity value that varies between two limits;

[1] www.flir.com.

[2] www.ptgrey.com.

these limits depend on depth of scene $d_{min} \leq d \leq d_{max}$. The cost value mentioned above could have different interpretations, in the current work it represents the degree of similarity between the two windows. Similarity is measured by means of the combination of two functions: mutual information and gradient information; high cost values represent similarities in the orientation of gradient vectors and content information.

By definition, mutual information I [25] is estimated from two windows that are extracted from the multispectral images and preprocessed as follows. Each pixel value inside these windows is scaled to a range [0, 1], and then quantized independently into Q levels. So, mutual information I of a pair of preprocessed windows w_l and w_r is estimated as:

$$I(w_l; w_r) = \sum_{a_i \in Q} \sum_{b_j \in Q} p_{w_l w_r}(a_i, b_j) \log \frac{p_{w_l w_r}(a_i, b_j)}{p_{w_l}(a_i) p_{w_r}(b_j)}, \tag{1}$$

where a_i and b_j are levels of the quantization; p_{w_l} and p_{w_r} are their respective marginal probability mass functions; and $p_{w_l w_r}$ is the joint probability mass function.

The join probability $p_{w_l w_r}$ is estimated in two steps. Firstly, from w_l and w_r a 2D histogram is computed, where every entry is obtained as follow:

$$p(a_i, b_j) = \frac{1}{N} \sum_{q=1}^{N} T\left[(a_i, b_j) = (w_l(q), w_r(q))\right], \tag{2}$$

where $T[\bullet]$ is 1 if the argument is true and 0 otherwise; and N is the size of the matching windows (note that both w_l and w_r has the same size). Next, a Parzen estimator is used following [26]. It assumes a Gaussian distribution g with standard deviation σ_g on every sample in the histogram $p(a_i, b_j)$, which was previously obtained. Then:

$$p_{w_l w_r}(a_i, b_j) = p(a_i, b_j) * g(a_i, b_j; \sigma_g). \tag{3}$$

The rest of probabilities, p_{w_l} and p_{w_r}, are determined by summing along each dimension of the previous joint probability (see [16] for more details).

In this problem the calibration and rectification stages (Section II-B) have a decisive impact, because not only the search for correspondences is restricted to one dimension, but also the contours and edges of the objects contained in the scene have a similar aspect, therefore increasing its probability of coincidence. As shown in [22], contours and boundaries are highly correlated in multispectral images (IR/VS). This is exploited in the current work to enrich mutual information by using the formulation presented in [27]. The gradient information is obtained by convolving the two-dimensional images (color and infrared) with a Gaussian derivative kernel of order n:

$$L_n(\mathbf{x}) = I_k(\mathbf{x}) * g_n(\mathbf{x}; \boldsymbol{\sigma}), \tag{4}$$

where $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$, I_k is the infrared (I_{IR}) or visible (I_{VS}) image, and $g_n(\mathbf{x}; \boldsymbol{\sigma})$ is the Gaussian derivative kernel of order n. If $n = 0$ the Gaussian function is obtained, otherwise its corresponding derivative kernel. In this section, since only gradient information is required, L_1 is computed. L_0 is also computed and used for mutual information estimation (1) in a scale space representation, as will be presented next. It could happens that gradient vectors appear in both modalities but with a phase difference near to 0 or π (phase or counter-phase) [4]. This fact is used to unveil possible matchings. So, let $\mathbf{x}$ and $\mathbf{x}'$ be the coordinates of two corresponding pixels that belong to w_l and w_r respectively. Then, their phase difference is defined as:

$$\theta(\mathbf{x}, \mathbf{x}') = \arccos\left(\frac{L_1(\mathbf{x}) \cdot L_1(\mathbf{x}')}{|L_1(\mathbf{x})|\,|L_1(\mathbf{x}')|}\right), \tag{5}$$

where $L_1(\mathbf{x}) \cdot L_1(\mathbf{x}')$ is the dot product of their gradient values in x and y direction. Following [24] and [27] the phase difference (5) is weighted by a function $\psi(\theta)$ that penalizes gradient vectors that are not in phase or counter-phase:

$$\psi(\theta) = \frac{\cos(2\theta) + 1}{2}. \tag{6}$$

The gradient information to be added to the mutual information is computed as follows:

$$G = \sum_{\mathbf{x} \in w_l, \mathbf{x}' \in w_r} \psi\left(\theta(\mathbf{x}, \mathbf{x}')\right) \min\left(|L_1(\mathbf{x})|, |L_1(\mathbf{x}')|\right). \tag{7}$$

Finally, the gradient information is combined with the mutual information through their product. Although other combinations are also possible, it has been shown in the literature that for this kind of multimodal fusion task the product is the best way to combine them. Furthermore, it does not require a user defined parameter to weight their contribution (e.g., [24], [27]). Hence, the product of I and G has been selected as an aggregation operator. It satisfies the properties of monotonicity, continuity and stability for linear transformations, while reducing the data into a unique representative value. Thus, the matching cost of two blocks: w_l and w_r is given by the next expression:

$$IG(w_l; w_r) = I(w_l; w_r) \cdot G(w_l; w_r). \tag{8}$$

D. Cost Aggregation

The matching cost function presented above is used in a scale space representation scheme. In the current work the scale space representation contains three levels obtained by convolving the original images with a Gaussian derivative kernel of different σ. Hence, (4) becomes:

$$L_n(\mathbf{x}; \boldsymbol{\sigma}_\mathbf{t}) = I_k(\mathbf{x}) * g_n(\mathbf{x}; \sigma_t). \tag{9}$$

Hereinafter, for a compactness reason, a given level of the scale space representation is referred to as σ_t.

In a given level σ_t, and for a pixel $\mathbf{x}$, a searching window centered on it is defined. This window, with a size of (wz × wz), is used to compute the matching cost in its corresponding searching space. Since rectified images are used this searching is restricted to a row on the other image. This generates a set of

IG values that need to be merged with the values coming from the other levels of the scale space representation.

The matching cost values IG computed at different scales are merged as follow:

$$IGSS(\mathbf{x};\sigma_t) = \lambda_t IG(\mathbf{x};\sigma_t) + (1-\lambda_t) IG(\mathbf{x};\sigma_{t-1}), \quad (10)$$

where λ_t is the confidence of current IG.

E. Disparity and Depth Computation

The process of disparity selection consists of two steps. Initially, the disparities with higher cost values are selected as correct with a classical winner take all criterion. In these cases, a correct match is determined by the positions d (image coordinate) where the cost function reaches the maximum value: $\arg\max_d\{IGSS(w_l(x,y); w_r(x+d,y))\}$. The disparity map obtained after this first step contains several wrong entries due to regions with low texture or no information. Note that the multispectral stereo matching case is more complicated than traditional (VS/VS) ones. The latter is due to the fact that, for instance, an object in the scene could appear textured in the visible spectrum, while it could have the same temperature all over its surface, therefore appear as a constant region in the infrared image, and vice versa.

As mentioned above, it is hard to select the correct d between several candidates with similar scores. Therefore, a second step to reject mismatching candidates is added. It consists in labelling as correct those correspondences with a cost score higher than a given threshold τ. The selection of this threshold is based on error rates (see Section III-A). Next, these reliable matchings are used for bounding the searching space in their surrounding. As it will be shown in the 3D maps, this helps to discard wrong matchings.

The τ parameter is included into our formulation for picking up only those pixels with large $IGSS$ values. Since the $IGSS$ cost function is reliable in textured regions, and those regions have higher $IGSS$ cost, τ is used as a threshold that split up the cost map into two groups: *i*) reliable matches and *ii*) unreliable matches. This parameter exploits the correlation between IR and VS edges.

Finally, a quadratic curve is used for a fine estimation of disparity values; this function fits a polynomial to points $\{d-1, d, d+1\}$ and its respective cost values. After computing the disparity of every point P in the images (color and infrared), their corresponding 3D positions (X,Y,Z) are obtained using a standard function for triangulation, which is included into the calibration toolbox [18].

III. Evaluation

The proposed stereo algorithm has been evaluated by adapting a technique recently presented for classical (VS/VS) semi-dense stereo matching algorithms. Furthermore, a well detailed multispectral dataset together with its corresponding ground truth is proposed. All this material (i.e., multispectral stereo dataset and ground truth images) is available through our website[3] for an automatic evaluation and comparisons of multispectral stereo algorithms. The next sections describe the quality metrics used for evaluating the performance of the proposed multispectral stereo matching algorithm.

A. Evaluation Methodology

In general, stereo algorithms have been evaluated following the methodology proposed in [28], which has become in a *de facto standard* in the stereo community. It presents two quality measures: *i*) RMS (root mean squared) error; and *ii*) percentage of bad matching pixels. In both cases, resulting disparity maps are compared to a known ground truth in a dense fashion. However, in uncontrolled scenarios, as outdoors, trying to get ground truth data as presented in [29] or [30] is not feasible, for that reason, we must evaluate our proposed algorithm following a semi-dense methodology.

The method presented in [17] capture both, error and sparsity in a single value, which is suitable for our dataset. So, we extend this framework to the multispetral case. The pairs: *error* and *sparsity* are plotted in a Receiver Operating Characteristics (ROC) curve as a unique value, letting visualize how performance is affected as more disparity values are taken. Remember that every disparity obtained by our method have a cost value associated, which depends on I and G. Therefore, regions with low information (low entropy) or without texture (gradient) could be rejected considering their cost. During the evaluation process the best τ parameter could be easily identified (see Section II-E).

In the current work, ROC curves have been used for evaluating the performance of the proposed multispectral stereo algorithm (Section II) independently of parameter settings (wz, Q, σ_t). The evaluation procedure is briefly detailed below following the original notation. The statistics about the quality of a semi-dense stereo algorithm should capture both: *i*) how often a matching error happens and *ii*) how often a matching is not found. These two values define the Error Rate (ER) and the Sparsity Rate (SR) respectively. In other words, the ER represents the percentage of incorrect correspondences:

$$ER = \frac{incorrect\_correspondences}{all\_matchable\_pixels}. \quad (11)$$

On the other hand, the SR is defined as the percentage of all missing correspondences over the set of matchable pixels:

$$SR = \frac{missing\_correspondences}{all\_matchable\_pixels}. \quad (12)$$

Note that these values are not computed over the whole set of pixels but over those pixels with a match in the ground truth. An illustration of ROC curves, for different scenarios, can be seen in Fig. 6; they will be explained in the experimental result section. In these representations there are four interesting points: the origin, which represents a dense and perfect matching algorithm; its opposite, where no correct matches are found; the (0,1) point corresponds to an algorithm that is dense but fully wrong; and finally, the (1,0) point that corresponds to a disparity map completely empty.

[3] http://www.cvc.uab.es/adas/datasets/cvc-multimodal-stereo.

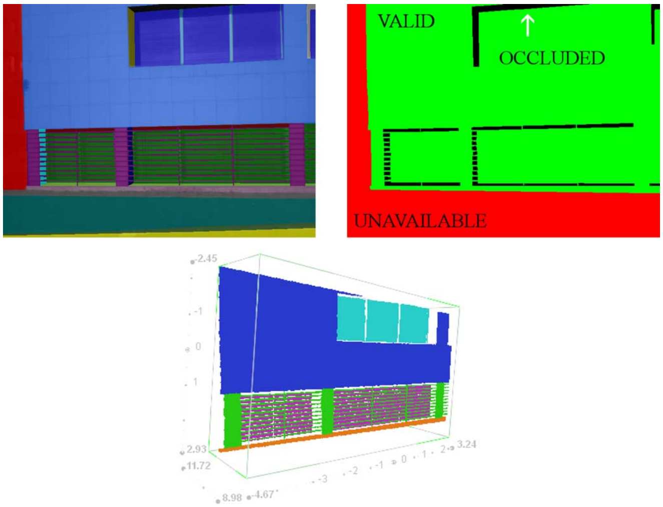


Fig. 4. (*top-left*) Facade image from the proposed dataset overlapped with a mask from the segmentation. (*top-right*) Mask of regions with occluded and no depth information. (*bottom*) Synthetic 3D representation generated from the visible stereo pair used as ground truth for evaluating the multispectral depth map.

B. Multispectral Datasets

A multispectral dataset has been generated for evaluating the different stages of the proposed algorithms. It contains multispectral images, ground truth disparity maps and ground truth depth maps. All this information was obtained as indicated below.

The dataset consists of four kinds of images, which are classified by their context and predominant geometry: *i*) roads; *ii*) facades; *iii*) smooth surfaces; and *iv*) OSU Color-Thermal dataset. The first three groups were acquired with the proposed stereo head and contain outdoor scenarios with one or multiple planes and smooth surfaces. The latter subset contains perfectly aligned IR and color images (i.e., without disparity). It was obtained from [31] and is publicly available.[4] These images are particularly interesting, since they are aligned the ground truth of disparity maps can be approximated by assuming a registration accuracy of about ± 2 pixels. Fig. 5 shows an illustration of the whole dataset.

The multispectral stereo images in the dataset have been enriched with ground truth disparity maps and ground truth depth maps semi-automatically generated. These ground truth were obtained by fitting planes to the 3D data points obtained from the Bumblebee stereo head. It works as follows. Firstly, a color image from the left Bumblebee camera is manually segmented into a set of planar regions (see Fig. 4(*top-left*)). Planar regions are easily identified since are the predominant surfaces in the considered outdoor scenarios. Then, every region is independently fitted with a plane using their corresponding 3D data points, by orthogonal regression using principal components analysis. Fig. 4(*bottom*) shows an illustration of the synthetic 3D representation containing different planes. Additionally, during this semi-automatic ground truth generation process, labels for occluded, valid and unavailable pixels are obtained (see Fig. 4(*top-right*)). These labels are needed for the evaluation methodology (Section III-A).

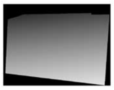
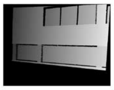
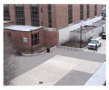
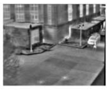
Dataset	Image			Pairs of
Name	*VS*	*IR*	Disparity & Depth	Images
Roads				2
Facades				2
Smooth surfaces				19
OSU Color-Thermal Dataset [31]			—	2

Fig. 5. Illustration of the four subsets of images contained in the proposed multispectral dataset.

Once the 3D planes for a given image are obtained, since they are referred to the VS camera, the corresponding data points are projected to the infrared camera. Thus, a ground truth disparity map is obtained. The fourth column of Fig. 5 shows some of these disparity maps and a sparse 3D representation.

In the case of smooth surfaces (e.g., third row in Fig. 5) no planes are fitted, and depth information provided by Bumblebee is used as a reference. Bumblebee software offers a trade off between density and accuracy of data points. Hence, in order to have a good representation, its parameters have been tuned so that 3D models are dense enough and contain few noisy data. Those models should not be considered as ground truth, strictly speaking, however we use them as a baseline for qualitative comparisons.

The evaluation by ROC curves compares row by row, a horizontal profile belonging to ground truth disparity map with its corresponding one obtained by the tested algorithm. A correct matching is assumed when the difference with respect to the correct value is smaller than or equal to 1 pixel. Note that only three sets of images (i.e., roads, facades and OSU) are used for the evaluation to avoid the problem of occlusion, which is slightly different in the VS/VS and VS/IR stereo rigs. Regarding the *roads* dataset, in all the image pairs there is a single plane hence there are not occluded areas; while in the *facades* dataset occluded areas are removed by generating a synthetic 3D model. Let us remember that OSU dataset was not obtained with our multispectral stereo rig; it is provided by [31] and contains perfectly aligned VS and IR images. The *smooth surfaces* dataset is not used during the evaluation since the differences between occluded areas in VS/VS and VS/IR stereo rigs could affect the results. Hence, the *smooth surfaces* dataset is just used for a qualitative validation of the proposed approach.

Fig. 4(*top-right*) shows three kind of regions identified in our dataset: *Occluded*; *Unavailable* (e.g., no textured or too far/close to the multispectral stereo head); and *Valid* regions. A region is valid when depth information is known or is possible to fit a plane with its defining pixels. Therefore, let V be the set of all pixels in ground truth with disparity information available; O be the occluded regions; B be the regions close to an occlusion,

[4] http://www.cse.ohio-state.edu/otcbvs-bench/.

by definition, this boundary is 5 pixels of wide; and finally, C be the candidate matches obtained by the evaluated algorithm.

Section III-A introduces the concepts of the two error metrics, ER and SR. Now, they are defined as a function of the following three terms. The operator $T[\bullet]$ is defined as in (2), and r is a point of (x, y) coordinates, both in the ground truth as well as in the disparity map obtained by the proposed algorithm. Notice that, ground truth and disparity map are referred to the same coordinate system; they can be overlapped and their coordinates are equivalents.

Mismatch (MI): a correspondence with a disparity value different from the ground truth value larger than one pixel:

$$MI(r) = T\left[|V(r) - C(r)| > 1\right], \tag{13}$$

this score considers pixels near to occlusions (B).

False Negative (FN): an unassigned correspondence where the correct match exists in the ground truth (i.e., a hole):

$$FN(r) = T\left[r \in V : r \notin C\right]. \tag{14}$$

False Positive (FP): an assigned correspondence in occluded areas:

$$FP(r) = T[r \in C : r \notin V]. \tag{15}$$

The ROC space is defined by the above functions, and from them ER and SR are derived; remember that they are used as vertical and horizontal axis respectively, in the ROC plots.

$$ER = \frac{1}{|V|}\sum_{r=1}^{R}\left(MI(r) + FP(r)\right), \tag{16}$$

where $r = 1, \ldots, R$ corresponds to the index of valid coordinates and $|V|$ is the number of valid pixels. Finally:

$$SR = \frac{1}{|V|}\sum_{r=1}^{R} FN(r). \tag{17}$$

In the ROC curves presented in Fig. 6, the sparsity rate parameter is varied as follows: the cost values of the candidate matches in C are sorted in descending order. Next, from this list, and by using a decreasing τ threshold, different values of the ROC curve are obtained. For instance, the first plotted element in the ROC curve corresponds to bottom right point, which is the maximum cost value achieved only for a few set of pixels. Then, by decreasing the τ threshold, all the other points that define the ROC curve are obtained. In other words, the more pixels are selected reducing the sparsity rate, the larger the resulting error rate.

IV. Experimental Results

This section presents experimental results obtained with different algorithm settings and scenes. The setting of parameters is obtained from two optimization steps. The first one is intended to find the best setting of: *i*) window size, *ii*) scale and *iii*) quantization levels, from the parameter space $P = \{\mathrm{wz} \times \sigma_t \times Q\}$. The second optimization step is devoted to find the best confident value $\Lambda = \{\lambda_1, \ldots \lambda_t\}$ used for propagating IG cost through consecutive levels ((10)). These two steps have been implemented as follows.

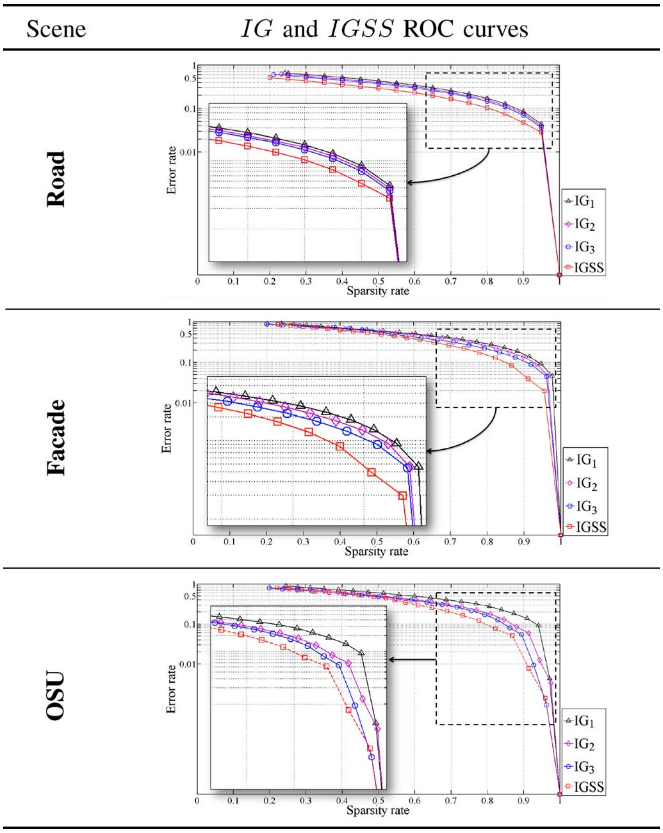


Fig. 6. Results obtained at different scales and with different settings ($IG_3(31, 1.5, 32)$, $IG_2(19, 1, 16)$ and $IG_1(7, 0.5, 8)$; as well as their merging, $IGSS$, with the proposed scale space representation).

Firstly, an efficiency measure (em) is defined to be used as a quantitative value for comparisons between different settings. Let $em = \int_0^1 ER\, dS\, R$ be the area under the error curve (ER) defined for all SR in the interval [0, 1], for a given setting of parameters. The parameter space P is sampled in a limited number of values defining a regular grid. Then, the best setting of parameters corresponds to the node of that grid where em reaches the minimum value. Since in the proposed approach a scale space representation is used, not only the setting with the minimum em value is considered, but the best p_i settings. Note that no prior information about the number of levels in the scale space representation is assumed. Hence, the family of parameter settings, with the lowest error, is obtained. This first optimization step is performed for each subset of the whole dataset. By analyzing the results is possible to find similarities between the best settings for the images in the evaluation dataset. Thus, it is possible to find relationships between the elements of the parameter space, particularly the relationship between the window size and the quantization level.

Then, the second optimization step finds the best set of $\Lambda = \{\lambda_1, \ldots, \lambda_t\}$ values for merging the IG costs corresponding to each of the p_i settings obtained above. Although initially a large family of p_i settings were considered, we experimentally found that three levels were enough to propagate the IG cost through

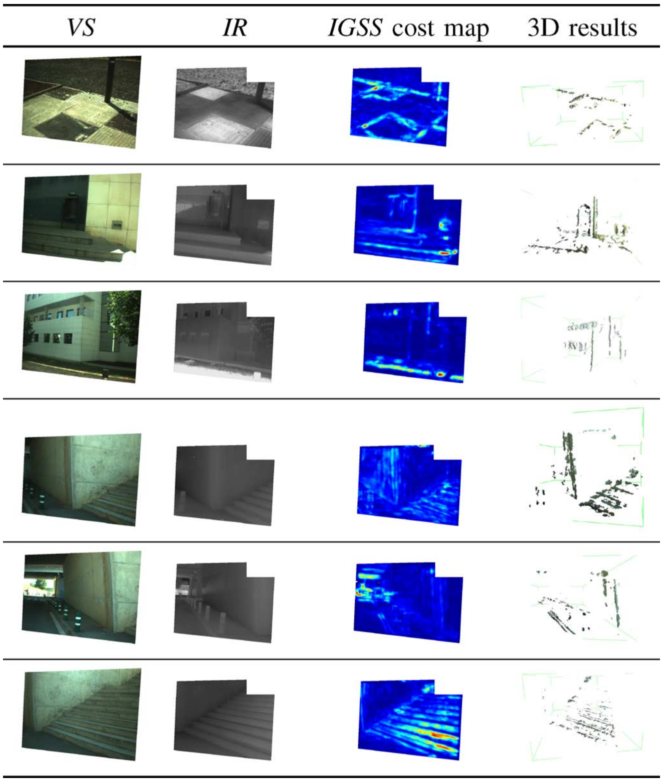


Fig. 7. Examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values).

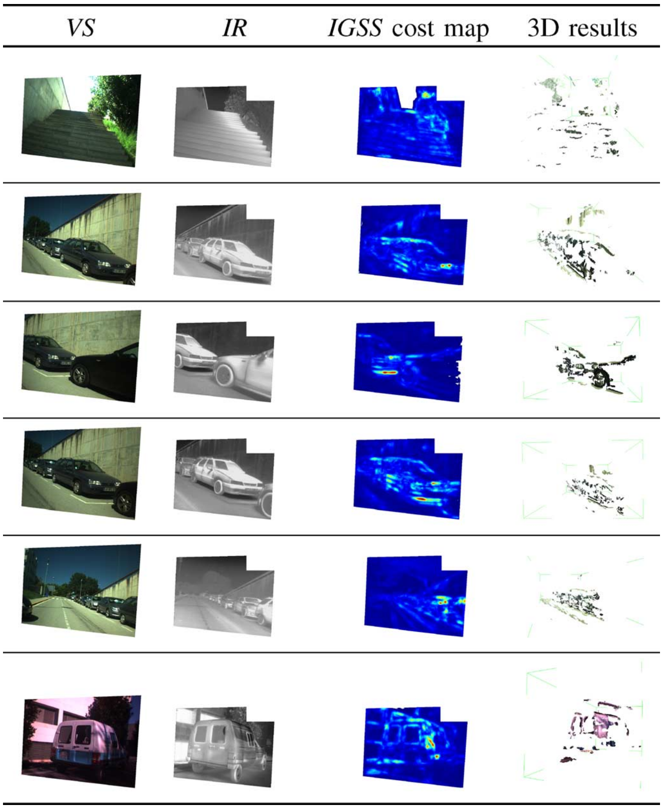


Fig. 8. Examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values).

the scale space representation. Hence, this second optimization process finds the best (λ_1, λ_2) using a similar approach.

The two optimization steps mentioned above are used to find the best combination of parameter settings. Initially, an exhaustive search in parameter space P is performed. The results are used to illustrate the behavior of ER and SR in each subset of dataset. Fig. 6 shows the three error curves corresponding to: road, facade, and OSU color-thermal. These curves depict the error and sparsity rate when the best settings are used in IG cost function, together with the improvement achieved by merging them $(IGSS)$. Finally, after finding the best settings for the whole dataset (including confidence parameters (λ_1, λ_2)) several sparse depth maps of real outdoor scenarios are presented (see right columns in Figs. 7 and 8).

The settings of parameters corresponding to the ROC curves presented in Fig. 6 were found with an exhaustive search in the following ranges: wz $=$ $\{7, 19, 31\}$, $\sigma_t = \{0.5, 1, \ldots 6\}$ and $Q = \{8, 16, 32, 64\}$. The obtained best set of parameters and propagation scheme is the following: $IG_3(p_3) \rightarrow IG_2(p_2) \rightarrow IG_1(p_1)$, where $p_3 = \{31, 1.5, 32\}$, $p_2 = \{19, 1, 16\}$ and $p_1 = \{7, 0.5, 8\}$. In our proposal, the windows sizes (wz parameter) decreases from 31 to 7 pixels, which looks like an inverted pyramid. This is to avoid smooth disparity maps, specially on edges, contours and boundaries, since the smaller windows (7 × 7) contributes in the last stage. On the other hand, we observed that information content decreases with scale, as previously reported in [32], but in our case faster at $\sigma_t = 2$. So, σ_t greater than this value decreases the correct matching score. This is due to the fact that gradient is not enough discriminative (G in (7)), and the windows tend to have low entropies (I in (1)). Therefore, the propagation of costs should be done with appropriate parameters since a wrong setting could increase the error rate. Regarding λ, the best settings, for the above parameter space $P = \{p_3, p_2, p_1\}$, is as follow $IG_3(p_3) \overset{\lambda_2=0.55}{\longrightarrow} IG_2(p_2) \overset{\lambda_1=0.65}{\longrightarrow} IG_1(p_1)$.

At this point, we must distinguish between I coming from the discrete version (2) and its smooth version obtained from (3). The difference lies in how the joint probability $p_{w_l w_r}$ is estimated. When the discrete joint probability is used, I results in a wave-like curve difficult to minimize. On the contrary, when the smooth $p_{w_l w_r}$ is considered a better behaved function is obtained, which helps us to find stable parameters. Parzen window estimation is done using three different Gaussian kernels (see (3)): $g(a_i, b_j; 3)$ for wz $=$ 7; $g(a_i, b_j; 7)$ for wz $=$ 19; and $g(a_i, b_j; 9)$ for wz $=$ 31.

As a conclusion from the plots presented above we can mention that the best result is obtained when the $IGSS$ is used. Improving the result at each stage, in a coarse to fine scheme. On average a 20% of correct matches, with less than 10% of ER, can be obtained with the proposed approach and by setting the parameters as indicated above. Another conclusion from Fig. 6 is that for a given sparsity rate always the best results (lowest error rate) is obtained after merging IG_i with the proposed scale space representation $IGSS$.

Figs. 7 and 8 show the results obtained with the proposed method. First and second columns correspond to the rectified images, visible and infrared spectrum respectively. Third column presents cost maps obtained after applying disparity selection method explained in

Section II-E. Each pixel in this representation corresponds to the maximum $IGSS$ value for a given d that maximize $\arg\max_d\{IGSS(w_l(x,y); w_r(x+d,y))\}$. Finally, the fourth column depicts the 3D sparse depth maps obtained from the correct matches. Sparse maps show how the multimodal correspondence between IR and VS can provide useful 3D information. Notice that the complexity of images used for validating the proposed approach, is also a challenge for a VS/VS stereo algorithm. However, the obtained results demonstrate the capability of our approach for finding correspondences in a wide range of radiometric differences, such as uncontrolled lighting conditions (sources), vignetting and shadows. Furthermore, the experimental results correspond to outdoor scenes with non-Lambertian surfaces and weakly textured regions. The construction of such a challenging dataset is motivated to push the limits of this novel technique, and provide insights of its application and research trends.

The results shown in the tables must be understood beyond the sparseness of 3D representations, or the accuracy with which the contours are recovered. For example, notice that the vehicles in the IR images appear quite poor textured, whereas in the VS images they appear textureless, however our approach can overcome this situation and provides a depth map free of mismatches over those regions (the same for the contrary case). This is consequence of the manner in which mutual and gradient information are combined. Thus, the multiplication of I and G reaches its maximum when a given correspondence is weighted as a correct one by both I and G cost functions (8). A more dense representation could be obtained by relaxing the τ threshold, but it will be affected by noisy data. Actually, this is a common trade off in stereo vision systems. It should be noticed that 3D representations presented in Fig. 7 and Fig. 8 provides not only the (X, Y, Z) and color components (r, g, b), as classical stereo systems, but also the thermal information corresponding to every 3D point.

The cost maps presented in Fig. 7 and Fig. 8 show that, in general, the similarity function introduced in Section II-D and II-C can match a window w_l with w_r extracted from a multispectral image pair with different accuracy. Our algorithm is designed to identify regions with high information content, such as edges and contours, and from them to obtain a 3D representation of the scene. Also, it penalizes mismatches in textureless areas, which are not reliable to find correspondences, for instance in image regions such as walls and floor. As can be appreciated on Fig. 7 and Fig. 8 (third column), higher cost values are concentrated on the edges, since in those regions a consensus between I and G is reached. Furthermore, it is possible to perceive the structure of the scene from these cost maps, which confirms the importance of discontinuities for relieving the ill-posedness of multimodal stereo. The strategy of cost propagation across a scale space representation enriches the $IGSS$, allowing to identify the correct disparity of a candidate set (Section II-E).

As a result from this section we can appreciate that although the current work is focused on recovering 3D information, we have confirmed that the $IGSS$ cost function overcomes mutual information and gradient-based approaches in multimodal template matching problems. This conclusion is supported by reviewing previous work [24], which uses a similar cost function. Since both evaluations (the current and previous one) use the same database (OSU Color-Thermal dataset [33]), we conclude that $IGSS$ is a valid similarity function for searching correspondences in multimodal video sequences. This conclusion could be also extended to the multimodal pedestrian tracking and detection problem. The previous statement is motivated by the fact that the work of Krotosky *et al.* (e.g., [10], [13]) is based only on the use of mutual information as a similarity function for matching pedestrian regions.

Finally, regarding the question formulated in Section I: "*is it possible to obtain 3D information from a multispectral stereo rig?*", we can say with safety that it is possible and it represents a promising research topic with several open issues.

V. Conclusions and Final Remarks

This paper presents a novel multimodal stereo rig build with a color and an infrared camera. The different stages for obtaining sparse depth maps are described. Furthermore, a ROC-based evaluation methodology is proposed for evaluating results from such a kind of multispectral stereo heads. It allows to analyze the behavior over a wide range of different parameter settings. Although the obtained results show a sparse representation, we should have in mind the challenge of finding correspondences in between these two separated spectral bands.

In summary, the main contributions of the current work are: (*i*) to present a study in an emerging topic as Multimodal Stereo IR/VS and achieves a sparse 3D representation from images coming from heterogeneous information sources; (*ii*) to propose a consistent criteria for making the multimodal correspondence; (*iii*) to establish a baseline for future comparisons; and (*iv*) to propose a framework that can be used as a test bed for evaluation purposes in this field.

Future work will be mainly focused on two aspects: (*i*) improving the disparity selection process by including Markov Random Fields, which allows to consider prior knowledge of the scene; and (*ii*) reformulating $IGSS$ ((8)) as a combination of two individual cost functions, which convert the cost function from a consensus scheme to a scheme where I and G contributes to a final matching score according to a set of assignment weights.

References

[1] A. Leykin and R. Hammoud, "Pedestrian tracking by fusion of thermal-visible surveillance videos," *Mach. Vis. Applicat.*, vol. 21, pp. 587–595, 2010.

[2] A. Fernández-Caballero, J. C. Castillo, J. Serrano-Cuerda, and S. M. Bascón, "Real-time human segmentation in infrared videos," *Expert Syst. With Applicat.*, vol. 38, no. 3, pp. 2577–2584, 2011.

[3] S.-H. Jung, J. Eledath, S. B. Johansson, and V. Mathevon, "Egomotion estimation in monocular infra-red image sequence for night vision applications," in *Proc. IEEE Workshop Applicat. Comput. Vis.*, February 2007.

[4] D. Firmenich, M. Brown, and S. Süsstrunk, "Multispectral interest points for RGB-NIR image registration," in *Proc. IEEE Int. Conf. Image Process.*, September 2011, pp. 181–184.

[5] G. Hermosillo and O. D. Faugeras, "Variational methods for multimodal image matching," *Int. J. Comput. Vis.*, vol. 50, pp. 329–343, 2002.

[6] E. P. Bennett, J. L. Mason, and L. McMillan, "Multispectral bilateral video fusion," *IEEE Trans. Image Process.*, vol. 16, no. 5, pp. 1185–1194, May 2007.

[7] Y. Y. Schechner and S. K. Nayar, "Generalized mosaicing: Wide field of view multispectral imaging," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 24, no. 10, pp. 1334–1348, Oct. 2002.

[8] D. A. Socolinsky and L. B. Wolff, "Multispectral image visualization through first-order fusion," *IEEE Trans. Image Process.*, vol. 11, no. 8, pp. 923–931, Aug. 2002.

[9] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognit.*, vol. 40, pp. 1771–1784, June 2007.

[10] S. J. Krotosky and M. M. Trivedi, "Person surveillance using visual and infrared imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1096–1105, Aug. 2008.

[11] R. Yang and Y. Chen, "Design of a 3-d infrared imaging system using structured light," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 2, pp. 608–617, Feb. 2011.

[12] S. Prakash, P. Y. Lee, and T. Caelli, "3d mapping of surface temperature using thermal stereo," in *Proc. Int. Conf. Control, Autom., Robot., Vis.*, Dec. 2006, pp. 1–4.

[13] S. J. Krotosky and M. M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 8, no. 4, pp. 619–629, Dec. 2007.

[14] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Comput. Vis. Image Understand.*, vol. 106, no. 2–3, pp. 270–287, 2007.

[15] G. Litos, X. Zabulis, and G. Triantafyllidis, "Synchronous image acquisition based on network synchronization," in *Proc. Workshop Comput. Vis. Pattern Recognit.*, Jun. 2006, p. 167.

[16] P. Viola and W. M. W. , III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, 1997.

[17] J. Kostlivá, J. Ĉech, and R. Ŝára, "Feasibility boundary in dense and semi-dense stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[18] J.-Y. Bouguet, Jul. 2010, Camera Calibration Toolbox for Matlab, [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc//index.html

[19] A. Kuijper, "On detecting all saddle points in 2D images," *Pattern Recognit. Lett.*, vol. 25, no. 15, pp. 1665–1672, Nov. 2004.

[20] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Mach. Vis. Applicat.*, vol. 12, no. 1, pp. 16–22, 2000.

[21] D. Scribner, P. Warren, and J. Schuler, "Extending color vision methods to bands beyond the visible," *Mach. Vis. Applicat.*, vol. 11, pp. 306–312, 2000.

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

[23] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1582–1599, Sep. 2009.

[24] F. Barrera, F. Lumbreras, and A. Sappa, "Multimodal template matching based on gradient and mutual information using scale-space," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, no. 10, pp. 2749–2752.

[25] T. M. Cover and J. A. Thomas*, Elements of Information Theory*. New York: Wiley-Interscience, 1991.

[26] J. Kim, V. Kolmogorov, and R. Zabih, "Visual correspondence using energy minimization and mutual information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2003, pp. 1033–1040.

[27] J. P. Pluim, J. B. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imaging*, vol. 19, no. 8, pp. 809–814, Aug. 2000.

[28] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, pp. 7–42, Apr. 2002.

[29] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[30] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2003, vol. 1, pp. 195–202.

[31] B. J. Davis and S. V. , "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understand.*, vol. 106, no. 2–3, pp. 162–182, 2007.

[32] A. Kuijper, "Mutual information aspects of scale space images," *Pattern Recognit.*, vol. 37, no. 12, pp. 2361–2373, 2004.

[33] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vision Image Understand.*, vol. 106, no. 2–3, pp. 162–182, 2007.

Fernando Barrera Campo received the electronics and telecommunication engineering degree from the University of Cauca, Colombia, in 2005, and the M.Sc. degree in computer vision and artificial intelligence from the Autonomous University of Barcelona, Spain, in 2009. He is currently a Ph.D. student at the Computer Vision Center and a member of the Advanced Driver Assistance Systems research group at Autonomous University of Barcelona, Spain. His research interests include stereo vision algorithms and three-dimensional reconstruction from multimodal sensor data.

Felipe Lumbreras Ruiz received the B.Sc. degree in physics from the Universitat de Barcelona in 1991 and the Ph.D. degree in computer science from the Universitat Autònoma de Barcelona, in 2001. He is currently an associate professor in the Computer Science Department at Universitat Autònoma de Barcelona and a member of the Computer Vision Center. His research interest includes wavelet and texture analysis, 3D reconstruction, and computer vision for automotive applications.

Angel Domingo Sappa (S'94–M'00) 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, 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 2D and 3D image processing. His current research focuses on stereoimage processing and analysis, 3D modeling, and dense optical flow estimation. He is a member of the IEEE.