Evaluation of Similarity Functions in Multimodal Stereo

Fernando Barrera[1], Felipe Lumbreras[1,2], and Angel D. Sappa[1]

[1] Computer Vision Center, Campus UAB, Bellaterra, Spain
[2] Computer Science Department, Universitat Autonóma de Barcelona, Spain
{jfbarrera,felipe,asappa}@cvc.uab.es

Abstract. This paper presents an evaluation framework for multimodal stereo matching, which allows to compare the performance of four similarity functions. Additionally, it presents details of a multimodal stereo head that supply thermal infrared and color images, as well as, aspects of its calibration and rectification. The pipeline includes a novel method for the disparity selection, which is suitable for evaluating the similarity functions. Finally, a benchmark for comparing different initializations of the proposed framework is presented. Similarity functions are based on mutual information, gradient orientation and scale space representations. Their evaluation is performed using two metrics: *i*) disparity error, and *ii*) number of correct matches on planar regions. In addition to the proposed evaluation, the current paper also shows that 3D sparse representations can be recovered from such a multimodal stereo head.

Keywords: Stereo vision, Infrared imaging, Multimodal imaging.

1 Introduction

Stereo matching is a classical problem in computer vision. Over the past years, a large number of researches have been focused on matching methods for binocular stereovision systems, which traditionally are made from color cameras (e.g., [1], [2], and [3]). However, a recent family of thermal infrared cameras have opened the possibility of combining them into a novel multimodal stereo rig. Thus, a new kind of stereo head that registers two spectral bands becomes a reality; visible (VS) and Long-Wavelength InfraRed (LWIR).

Initially, it may seem that disparity maps obtained from a multimodal stereo head (LWIR/VS) would lead to representations with a higher information content, in other words, not only depth but also temperature is provided. However, before obtaining such a kind of rich representation the correspondences between infrared and color images should be found. Note that these images are significantly different, making this action a challenging task. Furthermore, few approaches have been proposed in the literature and is difficult to identify the best option to develop a robust matching algorithm able to overcome infrared and color variations.

Krotosky et al. [4] introduce a matching algorithm for regions that contain human body silhouettes, both in thermal infrared and visible spectrum. They

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part I, LNCS 7324, pp. 320–329, 2012.

extract rectangular windows from these images. Next, their degree of similarity is measured by using mutual information. Although, this approach is valid for tracking people or depth estimation of pedestrians, is an application-oriented solution. Furthermore, it assumes that hot points correspond to persons, and only those regions should be matched.

Recently, Torabi et al. [5] performs a comparative evaluation of dense stereo correspondence algorithms in the multimodal field, under the same restrictions. They conclude that similarity functions, previously used in VS/VS stereo heads, such as Normalized Cross-Correlation (NCC) and Histograms of Oriented Gradients (HOG) [6] are highly sensitive to dissimilarities even in presence of edges. In contrast, mutual information and Local Self-Similarity (LSS) [7] are the most accurate and discriminative correspondence measures among the evaluated ones.

In the current work, an evaluation similar to Torabi et al. [5] is performed but without those restrictions (application domain and image regions). Hence, our results can be used to predict the behavior of a cost function in a general context. The evaluation starts by selecting the most suitable similarity functions for the multimodal stereo matching. The selection is based on the study of related works, not only on multimodal stereo but in similar problems where information from different modalities need to be merged. Similarity functions frequently used in VS/VS stereo, are not considered since they assume a linear correlation and our problem is clearly non linear. On the other hand, LLS [7] is excluded from the current evaluation since is an application oriented similarity measure.

The proposed evaluation framework includes four similarity functions. The first one is Mutual information (**I**), which has been successfully used in multimodal image registration, particularly in medical applications as well as in stereo systems with different lighting conditions (e.g., [3]). The next two functions are gradient information (**G**) and its combination with mutual information (**IG**) [8]. Finally, The gradient and mutual information in a scale-space representation (**IGSS**) is also evaluated [9]. Our evaluation differs from previous studies mainly in three aspects: *i*) the LWIR/VS stereo problem is tackled; *ii*) a large multimodal dataset is used for the evaluation; *iii*) it is not oriented to a specific application. Note that in all the cases the optimal parameters of the similarity functions are found in order to do a fair evaluation.

The multimodal dataset consists of a set of planar regions at different position and orientation, which allows us to quantify the accuracy of similarity functions. Moreover, it exploits an evaluation criterion used for comparing of dense stereo algorithm. Thus, it is obtained an error statistic on the whole dataset. The paper is organized as follows. Section 2 introduces the different similarity measures. Then, experimental results and evaluations are provided in Section 3. Finally, conclusions and final remarks are detailed in Section 4.

2 Similarity Functions

Recent works on computational stereo have shown that mutual information is a nonparametric cost function, which is capable to address non-linear correlated

signals [2]. However, in multimodal stereo problems the use of mutual information is not enough to achieve high rates of matching. Mainly, due to that images are compared only through its information content, ignoring shape information. In this section, four similarity functions for a multimodal stereo head are reviewed. First, mutual information [10] is analyzed; then, the use of gradient information is introduced; next, a scheme that combines them (i.e., mutual information and gradient) [8] is described; finally, it is introduced mutual and gradient information when they are propagated through a scale-space representation [9].

2.1 Mutual Information

Mutual information (**I**) measures the amount of information that one random variable contains about another. It is a useful concept where no prior relationships between the data are known. This is estimated in a local way, for two windows I_l and I_r, which are extracted from VS and $LWIR$ images respectively. These windows are centered on image coordinates: $\mathbf{x}$ and $\mathbf{x}'$; and have a size of wz pixels. Thus, **I** is defined as follows:

$$\mathbf{I}(I_l(\mathbf{x});\, I_r(\mathbf{x}')) = \sum_{a_i \in I_l} \sum_{b_j \in I_r} p_{I_l I_r}(a_i, b_j) \log \frac{p_{I_l I_r}(a_i, b_j)}{p_{I_l}(a_i)\, p_{I_r}(b_j)}, \tag{1}$$

where a_i and b_j are discretized pixel values into Q levels; $p_{I_l I_r}$ represents their joint probability mass function; and p_{I_l} and p_{I_r} are their respective marginal probability mass functions. Mutual information has a great advantage over other similarity functions that looking for identical pattern, as occurs in classical stereo (VS/VS). So, this is able to find linear and nonlinear correlations, taking into account the whole dependence structure of I_l and I_r. Figure 1 depicts almost locally that this condition is enough for matching windows as the indicated. However, our experiments have shown that its performance improves when more information is added. In the next section this issue is covered.

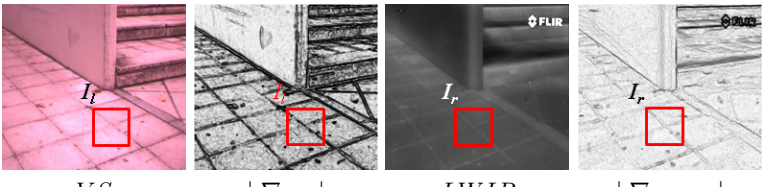

Fig. 1. Multimodal images and their corresponding gradient field

2.2 Gradient Information

Since the images are rectified, not only the search for correspondences is simplified to one dimension, but also the objects in the scene appear with a similar aspect (see Fig. 1). This is an important fact because the contours and edges are

regions with a high correlation value. Therefore, they have a high probability of being correctly matched [11]. The gradient information is obtained as follows:

$$\mathbf{G}(I_l(\mathbf{x});\, I_r(\mathbf{x}')) = \sum_{\mathbf{x}\in I_l,\, \mathbf{x}'\in I_r} w(\theta(\mathbf{x},\mathbf{x}'))\, \min(|\nabla_{I_l}(\mathbf{x})|\,|\nabla_{I_r}(\mathbf{x}')|), \tag{2}$$

where θ is the angle difference between two gradient vectors; $w(\theta)$ is a function that penalizes those gradient vectors that are not in phase or counter-phase; and $|\nabla|$ is the magnitude of the gradient vectors.

As indicated above, the information content of multimodal images is weakly correlated, except in the contours that appear in both images. Therefore, $\mathbf{I}$ could be enriched using the orientation of gradients in those regions [9].

2.3 Mutual and Gradient Information

The third similarity function evaluated is based on the combination of mutual and gradient information; it is defined as:

$$\mathbf{IG}(I_l(\mathbf{x});\, I_r(\mathbf{x}')) = \mathbf{I}(I_l;\, I_r) \cdot \mathbf{G}(I_l;\, I_r). \tag{3}$$

Previous works have shown that the gradient is not stable enough for multimodal matching [5] [8], since only the half range of its possible orientation is useful (it goes from zero to π). Therefore, $\mathbf{I}$ helps to $\mathbf{G}$ to overcome its loss of descriptiveness. Although there are different ways to combine them, their product has a noise cancellation effect, thus the cost values in a textureless region in the LWIR image and textured in the VS image are low (the same in the opposite case). This increases the reliability of correspondences.

2.4 Multiresolution Mutual and Gradient Information

Previous sections have highlighted the importance of using contours and edges in the matching process. Therefore, in this section a structural analysis of images is presented. This requires a scale-space representation, which is obtained applying local derivative operators. The aim is to recover significant information through a scale-space representation, and boost the accuracy of similarity functions at a fine level, using as feedback previous coarse levels. In order to do this, the similarity function presented above is adjusted to a multiresolution scheme. It works by taking the values of mutual and gradient information ($\mathbf{IG}$) at a certain scale t, and propagating it following a coarse to fine strategy, from a scale $t-1$ to t. The next equation is used to fuse these values:

$$\mathbf{IGSS}(I_l(\mathbf{x});\, I_r(\mathbf{x}');t) = \lambda\, \mathbf{IG}_t(I_l;\, I_r) + (1-\lambda)\mathbf{IG}_{t-1}(I_l;\, I_r), \tag{4}$$

2.5 Disparity Selection

The four similarity functions presented above (i.e., eqs. (1), (2), (3), and (4)) measure the degree of similarity between two windows $I_l(\mathbf{x})$ and $I_r(\mathbf{x}')$, but now the problem becomes on searching the right correspondences between $I_l(\mathbf{x})$ and all the possible $I_r(\mathbf{x}')$. The disparity selection process is tackled as a two step optimization problem, where the cost computed between a template and

all possible windows on the corresponding searching space is the variable to optimize. The disparity of each pixel is selected by the Winner-Takes-All (WTA) method. So, the correct match of a pixel is determined by the position d (image coordinate) where the following cost function reaches the maximum value:

$$\arg\max_{d} \{\mathbf{I}\left(I_l(x,\, y);\, I_r(x+d,\, y)\right)\}\,, \tag{5}$$

similarly for the rest of cost functions (eqs. (2), (3), and (4)).

The disparity map obtained in the first step contains mismatchings due to the costs not always has a global maximum (situation most frequent in stereo matching of VS and LWIR images than in just VS images). 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 τ threshold. The selection of threshold τ is evaluated in section 3. Next, these reliable matchings are used for bounding the searching space in their surrounding. This helps to discard wrong matching and decrease the sparsity of depth maps.

3 Evaluation

Before going in detail on the comparative study of similarity functions, the multimodal stereo rig setup and its calibration process are introduced. The multimodal stereo rig is built with a LWIR camera (PathFindIR[1]), and a color camera. The latter corresponds to the right camera of a commercial stereo vision system (Bumblebee[2]), which is used for validating the results, and it does not require in-field calibration. In summary, two stereo systems coexist (Fig. 2($left$)). The right camera coordinate system of Bumblebee is used as a reference for both stereo systems. In this way, disparity maps computed from the Bumblebee are valid for the multimodal stereo and used as an approximation of the structure of the scene.

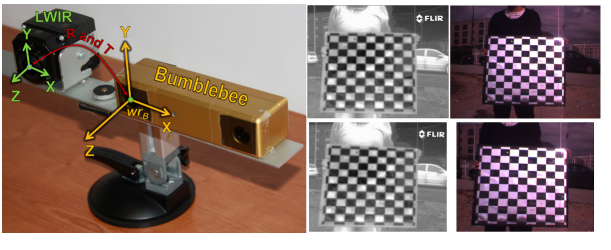

Fig. 2. (*left*) Multimodal stereo rig setup, together with the visible stereo rig. (*top*) Original stereo images of the checkerboard pattern. (*bottom*) Rectified images.

The multimodal stereo rig has been calibrated using Bouguet's toolbox [12]. The main challenge in this stage is to make visible the calibration pattern in

[1] [www.flir.com]
[2] [www.ptgrey.com]

both cameras. In order to do this, a special metallic checkerboard has been made. Figure 2(*top*) shows a pair of calibration images (LWIR and VS). Once the cameras have been calibrated, their intrinsic and extrinsic parameters are known, being possible not only the image rectification but also the estimation of 3D points from image matches. The images were rectified using the method proposed in [13], with an accuracy improvement due to the inclusion of distortion coefficients (radial and tangential) into their camera model. Rectified images are shown in Fig. 2(*bottom*).

The evaluation is performed on a set of images taken on real outdoor scenarios under uncontrolled conditions as lighting, temperature, and depth. It consists of 46 couples of images with dominant planar geometries, as shown in Fig. 3. These images were obtained from our multimodal stereo head and are used to evaluate the performance of the similarity functions.

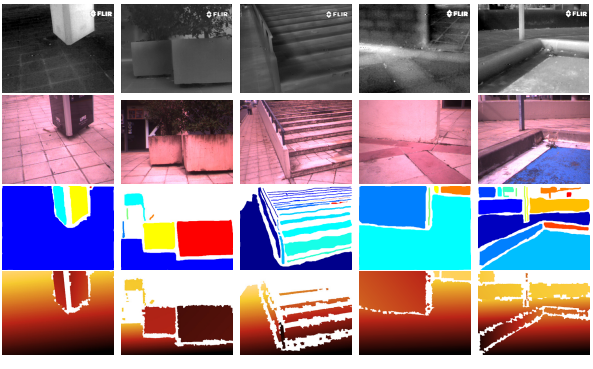

Fig. 3. Examples of images in the evaluation dataset: (1st row) LWIR images; (2nd row) color images; (3rd row) planar regions; and (4th row) disparity maps.

Since it is not possible to have an accurate ground truth data, an indirect method has been envisaged for evaluating the performance of similarity functions. This method consists in measuring the accuracy of disparity values computed by our framework on image regions that are planar surfaces. Actually, the evaluation is performed in the v-disparity space [14] (see Fig. 4 (*right*)). That is, a histogram of disparities in columns direction. The disparity maps are provided by the Bumblebee and are used for computing these representations.

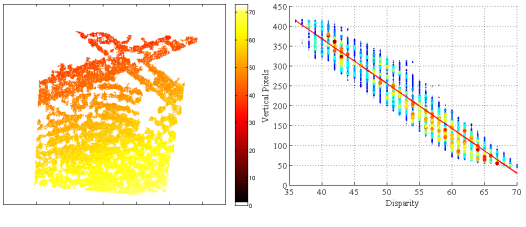

Fig. 4. Evaluation data: (*left*) Disparity map; (*right*) v-disparity representation

The interesting point of v-disparity space is that planes in the Euclidean space are mapped as straight lines. By identifying this straight line, the accuracy of the similarity functions can be evaluated. This works as follows: firstly, it identifies the planar regions in the evaluation images (see Fig. 3(*3rd row*)). Next, their corresponding contributions in v-disparity space are selected. Finally, a linear regression by least squares is applied only to this set of points, which provides the best fitting. In this way, the real position of a plane is estimated from noisy data. Figure 4 (*left*) shows the disparity map when an unique plane is recorded by the cameras; Fig. 4 (*right*) depicts its corresponding v-disparity. Notice that the number of rows in both plots is the same, but the disparity axis in v-disparity representation will depend on the position of the plane. In this plot, the straight line through points represents the *ideal* disparity values of the plane; it also shows the variance due to noise that motivates the proposed evaluation procedure.

Once the planar regions in the evaluation images have been identified, and their corresponding straight lines (ℓ) fitted in the v-disparity representation, an error function based on Root Mean Squared (RMS) is defined. This measures the orthogonal distance ($dist_{\top}$) between a disparity value obtained by our framework, and its corresponding ℓ (which depends on the region it belongs). Thus, the error is defined as:

$$R = \left(\frac{1}{N} \sum_{\mathbf{x} \in P} \left| dist_{\top} \left(d_C \left(\mathbf{x} \right), \ell \right) \right|^2 \right)^{\frac{1}{2}}, \tag{6}$$

where R is the RMS error for a given planar region; P is a planar region; d_C is a disparity value inside of this region; and N is the number of pixels belonging to that region.

In order to evaluate the performance of current framework, the most relevant parameters are varied as follow: $wz = \{7, 19, 31\}$; $\sigma = \{0.5, 1, 1.5, ..., 5.5, 6\}$

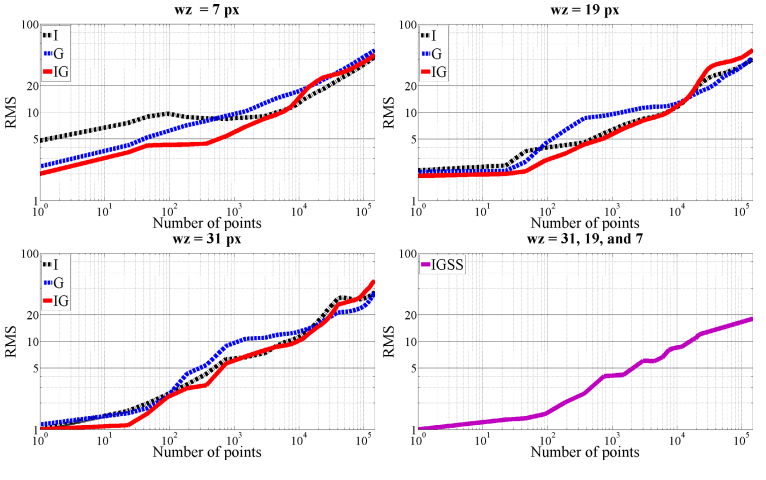

Fig. 5. Average accumulated RMS disparity errors sorted by window size

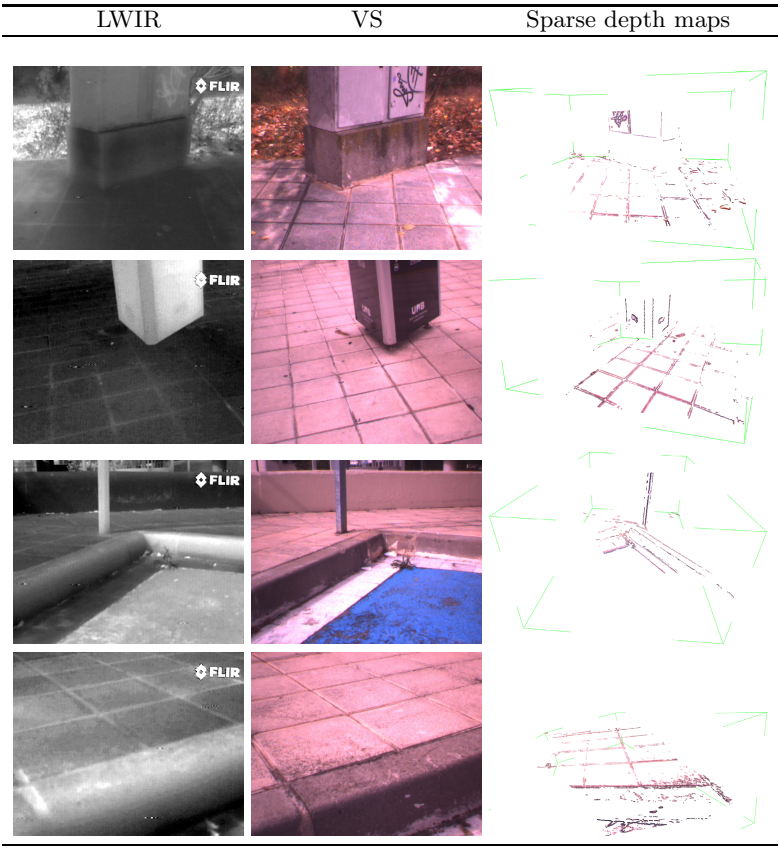

Fig. 6. Example of 3D result

and $Q = \{8, 16, 32, 64\}$. The σ parameter is the standard deviation of Gaussian derivative kernel used to obtain the t levels of the scale space representation.

Figure 5 shows the average accumulated error of **I**, **G**, **IG**, and **IGSS** for the evaluation dataset. The results are sorted into groups by windows sizes: {7}, {19}, {31}, and {31, 19, 7} pixels respectively. These plots depict the relationship between the cost values, obtained by a similarity function, and the error made by assuming that the maximum argument corresponds to the correct match (Sec. 2.5).

Every point on the plots presented in Fig. 5 represents the error of a set of matches, with a variable number of elements. The cost values are arranged in descending order, and the error is computed from the first matched couple, which has the maximum cost value, till the number of points indicated in the horizontal axis. For this reason, the range of the curve goes from 0 till the number of pixels in the images. This help us to visualize how the error increases as soon as more matches are accepted.

In all cases the combined use of **I** and **G** increases the number of good matches, proving that mutual and gradient information supply complementary information useful in matching process. Figure 5(*right*) shows how **IGSS** improves the previous similarity functions. A stable behavior on both, in error and number of correct matches, is noticeable. The best combination of parameters is $wz = \{31, 19, 7\}$, $Q = \{32, 16, 8\}$, and $\sigma = \{1.5, 1, 0.5\}$ (the values were arranged from coarse to fine).

The λ parameter in equation (4) weights the confidence of a **IG** cost, from a scale t, in comparison to another computed at previous scale. After an exhaustive search over the range spanned by λ parameter, it has been found that three scales are sufficient, and $\lambda = \{0.6, 0.5\}$ is the best combination for the **IGSS** similarity function (again, the values were arranged from coarse to fine).

Indirectly, Fig. 5 also depicts the relationship between error and τ parameter (Sec. 2.5). So, insofar as more pixels are matched more mistakes are done. The selection of the parameter τ depends on the application and the depth of the scene. For instance, for outdoor images with a depth between 15 and 20 meters, a τ corresponding to 45% of the pixels in the image is enough to get 3D representations as those shown in Fig. 6, giving an average error of about 5% (in the evaluation dataset).

Figure 6 shows the results obtained with our evaluation framework when **IGSS** is used. First and second columns correspond to the rectified images, thermal infrared and visible images respectively. Third column depicts the 3D sparse depth maps obtained.

4 Conclusions

This paper presents an evaluation framework that compares the performance of four multimodal similarity functions, which are evaluated with two metrics: RMS error and number of correct matches. The evaluation is performed in v-disparity space, which allows to consider the noise in the acquisition of the disparity maps. We have shown that adding gradient information (**G**) together with a scale-space analysis improves descriptive ability of a similarity function based on mutual information. Also, details of the parameter setting are presented (Q, wz, σ, τ, and λ) and how they affect the performance. Additionally, the different stages for obtaining sparse depth maps are described, from image acquisition till depth map computation. Finally, the results obtained from real environments show that **IGSS** is useful to find correspondence in multimodal images, in order to generate sparse 3D representations. Future work will be focused on improving the disparity selection process in order to obtain dense 3D representations.

Acknowledgements. This work was supported by the Spanish Government under Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and Projects TIN2011-25606 and TIN2011-29494-C03-02.

References

1. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in computational stereo. IEEE Trans. Pattern Anal. Mach. Intell. 25, 993–1008 (2003)
2. Hirschmuller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans. Pattern Anal. Mach. Intell. 31(9), 1582–1599 (2009)
3. Egnal, G.: Mutual information as a stereo correspondence measure. Technical report, University of Pennsylvania (2000)
4. Krotosky, S.J., Trivedi, M.M.: Mutual information based registration of multimodal stereo videos for person tracking. Computer Vision and Image Understanding 106, 270–287 (2007)
5. Torabi, A., Najafianrazavi, M., Bilodeau, G.A.: A comparative evaluation of multimodal dense stereo correspondence measures. In: IEEE International Symposium on Robotic and Sensors Environments, pp. 143 –148 (2011)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886 –893 (2005)
7. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
8. Pluim, J.P., Maintz, J.B., Viergever, M.A.: Image registration by maximization of combined mutual information and gradient information. IEEE Trans. on Medical Imaging 19(8), 809–814 (2000)
9. Barrera, F., Lumbreras, F., Sappa, A.: Multimodal template matching based on gradient and mutual information using scale-space. In: IEEE International Conference on Image Processing, pp. 2749–2752 (2010)
10. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley Interscience, New York (1991)
11. Morris, N., Avidan, S., Matusik, W., Pfister, H.: Statistics of infrared images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)
12. Bouguet, J.Y.: Camera calibration toolbox for matlab (2010), `http://www.vision.caltech.edu/bouguetj`
13. Fusiello, A., Trucco, E., Verri, A.: A compact algorithm for rectification of stereo pairs. Machine Vision and Applications 12(1), 16–22 (2000)
14. Labayrade, R., Aubert, D.: A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. In: IEEE Intelligent Vehicles Symposium, pp. 31–36 (2003)