COMBINATION OF CORRELATION MEASURES FOR DENSE STEREO MATCHING

Sylvie CHAMBON
Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR), France
chambon@ifsttar.fr

Alain CROUZIL
Institut de Recherche en Informatique de Toulouse (IRIT), France
crouzil@irit.fr

Keywords: Stereovision, matching, correlation, classic measures, robust statistics, fusion.

Abstract: In the context of dense stereo matching of pixels, we study the combination of different correlation measures. Considering the previous work about correlation measures, we use some measures that are the most significant in five kinds of measures based on: cross-correlation, classic statistics, image derivatives, non-parametric statistics and robust statistics. More precisely, this study validates the possible improvement of stereo-matching by combining complementary correlation measures and it also highlights the two measures that can be combined in order to take advantage of the different methods: Gradient Correlation measure (GC) and Smooth Median Absolute Deviation measure (SMAD). Finally, we introduce an algorithm of fusion that allows to combine automatically correlation measures.

1 INTRODUCTION

Finding homologous pixels in a stereo pair of images is one of the most important step in order to recover the 3D structure of a scene by stereovision. Many methods have been proposed in the literature where local methods are distinguished from global ones. More precisely, matching methods can be described with essential components, this term has been firstly introduced by (Scharstein and Szeliski, 2002). These components are: the matching cost, the optimization method, the introduction of multiple passes, i.e. to improve the matching performances, some approaches are based on several methods applied in sequence. This description leads to a four type categorization: local methods, global ones (without correlation measure), mixed method (global ones with a correlation measure) and the methods with multiple passes. Our purpose is to introduce a multipass algorithm based on combination of local methods.

Local methods are easy to implement, low time consuming, quite efficient, and consequently intensively used. Unfortunately, characterising how the existing correlation measures are effective, i.e. obtaining correct matching in different areas of the image, is still an open issue. In our work (Chambon and Crouzil, 2011) on local costs, the influence of different measures on the quality of stereo matching results have been studied, in particular, near occluded regions and, in (Chambon and Crouzil, 2004), we demonstrated that a measure based on a robust statistics tool combined with a cross correlation measure allows to obtain better performances than using a correlation measure alone. These results raise up three new questions:

- Which are the correlation measures that are the most complementary to cover all the matching difficulties?
- Is it advantageous to combine numerous measures and how many?
- Following up the previous questions, can we propose an algorithm that combines more than one measure to obtain a whole dense and correct matching (or disparity) map[1] and is it more efficient than the method based on a sole measure?

[1] A disparity (or matching) map represents for each pixel, the distance between the pixel and its correspondent. When the disparity map is represented by a grey level image, the clearer the pixel, the larger the distance is. Black pixels are occluded pixels.

Existing methods are briefly presented before the description of the data set used for validating the proposed method. Then, combination study is described leading to the proposal for matching algorithm based on merging the results obtained from various correlation measures. Finally, results are presented.

2 CORRELATION MEASURES

The principle of a local cost, i.e. a correlation measure, is to consider that two homologous pixels and their respective neighborhoods, are similar, from a photometric point of view. The main difficulties of these methods are: illumination changes, untextured areas and occlusions. Many measures have been introduced to tackle out these difficulties. Based on the results of 41 measures on a benchmark of 42 images, presented in (Chambon and Crouzil, 2011), we propose to study the complementarity of these measures, and, in particular, the best measures of each families.

Table 1: Notations used for the description of the measures.

I_w	The images with $w \in \{l,r\}$ (left and right).
$I_w^{i,j}$, $\mathbf{p}_w^{i,j}$	The grey level of the pixel $\mathbf{p}_w^{i,j}$ of coordinates (i,j) in image I_w is $I_w^{i,j}$. Moreover, $\mathbf{p}_r^{x,y}$ is the correspondant pixel of $\mathbf{p}_l^{i,j}$.
N_*	The number of pixels in the neighborhood: $N_f = (2N_v+1)\times(2N_h+1)$, $N_v, N_h \in \mathbb{N}^*$.
$\mathbf{f}_w$	The vector of grey levels of pixels in the correlation windows (in I_w): $\mathbf{f}_w = (\cdots I_w^{i+p,j+q} \cdots)^T = (\cdots f_w^k \cdots)^T$ where ^{T} is the matrix transposition operator and $p \in [-N_v;N_v]$, $q \in [-N_h;N_h]$.
$\overline{\mathbf{f}_w}$	The mean of the grey levels in $\mathbf{f}_w$.
f_w^k	The element k of vector $\mathbf{f}_w$.
L_P	The L_P norms: $\|\mathbf{f}_w\|_P = (\sum_{k=0}^{N_f-1} \|f_w^k\|^P)^{1/P}$ with $P \in \mathbb{N}^*$ and $\|\mathbf{f}_w\| = \|\mathbf{f}_w\|_2$.

In the following description, when no explicit reference is given, the reader should consult (Aschwanden and Guggenbül, 1992). We briefly present the notations in Table 1 and the five best measures of the different families that are considered.

(1) **Family 1:** *Cross correlation-based measures* – All these measures are based on a scalar product (Moravec, 1980) and NCC (Normalized Cross Correlation) is the most efficient one:

$$\mathrm{NCC}(\mathbf{f}_l,\mathbf{f}_r) = \frac{\mathbf{f}_l \cdot \mathbf{f}_r}{\|\mathbf{f}_l\|\|\mathbf{f}_r\|}. \quad (1)$$

(2) **Family 2:** *Classical statistics-based measures* – These types of measures can be used: measures based on a distance or/and that are locally centered, variance-based or fourth-order cumulant-based measures (Rziza and Aboutajdine, 2001). The best one is the LSAD (Locally scaled Sum of Absolute Differences) defined by:

$$\mathrm{LSAD}(\mathbf{f}_l,\mathbf{f}_r) = \|\mathbf{f}_l - \frac{\overline{\mathbf{f}_l}}{\overline{\mathbf{f}_r}}\mathbf{f}_r\|_1. \quad (2)$$

(3) **Family 3:** *Derivatives-based measures* – Instead of using grey levels, these measures employ the derivatives of the images at different orders (Seitz, 1989). Most of the existing measures use only the direction of the gradient vectors (Ullah et al., 2001), but, this kind of information can induce errors, in particular, with low norm gradient vectors whose direction is not reliable. In consequence, the most performant measure is based on the similarity of the image gradient vectors, GC (Gradient Correlation) (Crouzil et al., 1996). If the gradient vector at $\mathbf{p}_w^{i,j}$ in I_w is $\nabla I_w^{i,j}$ and the norm is denoted by $\|\nabla I_w^{i,j}\|$, the definition of GC is:

$$\mathrm{GC}(\mathbf{f}_l,\mathbf{f}_r) = \frac{\sum_A \|\nabla I_l^{i+p,j+q} - \nabla I_r^{v+p,w+q}\|}{\sum_A (\|\nabla I_l^{i+p,j+q}\| + \|\nabla I_r^{v+p,w+q}\|)}, \quad (3)$$

with $\sum_A = \sum_{p=-N_v}^{N_v} \sum_{q=-N_h}^{N_h}$.

(4) **Family 4:** *Non-parametric statistics-based measures* – They are based on the order of the grey levels inside the correlation window (Kaneko et al., 2002; Bhat and Nayar, 1998). Using the order of the grey levels allows these measures to be robust against noises and occlusions but, sometimes, it also gives an ambiguous result, i.e. the best correlation score is obtained for the wrong correspondant. The most performant measure of this family is a non-parametric one, CENSUS (Zabih and Woodfill, 1994). The similarity measure uses a transform that produces a bit chain which represents the pixels with an intensity lower than the central pixel:

$$\mathrm{R}_\tau(\mathbf{f}_w) = \bigotimes_{k\in[0;N_f-1]} \xi(f_w^{N_f/2}, f_w^k),$$

where $\xi(f_w^{N_f/2}, f_w^k) = 1$ if $f_w^k < f_w^{N_f/2}$ and 0 elsewhere. CENSUS is the sum of the Hamming distances, denoted by D_H, between the codes of each pixel of the correlation window:

$$\mathrm{CENSUS}(\mathbf{f}_l,\mathbf{f}_r) = \sum_{k=0}^{N_f-1} \mathrm{D}_H(\mathrm{R}_\tau(\mathbf{f}_l), \mathrm{R}_\tau(\mathbf{f}_r)). \quad (4)$$

(5) **Family 5:** *Robust measures* – We are particularly concerned with the occlusion problem which appears in the vicinity of a pixel near a depth discontinuity. In fact, some pixels lie on a first level of depth whereas the other pixels lie on a second level. It can disturb the matching process and introduce erroneous matches. To take this problem into account, robust statistics tools are introduced as correlation measures, like partial correlation (Lan and Mohr, 1997) or pseudo-norms (Delon and Rougé, 2004). The most efficient is SMAD, the Smooth Median Absolute Deviation (Rousseeuw and Croux, 1992):

$$\mathrm{SMAD}(\mathbf{f}_l,\mathbf{f}_r)=\sum_{k=0}^{h-1}(\mathbf{f}_l-\mathbf{f}_r-\mathrm{med}(\mathbf{f}_l-\mathbf{f}_r))^2_{k:N_f-1}, \quad (5)$$

where the ordered values of $\mathbf{f}_w$ are represented by: $(f_w)_{0:N_f-1}\leq\ldots\leq(f_w)_{N_f-1:N_f-1}$. It can be interpreted as a robust centered (median) and troncated distance and, in our experiments, $h=\frac{N_f}{2}$.

Robust and non-parametric measures (families 4 and 5) are efficient in the presence of noises and/or occlusions whereas the classic ones (families 1 and 2) obtain better results when there is no major problems. The derivatives measures have been designed to be more efficient in the presence of noises, but, most of the time, they are really less efficient than the other ones, except GC which seems to have better results than the others, in particular in low textured areas. Interested readers can find more details about all the measures in (Chambon and Crouzil, 2011).

3 EVALUATION PROTOCOL

To validate our approach, 42 images, with their ground truth or reference disparity maps, have been tested (see Figure 1 for examples): 1 random-dot stereogram, 2 synthetic pairs (*Murs*) and one real image pair [2], and, finally, 38 real pairs introduced by Scharstein and Szeliski (9 in 2002 (*Tsukuba*) (Scharstein and Szeliski, 2002), 2 in 2003 (*Cones*) (Scharstein and Szeliski, 2003), 6 in 2005 and 22 in 2006 (*Aloe*)). The last ones are the most complex scenes. The most consequent evaluation protocol to highlight the different performances of global methods is given by the authors of (Scharstein and Szeliski, 2002)[3]. Compared to their protocol, our comparison is based on all their 38 images instead of 4.

[2] http://www.irit.fr/~Benoit.Bocquillon/MYCVR/research.php

[3] http://vision.middlebury.edu/stereo/eval/

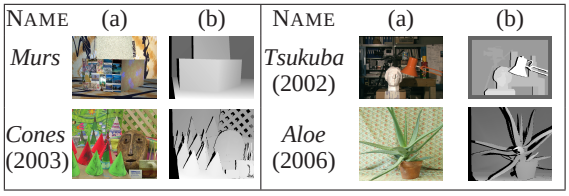

Figure 1: Examples of data used in our tests (left images (a) and disparities[1] (b)). Interested readers can find more explanations about the estimation of these reference maps (ground truth), both in the cited papers and in the cited web page of section 3 (active vision is used and/or some constraints about the geometry of the scene are introduced).

Many criteria can be used to evaluate the quality of the results based on ground truth (Chambon and Crouzil, 2004). However, for this evaluation, we use the percentage of erroneous matches, noted ER, and the evaluation of the complementarity of the results (also based on ER) because they are the two most important aspects to consider in order to evaluate the impact of the proposed fusion algorithm.

4 COMPLEMENTARITY STUDY

To evaluate the complementarity of similarity measures, we analyse the percentage of erroneous matches (ER) for each measure used alone, and for each combination, by supposing that the correct correspondent is always kept (when one of the measures that are combined finds the exact correspondent), see Table 2 for the combination of 2 measures and Figure 2 for the percentage of erroneous matches with more than 2 measures. We use these notations:

- M_i, with $i\in\{1;\ldots;N_m\}$, the N_m tested measures;
- $d_{th}(\mathbf{p}_l)$ the disparity of the pixel $\mathbf{p}_l$ given by the ground truth;
- $d_i(\mathbf{p}_l)$, the disparity given by the algorithm based on the correlation measure M_i;
- $d_{tc}^{N_m}(\mathbf{p}_l)$ the theoretical or optimal combination of N_m measures.

More formally, the optimal combinations of the results over N_m measures ($N_m\in\{2;\ldots;41\}$ because in our previous work 41 measures have been studied), denoted $d_{tc}^{N_m}$, is simply estimated by following this rule, for each pixel $\mathbf{p}_l$:

if $\exists\, i\in\{1;\ldots;N_m\}$ where $d_i(\mathbf{p}_l)=d_{th}(\mathbf{p}_l)$
then $d_{tc}^{N_m}=d_{th}(\mathbf{p}_l)$ and $d_{tc}^{N_m}$ is correct
else $d_{tc}^{N_m}$ is erroneous.

We have tested these configurations:

(C_1) $\mathbf{N_m=2}$: All the 41×41 combinations have been evaluated and it highlights the best combination:

GC and SMAD with only 14.13% for the mean percentage ER on the 42 images.

(C_2) $\mathbf{N_m = 41}$: All the 41 measures have been theoretically combined and the results show that the percentage ER can be decreased to 7.26%.

(C_3) $\mathbf{N_m} \in \{\mathbf{3}; \ldots; \mathbf{40}\}$: When we used the best combination GC-SMAD, any kind of measures can be added, the performances are quite equivalent. With 3 measures, the percentage ER decreases to about 13% and, then, it goes slowly to the minimum percentage ER (about 1% for each added measure) reached by the optimal combination of 41 measures. Moreover, when more than 10 measures are used, ER is close to this minimum.

First, the results show how the local matching with one correlation measure can be theoretically improved, and, second, which measures are the most complementary. In Table 2, we can remark that combining different measures can highly improve the results: on the whole image, from 7% of improvement (2 measures combined) to 17% (41 measures).

Table 2: Percentages of erroneous matches (ER) with each of the 5 best correlation measures and the best combinations of 2 correlation measures.

MEASURE	ER	MEASURE	ER
NCC	23.2	LSAD	23.3
GC	21	CENSUS	**20.2**
SMAD	27.9	GC-SMAD	**14.13**

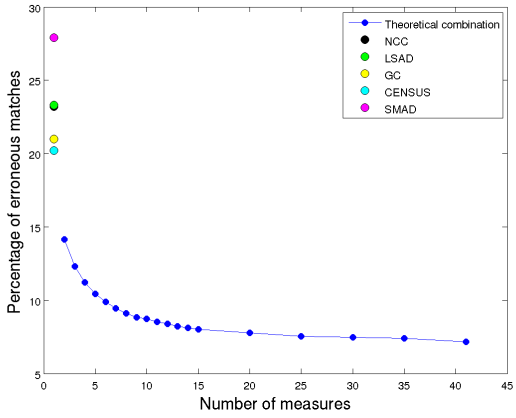

Figure 2: Percentage of erroneous matches versus the number of correlation measures theoretically combined. This graph illustrates the maximal number of measures that are interesting to combine (10) but it also highlights the biggest improvement obtained with only 2 measures.

The following analyses illustrate these results. Using four maps, see Figure 3, we propose to visualize:

(1) *The comparison of the two most complementary measures* – As expected by the definitions of these measures, this visualization illustrates that SMAD compensates for the weaknesses of GC in occlusion areas (or near occlusion areas) whereas GC compensates for the weaknesses of SMAD in non-occluded areas and, in particular, areas that are low textured, see (a) in Figure 3.

(2) *The areas with 1 correct correspondent over 5, 10 or 41 correspondents* – The most distinctive measure is GC, i.e. it is the most complementary measure to the other measures. SMAD is the second most complementary measure to the others, see (b) for 5, (c) for 10 and (d) for 41 in Figure 3. Moreover, with 10 different correlation measures, the results are quite near the results with 41 measures. And, our last conclusion is that combining more than 2 measures seems to be interesting because, most of the time, more than one measure obtains the correct correspondent. This last remark has inspired the fusion algorithm that is described in the next section.

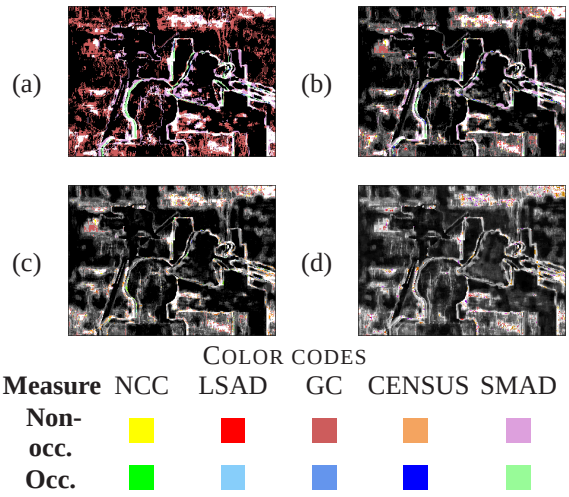

Figure 3: Study of complementarity (an example with image of Figure 1) – The pixels in grey levels correspond to pixels with more than 1 correct match over the N_m combinations. The darker the pixel, the higher the number of correct matches. The color codes are used when only one measure gives the correct match. Moreover, we discern the occluded areas (**Occ.**) from the non-occluded areas (**Non-occ.**). In (a), it shows that SMAD is efficient near occlusions whereas GC is more efficient than SMAD in low textured areas.

In conclusion, we have decided to present an algorithm that combines N_m different measures and we illustrate the interest of this kind of algorithm by using the two complementary measures: GC and SMAD.

5 ALGORITHM OF FUSION

In our first proposition of combination (Chambon and Crouzil, 2004), the algorithm was designed to take into account occlusions. In consequence, as expected, the results are good in non-occluded areas and also in occluded areas. The goal of this new algorithm is to improve this work by combining the advantages of each measure according to each kind of regions in order to take into account more difficulties, like low textured or noisy regions. Towards this goal, instead of detecting the occlusions, we work directly on how to merge the disparities by taking into account their variations in several matching maps (each map has been obtained with a different correlation measure).

Our method of fusion is based on two steps, the principle being to estimate a disparity map with each measure and then to merge the results applying the following two rules:

(1) If more than one disparity map give the same match, the correspondence is validated and this result is considered as reliable.

(2) In an "undetermined area" (i.e. rule (1) is not respected), the "most reliable" disparity is kept. The difficulty is to determine the most reliable. In this paper, we consider the disparities found in the neighborhood in the matching map of each considered measure.

Formally, these two rules can be defined as:

(1) *Initialization for each pixel* $\mathbf{p}_l$ – The term $d_f^{N_m}$ is the final disparity, after the fusion of N_m correlation measures.

If $\exists\, d \mid \left((d = \underset{e}{\operatorname{argmax}} \mathrm{M}_e(\mathbf{p}_l) \;\&\&\; (d \geq \frac{N_m}{2}) \right)$

then $d_f^{N_m}(\mathbf{p}_l) \leftarrow d$

else the disparity is undetermined.

We define :

$\mathrm{M}_e(\mathbf{p}_l) = \#\{i | d_i(\mathbf{p}_l) == e\}$.

(2) *Refinement* – For each pixel $\mathbf{p}_l$ without disparity, we estimate the ambiguity, denoted by A, of each possible disparity $d_i(\mathbf{p}_l)$. For the estimation of the function A, which represents how much the estimated disparity is reliable, we suppose that if most of the neighbors have the same disparity (in the same result obtained with the same correlation measure) the estimated disparity can be considered as sure. In consequence, for estimating A, we compare the studied disparity with the mean of the disparities in the neighborhood, denoted by $\mathcal{N}$. The disparity with the lowest ambiguity is kept only if this ambiguity is not important, i.e. higher than a given threshold ε.

For (each pixel $\mathbf{p}_l$)

$d = \underset{i \in \{1, N_m\}}{\operatorname{argmin}} \mathrm{A}(d_i(\mathbf{p}_l))$ with

$$\mathrm{A}(d_i(\mathbf{p}_l)) = |d_i(\mathbf{p}_l) - \frac{1}{\#\mathcal{N}(\mathbf{p}_l)} \sum_{k \in \mathcal{N}(\mathbf{p}_l)} d_i(\mathbf{p}_k)|$$

with $\mathcal{N}(\mathbf{p}_l)$ the neighborhood of $\mathbf{p}_l$[4].

If $(d < \varepsilon)$[5]

then $d_f^{N_m}(\mathbf{p}_l) \leftarrow d$

else $\mathbf{p}_l$ is occluded.

6 Matching results

For this part, the fusion algorithm has been tested with the fusion of the two most complementary measures: GC and SMAD. In order to try to detect occlusions and erroneous matches, we use the symmetry constraint that consists in estimating correspondences from the left image to the right image and then from the right to the left and in considering non-coherent matches as occluded pixels (these occluded pixels are shown in black in each disparity map). Table 3 shows the improvements of the percentage of erroneous matches obtained with the new algorithm of fusion on all the 42 tested images. The decreasing of this percentage is from 2.47 to 4.08 (with complex images), i.e. the images difficult to match because of the occlusion areas or the untextured areas. However, this improvement did not reach the theoretically maximal improvement that is showed in Table 2. Another way to appreciate the quality of the results is to look at the disparity maps that are given in Figure 4. The disparity maps obtained by fusion are the best ones because they contain less false negatives than the others. Moreover, the occlusion areas are better delimited (the contours are clean and contain no "holes").

As in the first step we have to estimate each disparity map induced by each measure, the execution time is the sum of the execution time of each correlation-based algorithm. The fusion algorithm does not take much time in comparison to the second step. In consequence, the higher the number of merged results, the higher the execution time and the execution time depends on the chosen measures. In our test, for example with *Tsukuba*, GC takes 17.6 *s* and SMAD 39.77 *s*, so finally, the fusion algorithm takes about 1 minute.

[4]The 8 neighbors have been taken into account.

[5]We have chosen $\varepsilon = 1$.

Table 3: Precentage of erroneous matches – **H** represents the images with untextured areas, like *Tsukuba* pair, **O**, the images with a lot of occlusions, like *Aloe* pair and **R**, the images with no major difficulties, like *Cones* pair (see Figure 4 for these images). The term Tc refers to the results obtained with a theoretical or optimal fusion, see Table 2. The percentage of erroneous matches with the new method is better than those obtained with the GC measure alone and in particular with complex scenes.

METHOD	**H+O**	**O**	**H**	**R**	Total
GC alone	25.6	17.5	19.6	15.9	20.9
Fusion	22.1	13.5	16.9	13.5	17.5
Tc	19.4	10.8	15.4	10.8	15.3

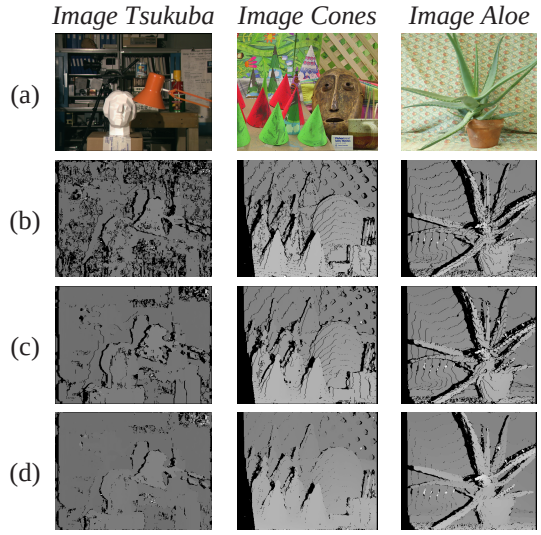

Figure 4: Disparity maps – (a), left image, (b) disparity map with SMAD, (c), with GC, (d), with FUSION. The fusion results present less false negatives, in particular for *Cones* and *Aloe*. The example of *Tsukuba* illustrates the limits of the method and the need to combine more than 2 measures.

7 CONCLUSION

In this paper, we proposed a study of the complementarity of correlation measures, illustrated with visualization maps, and we introduced a new way to combine complementary measures. Moreover, we highlight the most complementary measures: GC and SMAD. The tests on 42 images illustrate the improvement of performances of the new fusion algorithm compared to classic correlation matching, i.e. based on one correlation measure alone. These results are encouraging but also exhibit the limit of this approach that might lead to investigate the fusion approach based on a voting method in the neighborhood of the studied pixel or to distinguish the most reliable measures (in the first step of the algorithm). Moreover, we will study the influence of the number of measures involved in the proposed algorithm.

REFERENCES

Aschwanden, P. and Guggenbül, W. (1992). Experimental results from a comparative study on correlation type registration algorithms. In Förstner, W. and Ruwiedel, S., editors, *Robust computer vision: Quality of Vision Algorithms*, pages 268–282. Wichmann.

Bhat, D. and Nayar, S. (1998). Ordinal measures for image correspondence. *PAMI*, 20(4):415–423.

Chambon, S. and Crouzil, A. (2004). Towards correlation-based matching algorithms that are robust near occlusions. In *ICPR*, volume 3, pages 20–23.

Chambon, S. and Crouzil, A. (2011). Occlusions handling in dense stereo matching. *Pattern Recognition*. submitted.

Crouzil, A., Massip-Pailhes, L., and Castan, S. (1996). A new correlation criterion based on gradient fields similarity. In *ICPR*, volume 1, pages 632–636.

Delon, J. and Rougé, B. (2004). Analytic study of the stereoscopic correlation. Research report 2004-19, CMLA (ENS Cachan).

Kaneko, S., Murase, I., and Igarashi, S. (2002). Robust image registration by increment sign correlation. *Pattern Recognition*, 35(10):2223–2234.

Lan, Z. and Mohr, R. (1997). Robust location based partial correlation. Technical Report RR-3186, INRIA, France.

Moravec, H. (1980). *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. Phd thesis, Carnegie Mellon University.

Rousseeuw, P. and Croux, C. (1992). L_1-statistical analysis and related methods. In Dodge, Y., editor, *Explicit Scale Estimators with High Breakdown Point*, pages 77–92. Elsevier.

Rziza, M. and Aboutajdine, D. (2001). Dense disparity map estimation using cumulants. In *Conference on Telecommunications, ConfTele*.

Scharstein, D. and Szeliski, R. (2002). A taxomomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42.

Scharstein, D. and Szeliski, R. (2003). High-Accuracy Stereo Depth Maps Using Structured Light. In *CVPR*, volume 1, pages 195–202.

Seitz, P. (1989). Using local orientational information as image primitive for robust object recognition. In *Visual Communication and Image Processing IV*, volume SPIE–1199, pages 1630–1639.

Ullah, F., Kaneko, S., and Igarashi, S. (2001). Orientation Code Matching For Robust Object Search. *IEICE Transactions on Information and Systems*, E-84-D(8):999–1006.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV*, pages 151–158.