http://dx.doi.org/10.21611/qirt.2012.172

QIRT 2012

11th International Conference on
Quantitative InfraRed Thermography

Planar-Based Multispectral Stereo

by Fernando Barrera*, Felipe Lumbreras‡, Cristhian Aguilera† and Angel D. Sappa*

*Computer Vision Center, edifici O, campus UAB, Bellaterra, Barcelona, Spain, *{jfbarrera, assapa}@cvc.uab.es*
‡Computer Science Departament, edifici Q, campus UAB, Bellaterra, Barcelona, Spain, *felipe@cvc.uab.es*
†*Dept. of Electrical and Electronics Engineering, Collao 1202 University of Bio-Bio, Concepcion–Chile, cristhia@ubiobio.cl*

Abstract

This paper presents a framework for extracting dense disparity maps from a multispectral stereo head. It is based on the assumption of a piecewise planar scene modeling. This assumption implies that the surfaces of the given scene can be fitted through a set of predominant planes. The multispectral stereo head is constructed with a thermal infrared and a color camera. It is intended to explore novel stereo matching approaches that will allow the fusion of information from different sensors. The proposed framework consists of three stages. Firstly, an initial sparse disparity map is extracted by using an adapted multimodal cost function. Then, a set of plane hypotheses that describe the surfaces of the scene is obtained. Finally, the information collected in pervious stages is combined through a Markov Random Field, which is solved by the graph-cuts algorithm. Experimental results in outdoor scenarios are provided showing the validity of the proposed framework.

1. Introduction

Color and thermal infrared cameras are already coexisting in different applications; just as examples we can mention the video surveillance (e.g. [1], [2]) and driver assistance (e.g. [3], [4]) applications. However, in these applications color (hereinafter referred to as visible spectrum cameras: VS) and thermal infrared cameras provide information that is processed independently, and their result fused at the end. These kinds of systems are redundant, since each information stream has its own data and processing flow. In the current work, we propose an algorithm for recovering 3D data without dual processing of color and infrared images (see Fig. 1), which can improve the overall performance of above applications. Our main motivation and challenge is to explore the possibility of obtaining 3D information from such a multispectral system, more precisely we propose to use a stereo rig constructed with a visible (VS) and an infrared (LWIR) cameras. This challenge represents a step forward in the state-of-the-art of 3D multispectral community.

Regarding the extraction of 3D data from a stereo pair, a large amount of approaches can be found in the literature (e.g., [5], [6], [7], and [8]). Matching algorithms can be broadly classified into two categories, according to the minimization method used for finding correspondences between the images: local, where only pixel information is used (e.g., WTA: Winner Take All); or global when prior information is exploited (e.g., Markov Random Fields (MRF)). In general, the former algorithms result in sparse representations while the later in dense depth maps. In the current work, we present a hybrid approach that combines both schemes in order to overcome the poor correlation between LWIR and VS images. Furthermore, a pairwise potential function is included into MRF formulation which encourage planar surfaces in the disparity maps. Experimental results in real outdoor scenarios are provided showing the viability of the proposed approach.

2. Method

The proposed approach mainly consists of two stages (see Fig. 1); the first stage obtains an initial disparity map, which will be refined by a planar-wise MRF. This early representation of the scene is achieved by following a local window-based matching approach. So, a sparse but accurate disparity map between LWIR and VS images is computed. In the current work, a multimodal cost function that combines gradient and mutual information through a multiresolution context is used [9] and [10]. Then, a disparity value is assigned to each pixel by maximizing of this cost function. At this point, a maximization by a WTA (Winner Take All) criteria results in a noisy disparity map. Therefore, only those disparities with high cost value are taken into account, other are discarded.

The second stage begins by generating a set of plane hypotheses, which are obtained from the sparse disparity map and then used for obtaining a dense representation. Since, the decomposition of a sparse disparity map into a set of planar regions is an untreatable problem, a self-similarity constrain is imposed. Therefore, those neighbour pixels with similar aspect should belong to the same 3D surface. Note that this constrain is widely used in other areas of computer vision, particularly in applications that work with man-made environments (i.e., [11], [12], and [13]). In order to identify candidates planar regions the VS image is over-segmented into superpixels [14]. Next, these pieces are perceptually grouped by using a graph-based segmentation algorithm, as the one presented in [15]. This overlapping of

segmentations takes advantage of self-similarity present in the images for obtaining large regions, which can be modelled as planar regions.

Once all planar regions have been identified from the 2D image segmentation, an iterative RANSAC algorithm [16] is applied using information from the sparse disparity map to find the best plane to every region. This algorithm includes: *(i)* a robust planar model estimator based on orthogonal regression and principal components analysis; *(ii)* a voting scheme that considers the number of inliers; and *(iii)* a model selection using a best score criterion. Once all planes are fitted a space of plane hypotheses is generated. Finally, spurious and incoherent planes are removed, considering their normal vectors as proposed in [12].

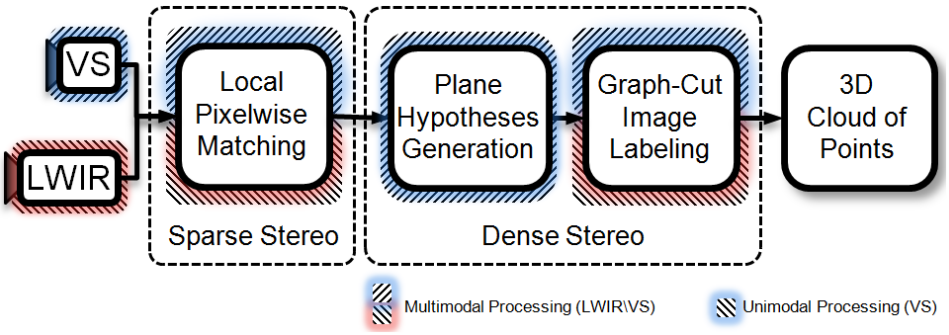


Fig.1. *Pipeline of the proposed multispectral stereo algorithm.*

2.1. Multimodal matching cost volume

The multimodal matching cost volume is computed from the multimodal images (I_{VS} stands for a VS image and I_{LWIR} for a LWIR image) following a local window based approach. We begin by fixing a window of size *wz* on a $p=(x, y)$ image coordinate in I_{VS}, while another window, with the same size, slides through $I_{LWIR}(x+d, y)$, where $d=\{d_{\min}, \ldots, d_{\max}\}$. Since these images are rectified, the searching space is bounded to a disparity range $d_{\min}<d<d_{\max}$. Thus, the cost volume referred in Eq. (1) as $C(p, d)$ is computationally represented by a multidimensional array, where every entry is indexed by a triplet of form (x, y, d), where (x, y) is a point p on the reference image (in this paper I_{VS}), and d represents the displacement of the matching window, which is centered on $I_{LWIR}(x+d, y)$.

The multimodal cost function shown in Eq. (1) is an adapted version of the one presented in [9] and [10]. We have redefined it as a weighted sum of mutual information and the similarity of gradient vectors within a pair of matching windows. So, the multimodal cost volume is defined as:

$$C(p,d)=\lambda C_{MI}(p,d)+(1-\lambda)C_{GI}(p,d), \tag{1}$$

where C_{MI} is mutual information of pixel values, and C_{GI} is the similarity degree of gradient vectors, the λ parameter represents the confidence of mutual information over gradient information. In order to increase the discriminative capability of the matching cost function a scale space representation is used. Hence, two stacks of images are generated for each multimodal image; one of them corresponds to a collection of blurred images while the other group contains gradient images (in scale space notation L_0 and L_1 respectively [17]). These representations are obtained by convolving an image (I_{VS} or I_{LWIR}) with a Gaussian kernel of order zero and one, while its standard deviation increases. Figure 2 presents a set of images involved in a scale space representation. Finally, both MI and GI should be computed at each level t of this hierarchy, and then aggregated into a unique value as depict the next equations:

$$C_{MI}(p,d)=[\alpha_0,\ldots,\alpha_t]\cdot\left[MI\left(\nabla_0^0(I(p,d))\right),\ldots,MI\left(\nabla_0^t(I(p,d))\right)\right], \tag{2}$$

$$C_{GI}(p,d)=[\beta_0,\ldots,\beta_t]\cdot\left[GI\left(\nabla_0^0(I(p,d))\right),\ldots,GI\left(\nabla_0^t(I(p,d))\right)\right], \tag{3}$$

$C_{MI}(p, d)$ is the resulting cost of propagating mutual information through of the hierarchy, from coarse to fine levels. It is expressed as a linear combination of all values of mutual information for a given position (p, d) in the stack $\nabla^t_0(I)$ of blurred images, together with a vector of weights that assigns a reliability value to every level. As was mentioned above, the MI operator provides a single value that measures the similarity degree of a pair of matching windows, considering only the pixel values. Gradient information (C_{GI}) is treated in an analogous manner.

http://dx.doi.org/10.21611/qirt.2012.172

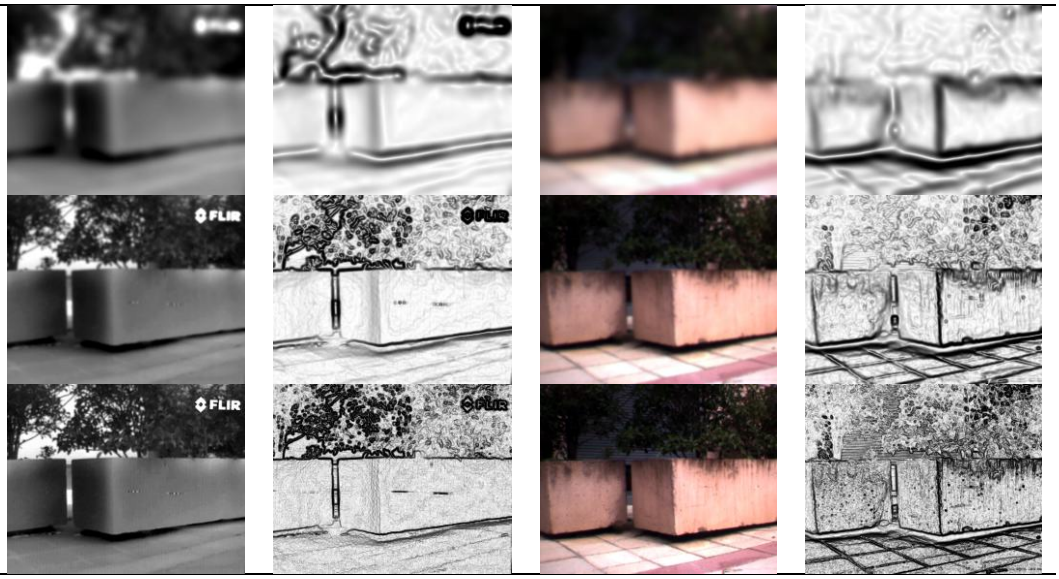
Scale	$\nabla_0(I_{LWIR})$	$\nabla_1(I_{LWIR})$	$\nabla_0(I_{VS})$	$\nabla_1(I_{VS})$
$t = 2$ $\sigma = 4$				
$t = 1$ $\sigma = 2$				
$t = 0$ $\sigma = 0.5$				

Fig.2. *Illustration of a set of images defining a scale space representation.*

Mutual information is defined in terms of entropies as:

$$MI(p,d) = h(p) + h(d) - h(p,d), \tag{4}$$

where $h(p)$ and $h(d)$ are entropies of two matching windows centered on image coordinate $I_{VS}(x,y)$ and $I_{LWIR}(x+d,y)$ respectively; $h(p,d)$ is their joint entropy. Thus, mutual information is formulated as a problem of Probability Distribution Functions (PDF) estimation. Note that it is only necessary to compute $h(p,d)$, since $h(p)$ and $h(d)$ are obtained from $h(p,d)$ [8]. We use a *nonparametric estimator* (*NP*) [18] for getting the joint PDF: $P_{p,d}(i_1,i_2)$. The later is a two dimensional matrix whose cells store the probability that an intensity i_1 corresponds to thermal infrared measuring i_2. Let us define the joint PDF as:

$$P_{p,d} = NP(p,d). \tag{5}$$

As shown in [7], the entropies in Eq. (4) can be estimated by a Parzen window method [19], and expressed as a sum of Gaussian distributions g with standard deviation ψ, as shown below:

$$h(p) = -\sum_{i_1} \log\left(P_p(i_1)\right) * g_\psi(i_1), \tag{6}$$

$$h(d) = -\sum_{i_2} \log\left(P_d(i_2)\right) * g_\psi(i_2), \tag{7}$$

$$h(p,d) = -\sum_{i_1,i_2} \log\left(P_{p,d}(i_1,i_2)\right) * g_\psi(i_1,i_2), \tag{8}$$

where $P_d(i_1) = \sum_{i_2} P_{p,d}(i_1,i_2)$ and $P_d(i_2) = \sum_{i_1} P_{p,d}(i_1,i_2)$ are the sum along each dimension of $P_{p,d}$. Finally, the gradient information is defined as:

$$GI(p,d) = \sum_{q,q'} w\left(\theta(q,q')\right) \min\left(|q|, |q'|\right), \tag{9}$$

where θ is the phase difference between two gradient vectors; $w(\theta)$ is a continue function that penalizes those angle differences out of the range (0, π); and |s| is the magnitude of the gradient vector. GI cost function operates on gradient images, see Eq. (3), thus q and q' represent gradients within the matching windows $I_{VS}(x,y)$ and $I_{LWIR}(x+d,y)$ at a certain scale t.

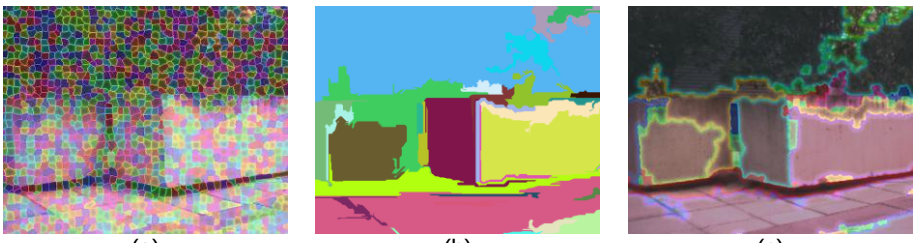
Once $C(p,d)$ has been computed, a Winner-Takes-All (WTA) method is used to select the best disparity for every point in the VS image; then, an initial sparse disparity map ($Dmap_0$) is obtained by filtering unreliable matches using the corresponding matching cost value ($C(p,d)>\tau$).

2.2. Plane based hypotheses generation

This stage consists of three steps, which results in a compact set of planar representations that will be used as labels in the final stage. The first step split up the given VS image into a set of perceptual regions. Since we are working with piecewise planar scenes, ideally, each region will correspond to a plane. Then, in the second step, a plane is fitted for each one of the regions obtained in the first step, by using the sparse disparity map computed in Section 2.1. Finally, in the third step, the large set of planes is compressed by extracting the dominant planes in the scene.

2.2.1. Split and merge segmentation

In order to overcome the limited information supplied by the initial disparity map, which prevents a correct detection of planar regions, a strategy for partitioning the images into approximately planar regions is adopted. The algorithm works as follows. Initially, the visible image I_{VS} is split up into s_i superpixels [14], which preserve edges and is adjusted to the local structure of the scene. Additionally, the original I_{VS} is also segmented into larger regions P that somehow capture perceptual aspects of the scene [15]. Finally, the s_i superpixels are clustered using as a criteria the perceptual regions p_i, resulting in a set of regions (R). The selection of [15] as a merging criterion is due to the fact that the images depict man-made structures, which can be efficiently segmented using an algorithm inspired on perceptual grouping. Furthermore, this algorithm puts special emphasis on edge variability, which in the current work is important since it reveals the orientation of surfaces.



(a) (b) (c)

Fig.3. *Split and merge segmentation example:* (a) *superpixels* S*;* (b) *perceptual regions* P*; and* (c) *resulting candidate planar regions* R*.*

2.2.2. Planar hypotheses generation

Once the sparse disparity map ($Dmap_0$) has been computed and the color image segmented into r_i regions, a set of hypotheses of planar regions to describe the surfaces in the scene are imposed. So, for every region $r_i \in R$ a RANSAC like algorithm [16] is employed to estimate the plane parameters. Note that the planar region estimator operates in the disparity space (x, y, d), which is a difference with respect to previous approaches that work on depth maps represented in the Euclidean space (e.g., [11] and [13]).

Let us remember that since a Manhattan world assumption is used, regions obtained from the segmentation in the color image are directly related with planar regions to be obtained in the disparity space. Since the accuracy of the final stage (i.e., piecewise planar labelling) depends on the confidence of the planar hypotheses, a robust random sample consensus paradigm has been used for estimating the free parameters of the model (plane). This method is capable to find local models from noisy cloud of data; previous works have demonstrated that this kind of algorithm overcomes least squared based techniques, since they are sensitive to outliers [20]. As mentioned above, the plane parameters for a given region r_i are obtained, with a RANSAC like algorithm, only if the region contains three or more valid disparities ($Dmap_0(r_i)$).

Once all planes have been fitted, a postprocessing stage is performed to merge planar patches defined by similar parameters. This postprocessing is performed to simplify the number of planar hypotheses. Note that the planes have been obtained in a local way, and then the number of planar hypotheses could be as large as the number of regions in R. Hence, the goal of this postprocessing stage is to reduce the number of planar hypotheses up to a minimum value so that the structure of the scene is still preserved. The plane linking stage is based on a distance ($dist_\pi$) computed from two planar patches, which was initially proposed in [21]. It is defined as follow:

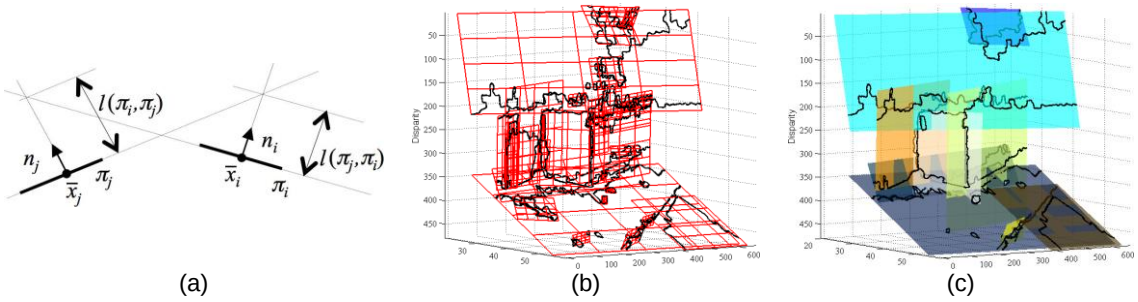
$$dist_{\Pi}(\pi_i, \pi_j) = l(\pi_i, \pi_j) + l(\pi_j, \pi_i), \tag{10}$$

$$l(\pi_i, \pi_j) = \frac{(\bar{x}_j - \bar{x}_i) \cdot n_j}{n_i \cdot n_j}. \tag{11}$$

The Eq. (11) corresponds to the length of the segment defined by $\overline{x_j}$ and the intersection of n_j, passing through $\overline{x_j}$ with π_i. In order to make it clear, a 2D representation of the segment lengths used for computing Eq. (10) is given in Fig. 4(a).

The previous planes distance (Eq. (10)) is used as a similarity function for merging a pair of planar patches. Hence, two planar patches are fused into a single one if ($dist_{\pi}(\pi_i,\pi_j) \leq \tau_{link}$). Once all possible combinations have been evaluated (only connected neighbor regions are considered) a new relabelled set R is obtained and the RANSAC algorithm is called again until convergence is reached.

Figure 4(b) and 4(c) show the planar hypotheses obtained before/after merging planar patches with similar parameters and filtering the noisy ones. The original set contains 38 hypotheses (see Fig. 4(b)), while the one presented in Fig. 4(c) is defined by only 9 hypotheses. They were obtained after six iterations of the plane linking stage.



(a) (b) (c)

Fig.4. *Planar hypotheses generation:* (a) *illustration of distances between two planes;* (b) *initial set of planar hypotheses from the segmentation presented in Fig. 3*(c)*; and* (c) *planar hypotheses resulting after six iterations of postprocessing algorithm (9 planes).*

2.3. Piecewise superpixel labeling

The set of planar hypotheses obtained above are now converted into labels for reformulating the disparity computation as a global minimization problem. It allows to take into account contextual constraints in order to achieve a dense disparity representation from multispectral information. The global minimization problem is based on the local correlation indicators computed in previous sections (i.e., mutual and gradient information boosted by the scale space representation). In this section, former indicators that were extracted at a level of pixels, are now interpreted as projections of planar surfaces. This helps to constrain the searching space to a few candidates, while spatial coherence of disparity values is hold. Notice that an extra planar hypothesis denoted as π_∞ is added to Π. It represents all those regions out of the stereo range (e.g., sky or distant surfaces).

The final step is to perform a piecewise superpixel labeling of reference image (VS), which assigns to each of the superpixels in S one of the plane hypotheses. Once every superpixel in the image has been labelled, a dense representation of disparities of VS and LWIR is obtained. Note that in the current section the set of plane hypotheses computed above are used as labels.

The matching of planar regions that belongs to different modalities (LWIR/VS) is now formulated as an energy minimization problem in the superpixel domain. Thus, an MRF is defined and solved by a graph cuts framework [22]. The goal of this section is to obtain a label f that assigns a given superpixel s to a plane of plane hypotheses. This label minimizes a global energy function E, which consists of a data term D_s that compares the current label with the observed data, and a pairwise smoothness term V_{st}. This energy function is defined as:

$$E = \sum_{s} D_s(f_s) + \sum_{s,t \in \mathrm{N}} \lambda_{smooth} V_{st}(f_s, f_t) \tag{12}$$

where S is the set of all superpixels; D_s is the data term that measures how well a plane hypothesis explains the disparity value for a given s region; $V_s(f_s,f_t)$ is a smoothness prior computed in surrounding regions to superpixel s; N represents that neighborhood; f_s, f_t are the current labels for superpixels s and t respectively; and λ_{smooth} is a constant value used for normalization. Similarly to the work proposed in the VS/VS field [13], in the current work the D_s function is defined as follows:

$$D_s(f_s)=\begin{cases}\min\left(C_\pi(f_s),C_{\max}\right) & \text{if } f_s\in\{\pi_1,\pi_2\ldots\pi_n,\pi_\infty\}\\ \min\left(C_\pi(f_s),C_{\max}\right)+c_{bias} & \text{if } f_s=non-plane\\ 0.9\cdot C_{\max} & \text{if } f_s=discard\end{cases} \quad (13)$$

where $C_\pi(f_s)$ is the cost of assigning a plane hypothesis (label f_s) to superpixel s. This cost is defined as follows:

$$C_\pi(f_s)=\sum_{p\in s}C(p,d), \quad (14)$$

where d is the disparity value obtained by evaluating p in the current plane hypothesis Π ($d=c_1x+c_2y+c_3$). This cost is equal to the aggregation of costs spanned by the plane f_s in the $C(p,d)$ volume (Section 2.1). Equation (13) includes a constant value that is denoted as C_{max}, which is used for: *i)* truncating the C_π cost; *ii*) penalizing inconsistent plane hypothesis for a certain region s (e.g., plane hypothesis that generates disparity values outside of the cost volume); *iii*) allowing that a given region changes its label by the one of its neighbor.

The smoothness term is defined as:

$$V_{st}(f_s,f_t)=\underbrace{\frac{1}{\left(\gamma\left|\ \nabla_1(I_{VS})\ \right|+1\right)}}_{g}\begin{cases}0 & \text{if } f_s=f_t\\ d_{\max} & \text{if } f_s \text{ or } f_t \text{ is not a plane}\\ dist_\Pi(f_s,f_t) & \text{otherwise}\end{cases} \quad (15)$$

where: $dist_\Pi$ is defined in Eq. (10); d_{max} is a constant value that penalizes discontinuities; g is a weighting function with domain in the gradient magnitude of VS image ($|\nabla_1(I_{VS})|$); this g function is evaluated at the midpoint of the segment that links two superpixel' centroid. Finally, the energy function defined in Eq. (12) is minimized with the graph cut framework presented in [22]. Figure 5(c) shows the disparity map of the case study used as an illustration through the manuscript; its corresponding textured 3D map is presented in Fig. 5(e).

In general, urban environments contain non-planar surfaces that should be detected before graph-cuts labeling. These surfaces mainly correspond to bushes, trees and grass, which appear in images overlapping building's structure; they are hard to fit by a planar model. Therefore, as pointed out by Gallup et al. [13] a non-planar region classifier can be used for identifying these surfaces. Since, the classification problem essentially consists in determining if a given patch corresponds to a planar or to a non-planar surface; a simple two-class algorithm can be used.

In the current work we introduce two modifications to the approach proposed by Gallup et al. [13]. Firstly, the feature vector that describes every patch is similar to the one presented in [23]. Thus, features such as color, texture, and location are computed from each patch. On the other hand, instead of obtaining the class membership probability of a patch by a k-nearest-neighbor method, we use an Adaboost implementation based on decision trees [24], which directly returns a confidence value that is assumed as a class membership probability.

All the images in the dataset are split up into patches of 15x15 pixels, and then a feature vector is obtained from each patch. We collect a total of 1200 patches for each class, which were extracted from 20 images and manually classified into planar or non-planar surfaces. Then, they were divided into two groups, one for training and the other one is used for validation. After analyzing the classification error with respect to the number of weak classifiers, we concluded that a scheme with 100 weak classifiers is enough for the trade-off between accuracy and computational cost. From the classification results we can conclude that the weight vector provided by the decision tree of the Adaboost classifier, confirms that the texture is a discriminative feature for classifying non-planar region. Hence, this information is used to detect unstructured objects, which in general in urban scenes correspond to non man-made elements.

Finally, the data term in Eq. (13) is complemented by the class membership probability coming from the Adaboost classifier as follows:

$$D'_s(f_s)=D_s(f_s)+\lambda_{class}\begin{cases}1-a & \text{if } f_s\in\{\pi_1,\pi_2,\ \ldots,\ \pi_n,\pi_\infty\}\\ a & \text{if } f_s=non-plane\\ 0 & \text{if } f_s=discard.\end{cases} \quad (16)$$

where a is the class membership probability; and λ_{class} is a constant that weights the contribution of non-planar classifier. Figure 5(a) shows an illustration of a probability map of non-planar regions for the image used as a case study through the manuscript; in Fig. 5(d) patches with a probability value higher than 0.5 are depicted. The labelled regions and

disparity map resulting from MRF are presented in Fig. 5(b) and Fig. 5(c). Finally, corresponding 3D representations are given in Fig. 5(e).

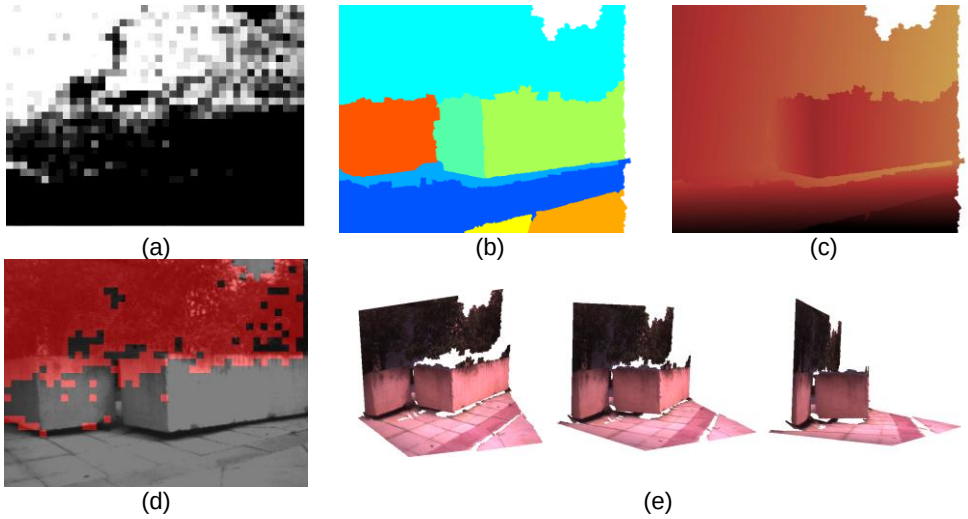


Fig.5. *MRF labelling:* (a) *Adaboost classifier probability map;* (b) *labels resulting from MRF;* (c) *corresponding disparity map;* (d) *non-planar regions; and* (e) *final 3D representations.*

3. Experimental results

Before presenting the results obtained with the proposed approach a brief description of the multimodal stereo system used to acquire the I_{VS} and I_{LWIR} images is presented. Additionally, details about the evaluation dataset are also provided. The multimodal stereo head consists of a pair of cameras (VS/LWIR) separated by a baseline of 12 cm and a non-verged geometry. This configuration is obtained by adjusting the pose of the cameras till their z coordinate axes are parallel, and perpendicular to the baseline. Hence, the images provided by the multimodal stereo rig are pre-aligned, somehow ensuring their right rectification. Thermal infrared images are obtained with a *Long-Wavelength InfraRed* camera (PathFindIR from Flir) while color ones with a standard color camera based on the Sony ICX084 sensor, which has a focal length of 6 *mm*. The former detects radiation in the range 8-14 μm (LWIR band), whereas the color camera responds to wavelengths from about 390 to 750 nm (Visible Spectrum).

As mentioned above, the cameras have been aligned before starting the calibration process. This action ensures that the needed projective transformations for image rectification are smooth, since the image planes position is approximately coplanar. Once each camera has been calibrated [25], and its intrinsic parameters are known, the next step is to estimate the geometry of multispectral stereo rig. Since the current work is focused in the generation of 3D models up to scale, it is only necessary to estimate the epipolar geometry through of the fundamental matrix F.

Image rectification is a critical issue since the proposed algorithm assumes that all epipolar lines in the multispectral images are horizontally aligned. Despite the accuracy with which F is estimated, it is still necessary to use a rectification method that takes into account the dissimilarity of intrinsic parameters of cameras. Therefore, the rectification method presented in [26] is adopted, which compute a pair of projective transformations, one for each camera. It rectifies the images while preserving the aspect of image content. This method reduces the loss and the creation of pixels due to resampling effect.

In order to evaluate the proposed method a set multispectral images has been collected. The captured data depicts a variety of urban scenes, which includes: buildings, sidewalk, trees, and vehicles, among others. Although this dataset shows a large collection of planar surfaces with different orientations, other types of non-planar surfaces are also captured.

The proposed approach has been validated using a large data set of outdoor scenarios. Dense disparity maps were obtained by setting the different parameters as indicated next; the different values were empirically obtained and the same setting is used in all the scenarios. The initial Dmap_0 is obtained by defining d_{min}=0 and d_{max}=120. The scale space representation contains three levels and the values used for propagating mutual and gradient information through the different levels: $[\alpha_0, \alpha_1, \alpha_2]^T$ = [0.2, 0.3, 0.5]^T and $[\beta_0, \beta_1, \beta_2,]$ = [0.2, 0.3, 0.5]^T; threshold τ is set as 10% of the maximum cost value; finally, mutual and gradient information in Eq. (1) are fused defining λ = 0.5. The two values related with the planar hypothesis generation were set as follow: τ_{RANSAC}=0.2 and τ_{link}=2.5. The values given by default in the graph cut implementation provided by [13] were used for the global minimization.

Figures 6, 7 and 8 ((a) and (b)) show three multispectral pairs used to evaluate the performance of the proposed approach. Results from each stage are presented: (c) shows the planar hypotheses used as input of the global minimization stage; (d) corresponds to the probability map obtained from Adaboost classifier; (e) labels resulting after

combining (c) and (d) by the graph-cuts algorithm; (f) disparity map extracted from (e); (g) different 3D views obtained from the resulting dense disparity map. In summary, these three scenes show that under certain restrictions multispectral images can be used to extract dense disparity information. This information can be directly converted into a 3D representation describing the geometry of the scene.

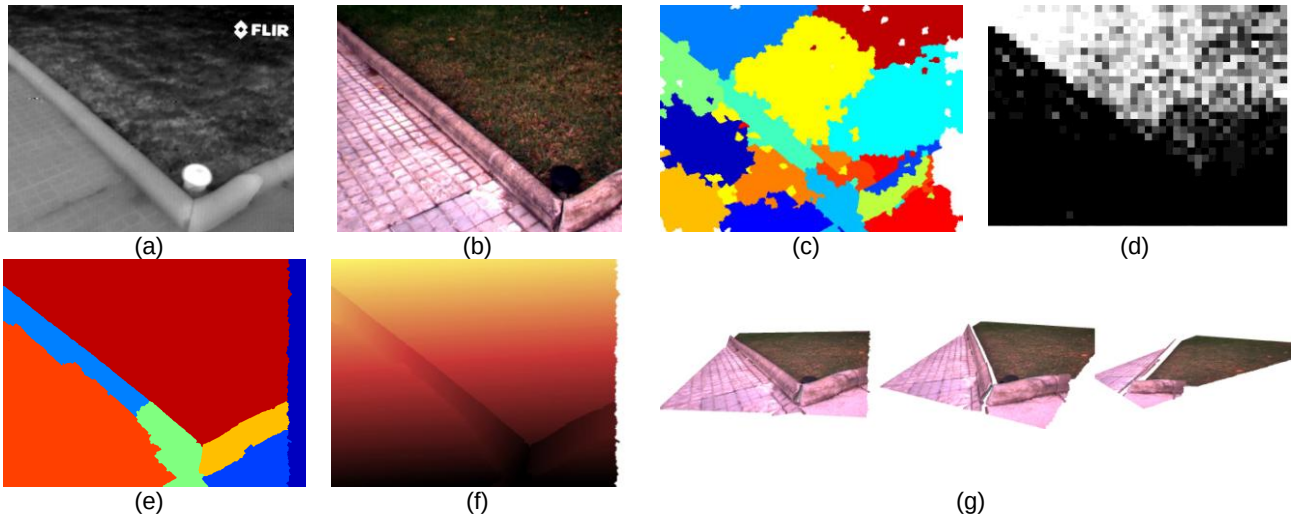


Fig.6. *Scene 1:* (a) *and* (b) *pair of multispectral images;* (c) *set of planar hypotheses;* (d) *planar and non-planar regions probability map;* (e) *and* (f) *labels and disparity map resulting from MRF;* (g) *3D views.*

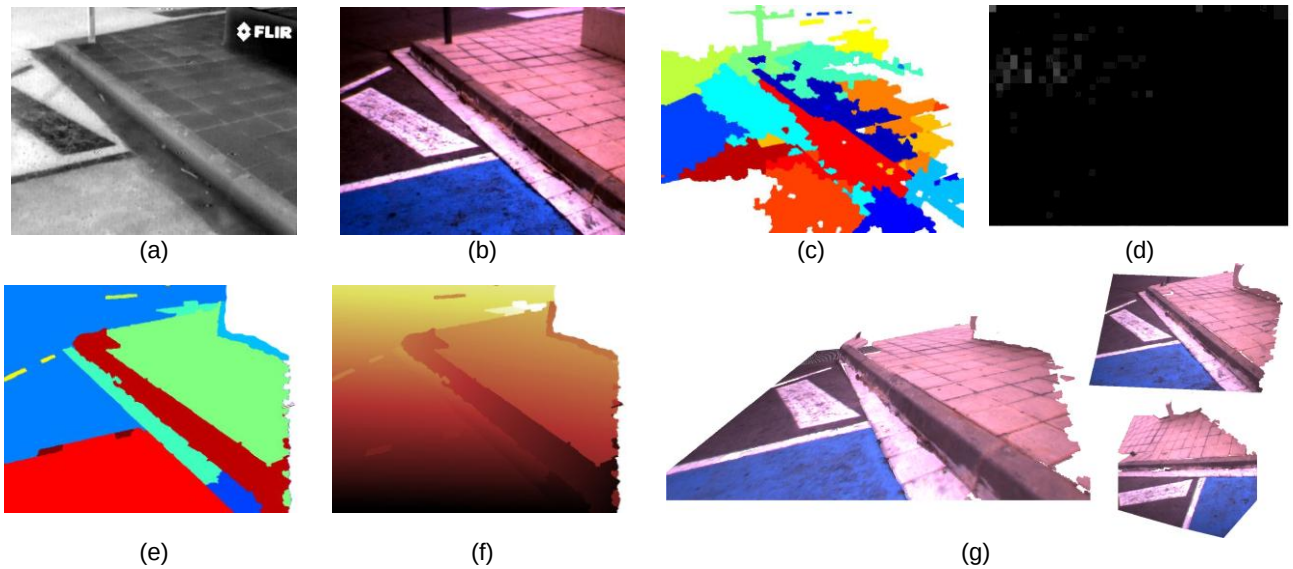


Fig.7. *Scene 2:* (a) *and* (b) *pair of multispectral images;* (c) *set of planar hypotheses;* (d) *planar and non-planar regions probability map;* (d) *and* (f) *labels and disparity map resulting from MRF;* (g) *3D views.*

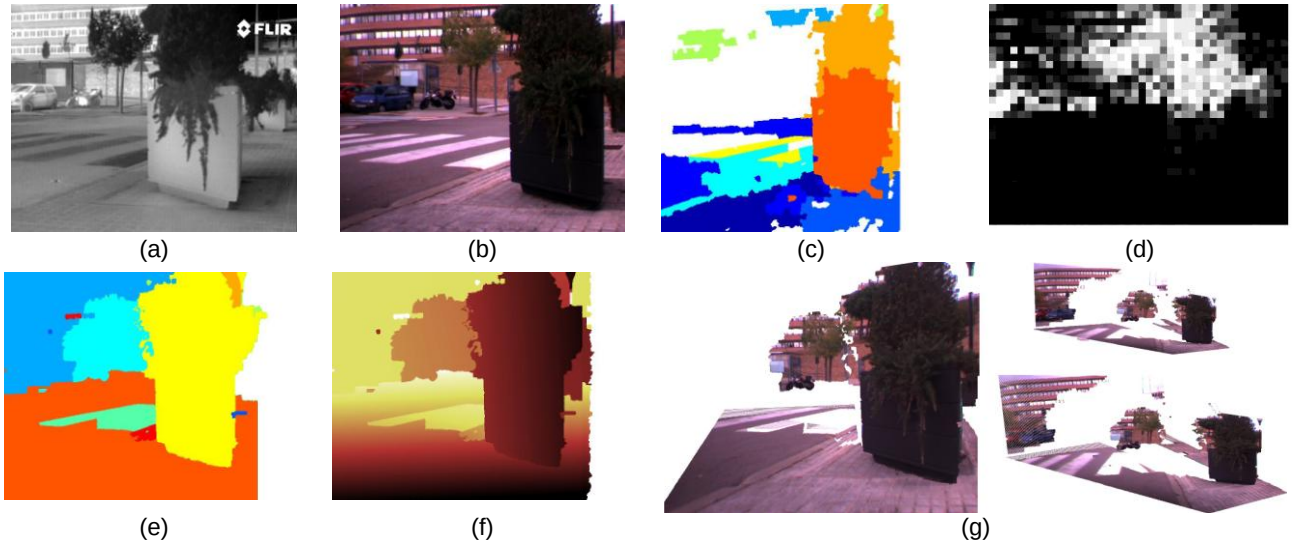


Fig.8. *Scene 3:* (a) *and* (b) *pair of multispectral images;* (c) *set of planar hypotheses;* (d) *planar and non-planar regions probability map;* (d) *and* (f) *labels and disparity map resulting from MRF;* (g) *3D views.*

4. Conclusions and Further Work

The current work presents a framework for extracting dense disparity maps from multispectral stereo images. The different stages are described together with the LWIR and VS stereo head. The proposed approach represents a step forward in the extraction of 3D information from multispectral stereo images; results obtained from this research can benefit those fields where visible and infrared images coexist.

This work has shown that under certain restrictions the scene structure can be inferred from a small set of good correspondences, despite the low correlation between LWIR and VS images. The similarity function based on mutual and gradient information significantly increases the number of good matches up to overcome the minimum number of correspondences needed to obtain an accurate representation of the scene. Although its performance highly depends on parameter setting, both sparse and dense disparity maps can be obtained.

The proposed energy function allows obtaining dense representations by modelling the scene as a piece-wise planar surface. Furthermore, the non-planar regions, labelled by the Adaboost classifier, are also considered during the minimization resulting in a robust solution even though their presence in the given scene. The energy function is minimized through the graph-cuts algorithm. This formulation drives to two different representations. On the one hand, it allows obtaining the dominant planar regions of the given image; on the other hand, it computes the sough dense disparity map. Dominant planar regions are obtained through graph-cuts using as a prior a piece-wise planar representation that evolves during the minimization process.

Future work will be mainly focussed on the extraction of ground truth data to quantitatively validate the obtained results. Furthermore, the use of other cost functions will be explored.

5. Acknowledgement

This work was partially supported by the Spanish Government under Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and Projects: TIN2011-25606 and TIN2011-29494-C03-02. The third author was supported by MECESUP2 Postdoctoral program and Regular Project 100710 3/R of the University of Bio-Bio, Chile.

REFERENCES

[1] Leykin A. and Hammoud R., “Pedestrian tracking by fusion of thermal-visible surveillance videos”. Mach. Vision and Appl., vol. 21, pp. 587-595, 2010.
[2] Fernández-Caballero A., Castillo J. C., Serrano-Cuerda J., and Bascón S. M., “Real-time human segmentation in infrared videos”. Expert Systems with Appl., vol. 38, no. 3, pp. 2577-2584, 2011.
[3] Jung S.H., Eledath J., Johansson S. B., and Mathevon V., “Egomotion estimation in monocular infrared image sequence for night vision applications”. IEEE Workshop Appl. of Computer Vision, Austin (USA), 2007.
[4] Krotosky S. J., and Trivedi M. M., “On color-, infrared-, and multimodal-stereo approaches to pedestrian detection”. IEEE Trans. Intell. Transportation Systems, vol. 8, no. 4, pp. 619-629, 2007.

[5] Brown M.Z., Burschka D., and Hager G.D., “Advances in computational stereo”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 8, pp. 993-1008, 2003.
[6] Scharstein D., and Szeliski R., “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”. Int'l. J. of Computer Vision, vol. 47, pp.7-42, 2002.
[7] Kim J., Kolmogorov, V., and Zabih, R., “Visual correspondence using energy minimization and mutual information”. IEEE Int’l Conf. on Computer Vision, Nice (France), pp. 1033-1040, 2003.
[8] Hirschmuller, H., and Scharstein, D., “Evaluation of Stereo Matching Costs on Images with Radiometric Differences”. IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, pp. 1582-1599, 2009.
[9] Barrera F., Lumbreras F., and Sappa A., “Multimodal template matching based on gradient and mutual information using scale-space”, IEEE Int'l. Conf. Image Process., HK (China), pp. 2749-2752, 2010.
[10] Barrera F., Lumbreras F., and Sappa A., “Evaluation of similarity functions in multimodal stereo”. Int'l Conf. on Image Anal. and Recog., Aveiro (Portugal), 2012.
[11] Furukawa, Y., Curless, B., Seitz, S.M., and Szeliski, R., “Manhattan-world stereo”. IEEE Int'l Conf. on Computer Vision and Pattern Recog., Miami (USA), pp. 1422-1429, 2009.
[12] Bleyer M., and Gelautz M., “A layered stereo matching algorithm using image segmentation and global visibility constraints”. Int'l. J. of Photogrammetry and Remote Sensing, vol. 59, no. 3, pp. 128-150, 2005.
[13] Gallup D., Frahm J.M., and Pollefeys M., “Piecewise planar and non-planar stereo for urban scene reconstruction”. IEEE Int'l. Conf. on Computer Vision and Pattern Recog., pp. 1418-1425, SF (USA), 2010.
[14] Levinshtein, A., Stere, A., Kutulakos, K.N., Fleet, D.J., and Dickinson, S.J., “TurboPixels: Fast Superpixels Using Geometric Flows”, IEEE Trans. on Pattern Anal. Mach. Intell., vol.31, no. 12, pp. 2290-2297, 2009.
[15] Felzenszwalb P. F., and Huttenlocher D. P., “Efficient graph based image segmentation”. Int'l. J. of Computer Vision, vol. 59, pp. 167-181, 2004.
[16] Fischler M., and Bolles R., “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”. Comm. of the ACM, vol. 24, pp. 381-395, 1981.
[17] Lindeberg T., “Scale-space theory in computer vision”. Kluwer academic publishers, Dordrecht (Netherlands), 1994.
[18] Dowson, N. and Kadir, T. and Bowden, R. "Estimating the Joint Statistics of Images Using Nonparametric Windows with Application to Registration Using Mutual Information". IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, pp. 1841-1857, 2008.
[19] Viola P., and Wells W., “Alignment by Maximization of Mutual Information”. Int'l. J. of Computer Vision, vol. 24, pp. 137-154, 1997.
[20] Torr P. H. S., and Zisserman A., “MLESAC: A New Robust Estimator with Application to Estimating Image Geometry”. Computer Vision and Image Understanding, vol. 78, pp. 138-156, 2000.
[21] Tao, H., Sawhney S., and Kumar R. “A Global Matching Framework for Stereo Computation”. IEEE Int. Conf. on Computer Vision, Los Alamitos (USA), pp. 532-539, 2001.
[22] Boykov Y., Veksler O., Zabih R., “Fast approximate energy minimization via graph cuts”, IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222-1239, 2001.
[23] Hoiem D., Efros A., and Hebert M., “Geometric Context from a Single Image”. IEEE Int. Conf. on Computer Vision, Beijing (China), pp. 654-661, 2005.
[24] GML AdaBoost toolbox, in http://www.inf.ethz.ch/personal/vezhneva/
[25] Camera calibration toolbox for matlab, in http://www.vision.caltech.edu/bouguetj/calib_doc/
[26] Mallon J., and Whelan P., “Projective rectification from the fundamental matrix”. Image and Vision Computing, vol. 7, pp. 643-650, 2005.