Pattern Recognition Letters 34 (2013) 52–61

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Multispectral piecewise planar stereo using Manhattan-world assumption

Fernando Barrera [a,*], Felipe Lumbreras [a,b], Angel D. Sappa [a]

[a] Computer Vision Center, Autonomous University of Barcelona, 08193 Bellaterra, Barcelona, Spain
[b] Dept. Computer Science, Autonomous University of Barcelona, 08193 Bellaterra, Barcelona, Spain

ARTICLE INFO

Article history:
Available online 30 August 2012

Keywords:
Multispectral stereo rig
Dense disparity maps from multispectral stereo
Color and infrared images

ABSTRACT

This paper proposes a new framework for extracting dense disparity maps from a multispectral stereo rig. The system is constructed with an infrared and a color camera. It is intended to explore novel multispectral stereo matching approaches that will allow further extraction of semantic information. The proposed framework consists of three stages. Firstly, an initial sparse disparity map is generated by using a cost function based on feature matching in a multiresolution scheme. Then, by looking at the color image, a set of planar hypotheses is defined to describe the surfaces on the scene. Finally, the previous stages are combined by reformulating the disparity computation as a global minimization problem. The paper has two main contributions. The first contribution combines mutual information with a shape descriptor based on gradient in a multiresolution scheme. The second contribution, which is based on the Manhattan-world assumption, extracts a dense disparity representation using the *graph cut* algorithm. Experimental results in outdoor scenarios are provided showing the validity of the proposed framework.

1. Introduction

The development of multispectral systems has been an attractive research topic in the computer vision field during the last decade; mainly because they provide a rich representation of the scene by means of a collection of images taken by different sensors. These systems have grown to become a significant tool for dealing with a wide range of problems, for instance: remote sensing, navigation, surveillance, medical imaging, among others. However, in the 3D information recovery domain the potentiality and capability of such systems are still not clear. In the current work, a multispectral stereo matching algorithm for extracting dense disparity maps from thermal infrared and color[1] images is presented. These images are acquired with a Long Wave Infra-Red band camera (LWIR) and a color camera (VS) respectively.

Thermal infrared/visible multispectral 3D representations can be broadly divided into two categories according to the role performed by the LWIR camera. The first category includes systems that combine thermal infrared cameras with well-studied techniques for extracting 3D, such as stereo-vision systems (VS/VS) or structured light. These systems are responsible for providing depth information, which is then enriched with the thermal measurement (e.g., Trivedi et al., 2004; Yang and Chen, 2011). Although, a valid multispectral representation of the given scene is achieved, the thermal information is treated as a *complementary* source. That is, only mapping thermal infrared information into the resulting 3D representation. On the contrary, the second category includes those approaches where thermal and visible information is matched for extracting a sparse 3D representation (e.g., Krotosky and Trivedi, 2007; Barrera et al., 2010). In other words, the information is used in a *collaborative* framework.

Up to our knowledge dense disparity maps only from the first category have been reported in the literature. However, the increasing number of systems where LWIR and visible cameras coexist leads us to state the following questions: "is it possible to obtain *dense* disparity maps from a multispectral stereo head defined with a camera working in the visible and another in the thermal infrared spectral band?". From an efficiency point of view we wonder whether these *complementary* sensors could be used in a *cooperative* framework that allows to exploit thermal and visible information for extracting a 3D representation.

The structure of the paper is the following. A review of related work on multispectral stereo algorithms is presented in Section 2. Then, the steps of the proposed algorithm are presented in Section 3. Technical details of multimodal stereo head used for evaluating the proposed approach are presented in Section 4, together with details of the generated data set and the obtained experimental results. Conclusions and final remarks are given in Section 5.

2. Related work

The extraction of 3D information from multispectral stereo heads (LWIR/VS) has attracted the interest of researchers in

* Corresponding author. Fax: +34 93 581 1670.
E-mail addresses: jfbarrera@cvc.uab.es (F. Barrera), felipe@cvc.uab.es (F. Lumbreras), asappa@cvc.uab.es (A.D. Sappa).

[1] For interpretation of color in Figs. 2, 3, and 6–9, the reader is referred to the web version of this article.

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.08.009

different computer vision applications, for examples: human detection (Han and Bhanu, 2007), video surveillance (Krotosky and Trivedi, 2008), and 3D mapping of surface temperature (Yang and Chen, 2011; Prakash et al., 2006). Recently, a comparison of two stereo systems is presented in (Krotosky and Trivedi, 2007), one working in the visible spectrum (composed of two color cameras) and the other in the infrared spectrum (using two LWIR cameras). Since that 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 a LWIR camera. In this framework, infrared information is not used for stereoscopy but just for mapping LWIR information over the 3D points computed from the VS/VS stereo head. This allows to develop robust approaches for video surveillance applications (e.g., Krotosky and Trivedi, 2008).

On the contrary to the previous approaches, a multimodal stereo head constructed with just two cameras, an infrared and a color one is proposed in (Krotosky and Trivedi, 2007). In this case the challenge is to match 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 (Viola and Wells, 1997). Although, these assumptions are valid, they restrict the scope of applications to those scenarios containing pedestrians. Furthermore, it should be noted that this approach is able to extract 3D information only on those pixels defining the surface of the pedestrian's body.

A more general solution should be envisaged, allowing such a kind of multispectral stereo head to be used in different applications. In other words, the matching should not be constrained to regions containing some predefined characteristic (e.g., human body silhouettes). Note that formulating the solution in a general framework is mandatory in order to extract dense disparity maps.

Up to our knowledge, none of the previous multispectral stereo algorithms for thermal infrared and visible images are able to obtain dense representations. Although the proposed framework is based on a Manhattan world assumption, which could be seen as a constraint, it should be noted that piecewise planar representations are valid in most of man-made environments (Coughlan and Yuille, 1999).

3. Piecewise planar stereo

The proposed approach consists of three stages. Firstly, it starts by estimating a sparse but accurate disparity map of the scene. Then, in the second stage the initial map is represented by means of a set of planes. Finally, a dense disparity map is obtained by a piecewise planar labeling framework. These stages are detailed next; an illustration is provided in Fig. 1. For the sake of presentation simplicity, throughout this paper thermal infrared images I_{LWIR} will be referred to as infrared images, while color images will be referred indistinctly as color or visible images I_{VS} (VS: visible spectrum).

3.1. Initial disparity map

The goal of this section is to compute an initial disparity map, which will be fitted by a set of planes. This disparity map is obtained from a matching cost volume following a local window based approach, in which a Winner Take All (WTA) strategy is used for disparities selection. The main challenge of this first stage are both, to get a large number of good correspondences and to have a high accuracy in their locations.

In order to address the matching problem, we propose to extend a cost function based on mutual information by enriching it with gradient information in a scale space representation (Barrera et al., 2010). A motivation of this proposal is shown in the two leftmost illustrations in Fig. 2, LWIR does not match at a pixel level with VS; so classical stereo strategies cannot be directly applied. On the contrary, mutual information, as shown in similar multispectral problems (e.g., Pluim et al., 2001; Fookes et al., 2004), can be used in this case. Furthermore, in the same figure we can see that edges seem to be a relevant feature present in both modalities; this motivates us to include this kind of information in the proposed solution. Finally, a scale space representation adds robustness and spread local matches from coarser to finer scales increasing the number of final correspondences. In order to tackle the second challenge mentioned above, related with the accuracy of the locations, disparity values are obtained by local quadratic interpolations.

The initial disparity map is obtained by using a matching cost function inspired by Barrera et al. (2010); actually, a slight modification is introduced to improve the number of correspondences. This assumes that images are rectified, therefore the searching space is constrained to one dimension. So, let: $\boldsymbol{p} = (x, y)$ be a given pixel in the color image I_{VS}; $\boldsymbol{q} = (x + d, y)$ be its expected correspondences in the infrared thermal image I_{LWIR}; and d be the disparity value. The cost of corresponding two windows centered on points $\boldsymbol{p}$ and $\boldsymbol{q}$ is obtained as follows:

$$C(\lambda, \boldsymbol{p}, d) = \lambda C_{MI_{ss}}(\boldsymbol{p}, d) + (1 - \lambda) C_{GI_{ss}}(\boldsymbol{p}, d), \tag{1}$$

where $C_{MI_{ss}}$ and $C_{GI_{ss}}$ are the cost terms based on the mutual and gradient information in a scale space representation, which will be detailed next; and $d = \{d_{min}, \ldots, d_{max}\}$.

By definition in information theory, mutual information (MI) measures the information content in common between two random sampled signals (I_1 and I_2), considering them as a collection of symbols that are drawn in a random manner (Cover and Thomas, 1991). However, from the point of view of our problem I_1 and I_2 are a pair of windows centered on $I_{VS}(\boldsymbol{p})$ and $I_{LWIR}(\boldsymbol{q})$ respectively, which encode energy measurements in visible and thermal infrared bands. Similarly, we propose to use MI as a cost function that assigns a value depending on its information content; in other words, probabilities of symbols. Formally, $MI(I_1, I_2)$ is defined in terms of individual entropies $h(\cdot)$ and joint entropy $h(\cdot, \cdot)$ as:

$$MI(I_1, I_2) = h(I_1) + h(I_2) - h(I_1, I_2). \tag{2}$$

Alternatively, the above equation can be expressed in its continuous form as integrals of the marginal probability distribution functions (PDFs) and joint PDF of pixel values i_1 and i_2 into I_1 and I_2 respectively, then:

$$h(I) = -\int_0^1 P_I(i) \log P_I(i) di, \tag{3}$$

$$h(I_1, I_2) = -\int_0^1 \int_0^1 P_{I_1,I_2}(i_1, i_2) \log P_{I_1,I_2}(i_1, i_2) di_1 di_2, \tag{4}$$

where P_{I_1,I_2} represents the joint PDF and P_I the marginal PDFs. Kim et al. Kim et al. (2003) approximate these PDF and PDFs by a Parzen window density estimation, which is a sum of Gaussian distributions g, with mean μ and covariance ψ (a detailed explanation can be found in (Viola and Wells, 1997)). In the current work a *nonparametric estimator* (*NP*) (Dowson et al., 2008) is used for computing the joint PDF, instead of using a Parzen estimator. In this way, we avoid dependencies in the selection of parameters: μ and ψ of the Parzen estimator and the parameter needed for binning I_1 and I_2. Notice that a joint PDF is a two dimensional histogram, where rows and columns represent symbols from I_1 and I_2. In our scope, these symbols come from pixel values of multispectral images; however,

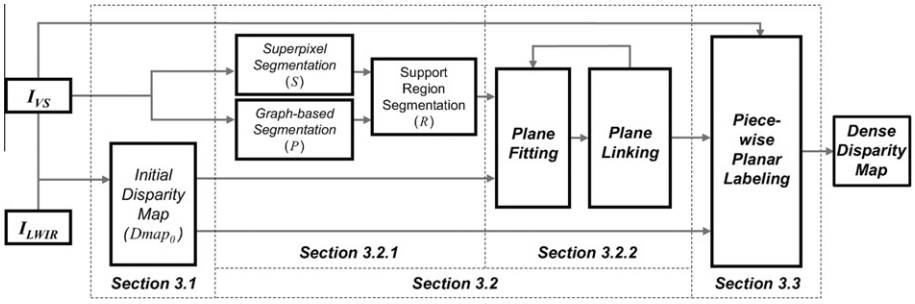

Fig. 1. Illustration of the algorithm's stages.

Fig. 2. Inputs and output images of first stage: (a) rectified infrared image I_{LWIR}; (b) rectified visible image I_{VS}; and (c) initial sparse disparity map $Dmap_0$.

since thermal infrared measurements tend to be concentrated in few bins, particularly in outdoor scenes where the temperature remains uniform (thermal equilibrium), the contribution of the estimator used in the current work is significant because it does not require a parameter tuning for binning I_1 and I_2 as Barrera et al. (2010). Hence, the joint PDF is obtained as:

$$P_{I_1,I_2}(i_1,i_2) = NP(I_1,I_2). \tag{5}$$

Once P_{I_1,I_2} is obtained, the joint entropy term, $h(I_1,I_2)$ in Eq. (4), is computed as follows:

$$h(I_1,I_2) = -\sum_{i_1\in I_1, i_2\in I_2} \log(P_{I_1,I_2}(i_1,i_2)) * g_\psi(i_1,i_2), \tag{6}$$

where g_ψ is a Gaussian kernel needed to approximate the continuous form in Eq. (4) from to its equivalent discrete (see Kim et al., 2003; Hirschmuller, 2008 for more details). In practice, we found that using a small kernel of 3×3 pixels is enough for achieving good approximations, despite of the few samples in I_1 and I_2. Finally, marginal PDFs, corresponding to $P_{I_1}(i)$ and $P_{I_2}(i)$ in Eq. (3), are computed in a similar way to the joint probability but along each dimension of P_{I_1,I_2}. Thus, $P_{I_1}(i_1) = \sum_{i_2\in I_2} P_{I_1,I_2}(i_1,i_2)$ and $P_{I_2}(i_2) = \sum_{i_1\in I_1} P_{I_1,I_2}(i_1,i_2)$. Then:

$$h(I) = -\sum_i \log(P_I(i)) * g_\psi(i), \tag{7}$$

where $P_I(i)$ represents $P_{I_1}(i_1)$ or $P_{I_2}(i_2)$, which are one dimensional vectors. The matching cost volume of mutual information for the whole image is:

$$C_{MI}(\boldsymbol{p},d) = h_{I_{VS}}(\boldsymbol{p}) + h_{I_{LWIR}}(\boldsymbol{q}) - h_{I_{VS},I_{LWIR}}(\boldsymbol{p},\boldsymbol{q}), \tag{8}$$

remember that $\boldsymbol{q} = (x+d,y)$ are the expected correspondences of $\boldsymbol{p}$ in I_{LWIR} computed from $d = \{d_{min},\ldots,d_{max}\}$.

Mutual information is able to find linear and nonlinear correlations between a pair of windows, taking into account the whole dependence structure of variables. However, since local image structures provide rich information that could be also exploited, we introduce a term based on gradient information (*GI*). Thus, this new term is intended to contribute to the matching score in textured regions comparing the orientation of gradient vectors. It is based on the observation that gradient vectors with similar orientations unveil potential matches. The *GI* is defined by the product of two elements; the first one measures the similarity in the orientation of gradient vectors; while the second one is a factor that weights this similarity value. Analogously to *MI*, the gradient information is extracted from I_1 and I_2; it is defined as follows:

$$GI(I_1,I_2) = \sum_{\boldsymbol{x},\boldsymbol{x}'} w(\theta(\nabla_1(\boldsymbol{x}),\nabla_2(\boldsymbol{x}'))) \min(|\nabla_1(\boldsymbol{x})||\nabla_2(\boldsymbol{x}')|), \tag{9}$$

where: $\nabla_1(\cdot)$ is the gradient vector field of I_1; $\boldsymbol{x}$ is a coordinate refereed to this vector field (same for $\nabla_2(\cdot)$, where $\boldsymbol{x}'\in I_2$); $|\cdot|$ is the norm; $\theta(\boldsymbol{x},\boldsymbol{x}')$ is the angle between them; and $w(\theta)$ is a function that penalizes gradient orientation out of phase or counter phase: $w(\theta) = (\cos(2\theta)+1)/2$. The gradient information is computed similarly to *MI* on two windows centered on $\nabla(I_{VS}(\boldsymbol{p}))$ and $\nabla(I_{LWIR}(\boldsymbol{q}))$, thus the cost volume $C_{GI}(\boldsymbol{p},d)$ is obtained by sliding them through the searching space defined by each $\boldsymbol{p}$ on the reference image.

It has been reported in the literature (e.g., Fookes et al., 2004; Pluim et al., 2001; Barrera et al., 2010) that an additional improvement can be obtained by using a scale space representation that propagates cost values between levels. It starts in the coarsest level and ends in the finest one (from level t to 0) as is depicted next:

$$C_{MI_{ss}}(\boldsymbol{p},\boldsymbol{q}) = [\alpha_0,\ldots,\alpha_t]^T \cdot C_{MI}(L_0^t(\boldsymbol{p},\boldsymbol{q})), \tag{10}$$

$$C_{GI_{ss}}(\boldsymbol{p},\boldsymbol{q}) = [\beta_0,\ldots,\beta_t]^T \cdot C_{GI}(L_1^t(\boldsymbol{p},\boldsymbol{q})), \tag{11}$$

where t is an index that refers to the level in the scale space ($t\in\mathbb{N}$); L_0^t and L_1^t are scale space representations given by convolution of a image with a Gaussian kernel of standard deviation (σ), which is progressively increased until obtaining an image stack. Two Gaussian derivative kernels of order 0 and 1 are used to generate blurred and gradient stacks. The α_i and β_i are weights for the linear combination of the results from the different levels of the stack. From $C_{MI_{ss}}$ and $C_{GI_{ss}}$ a volume of cost values is obtained from the cost function presented in Eq. (1). Finally, an initial sparse disparity map ($Dmap_0$)

is extracted using a bounded WTA strategy. This bounded WTA only considers disparity values in Eq. (1) resulting from $C(\lambda, \boldsymbol{p}, d) \geqslant \tau$.

Fig. 2(c) shows an illustration of the sparse disparity map resulting from this first stage (output). The multispectral input images are rectified during the calibration stage (see Section 4 for more details).

3.2. Plane based hypotheses

In this section the given color image is segmented into a set of regions. Then, each region is represented by a single plane, using the information from the initial sparse disparity map. These planes are used in the final stage as labels for computing the dense disparity map we are looking for.

3.2.1. Support region segmentation

This step involves the combination of two segmentation algorithms (i.e., Levinshtein et al., 2009; Felzenszwalb and Huttenlocher, 2004), which are applied to I_{VS} image for obtaining regions that preserve the objects boundaries in the scene. These segmentation algorithms are used in a split-and-merge scheme, in order to unveil potential planar regions. So, the image is decomposed into small regions (superpixels) that later on are connected, following a perceptual criterion. It should be noted that this combination of algorithms is motivated by the application domain (man-made environments).

The segmentation into support regions begins by splinting the given I_{VS} into a large set of small regions, referred to as *superpixels* (Levinshtein et al., 2009): $S = \{s_0, s_1, \ldots, s_m\}$. Without loss of generality, we assume that disparity values inside a superpixel s_i can be accurately fitted by a plane; this assumption is met as long as I_{VS} is oversegmented. Hence, a trade-off between size of superpixels and fitting error should be found (in the current work 1000 regions were used). Large regions have a high probability of covering more than a single planar surface, on the contrary, smaller ones provide few samples to make a proper estimation of the geometry of the selected region. Then, in order to extract perceptually meaningful regions, the segmentation algorithm proposed in (Felzenszwalb and Huttenlocher, 2004) is applied to I_{VS}. This results in a set of P partitions of the reference image: $P = \{p_0, p_1, \ldots, p_n\}$. Finally, the results from superpixels (S) belonging to the same perceptual region (P) are connected giving rise to the support regions $R = \{r_0, r_1, \ldots, r_n\}$ we were looking for (details on the two segmentation algorithms can be found in (Levinshtein et al., 2009; Felzenszwalb and Huttenlocher, 2004) respectively):

$$r_i = \bigcup_{j \in \Omega_i} s_j, \quad \Omega_i = \{j | s_j \cap p_i \geqslant s_j \cap p_k,\ k \neq i\} \tag{12}$$

where Ω_i are the indexes of those superpixels with a maximum overlaping with the given perceptual region p_i. Fig. 3 shows an illustration of the results from the two segmentation algorithms, S and P, as well as their fusion R.

3.2.2. Planar hypothesis 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 is imposed. So, for every region $r_i \in R$ a RANSAC like algorithm (Fischler and Bolles, 1981) is employed to estimate a pair $\langle \widehat{n}, \overline{\boldsymbol{x}} \rangle$, where $\widehat{n}$ is the normal vector and $\overline{\boldsymbol{x}}$ is the mean value coordinates of the points used for fitting this plane. Note that the planar region estimator operates in the disparity space (x, y, d), which is different to previous approaches that work on depth maps represented in the Euclidean space (e.g., Gallup et al., 2010; Sinha et al., 2009).

A RANSAC based plane estimator is chosen since the accuracy of the sought disparity maps depends directly on the confidence of the planar hypotheses. By definition, these methods are capable to find local models from noisy cloud of data; for instance, previous works have demonstrated that this kind of algorithms overcomes least squared based techniques, since they are less sensitive to outliers (Torr and Zisserman, 2000). It should be mentioned that only those regions r_i that contain three or more valid disparities ($Dmap_0(r_i)$) are considered during this fitting step.

Once RANSAC algorithm has been applied to all the regions in R, 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, 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 $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$ 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 (Tao et al., 2001). It is defined as follow:

$$dist_\Pi(\pi_i, \pi_j) = l(\pi_i, \pi_j) + l(\pi_j, \pi_i), \tag{13}$$

$$l(\pi_i, \pi_j) = \frac{(\overline{x_j} - \overline{x_i}) \cdot \widehat{n}_j}{\widehat{n}_i \cdot \widehat{n}_j}. \tag{14}$$

The Eq. (14) corresponds to the length of the segment defined by $\overline{x_j}$ and the intersection of $\widehat{n}_j$, 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. (13) is given in Fig. 4.

The previous planes distance (Eq. (13)) is used as a similarity function for merging a pair of planar patches. Hence, two planar patches are fused into a single one if ($dist_\Pi(\pi_i, \pi_j) \leqslant \tau_{link}$). Once all possible combinations have been evaluated (only connected neighbor regions are considered) a new relabeled R is obtained and the RANSAC algorithm is called again until convergence is reached.

Finally, once there are not more planes to be joined, a noisy planar hypothesis removal is performed. It is based on detecting the predominant planar orientations, using a PCA over all normal vectors ($\widehat{n}$). This filtering stage tends to remove planes with an orientation $\widehat{n}_i$ far away from the principal directions. This results in a compact set of planar hypotheses $\Pi = \{\pi_1, \pi_2, \ldots, \pi_c\}$; it is expected that the number of hypotheses has been reduced: $c \ll n$. Fig. 5(b) shows the planar hypotheses obtained after merging planar patches with similar parameters and filtering the noisy ones. The original set contains 179 hypotheses (see Fig. 5(a)), while the one presented in Fig. 5(b) is defined by only 14 hypotheses. They were obtained after four iterations of the plane linking stage.

3.3. Piecewise planar 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_∞ that represents all those regions out of the stereo range is added to Π (e.g., sky or distant surfaces).

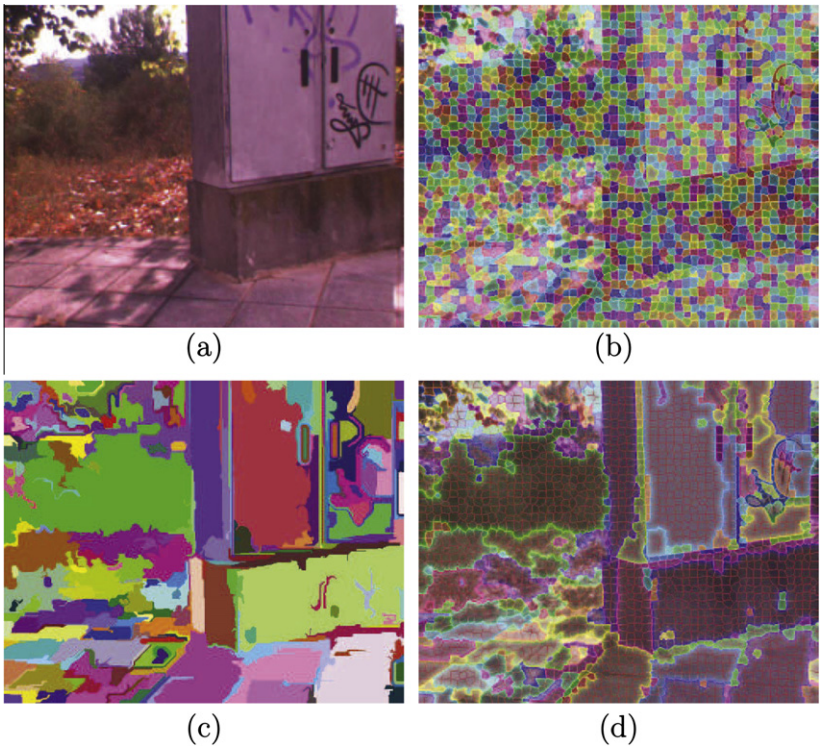

Fig. 3. Illustration of the support region segmentation: (a) original I_{VS}; (b) superpixels (S) obtained from Levinshtein et al. (2009); (c) perceptual regions (P) from Felzenszwalb and Huttenlocher (2004); and (d) support regions (R) obtained by fusing (b) and (c).

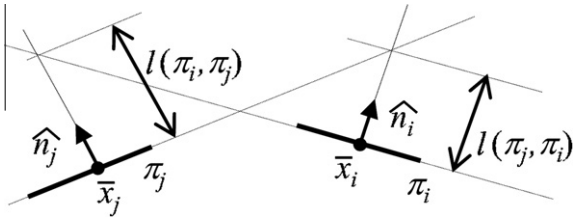

Fig. 4. 2D illustration of segment lengths used to compute $dist_{\Pi}(\pi_i, \pi_j)$.

The Markov Random Field theory provides a framework to relate local correlation indicators together with contextual constraints. These two elements are used to define the energy function. Then, in the current work, this function is minimized through the classical graph cuts (Boykov et al., 2001). It works by defining a regular grid where every node represents a pixel in the image; these nodes are then connected to a set of additional nodes corresponding to the planar hypotheses Π. Hence, a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ represents the vertices and $\mathcal{E}$ represents the edges of the graph, is obtained. Then, the graph cut algorithm searches in $\mathcal{G}$ for the best set of labels (f) that minimizes the following energy function:

$$E(f) = \sum_{p \in \mathcal{P}} D_p(f_p) + \sum_{p,q \in \mathcal{N}} \lambda_{smooth} V_{pq}(f_p, f_q), \tag{15}$$

where $\mathcal{P}$ is the set of pixels in the image; D_p is the data term that measures how well a planar hypothesis explains a disparity value for a given pixel p; $V_{pq}(f_p, f_q)$ is a smoothness prior computed in a neighborhood $\mathcal{N}$ (in the current work the first-order Markov Random Field is considered, in other words a neighborhood of four connections); f_p, f_q are the current labels for pixels p and q respectively;

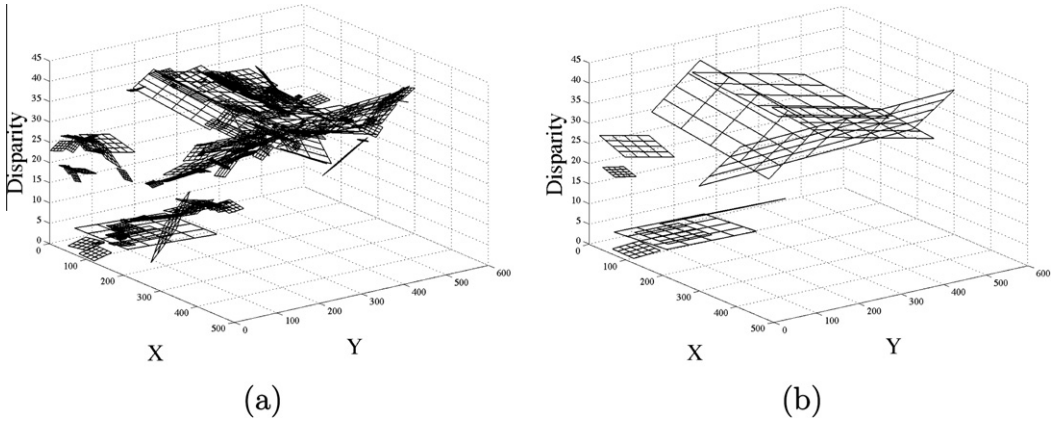

Fig. 5. Planar hypotheses simplification: (a) original set of planar hypotheses from the segmentation presented in Fig. 3 (179 planes) and (b) planar hypotheses resulting after four iterations of the postprocessing stage (14 planes).

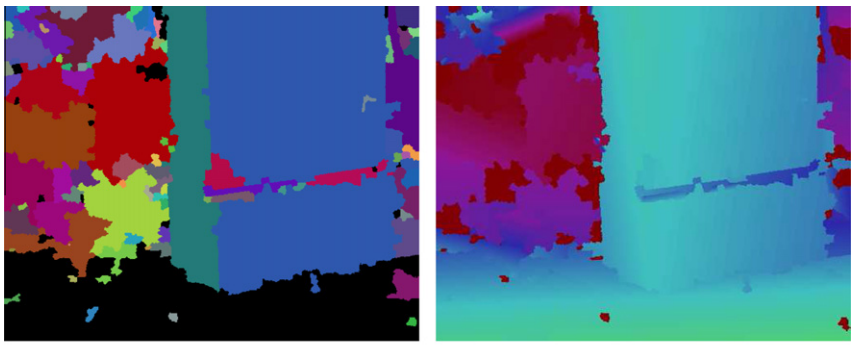

Fig. 6. Graph cuts outcomes: (a) labeled regions from graph cuts and (b) dense disparity map obtained with the proposed approach.

and λ_{smooth} is a weighting factor for the regularization term. The D_p function is defined as follows:

$$D_p(f_p) = \begin{cases} \min(C(f_p), C_{max}) & \text{if } f_p \in \{\pi_1, \pi_2, \ldots, \pi_n\}, \\ 0.9 \cdot C_{max} & \text{if } f_p \in \{\pi_\infty\}, \end{cases} \tag{16}$$

the cost assigned to a pixel p represents the degree of membership of p to a given plane π_i. This cost is obtained from $C(\lambda, \boldsymbol{p}, d)$ (see Eq. (1)), where d corresponds to the hypothetical disparity obtained if p is assigned to the plane π_i; if certain hypothesis π_i produces an inconsistent d, for instance a value outside of the searching space, that p is penalized with a maximum cost C_{max}. Finally, the smoothness term V_{pq} is defined as:

$$V_{pq}(f_p, f_q) = \nabla \cdot \begin{cases} 0 & \text{if } f_p = f_q, \\ d_{max} & \text{if } f_p \text{ or } f_q \in \pi_\infty, \\ d(f_p, f_q) & \text{otherwise}, \end{cases} \tag{17}$$

∇ is the gradient of VS image; and $d(f_p, f_q)$ is the Euclidean distance between p and q derived from the planes they belong to. The minimization of Eq. (15) assigns every pixel (p) of the image to a planar hypothesis π_i (see Fig. 6(a)). Then, from this membership the corresponding disparity value is obtained by computing the intersection of a ray passing through p with the assigned plane π_i. Fig. 6(b) shows the dense disparity map corresponding to the illustration used as a case study in previous sections.

4. Experimental results

This section first describes the multispectral stereo head, details about its geometry and calibration are also provided. Then, the proposed multispectral data set and the evaluation framework are presented. Finally, resulting dense disparity maps together with their analyses are given.

The multispectral stereo head consists of a pair of cameras 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 multispectral stereo head are pre-aligned, ensuring their right rectification. Thermal infrared images are obtained with a *Long-Wavelength InfraRed* camera (PathFindIR from Flir[2]) while color ones with a standard Sony ICX084 camera, which has a focal length of 6 mm.

Multispectral stereo camera calibration is considerably more complex than the classical VS/VS, because the LWIR sensor measures heat variations. Therefore, a calibration pattern ideally should have two different temperatures for generating contrast images. In practice, this is not feasible. Furthermore, the effect of thermal diffusion between the calibration pattern and air causes both smooth step edges and distorted corners in infrared images, which are not perceived at a glance. In order to avoid these problems we calibrate the multispectral head in an outdoor scenario using a metallic checkerboard. In this way, sun rays are reflected in white rectangles and absorbed in the black ones, this procedure enhances the contrast of image and helps the detection of calibration points. Although the problem of blurred calibration points is partially solved by the lighting reflection/absorption technique, a saddle point detector is considered instead of a classical corner detector to obtain more robust results.

As mentioned above, the cameras have been aligned before starting the calibration process. This action ensures that the needed projective transformations for their rectification are smooth (the image planes' position are approximately coplanar). Once each camera has been calibrated, and its intrinsic parameters are known, the next step is to estimate the geometry of multispectral stereo rig. Since the current work is focused on the generation of dense disparity maps, it is only necessary to estimate the epipolar geometry (fundamental matrix $\boldsymbol{F}$). Then, with this matrix, the next step is to rectify the multispectral images.

The 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 $\boldsymbol{F}$ was estimated, it is essential to use a rectification method that takes into account the large dissimilarity of intrinsic parameters of the cameras. In the current work the method proposed in (Mallon and Whelan, 2005) has been used. This reduces the loss and creation of pixels due to projective transformations during the rectification process (resampling effect), while preserving the aspect of image content.

In order to evaluate the proposed method a set of multispectral images has been collected. The captured data depict a variety of urban scenes, which includes: buildings, sidewalk, trees, vehicles, among others (see Fig. 7). It consists of 116 scenes. Every scene includes their corresponding thermal infrared and color images, which were acquired by using the proposed multispectral stereo head, and processed till get a rectified pair. Moreover, this includes both a disparity map of the scene and a hand-annotated map of planar regions.

The disparity maps are provided by a VS/VS stereo vision system: Point Grey Bumblebee (PGB). Note that, for the sake of simplicity, the multispectral stereo head was presented as an independent system, however it uses one of the cameras that belongs to PGB—the right one. In other words, the camera refereed to as VS in the current section corresponds to the right camera of PGB. This stereo rig setup was selected because it is efficient in terms of hardware and software.

[2] www.flir.com.

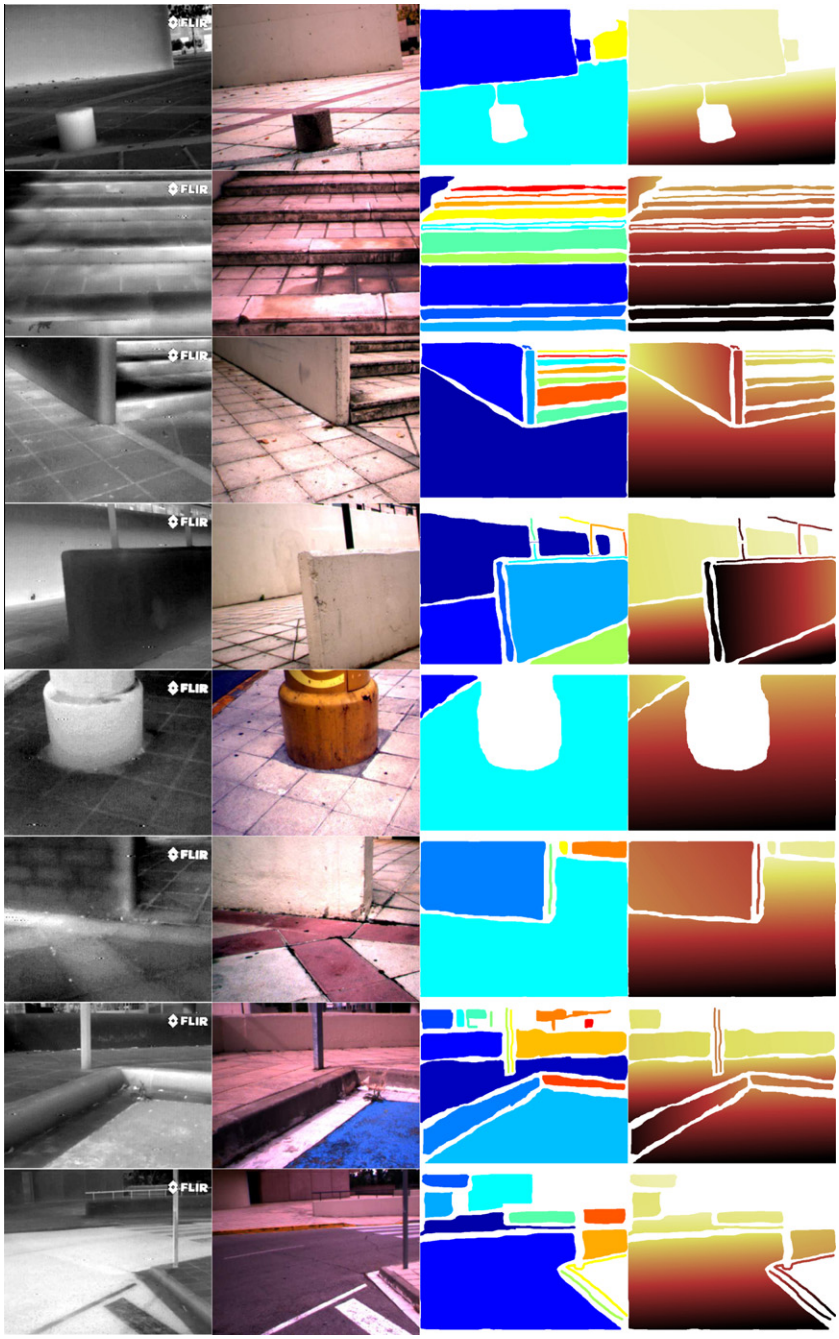

Fig. 7. Examples of evaluation data set; each column corresponds to: (*1st*) I_{LWIR} images; (*2nd*) I_{VS} images; (*3rd*) maps of planar regions; and (*4th*) synthesized disparity maps used as ground truth for evaluating the proposed approach.

Additionally, our data set also includes a group of maps that indicate connected planar regions. These maps have been hand-labeled taking into account the geometry of the surfaces, thus a unique label is assigned to each region and it identifies the pixes that belongs to the same plane. Fig. 7 shows some of the images used for validating the proposed approach. I_{LWIR} and I_{VS} images are rectified; in both cases the size of resulting images is 506×408 pixels. Hand-labeled and disparity maps are given in their original format 640×480 pixels. Since the disparity maps provided by PGB are only accurate in textured regions, we have used a hand-labeled planar regions for obtaining dense and accurate representations, particularly in textureless and noisy regions. To address this problem, we fit a plane to each hand-annotated region, through of image coordinates and corresponding disparity values. The disparity maps resulting from this user supervised labeling process are shown in Fig. 7 (*4th column*).

The proposed approach has been validated using the data set presented above. 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 $d_{max} = 64$. The scale space representation contains three levels and the values used for propagating mutual and gradient information through the different levels: $[\alpha_0, \ldots, \alpha_t] = [0.2, 0.3, 0.5]$ and $[\beta_0, \ldots, \beta_t] = [0.2, 0.3, 0.5]$; threshold τ is set as 10% of the maximum cost value; finally, mutual and gradient information in Eq. (1) are fused defining $\lambda = 0.65$. The two values related with the planar hypothesis generation were set as follow: $\tau_{RANSAC} = 0.2$

and $\tau_{link} = 2.5$. The values given by default in the graph cut implementation provided by Gallup et al. (2010) were used for the global minimization.

Figs. 8 and 9 show the results obtained for eight different scenes. The initial multispectral stereo images are provided in the first and second columns of Fig. 7. The results are grouped by scenes and they show the output of each step of proposed algorithm. So, every scene has associated four images: (*top-left*) corresponds to support regions R, which split up the I_{VS} image into planar regions; (*top-right*) is an illustration of the planar hypotheses Π; (*bottom-left*) shows the labeled regions obtained by graph cuts; and (*bottom-right*) is the final disparity map. Notice that Π is the set of labels used during the minimization step, and the disparity map is obtained by using the plane parameters corresponding to each label. On the other hand, it can be appreciate how the minimization stage is able to filter out small regions and propagates information across the neighbors (see (*bottom-left*) illustrations in the different scenes and compare them with their corresponding (*top-right*) images in Figs. 8 and 9).

Fig. 9 (scene 5) shows that the proposed approach can obtain dense disparity maps even in non-planar regions. In this illustration a large cylinder is approximated by two planar patches. The number of planar patches depends on the value used for setting the τ_{link} parameter. Even in this challenging case the proposed algorithm is capable of finding a set of planar hypotheses, and converges toward an optimal solution that preserve the appearance of the scene.

The accuracy of proposed algorithm is evaluated by using two metrics. They are frequently employed as quantitative evaluation criteria for stereo matching algorithms. Initially, the absolute mean error (E_{abs}) is computed in a global manner for a given disparity map as follows:

$$E_{abs} = \frac{1}{N}\sum_{j=1}^{N}|d_C(j) - d_T(j)|, \tag{18}$$

where d_C is the disparity map computed by the proposed algorithm, d_T is the ground truth, and N is number of evaluated points. Since our data set offers reliable ground truth only in those points that lie on a planar region, the error measurement is limited to those image coordinates that have a valid ground truth data and disparity value. Notice that the label π_∞ used during the minimization step corresponds to non disparity, for this reason these image coordinates are excluded from the evaluation. The main drawback of using E_{abs} as an evaluation metric lies on the fact that it does not distinguish between few disparity estimations with large errors and lot disparity estimations with small errors. Furthermore, it does not take into account that a small disparity value corresponds to a large depth value, and therefore its contribution to the global error should be different, for instance, in comparison to a large disparity (small distance). Hence, in order to take into account this effect, the mean relative error (E_{rel}) is also used. It is computed as follows:

$$E_{rel} = \frac{1}{N}\sum_{j=1}^{N}\frac{|d_C(j) - d_T(j)|}{d_T(j)}. \tag{19}$$

E_{abs} and E_{rel} are computed from the 8 scenes that are used as case studies (see Figs. 8 and 9); their corresponding error scores are presented in Table 1. The E_{rel} in the scenes 1 and 4 are considerable larger than the rest of scenes in the data set. In both cases these large values result for the wrong matchings due to the lack of texture in the predominant geometries (a vertical non-textured wall).

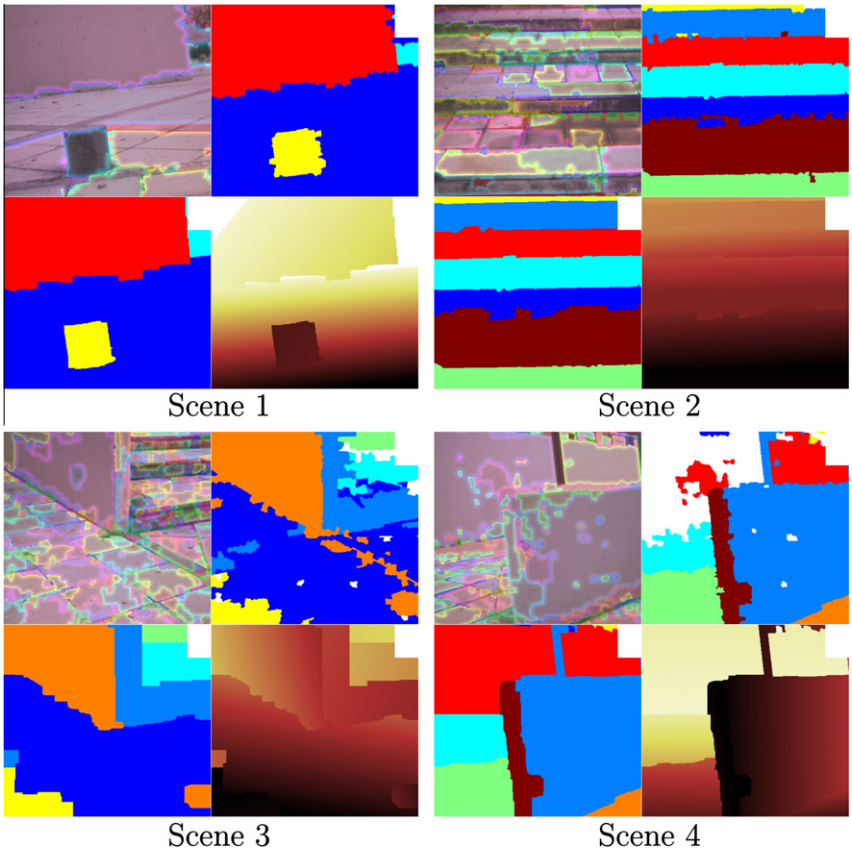

Fig. 8. Experimental results from different stages of the proposed approach; in each scene the illustrations correspond to: (*top-left*) support region R; (*top-right*) planar hypotheses Π; (*bottom-left*) labeled regions from graph cuts; and (*bottom-right*) final disparity map.

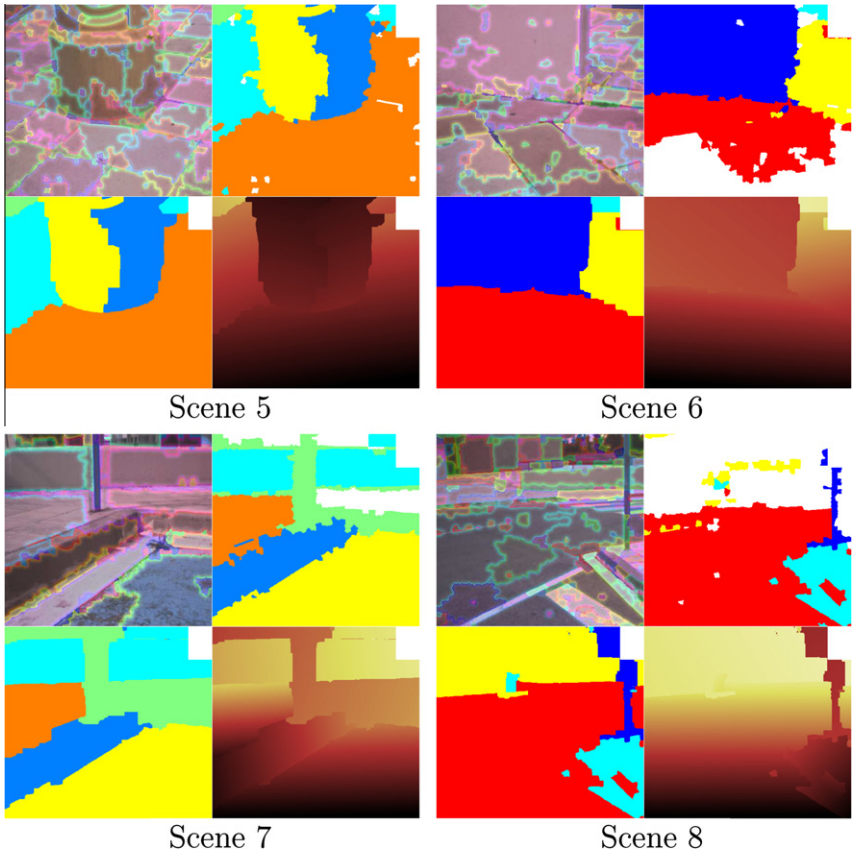

Fig. 9. Experimental results from different stages of the proposed approach; in each scene the illustrations correspond to: (*top-left*) support region R; (*top-right*) planar hypotheses Π; (*bottom-left*) labeled regions from graph cuts; and (*bottom-right*) final disparity map.

Table 1
Global E_{abs} and E_{rel} of the case studies presented in Figs. 8 and 9.

Scene	1	2	3	4	5	6	7	8
E_{abs}	0.635	0.443	0.582	0.629	0.484	0.387	0.465	0.505
E_{rel}	0.167	0.016	0.096	0.144	0.042	0.060	0.062	0.057

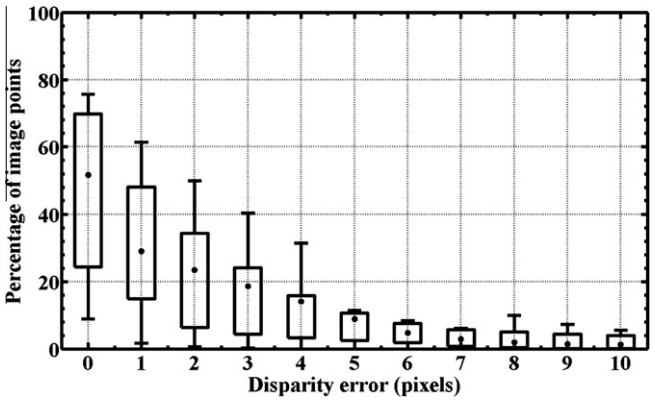

Fig. 10. Average accuracy of the results obtained with the proposed approach computed from the whole data set.

Fig. 10 shows the average accuracy of the obtained dense disparity maps, when all the scenes in our data set are considered. For each scene an accuracy histogram is computed by using its corresponding ground truth map. The histogram counts the number of points for a given disparity error, spanning from 0 till 10 pixels.

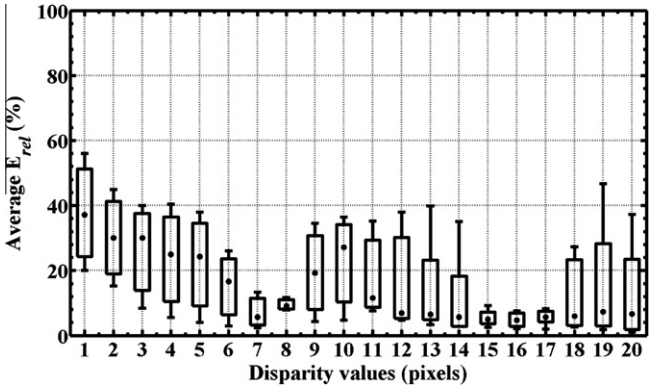

Fig. 11. Average E_{rel} of the results obtained with the proposed approach computed from the whole data set.

Then, from all these histograms a single plot is computed showing the variability of results (see Fig. 10). In this plot the central mark of each box corresponds to its median value and the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme cases.

Fig. 11 shows the average E_{rel} computed from the whole data set. The E_{rel} measurements are restricted to discrete values from 1 to 20 for the sake of visualization. This plot is similar to the previous one and presents the box plot of the average relative error when all the images into the evaluation data set are considered. Note how the mean E_{rel} decreases to values below 10% when the disparity is higher than 10 pixels. On the other hand, disparity values smaller or equal than 10 pixels correspond to distant points

(several meters away from the stereo rig), which are out of the calibration range of the current work.

The results presented above answer the question that was formulated in the Introduction (Section 1), which motivated the current work. They 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. This will allow for instance to extract semantic relationships between the objects in the scene.

5. Conclusions

The current work presents a novel framework for extracting dense disparity maps from multispectral stereo images, each one of its stages is described as well as the image rectification and camera calibration. The results obtained from this research can benefit those fields where visible and thermal infrared cameras coexist. The main contribution of current work are as follow: (i) it introduces a cost function for obtaining multispectral matching, exploiting mutual and gradient information in a scale space representation; (ii) it proposes a global minimization scheme, which is based on the Manhattan-world assumption, to extract dense disparity maps. Finally, although not a theoretical contribution, a large data set of multispectral stereo images has been generated and is freely available by contacting the authors.

We have shown that under certain restrictions is possible to obtain accurate disparity maps, however the low correlation between thermal infrared and visible images restricts its usefulness in complex environments, being this still an open issue. Future work will be mainly focused on the extraction of a ground truth data, which should includes depth information both of planar and non-planar regions. Additionally, different interest regions such as occlusion and discontinuities would have to be identified, as happen in the (VS/VS) evaluation frameworks for dense stereo algorithm.

Acknowledgment

This work has been partially supported by the projects TIN2011-29494-C03-02 and TIN2011-25606; and research programme Consolider-Ingenio 2010: MIPRCV (CSD2007-00018).

References

Barrera, F., Lumbreras, F., Sappa, A., 2010. Multimodal template matching based on gradient and mutual information using scale-space. In: IEEE Internat. Conf. on Image Processing, Hong Kong, pp. 2749–2752.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Machine Intell. 23 (11), 1222–1239.

Coughlan, J., Yuille, A., 1999. Manhattan world: compass direction from a single image by bayesian inference. In: IEEE Internat. Conf. on Computer Vision, vol. 2, Kerkyra, Corfu, Greece, pp. 941–947.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. Wiley-Interscience, New York, NY, USA.

Dowson, N., Kadir, T., Bowden, R., 2008. Estimating the joint statistics of images using nonparametric windows with application to registration using mutual information. IEEE Trans. Pattern Anal. Machine Intell. 30 (10), 1841–1857.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Internat. J. Comput. Vision 59, 167–181.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24 (6), 381–395.

Fookes, C., Maeder, A.J., Sridharan, S., Cook, J., 2004. Multi-spectral stereo image matching using mutual information. In: IEEE Internat. Symp. on 3D Data Processing, Visualization and Transmission, Thessaloniki, Greece, pp. 961–968.

Gallup, D., Frahm, J.-M., Pollefeys, M., 2010. Piecewise planar and non-planar stereo for urban scene reconstruction. In: IEEE Internat. Conf. on Computer Vision and Pattern Recognition, San Francisco, CA, USA, pp. 1418–1425.

Han, J., Bhanu, B., 2007. Fusion of color and infrared video for moving human detection. Pattern Recognition 40, 1771–1784.

Hirschmuller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Machine Intell. 30 (2), 328–341.

Kim, J., Kolmogorov, V., Zabih, R., 2003. Visual correspondence using energy minimization and mutual information. In: IEEE Internat. Conf. on Computer Vision, vol. 2, Nice, France, pp. 1033–1040.

Krotosky, S.J., Trivedi, M.M., 2007. Mutual information based registration of multimodal stereo videos for person tracking. Computer Vision and Image Understanding 106 (2–3), 270–287.

Krotosky, S.J., Trivedi, M.M., 2007. on color-infrared-and multimodal-stereo approaches to pedestrian detection. IEEE Trans. Intell. Transport. Syst. 8 (4), 619–629.

Krotosky, S.J., Trivedi, M.M., 2008. Person surveillance using visual and infrared imagery. IEEE Trans. Circuits Syst. Video Technol. 18 (8), 1096–1105.

Levinshtein, A., Stere, A., Kutulakos, K., Fleet, D., Dickinson, S., Siddiqi, K., 2009. Turbopixels: Fast superpixels using geometric flows. IEEE Trans. Pattern Anal. Machine Intell. 31 (12), 2290–2297.

Mallon, J., Whelan, P.F., 2005. Projective rectification from the fundamental matrix. Image Vision Comput. 23 (7), 643–650.

Pluim, J., Maintz, J., Viergever, M., 2001. Mutual information matching in multiresolution contexts. Image Vision Comput. 19 (1–2), 45–52.

Prakash, S., Lee, P.Y., Caelli, T., 2006. 3d mapping of surface temperature using thermal stereo. In: Internat. Conf. on Control, Automation, Robotics and Vision, pp. 1–4.

Sinha, S.N., Steedly, D., Szeliski, R., 2009. Piecewise planar stereo for image-based rendering. In: IEEE Internat. Conf. on Computer Vision, Kyoto, Japan, pp. 1881–1888.

Tao, H., Sawhney, H.S., Kumar, R., 2001. A global matching framework for stereo computation. IEEE Internat. Conf. on Computer Vision 1, 532–539.

Torr, P.H.S., Zisserman, A., 2000. Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78, 138–156.

Trivedi, M.M., Cheng, S., Childers, E., Krotosky, S., 2004. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. IEEE Trans. Veh. Technol. 53 (6), 1698–1712.

Viola, P., Wells III, W.M., 1997. Alignment by maximization of mutual information. Internat. J. Comput. Vision 24, 137–154.

Yang, R., Chen, Y., 2011. Design of a 3-d infrared imaging system using structured light. IEEE Trans. Instrum. Meas. 60 (2), 608–617.