AN EVALUATION METHODOLOGY FOR STEREO CORRESPONDENCE ALGORITHMS

Ivan Cabezas[1], Maria Trujillo[1] and Margaret Florian[2]
[1]*School of Systems Engineering and Computer Sciences, Universidad del Valle, Ciudadela Universitaria, Cali, Colombia*
[2]*Department of Systems Delivery, Ayax Systems, Cali, Colombia*

Keywords: Computer Vision, Disparity Estimation, Error Measures, Quantitative Evaluation Methodologies, Stereo Correspondence.

Abstract: A comparison of stereo correspondence algorithms can be conducted by a quantitative evaluation of disparity maps. Among the existing evaluation methodologies, the Middlebury's methodology is commonly used. However, the Middlebury's methodology has shortcomings in the evaluation model and the error measure. These shortcomings may bias the evaluation results, and make a fair judgment about algorithms accuracy difficult. An alternative, the A* methodology is based on a multiobjective optimisation model that only provides a subset of algorithms with comparable accuracy. In this paper, a quantitative evaluation of disparity maps is proposed. It performs an exhaustive assessment of the entire set of algorithms. As innovative aspect, evaluation results are shown and analysed as disjoint groups of stereo correspondence algorithms with comparable accuracy. This innovation is obtained by a partitioning and grouping algorithm. On the other hand, the used error measure offers advantages over the error measure used in the Middlebury's methodology. The experimental validation is based on the Middlebury's test-bed and algorithms repository. The obtained results show seven groups with different accuracies. Moreover, the top-ranked stereo correspondence algorithms by the Middlebury's methodology are not necessarily the most accurate in the proposed methodology.

1 INTRODUCTION

The research and development process on stereo correspondence algorithms requires of an objective assessment of results. In fact, an evaluation methodology should be followed in order to perform a fair comparison among different algorithms (Szeliski, 1999), tune the parameters of an algorithm within a particular context (Kelly et al., 2007), identify algorithm's advantages and drawbacks (Kostlivá et al., 2007), determine the impact of specific procedures and components (Hirschmüller and Scharstein, 2009; Bleyer and Chambon, 2010), support decision for researchers and practitioners (Cabezas and Trujillo, 2011), and, in general, to measure the progress on the field (Szeliski and Zabih, 2000). Among the quantitative evaluation methodologies available in the literature, the Middlebury's methodology (Scharstein and Szeliski, 2002; 2011) is widely used. This methodology is based on a test-bed composed by four images of different geometric characteristics. Three different error criteria are defined in relation to challenging image regions. The percentage of Bad Matched Pixels (BMP) is used as error measure. It is gathered according to error criteria by comparing the estimated disparity maps against ground-truth data, using an error threshold. A rank is assigned to algorithms under evaluation, according to error scores and error criteria. A final ranking is computed by averaging previously established ranks. In this way, the evaluation model of Middlebury's methodology relates ranks to weights. The model assumes that the algorithm of minimum weight has the best accuracy. However, the Middlebury's methodology has some shortcomings, such as the use of the BMP measure along with the evaluation model. The BMP measure is a binary function using a threshold, where the treshold selection may impact on the evaluation results. Once the error in estimation exceeds the threshold, the error magnitude is ignored. Moreover, the same magnitude errors may cause depth reconstruction errors of different magnitude (Van der Mark and

Cabezas I., Trujillo M. and Florian M..
AN EVALUATION METHODOLOGY FOR STEREO CORRESPONDENCE ALGORITHMS.
DOI: 10.5220/0003850801540163
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 154-163
ISBN: 978-989-8565-04-4

Gavrila, 2006; Gallup et al., 2008; Cabezas et al., 2011). Nevertheless, the BMP measure does not consider this fact. Consequently, the BMP measure may not be suited to properly evaluate the accuracy of a disparity map (Cabezas et al., 2011). In regard to the evaluation model, two or more algorithms may have the same error score using an error criterion. In this case, the rank assigned to these algorithms became arbitrary. This fact may impact on the final ranking. Adittionally, different algorithms may have the same average ranking. Nevertheless, it does not mean that these algorithms perform similarly on test-bed images. Moreover, althought it is possible to determine a set of top ranking algorithms based on the Middlebury's methodology, the cardinality of this set is a free parameter. This fact may lead to discrepancies or controversy among researchers about the *state-of-the-art* on the field. Thus, the above shortcommings may introduce bias to evaluation results, as well as they may impact the interpretation on the *state-of-the-art* of stereo correspondence algorithms in the research comunity.

In the $\mathbf{A}^*$ evaluation model proposed in (Cabezas and Trujillo, 2011), the composition of the most accurate set of algorithms is determined withouth ambiguity. However, this evaluation model fails in considering an evaluation scenario on which a user may be interested in an exhaustive evlauation of the entire set of algorithms, and not only in determining which algorithms are the most acccurate overall.

In this paper, an extension to the $\mathbf{A}^*$ evaluation model presented in (Cabezas and Trujillo, 2011) is introduced. The extension is based on iteratively evaluating the entire set of stereo correspondence algorithms by computing groups of algorithms with comparable accuracy. The composition of each group is unambiguously determined based on error scores and the Pareto Dominance relation (Van Veldhuizen et al., 2003). Additionally, the error measure proposed in (Cabezas et al., 2011) is used in the presented evaluation. The obtained evaluation results show that the extended methodology, in conjuction with the used measure, allow a better understanding and analysis of the accuracy of stereo correspondence algorithms. The imagery test-bed, and the error criteria of Middlebury's methodology, as well as a set of 112 stereo correspondence algorithms (Scharstein and Szeliski, 2011), were used in the experimental validation. As evaluation results, seven different groups of algorithms were obtained. Each group is associated to a different accuracy. In particular, the most acurate group is composed by nine algorithms. Among these algorithms, two of them apply local optimisation strategies, whilst the seven remains apply global optimisation strategies. Most of these global algorithms are based on Graph Cuts (GC) (Kolmogorov and Zabih, 2001).

2 RELATED WORK

A quantitative comparison of stereo correspondence algorithms is presented in (Szeliski and Zabih, 2000). The evaluation considers both, the comparison against ground-truth data, using the BMP measure, and the prediction-error approach of (Szeliski, 1999). In this work, these two approaches were applied separately over data, and it is concluded that consistent results were obtained between the two approaches, whilst certain types of errors are emphasised by each approach.

Accuracy and density are considered as the two main properties of a disparity map in (Kostlivá et al., 2007). Two errors measures are defined based on these properties: the error rate (i.e. mismatches and false positives) and the sparsity rate (i.e. false negatives). A Receiver Operating Characteristics (ROC) analysis is adopted upon the defined errors. An *is better* relation is defined based on the ROC curve. A particular algorithm's parameter setting *is better* than another if it produces a more accurate and denser result. Nevertheless, the ROC curve can be computed using only one set of stereo images. Thus, the evaluation turns probabilistic when the imagery test-bed involves more stereo images. Additionally, the evaluation model of this methodology requires a weight setting in relation to the importance of each stereo image considered in the test-bed. Moreover, error rate is computed using a threshold. Thus, it may have the same drawbacks than the BMP measure.

A cluster ranking evaluation method is proposed in (Neilson and Yang, 2008). It consists on using statistical inference techniques to rank the accuracy of stereo correspondence algorithms over a single stereo image, and the posterior combination of rankings from multiple stereo images, to obtain a final ranking. This work is focused on comparing matching costs using a hierarchical Belief Propagation (BP) algorithm (Felzenszwalb and Huttenlocher, 2004). Additionally, different significance tests should be applied when the test-bed involves several stereo images. Moreover, a greedy clustering algorithm is used. The clustering algorithm computes iteratively the final ranks as the average of several ranks in a partition. In this way, the assigned rank may be a real number which lacks of a concise interpretation. In this work, the BMP measure is used in order to determine estimation errors.

Stereo correspondence algorithms of real-time performance and limited requirements in terms of memory are evaluated in (Tombari et al., 2010). The evaluation involves both: accuracy and computational efficiency. The complement of the BMP measure is used as the accuracy measure, whilst the amount of estimated disparities per second is used for measuring computational efficiency. However, the fact that estimation accuracy and computational efficiency are opposite goals is not taken into account. Moreover, the evaluation model of Middlebury's methodology is used in this work. Thus, the averaged rankings values may lack of a concise meaning.

The Middlebury's methodology was introduced in (Scharstein and Szeliski, 2002). Additionally, a website with an online ranking is kept updated in (Scharstein and Szeliski, 2011). This evaluation methodology uses a ground-truth imagery test-bed of four stereo images: Tsukuba, Venus, Teddy and Cones, (Scharstein and Szeliski, 2003). Three error criteria are defined in relation to image regions: *nonocc*, *all* and *disc*. The *nonocc* criterion is associated to image points in non-occluded regions. The *all* criterion involves points in the whole image. The *disc* criterion is associated to image points in discontinuity regions or neighbourhoods of depth boundaries. The BMP measure is gathered on these image regions. A threshold δ is defined by the user. A threshold value equal to 1 pixel is commonly used. Nevertheless, the BMP measures the quantity of errors. It may conceal estimation errors of a large magnitude, and, at the same time, it may penalise errors of small impact in the final 3D reconstruction. On the other hand, the evaluation model of the Middlebury's methodology can be seen as a linear combination of ranks, where a real value is associated to the accuracy of an algorithm. However, there are several processes of non-linear nature involved in the 3D reconstruction from stereo images. This fact may raise concerns about the convenience of evaluating the disparity estimation process by a linear approach (Cabezas and Trujillo, 2011), and has to be considered in addition to the other weaknesses already identified in the first section of this paper.

The $\mathbf{A}^*$ evaluation methodology is introduced in (Cabezas and Trujillo, 2011). In this work, the evaluation of disparity maps is addressed as a Multiobjective Optimisation Problem (MOP). The evaluation model is based on the Pareto Dominance relation (Van Velduizen et al., 2003). It computes a proper subset A^* from the set of stereo correspondence algorithms under evaluation. The set A^* is composed by the algorithms which associated error score vectors are not better, neither worst, among them. It is argued that the proposed evaluation model can be used to compute more sets or disjoint groups of algorithms with comparable accuracy. However, this capability is not demonstrated neither discussed. Moreover, the formal definition of the set A^* is presented, but its computation from an algorithmic point of view is not discussed. In this work the BMP is used as the error measure.

In regard to error measures, the mean absolute error, the Mean Square Error (MSE), the Root Mean Square Error (RMSE), and the Mean Relative Error (MRE) have been used for ground-truth based evaluation (Van der Mark and Gavrila, 2006). However, the MSE and the RMSE measures ignore the relation between depth and disparity and penalise, in the same way, all errors regardless of the true depth, whilst the formulation of the MRE measure presented in (Van der Mark and Gavrila, 2006) is not suited to be used along with the concept of error criteria.

Table 1: Error measure scores using the Tsukuba stereo image, and SymBP+occ and EnhancedBP algorithms.

		SymBP+occ	EnhancedBP
BMP	nonocc	0,966	0,945
	all	1,755	1,736
	disc	5,086	5,048
SZE	nonocc	568,767	809,588
	all	636,391	876,363
	disc	127,504	142,798

The Sigma-Z-Error measure (SZE) is proposed in (Cabezas et al., 2011). This ground-truth based measure considers the inverse relation between depth and disparity, as well as the magnitude of estimation errors. It is focused on measuring the impact of disparity estimation errors in the Z axis. The measure allows a better judging of algorithm's accuracy, since two algorithms may have a similar quantity of errors using the BMP measure. It is illustrated in Table 1, using the error scores of the BMP measure and the SZE measure, the Tsukuba stereo image, and the EnhancedBP (Larsen et al., 2007) and SymBP+Occ (Sun et al, 2005) algorithms. It can be observed that the BMP error scores are similar. In fact, the BMP scores indicate that the EnhancedBP algorithm is, by a slight difference, more accurate than the SymBP+Occ algorithm. In contrast, the SZE error scores indicate a considerable difference in the accuracy of the estimated maps, as well as the SymBP+Occ algorithm is more accurate than the EnhancedBP algorithm. A larger comparison between the BMP and the SZE measures can be found in (Cabezas et al., 2011).

3 BACKGROUND

The formalisation of different aspects related to the quantitative evaluation of disparity maps, as well as to the Pareto Dominance relation, is presented in this section, for the sake of completeness.

3.1 Ground-truth based Evaluation of Stereo Correspondence Algorithms

Let A be a non-empty set of stereo correspondence algorithms under evaluation, as follows:

$$\mathrm{A} = \left\{ \mathrm{a} \in \mathrm{A} \,\middle|\, \mathrm{a}\colon (\,\mathrm{I_{stereo}}) \rightarrow \mathrm{D_{estimated_{(a)}}} \right\}, \quad (1)$$

where $\mathrm{I_{stereo}}$ is a non-empty set of stereo images (i.e. the imagery test-bed), and $\mathrm{D_{estimated_{(a)}}}$is the set of estimated disparity maps obtained by a particular stereo correspondence algorithm.

Let $\mathrm{D_{estimated}}$ be the set of estimated disparity maps to be compared, defined as:

$$\mathrm{D_{estimated}} = \left\{ \mathrm{D_{estimated_{(a)}}} \in \mathrm{D_{estimated}} \,\middle|\, \forall \mathrm{a} \in \mathrm{A}\colon \exists\, \mathrm{D_{estimated_{(a)}}} \right\}. \quad (2)$$

The base of the ground-truth based evaluation process is the comparison of the set $\mathrm{D_{estimated}}$ against the ground-truth data, considering different elements that compose an evaluation methodology. This can be formalised as follows. Let g be a function such that:

$$\mathrm{g}\colon \left(\mathrm{D_{estimated_{(a)}}} \mathrm{x} \mathrm{D_{true}} \mathrm{x} \mathrm{R_{criteria}} \mathrm{x} \mathrm{E_{measures}} \right) \rightarrow \mathrm{V_a}\,, \quad (3)$$

where $\mathrm{D_{true}}$ is the set of ground-truth data, $\mathrm{R_{criteria}}$ is the set of errors criteria, $\mathrm{E_{measures}}$ is the set of error measures, and $\mathrm{V_a} \in \mathbb{R}^{\mathrm{k}}$ is a vector of error scores. The magnitude of k is determined by the cardinality of the sets $\mathrm{D_{true}}$, $\mathrm{R_{criteria}}$, and $\mathrm{E_{measures}}$.

Let V be the set obtained by applying the function g to the set $\mathrm{D_{estimated}}$:

$$\mathrm{V} = \left\{ \mathrm{V_a} \in \mathrm{V} \,\middle|\, \forall\, \mathrm{D_{estimated_{(a)}}} \in \mathrm{D_{estimated}}\colon \exists\, \mathrm{V_a} \right\} \quad (4)$$

The evaluation model of the Middlebury's methodology assigns a ranking to each algorithm under evaluation, based on the error scores of the estimated disparity maps. This ranking is a real value. The evaluation model is formalised as:

$$\forall\, \mathrm{V}_a \in \mathrm{V}\colon \exists\, \mathrm{r} \,\middle|\, \mathrm{r}\colon (\mathrm{V_{(a)}}) \rightarrow \mathbb{R}\,. \quad (5)$$

On the other hand, the evaluation model of the $\mathbf{A}^*$ methodology operates by defining a partition over the set A. It is formalised as follows. Let d be a function such as:

$$\mathrm{d}\colon (\mathrm{A}) \rightarrow \mathrm{A}' \cup \mathrm{A}^* \,\middle|\, \mathrm{A}' \cup \mathrm{A}^* \subseteq \mathrm{A} \,\wedge\, \mathrm{A}' \cap \mathrm{A}^* = \{\emptyset\}\,, \quad (6)$$

subject to:

$$\nexists\, \mathrm{A}'_{\mathrm{a}} \in \mathrm{A}' \,\middle|\, \mathrm{A}'_{\mathrm{a}} \prec \mathrm{A}^*_{\mathrm{a}} \in \mathrm{A}^*\,, \quad (7)$$

where the symbol "≺" denotes the Pareto Dominance relation.

Additionally , the $\mathbf{A}^*$ methodology defines an interpretation of results based on the cardinality of the set A^*, which is stated as follows:

$$\begin{cases} \text{if } \left|\mathrm{A}^*\right| = 1, & \text{then } \mathit{superior\ accuracy} \\ \text{if } \left|\mathrm{A}^*\right| > 1, & \text{then } \mathit{comparable\ accuracy} \end{cases}, \quad (8)$$

where *superior accuracy* means that there exists a unique stereo correspondence algorithm with a superior accuracy, and *comparable accuracy* means that there exists a set of algorithms with a comparable accuracy among them (i.e. producing disparity maps which associated error vectors are not better neither worst among them), see (14). The interpretation of results, in both of the above cases, is performed in regard to the imagery test-bed considered. It cannot be extrapolated to other images, neither be generalised to all possible capturing conditions.

3.2 Evaluation of Disparity Maps based on the Pareto Dominance

In general, a MOP involves two different spaces: a decision space and an objective space (Van Veldhuizen et al., 2003). The nature of these spaces may depend on the nature of the particular MOP. The evaluation of disparity maps is viewed as a MOP in (Cabezas and Trujillo, 2011). In this work, the decision space is defined as the set of stereo correspondence algorithms under evaluation, and the objective space is defined as a set composed by vectors of error measures scores. In particular, let p and q be elements of the decision space: $\mathrm{p}, \mathrm{q} \in \mathrm{A} \,\wedge\, \mathrm{p} \neq \mathrm{q}$. Let $\mathrm{V_p}$ and $\mathrm{V_q}$ be a pair of vectors belonging to the objective space, defined based on (1) and (3). Then, the following relations between V_p and V_q can be considered, without loss of generalisation:

$$V_p = V_q \text{ iff } \forall\, i \in \{1,2,\dots,k\}: V_{p_i} = V_{q_i}. \tag{9}$$

$$V_p \leq V_q \text{ iff } \forall\, i \in \{1,2,\dots,k\}: V_{p_i} \leq V_{q_i}. \tag{10}$$

$$V_p < V_q \text{ iff } \forall\, i \in \{1,2,\dots,k\}: V_{p_i} < V_{q_i} \wedge V_p \neq V_q. \tag{11}$$

In the context of stereo correspondence algorithms comparison by quantitative evaluation of disparity maps, for any two elements in the decision space, three possible relations are considered:

$$p \prec q \text{ (p dominates q) iff } V_p < V_q. \tag{12}$$

$$p \preccurlyeq q \text{ (p weakly dominates q) iff } V_p \leq V_q. \tag{13}$$

$$p \sim q \text{ (p is comparable to q) iff } V_p \nleq V_q \wedge V_q \nleq V_p. \tag{14}$$

The relations above gives rise to the computation of sets A' and A^* in (6).

3.3 Error Measures

The BMP error measure is computed by counting the disparity estimation errors that exceeds the threshold δ. It is formulated as follows:

$$\varepsilon(x,y) = \begin{cases} 1 \text{ if} |D_{\text{true}}(x,y) - D_{\text{estimated}}(x,y)| > \delta \\ 0 \text{ if} |D_{\text{true}}(x,y) - D_{\text{estimated}}(x,y)| \leq \delta \end{cases} \tag{15}$$

$$\text{BMP} = \frac{1}{N}\sum_{(x,y)}^{N} \varepsilon(x,y), \tag{16}$$

where N is the total of pixels and δ is the error threshold.

The SZE error measure is an unbounded metric. It is computed by summing the differences between the true depth and the estimated depth over the entire estimated map (Cabezas et al., 2011). It is formulated as follows:

$$\text{SZE} = \sum_{(x,y)} \left| \frac{f * B}{D_{\text{true}}(x,y) + \mu} - \frac{f * B}{D_{\text{estimated}}(x,y) + \mu} \right|, \tag{17}$$

where f is the focal length, B is the distance between optical centres, and μ is a small constant value which avoids the instability caused by missing disparity estimations.

4 GROUPING STEREO ALGORITHMS

The methodology of (Cabezas and Trujillo, 2011) offers theoretical advantages over conventional evaluation methodologies for stereo correspondence algorithms. In particular, it avoids a subjective interpretation of evaluation results, since it reformulates the problem as a MOP, and it is based on the cardinality of the set A^*. However, it fails in considering an evaluation scenario on which a user may be interested in an exhaustive evaluation of the entire set of algorithms, and not only in determining which ones are the most accurate. In practice, this scenario may rise very often. For instance, when a user is interested in a particular algorithm which is not in the set A^*, but belonging to the set A'. This situation can be tackled by introducing an algorithm devised to iteratively compute a partition of the set A into the sets A' and A^*. The composition of these sets is determined based on the three possibilities in regard to the Pareto Dominance relation in (12), (13) and (14).

The `partitioningAndGrouping` algorithm assigns to each computed set A^* an ordinal label related to the partition established in the iteration. In particular, the first partition –the set A_1^*– is composed by the most accurate stereo correspondence algorithms, among the evaluated algorithms. Moreover, all algorithms in a partition with an n-label of accuracy are dominated by at least one algorithm in a partition with an m-label of accuracy, subject to m<n. This is:

$$\forall q \in A_n^* \; \exists\, p \in A_m^* \; \middle|\; p \prec q, \tag{18}$$

subject to:

$$m < n. \tag{19}$$

The `partitioningAndGrouping` algorithm is presented below for the sake of completeness in an object oriented pseudo-code as follows:

```
partitioningAndGrouping(void){
 // A is the set under evaluation
 A = Set( );
 A.load("Algorithms.dat");
 //p and q are two particular
 //algorithms
 p = Element ( );
 q = Element ( );
 //auxiliary variables
 //A' and A* are empty sets
 A' = Set( );
 A* = Set( );
 levelCount = Int (1);
 p_DomFlag = Boolean (false);
 q_DomFlag = Boolean (false);
 otherwiseFlag = Boolean (false);
 do{
    //a member of A is introduced in A*
    A*.push( A.pop( ) );
    //A iteration invariant
    while ( !A.isEmpty( ) ){
       p_DomFlag=false;
       q_DomFlag=false;
       otherwiseFlag=false;
       p= A.pop( );
       // A* iteration invariant
       for each element in A* {
          q = A*.next( );
          if (q≺p){
             A′.push(p);
             q_DomFlag=false;
          }
          elseif(p≼q || p~q){
             otherwiseFlag=true;
          }
          elseif(p≺q){
             A′.push(q);
             A*.remove(q.id);
             p_DomFlag=true;
          }
       }//for each
       if (!q_DomFlag &&
            (otherwiseFlag ||
             p_DomFlag)){
               A*.push(p);
       }// end if
    }//while
    A*.save("A*Label_"+labelCount);
    labelCount++;
    //the sets are updated
    A = A′;
    A′.removeAll( );
    A*.removeAll( );
 }while(!A.isEmpty( ));
}; // end method
```

The key step of the `partitioningAndGrouping` algorithm consists in computing a partition: the set A is updated with the elements of the set A', and A^* is recomputed. The algorithm stops once the set A became an empty set. In this way, the entire set of algorithms under evaluation is grouped according to the Pareto Dominance relation.

5 EVALUATION RESULTS

The set of 112 stereo correspondence algorithms reported in the online ranking of Middlebury's evaluation methodology (Scharstein and Szeliski, 2011) is used for evaluating the proposed methodology. All algorithms and estimated disparity maps were taken as the input into the proposed methodology. As output, seven groups of algorithms were obtained, using the SZE error measure. The cardinality of the obtained groups, as well as the percentage of each group in relation to the total of algorithms, is shown in Table 2.

Table 2: Obtained **A***-Groups using algorithms at Middlebury's web site.

A^* Group	Cardinality	Percentage
1	9	8 %
2	47	42 %
3	19	17 %
4	19	17 %
5	6	5.3 %
6	9	8 %
7	3	2.7 %
Total	112	100%

The algorithms in the group 1 are of special interest, since they produce the most accurate disparity maps for the considered test-bed. Table 3 contains the SZE scores of selected algorithms among the seven groups. Among these, the GC+SegmBorder algorithm belongs to the first group, the ObjectStereo algorithms belongs to the second group, the RTAdaptWgt algorithm belongs to the third group and so on. It can be observed that the GC+SegmBorder stereo correspondence algorithm has superior performance, according to evaluation criteria.

More evaluation results are presented in the following subsections. The algorithms in the first group of accuracy are briefly described in the subsection 5.1. The results obtained by using the proposed methodology are contrasted against the results obtained by using the Middlebury's and the conventional $\mathbf{A}^*$ methodologies in the subsection 5.2. The algorithms belonging to the groups 2 to 7 are listed in the subsection 5.3.

5.1 A Brief Description of the First Group of Algorithms

The first group of algorithms and associated SZE scores are shown in Table 4. Among these nine algorithms, seven algorithms can be classified as global algorithms, whilst two of them can be classified as local algorithms according to the taxonomy in (Scharstein and Szeliski, 2002). This result implies that there exist local algorithms with comparable performance of global algorithms.

Among the global algorithms, the GC+occ, MultiCamGC (Kolmogorov and Zabih; 2001, 2002), and MultiResGC (Papadakis and Caselles, 2010) algorithms are based on GC. Although the FeatureGC and GC+SegmBorder algorithms, are, as far to the authors known, unpublished (Dec, 2011), being reasonable to assume that they are based on the GC. The GC+occ and MultiCamGC algorithms consider an energy function using three terms: image intensities similarity, disparity smoothness, and occlusion handling and uniqueness constraint fulfilment. The MultiResGC algorithm considers the above criteria, using an energy function with four terms.

The DoubleBP algorithm (Yang et al., 2008), is based on the BP. In particular, a hierarchical BP is applied twice, where occlusions and textureless areas are first identified and filled using neighbouring values.

The Segm+Visib algorithm (Bleyer and Gelautz, 2004) is based on colour segmentation. Segments are grouped into planar layers. The textureless areas and the depth discontinuities are handled by segmentation, whilst occlusions are detected by the layers assignment.

In regard to the local algorithms, the DistinctSM algorithm (Yoon and Kweon, 2007) uses a similarity measure based on the distinctiveness of image points and the dissimilarity between them. The idea is: the more similar and distinctive two points are, thus the probability of a correct match is larger (Manduchi and Tomassi, 1999). In this algorithm, occluded points are not modelled explicitly.

In the algorithm PatchMatch (Bleyer et al., 2011), the concern is about when the window captures a slanted surface. An adaptation of the work of (Barnes et al., 2009) is used to find a slanted support plane at each pixel. It is pointed out in (Bleyer et al., 2011) that according to the BMP measure, the algorithm PatchMatch has an outstanding performance in the *nonocc* criterion of the Teddy stereo image. However, the SZE scores indicate that the performance of the algorithm in the *nonocc* criterion is not particularly outstanding. In contrast, the SZE scores of the Venus image, composed in essence by slanted planes, are considerably better than the scores obtained by most of the others algorithms, as can be observed in Table 4.

5.2 Obtained Results vs. Related Works Results

The proposed methodology – **A*** Groups – performs an exhaustive evaluation of all algorithms, by computing groups of comparable accuracy. In this way, the proposed methodology allows a complete and objective interpretation of evaluation results. In contrast, the results presented in (Cabezas and Trujillo, 2011), only identifies the algorithms in the first group, without providing useful information to the user about the rest of algorithms. In practice, and for evaluation purposes, this difference may be quite

Table 3: Selected SZE scores of algorithms from different groups.

	Tskuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
GC+SegmBorder	212.0	242.8	124.6	30.8	46.8	24.0	39.4	62.89	20.3	50.4	64.9	24.3
ObjectStereo	578,7	618,5	136,2	885,3	912,4	107,9	141,6	215,1	35,8	73,9	117,9	36,2
RTAdaptWgt	651,0	705,9	151,0	1078,3	1131,7	109,3	186,3	246,92	47,7	83,0	144,8	45,9
RealtimeBP	748,0	931,6	188,0	1114,1	1223,9	158,7	205,0	311,7	64,7	92,6	198,7	56,8
OptimizedDP	829,8	963,6	190,7	1274,4	1424,0	168,0	213,1	366,8	68,5	95,3	211,0	59,5
DP	901,4	989,5	203,0	2052,9	2206,3	258,6	239,4	721,8	82,8	146,3	524,4	92,6
MI-nonpara	1149,3	1301,9	282,4	2227,0	2560,4	378,5	1233,6	2903,4	139,6	184,3	1815,9	103,4

Table 4: SZE values using algorithms in the first group of accuracy.

	Tskuba			Venus			Teddy			Cones		
	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc
DoubleBP	658.8	703.0	116.3	1062.7	1115.3	95.8	102.3	155.6	28.4	71.3	341.5	37.8
PatchMatch	538.5	568.8	168.4	539.1	571.0	85.4	115.4	498.1	66.0	49.9	261.8	32.8
GC+SegmBorder	212.0	242.8	124.6	30.8	46.8	24.0	39.4	62.89	20.3	50.4	64.9	24.3
FeatureGC	212.2	257.9	102.9	1013.6	1045.1	97.6	98.6	185.9	36.9	77.4	130.1	46.0
Segm+visib	388.1	414.8	122.4	1088.0	1131.2	124.3	63.7	87.8	27.8	67.0	127.8	43.2
MultiResGC	411.9	451.5	108.4	1080.5	1137.3	113.5	121.4	170.6	43.7	92.7	154.9	49.8
DistinctSM	363.2	411.9	115.3	1050.9	1103.0	103.0	143.0	191.8	52.2	73.0	121.7	37.0
GC+occ	190.9	266.1	116.8	1319.6	1455.2	168.8	469.5	951.6	131.6	301.4	792.9	163.8
MultiCamGC	192.7	266.1	113.7	1201.0	1269.6	107.1	448.0	679.9	108.3	218.2	411.5	102.3

significant. On the other hand, the evaluation results based on the ranking computed by the Middlebury's methodology may lead to a subjective interpretation.

Table 5: List of the algorithms in the first **A***-group vs. Middleburry's rank.

Algorithm	Middleburry's Rank
DoubleBP	4
PatchMatch	11
GC+SegmBorder	13
FeatureGC	18
Segm+visib	29
MultiResGC	30
DistinctSM	34
GC+occ	67
MultiCamGC	68

Table 6: Middleburry's results vs. obtained results.

Algorithm	Middleburry's Rank	$\mathbf{A}^*$-Group
ADCensus	1	2
AdaptingBP	2	2
CoopRegion	3	2
DoubleBP	4	1
RDP	5	2
OutlierConf	6	2
SubPixDoubleBP	7	2
SurfaceStereo	8	2
WarpMat	9	2
ObjectStereo	10	2
PatchMatch	11	1
Undr+OvrSeg	12	2
GC+SegmBorder	13	1
InfoPermeable	14	2
CostFilter	15	2

The algorithms in the first group, and the ranks assigned by the Middlebury's methodology are shown in Table 5. It can be observed that the assigned ranks are distant among them. This indicates that the SZE error measure and the evaluation model of the proposed methodology show different evaluation properties to those of the BMP measure and Middlebury's evaluation model.

On the other hand, the top 15 ranked algorithms by the Middlebury's evaluation and the assigned group by the proposed methodology are shown in Table 6. It can be observed that most of algorithms are in the group 2, using the Pareto Dominance relation. Thus, Middlebury's ranking may be failing in identifying the most accurate algorithms. These discrepancies can be explained by the fact that accurate depth estimation implies accurate disparity estimation, but a small percentage of disparity errors (i.e. under an error threshold) do not necessarily imply accurate estimation in terms of depth.

5.3 Exhaustive Evaluation of Stereo Correspondence Algorithms

The algorithms in the groups 2 to 7 are listed in Table 7. These results may be of interest for researchers that had reported a stereo correspondence algorithm to the Middlebury's online website. It is worth to note that the selection of test-bed images, error criteria, error measures and algorithms in the evaluation process will impact on the groups' composition and cardinality.

6 CONCLUSIONS

In this paper, an evaluation methodology of stereo correspondence algorithms based on the Pareto Dominance relation is extended by the introduction of the `partitioningAndGrouping` algorithm. The resulting methodology is termed as $\mathbf{A}^*$ Groups. As a distinctive property, the $\mathbf{A}^*$ Groups methodology allows to perform an exhaustive evaluation and an objective interpretation of results. Innovative evaluation results were obtained using the SZE error measure, which is based on the inverse relation between depth and disparity, and considers the error magnitude. The obtained evaluation results indicate: there is not a single algorithm of superior performance. There are local stereo algorithms with adaptive support strategies of comparable performance to global stereo algorithms according to the Pareto Dominance relation. Algorithms based on GC, BP and colour segmentation are the most accurate algorithms using global optimisation strategies.

Table 7: Algorithms and **A***-groups.

Algorithm	**A*-Group**
ADCensus	2
AdaptingBP	2
CoopRegion	2
RDP	2
OutlierConf	2
SubPixDoubleBP	2
SurfaceStereo	2
WarpMat	2
ObjectStereo	2
Undr+OvrSeg	2
InfoPermeable	2
CostFilter	2
GlobalGCP	2
AdaptOvrSegBP	2
P-LinearS	2
PlaneFitBP	2
SymBP+occ	2
PUTv3	3
GradAdaptWgt	3
RT-ColorAW	3
MultiCue	3
HistoAggr	3
BPcompressed	3
FastAggreg	3
Layered	3
ESAW	3
ConvexTV	3
TensorVoting	3
HRMBIL	3
ReliabilityDP	3
HBpStereoGpu	3
AdaptDispCalib	4
CurveletSupWgt	4
FastBilateral	4

Table 7: Algorithms and **A***-groups. (cont.)

ASSM	2
ConfSuppWin	2
GeoDif	2
C-SemiGlob	2
IterAdaptWgt	2
RandomVote	2
SO+borders	2
Bipartite	2
MVSegBP	2
OverSegmBP	2
LocallyConsist	2
SegmentSupport	2
VSW	2
SegTreeDP	2
AdaptWeight	2
InteriorPtLP	2
ImproveSubPix	2
BP+DirectedDiff	2
SemiGlob	2
RealTimeABW	2
PlaneFitSGM	2
2OP+occ	2
VarMSOH	2
Unsupervised	2
SNCC	2
StereoSONN	2
RealtimeVar	2
GenModel	2
RTCensus	2
GC	2
GeoSup	3
RTAdaptWgt	3
CostAggr+occ	3
RegionTreeDP	3
EnhancedBP	3
CostRelaxAW	4
RealtimeBFV	4
VariableCross	4
RealtimeBP	4
CCH+SegAggr	4
AdaptPolygon	4
RealTimeGPU	4
CostRelax	4
AdaptDomainBP	4
TreeDP	4
CSBP	4
DCBGrid	4
H-Cut	4
SAD-IGMCT	4
FLTG-DDE	4
PhaseBased	4
OptimizedDP	5
TwoWin	5
DOUS-Refine	5
BP+MLH	5
IMCT	5
PhaseDiff	5
BioPsyASW	6
DP	6
DPVI	6
2DPOC	6
RegionalSup	6
SSD+MF	6
SO	6
STICA	6
Infection	6
MI-nonpara	7
LCDM+AdaptWgt	7
Rank+ASW	7

REFERENCES

Barnes, M., Shechtman, C., & Finkelstein, E., 2009. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics*, 28(3). pp. 1–11.

Bleyer, M., & Chambon, S., 2010. Does Color Really Help in Dense Stereo Matching?. In *Proceedings of the International Symposium 3D Data Processing, Visualization and Transmission*, pp. 1–8.

Bleyer M. & Gelautz M. 2004. A layered stereo algorithm using image segmentation and global visibility constraints. In *Proceedings International Conference Image Processing*. IEEE Computer Society pp. 2997–3000.

Bleyer, M., Rhemann, C., & Rother, C., 2011. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In *Proceedings of the British Machine Vision Conference*, pp. 1–11.

Cabezas, I., Padilla V., & Trujillo, M., 2011. A measure for accuracy disparity maps evaluation. In *Proceedings of the Iberoamerican Congress on Pattern Recognition*. Springer-Verlag, pp. 223–231.

Cabezas, I., & Trujillo, M., 2011. A Non-Linear Quantitative Evaluation Approach for Disparity Estimation. In *Proceedings of the International Conference on Computer Vision, Theory and Applications*, pp. 704–709.

Felzenszwalb, P., & Huttenlocher, D., 2004. Efficient belief propagation for early vision. In *Proceedings of Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. 261–268.

Gallup, D., Frahm, J., Mordohai, & P., Pollefeys, M., 2008. Variable Baseline/Resolution Stereo. In *Proceedings of Computer Vision and Pattern Recognition*, IEEE Computer Society, pp. 1–8.

Hirschmüller, H., & Scharstein, D., 2009. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9). IEEE Computer Society, pp. 1582–1599.

Kelly, P., O'Connor N., & Smeaton A., 2007. A Framework for Evaluating Stereo-Based Pedestrian Detection Techniques. In *IEEE Transactions Circuits and Systems for Video Technology*, 18(8). IEEE Computer Society. IEEE, pp. 1163–1167.

Kolmogorov, V., & Zabih, R., 2001. Computing visual correspondence with occlusions using graph cuts. In *Proceedings of Eighth International Conference on Computer Vision*, IEEE Computer Society, pp. 508–515.

Kolmogorov, V., & Zabih, R., 2002. Multi-camera scene reconstruction via graph cuts. In *Proceedings of European Conference on Computer Vision*, Springer Verlag, pp. 82–96.

Kostlivá, J., Cech, J., & Sara, R., 2007. Feasibility Boundary in Dense and Semi-Dense Stereo Matching. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. 1–8.

Larsen, E., Mordohai, P., Pollefeys, M., & Fuchs, H., 2007. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *International Conference on Computer Vision*. IEEE Computer Society, pp. 282–298.

Manduchi, R., & Tomasi, C., 1999. Distinctiveness maps for image matching. In *International Conference on Image Analysis and Processing*. pp. 26–31.

Neilson, D., & Yang, Y., 2008. Evaluation of Constructable Match Cost Measures for Stereo Correspondence using Cluster Ranking. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. 1–8.

Papadakis, N., & Caselles V., 2010. Multi-label Depth Estimation for Graph Cuts Stereo Problems. In *Journal of Mathematical Imaging and Vision*, 38(1). Kluwer Academic Publishers, pp. 70–82.

Scharstein, D., & Szeliski, R., 2011. Middlebury Stereo Evaluation - Version 2. Retrieved October 24th, 2011, from: http://vision.middlebury.edu/stereo/eval/.

Scharstein, D., & Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. In *International Journal of Computer Vision*, Volume 47, pp. 7–42.

Scharstein, D., & Szeliski, R., 2003. High-accuracy Stereo Depth Maps using Structured Light. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, pp. I–195–I–202.

Sun J., Li, Y., Kang, S., & Shum H., 2005. Symmetric stereo matching for occlusion handling. In C*omputer Vision and Pattern Recognition*. IEEE Computer Society, pp. 399–406.

Szeliski, R., 1999. Prediction Error as a Quality Metric for Motion and Stereo. In *International Conference on Computer Vision*, Volume 2. IEEE Computer Society, pp. 781–788.

Szeliski, R., & Zabih, R., 2000. An Experimental Comparison of Stereo Algorithms. In *Proceedings of the International Workshop on Vision Algorithms*. Springer-Verlag, pp. 1–19.

Tombari, F., Mattoccia, S., & Di Stefano, L., 2010. Stereo for Robots: Quantitative Evaluation of Efficient and Low-memory Dense Stereo Algorithms. In *Proceedings of International Conference Control Automation Robotics and Vision*. IEEE Computer Society, pp. 1231–1238.

Van der Mark, W., & Gavrila, D., 2006. Real-time Dense Stereo for Intelligent Vehicles. In *IEEE Transactions on on Intelligent Transportation Systems*, 7(1). IEEE Computer Society, pp. 38–50.

Van Veldhuizen, D., Zydallis, J., & Lamont, G., 2003. Considerations in engineering parallel multiobjective evolutionary algorithms. In *IEEE Transactions on Evolutionary Computation*, 7(2). IEEE Computer Society, pp. 144–173.

Yang, Q., Wang, L., Yang, R., Stewénius, H., & Nistér, D., 2008. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In *Transactions on Pattern Analysis and Machine Intelligence*, 31(3). IEEE Computer Society, pp. 492–504.

Yoon, K., & Kweon, I., 2007. Stereo matching with the distinctive similarity measure. In *International Conference on Computer Vision*. IEEE Computer Society, pp. 1–7.