Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching

Aristotle Spyropoulos[1] Nikos Komodakis[2] Philippos Mordohai[1]
[1]Stevens Institute of Technology [2]Ecole des Ponts ParisTech
{ASpyropo, Philippos.Mordohai}@stevens.edu Nikos.Komodakis@enpc.fr

Abstract

While machine learning has been instrumental to the ongoing progress in most areas of computer vision, it has not been applied to the problem of stereo matching with similar frequency or success. We present a supervised learning approach for predicting the correctness of stereo matches based on a random forest and a set of features that capture various forms of information about each pixel.We show highly competitive results in predicting the correctness of matches and in confidence estimation, which allows us to rank pixels according to the reliability of their assigned disparities. Moreover, we show how these confidence values can be used to improve the accuracy of disparity maps by integrating them with an MRF-based stereo algorithm. This is an important distinction from current literature that has mainly focused on sparsification by removing potentially erroneous disparities to generate quasi-dense disparity maps.

1. Introduction

Stereo matching is an inverse problem and, as such, it is notoriously prone to errors, mostly due to occlusion, lack of texture and repeated structures. Since the common causes of the errors are well known, one would expect that learning methods could have been used to detect them. Helpful cues are available in the neighborhood of a pixel as well as in information generated during the matching process. Surprisingly, very few publications have attempted to tackle stereo matching from a learning perspective [4, 12, 13] and they have not gained much traction. Very recently, Haeusler et al. [7] presented an approach for learning a confidence measure from several features, some of which are similar to those proposed by us, since both approaches rely on [9] for feature selection. Haeusler et al. also use a random forest for classification, but, unlike this paper, they do not propose ways of leveraging the estimated confidence to generate dense disparity maps of higher accuracy.

What separates our approach from recent literature on confidence estimation [20, 6, 9, 21, 7], regardless of the use of learning, is that the main objective of these methods is sparsification. They can indeed generate disparity maps with progressively fewer errors by removing matches starting from the least reliable ones. What has not been shown, however, is how this capability can be used to correct the initially wrong matches. We present such an approach in this paper.

Given a training set of stereo pairs with ground truth disparity, the goal of this paper is to answer the following questions without making scene-specific assumptions:

Is it possible to predict whether a stereo correspondence is right or wrong based on features extracted from the stereo pair for that pixel and a trained classifier?

Is it possible to use these predictions to improve the disparity map?

Our results show that the answer is affirmative in both cases. Figure 1 shows the inputs to our algorithm: an image and a Winner-Take-All (WTA) disparity map, as well as its outputs: a correctness prediction map and an improved disparity map after Markov Random Field (MRF) optimization. The matching cost volume is an additional input not shown here.

To answer the first question, we formulate a binary clas-

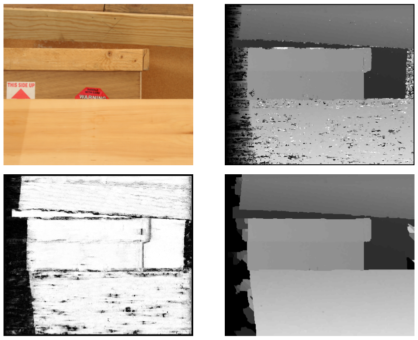


Figure 1. Top row: Input image and WTA disparity map using NCC for Wood2 [22]. Bottom row: prediction map, in which bright intensities correspond to WTA matches that are likely to be correct, and final disparity after MRF optimization.

sification problem and tackle it using a random forest (RF) classifier [3]. We argue that this problem is more fundamental than confidence estimation without the ability to decide on correctness [9, 20] or selection of a hypothesis among a set generated by a mixture of experts [13, 16]. Ranking stereo matches according to confidence accurately is valuable but does not imply the capability to determine which of the matches are correct, since the error rate may fluctuate from image to image making the selection of a threshold hard without knowledge of the priors. As shown in Section 5, we are able to predict the correctness of matches on disparity maps with very different error rates at nearly optimal rates. Haeusler et al. [7] have been able to show very good results on a similar task on the KITTI benchmark [5].

Before summarizing the contributions of our method, let us remark that we made every effort to keep it generic. Customizing our approach to a specific domain would allow us to introduce task-specific features, likely resulting in even higher accuracy. For example, if the task was driver assistance [5], accuracy would benefit from features such as image coordinates that provide information on which parts of the scene are likely to be road, buildings or sky. We leave this extension for future work. Our current contributions are:

- an algorithm that achieves high accuracy in predicting the correctness of stereo matching given training data,
- a diverse set of features that enable classification,
- a technique for detecting ground control points and for inserting them as soft constraints into an MRF-based optimizer, leading to improved disparity maps.

We show results on the extended Middlebury benchmark [22] that contains 27 image pairs with ground truth, including comparisons with numerous baselines.

2. Related Work

For a survey of stereo methods we refer readers to [23] and its companion website. Here we focus on research that aims at inferring the correctness of correspondences using learning, or at detecting ground control points (GCPs).

Early work on applying machine learning to stereo includes that of Lew et al. [14] who presented an approach for selecting a set of features that form an effective descriptor for stereo matching. Cruz et al. [4] addressed the problem of determining whether a match in edge-based stereo was correct or not. Classification relies on four features extracted by filtering the images and uses a perceptron to determine which feature mappings from the left to the right image are indications of correct matching. This approach, however, does not address challenges in textureless regions, since it is only applied to edge pixels, and also does not model mismatches due to repeated structures.

Kong and Tao [12] used non-parametric techniques to learn the probability of a potential match to belong in three categories: correct, wrong due to foreground over-extension or wrong for other reasons. They used features extracted from image appearance and matching cost estimates, while final disparity assignments to fronto-parallel superpixels were made via simulated annealing on an MRF. The integration of the correctness probabilities into the MRF improved accuracy on the Middlebury benchmark, but the accuracy of the stand-alone classifier was not reported in the paper. This approach was extended [13] to select among 36 experts in the form of different normalized cross-correlation (NCC) matching windows using similar features and optimization technique. Motten et al. [17] presented a classifier using decision trees implemented on FPGA for selecting among multiple disparity hypotheses generated by trinocular stereo. Sabater et al. [21] introduced an a contrario approach for validating the correctness of stereo matches. A user-specified acceptable number of false matches determines the density of the final disparity map.

We would be remiss if we did not include the work of Mac Aodha et al. [16] on optical flow, which shares some characteristics with ours, such as an emphasis on being applicable to general scenes and operating on individual pixels. A multi-class classifier that selects among four state of the art methods is used to learn the posterior of each expert being correct. The estimated posteriors are then used as confidence measures. Other recent research on confidence estimation, from which we draw inspiration and borrow features, includes the work of Reynolds et al. [20] on time-of-flight data and of Hu and Mordohai [9] on stereo. Haeusler and Klette [6] also considered several confidence measures, as well as the product of all measures, demonstrating good performance in sparsification. Pfeiffer et al. [19] integrated three confidence measures into a mid-level representation for 3D reconstruction and showed that Bayesian reasoning outperforms sparsification by thresholding.

Contrasted with methods for selecting among a set of experts, such as those of Kong and Tao for stereo [13] and Mac Aodha et al. for optical flow [16], our research addresses the more fundamental problem of verifying whether a prediction from a single expert is correct. In that sense, it is similar to the work of Haeusler et al. [7] who also make predictions about the correctness of the outputs of the semi-global matching algorithm.

Methods for selecting GCPs typically rely on heuristics that are strongly correlated with correctness, but make hard decisions based on multiple thresholds. Bobick and Intile [2] imposed several constraints on GCPs: lower cost than all competing matches in both images, low matching cost, sufficient image texture and presence of nearby GCPs to suppress outliers. Kim et al. [10] use left-right consistency (LRC) and comparison of the matching cost against a threshold for selecting GCPs. Wang and Yang [25] pick GCPs by running three different Winner-Take-All (WTA) stereo algorithms and require that the disparities be consistent among all the matchers in each image, as well as left-right consistent. Sun et al. [24] used LRC and the ratio of the best to the second best matching cost in a disparity propagation framework. Our approach integrates numerous criteria in a principled way via supervised learning and learns how to make decisions based on labeled data rather than intuition. One of the byproducts of this approach is the much higher density of GCPs without loss of accuracy, which is at 99.7% on our data.

3. Method Overview

In this section, we briefly describe the steps of our algorithm. Initially, eight features are extracted for all pixels with assigned disparity values in all images of the training set (Section 4). In the *training phase*, a random forest (RF) classifier is trained on individual pixels to predict whether their assigned disparities are correct. In the *testing phase*, the same features are extracted for all pixels of a test image and the classifier generates a prediction for their correctness. The effectiveness of the classifier is evaluated in Section 5 where we measure the accuracy of the predictions, as well as the ability of our method to rank pixels correctly in order of decreasing reliability. A comparison against the strongest individual features shows that the RF easily outperforms them and approaches optimal performance.

The predictions of the RF can be used to select ground control points (GCPs) which are of very high accuracy and high density (Section 6) compared to baseline GCP selection methods. Finally, the GCPs are integrated as soft constraints into an MRF optimizer to improve the input Winner-Take-All (WTA) disparity maps. Our results in Section 7 clearly demonstrate that it is possible to improve the accuracy of binocular stereo by learning from features extracted from images, disparity maps and matching cost volumes.

4. Features and Learning

In this section, we present the rationale behind the features and learning algorithm we selected. This set of features is by no means exhaustive, but it aims at extracting useful information from various sources including the cost curve for each pixel and the pixel's neighbors in the disparity map. The label for each pixel indicates whether the disparity with the minimum cost that would have been assigned to it by a WTA stereo algorithm, is correct or not. The usual definition of correctness (disparity error less than or equal to one [23]) is used.

Before describing the features, we introduce some notation. Given a pair of rectified images, we compute the *cost volume* $c(x_L, x_R, y)$ that contains a cost value for each possible match from a pixel in the left image (x_L, y) to a pixel in the right image (x_R, y). Disparity is defined conventionally as $d = x_L - x_R$ and we assume that the minimum and maximum values it can take, d_{min} and d_{max}, are externally provided. For convenience, we define the disparity of a pixel in the right image to be equal to d, $d_R = x_L - x_R$. Values in the cost volume for matches beyond the disparity range are flagged as invalid and ignored in all computations. If a similarity, instead of a cost function, is used to assess matches, we negate its output to convert it to cost. The *cost curve* of a pixel is the set of cost values for all allowable disparities for the pixel. We use c_1 and c_2 for the minimum and second minimum values of the cost curve, without requiring c_2 to be a local minimum. The disparity value corresponding to c_1 is denoted by d_1.

We used the following eight features for the experiments in this paper. Four of them were considered individually as confidence measures in [9].

Cost. This is the minimum matching cost over all disparities for a given pixel and captures the fact that low cost often corresponds to high likelihood of correct matching.

Distance from Border (DB). This feature measures the distance in pixels from the nearest image border. It is based on the assumption that pixels near the borders are likely to be outside the field of view of the other camera and that causes mismatches. We experimented with four separate features measuring the distance from the left, right, top and bottom borders, but no improvement was observed.

Maximum Margin (MMN). This feature measures the difference between the two smallest cost values, c_1 and c_2, of a pixel [9]. The rationale here is that a large difference may indicate an unambiguous disparity assignment.

Attainable Maximum Likelihood (AML). This feature is based on the conversion of the cost curve to a probability density function over disparity. It has been shown that subtracting the minimum cost $c_1(x_L, y)$ from all cost values leads to higher discriminative power [9]. AML is defined as follows.

$$f_{\text{AML}}(x_L, y) = \frac{1}{\sum_{x_R} e^{-\frac{(c(x_L, x_R, y) - c_1(x_L, y))^2}{2\sigma^2_{AML}}}} \qquad (1)$$

Left-Right Consistency (LRC). A good indicator of the correctness of a match from the left to the right image is whether it is confirmed in the opposite direction. LRC, here, is a binary feature set to 0 when the absolute value of the difference between the disparity d at pixel (x_L, y) in the left image and the disparity at pixel $(x_L - d, y)$ in the right image is less than or equal to 1. LRC is 1 when the difference is greater than 1.

Left-Right Difference (LRD). This confidence measure [9] favors a large margin between the two smallest minima of the cost for pixel (x_L, y) in the left image and also consistency of the minimum costs across the two images.

$$f_{\text{LRD}}(x_L, y) = \frac{c_2(x_L, y) - c_1(x_L, y)}{|c_1(x_L, y) - min_{x'}\{c(x', x_L - d, y)\}|} \tag{2}$$

The intuition is that truly corresponding pixels should result in similar cost values and thus a small denominator. LRD can be small for two reasons: if the margin is small, or if the margin $c_2 - c_1$ is large, but the pixel has been mismatched causing the denominator to be large.

Distance from Discontinuity (DD). Pixels near depth discontinuities are likely to be mismatched. Since we do not know the true discontinuities, we use the WTA disparity estimates as a proxy and declare as discontinuous any pixel whose disparity is not equal to all of its four neighbors. DD then is equal to the horizontal distance from each pixel to the nearest discontinuity.

Difference with Median Disparity (MED). Pixels with disparity values that are consistent with their neighborhood are more likely to be correct. We capture this by computing the median disparity in a 5×5 window centered at each pixel and taking the absolute value of the difference between the median and the pixel's own disparity. This difference is truncated at 2 in our current implementation.

We experimented with some other features, but they did not appear to contribute towards higher prediction accuracy. We were not able to extract useful information from image appearance using gradient or color variance-based features. We speculate that the reason is that large gradients are associated with discontinuities that have large mismatch probability, but also with highly textured pixels that can be reliably matched. We also tried a feature that indicates whether a pixel is occluded according to current disparity estimates, but it also appears to offer little additional benefit. Other features from [9] are either weak predictors or strongly correlated with the ones above. Haeusler et al. [7] have used eight features, two of which are similar to AML and LRC, as well as the variance of the disparity map which bears some similarity to DD. They used horizontal intensity gradients features, but they had low importance scores.

Random Forest. Our feature design was not done with any learning algorithm in mind, an approach that allowed us to experiment with different options. We have selected a random forest [3] among alternatives, such as linear and nonlinear Support Vector Machines which performed worse in our tests. We believe that the non-parametric nature of the random forest and its resilience to noisy labels make it a good fit for our data. Boosting, which we did not attempt, may have also been successful. We trained the random forest in regression mode, using binary labels indicating whether the disparity assigned to a pixel is correct, in order to obtain a soft prediction Y for the correctness of each pixel. The predictions can be viewed as confidence measures. They can be used to rank disparity assignments, or they can be thresholded to classify them. Since we cannot expect to know whether a pixel is occluded during testing, we included the occluded pixels in the training set without distinguishing them from non-occluded pixels. The ground truth labels for the occluded pixels were treated identically to those of the non-occluded ones.

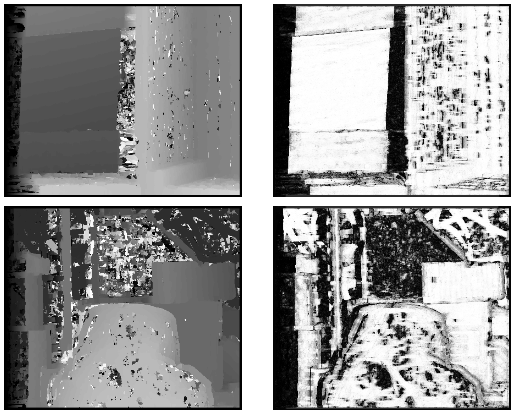

(a) WTA disparity (b) RF Prediction

Figure 2. Input WTA disparity maps and RF predictions for Wood1 and Lampshade1. Notice the low predictions (dark pixels) for occluded regions and other errors.

5. Experimental Validation of Correctness Prediction and Confidence Estimation

In this section, we present results that show the ability of our approach to classify and rank matches without modifying them. The output of WTA stereo is used as-is in this section. We use the extended Middlebury benchmark (2005 and 2006 datasets) [22] that includes *27 stereo pairs*. All experiments were performed on cost volumes computed using normalized cross-correlation (NCC) in 5×5 windows and negating the NCC values to obtain costs for disparity values from 0 to 85. The choice of matching function and window size is not optimized in any sense, but produces reasonable results. σ_{AML} in (1) was set to 0.2. We trained random forests comprising 50 trees in regression mode using the Matlab TreeBagger package. Three-fold cross-validation was used throughout by training a random forest on 18 stereo pairs and testing on the 9 remaining pairs. Figure 2 contains two noisy examples to show the ability of the RF to assign low prediction scores to unreliable pixels.

It is important to distinguish between **disparity errors**, which are defined as pixels with incorrect disparities, and **prediction errors**, which are errors made by our classifier by considering a disparity assignment as incorrect, when it was correct and vice versa.

In Table 1, we report the *prediction accuracy* of our classifier on the 27 stereo pairs. We classify disparity assignments of WTA stereo by thresholding the prediction Y of the random forest at 0.5. Note that our method is effective for disparity maps with both low and high error rates. See for example Books and Lampshade2 which have a prediction error of approximately 11%, while the disparity error of the WTA disparity maps is 22% and 32% respectively. Low sensitivity to input variability differentiates our work from confidence estimation methods which may be able to rank matches accurately, but are unable to determine which ones are correct without knowledge of the disparity error rate. The overall prediction error for pixels with correct disparity is 4.5% and for pixels with incorrect disparity it is 22.8%, for a combined prediction error of 8.4%.

Following recent publications on evaluating the confidence of stereo [9], time-of-flight data [20] and optical flow [16], we evaluated the accuracy of the ranking of disparity assignments using receiver operating characteristic (ROC) curves of error rate as a function of disparity map density. We ranked all matches in decreasing order of prediction and produced disparity maps of increasing density by selecting pixels according to rank. The area under the curve (AUC) quantifies the ability of a confidence measure to predict correct matches. Better confidence measures result in lower AUC values. The optimal AUC can be obtained by selecting all correct matches first and is equal to $A_{opt} = \int_{1-\varepsilon}^{1} \frac{d_m-(1-\varepsilon)}{d_m} dd_m = \varepsilon + (1-\varepsilon)ln(1-\varepsilon)$, where ε is the disparity error rate [9]. The average optimal AUC over all 27 pairs is 0.0336. The average AUC value for RF is 0.043, which is very close to the optimal. The AUC is much higher for the baselines: 0.106 for NCC, 0.085 for AML, and 0.078 for LRD. *Our method is superior to all other methods on every stereo pair*, while its average AUC is roughly one half of that of the baseline methods. Figure 3 shows the AUC obtained by each method for all images.

	Correct Disparity		Incorrect Disparity	
Image	Y < 0.5	Y ≥ 0.5	Y < 0.5	Y ≥ 0.5
Aloe	4,377	106,143	16,113	5,805
Baby1	1,934	119,735	10,074	3,210
Books	7,612	108,181	21,335	8,824
Cloth1	554	130,283	5,993	174
Lampshade1	9,539	82,016	33,005	8,847
Lampshade2	7,456	84,364	32,910	7,501
Wood1	3,052	125,435	11,711	3,843
...	...	...	...	...
TOTAL	130,142	2,756,764	601,110	177,227
ACCURACY		95.49%	77.23%	

Table 1. Prediction accuracy of our classifier on WTA disparity assignments for non-occluded pixels by thresholding the prediction at 0.5. The second and third column correspond to correctly classified pixels in each class, while the first and fourth to misclassifications. We show raw pixel numbers here to highlight the inhomogeneity of the disparity error rate across images. The last row shows the prediction accuracy for pixels with correct and incorrect disparities over all 27 stereo pairs. *The overall accuracy of the classifier is 91.6%.*

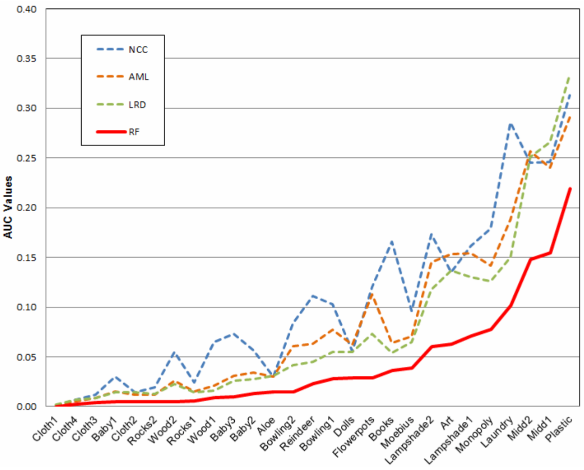


Figure 3. AUC values obtained by sorting the disparity assignments according to NCC, AML, LRD and the RF prediction (solid red curve). Disparity maps have been sorted in order of increasing AUC to aid visualization. *Our method achieves the minimum AUC for every stereo pair.*

6. Detection of Ground Control Points

In this section, we present an approach for selecting ground control points (GCPs), which will be used in the next section to improve WTA disparity maps via global optimization. Consistent with earlier definitions, a GCP here is defined as a pixel with a disparity assignment that is assumed to be very reliable and, therefore, can be used to influence neighboring pixels. We present a principled way of detecting such points using the RF predictions of the previous sections. Quantitative results in Section 7 demonstrate that our approach succeeds in the main challenge when selecting GCPs: *the trade-off between density and accuracy*. If GCPs are not accurate and contain many pixels with wrong disparities, these errors will be propagated to neighboring pixels and can have a strong negative effect on overall accuracy. See, for example, some of the results produced by the baseline methods in Fig. 5. On the other hand, if GCP detection is overly conservative, the small number of selected GCPs has little effect on overall accuracy, since they do not appear in uncertain regions of the images.

The goal is to achieve the highest possible density of GCPs while including a very small fraction of wrong matches in the set. Since the random forest has proven very effective in ranking disparity assignments in order of reliability, we chose GCPs by learning a threshold on the RF prediction that resulted in the highest overall disparity accuracy after MRF optimization. The threshold was learned using cross-validation. It was set to 0.7 and remained constant throughout all experiments.

We compared the GCPs selected by our approach with several alternatives, both in terms of density and accuracy of the GCPs (Table 2) and in terms of accuracy of the resulting, MRF-optimized disparity maps (Section 7). GCPs in Table 2 were selected by choosing pixels that exceeded a threshold in NCC, LRC, LRD or RF prediction. All thresholds were determined by cross-validation. The RF predictions are clearly superior in terms of final disparity map accuracy, but also in terms of GCP accuracy. In fact, the very small fraction of errors in the GCPs is what enables our method to outperform the baselines after MRF optimization.

Our method was successful in addressing a major challenge in GCP selection: on one hand, stereo pairs, for which WTA stereo works well, often have their accuracy degraded by regularization which may over-smooth details, while, on the other hand, stereo pairs for which WTA stereo performs poorly require more regularization and small GCP sets to avoid including errors in them. The RF scores are more flexible in automatically adapting to the inherent difficulty of each stereo pair. The density of GCPs is above 92% for the easy Cloth images and below 50% for harder images, such as Midd1, Midd2 and Plastic. Baseline methods lack this flexibility.

Despite the accuracy of detected GCPs, we chose not to impose them as hard constraints on the MRF. Among several alternatives, we decided on the following that was proven to be superior experimentally. When the random forest predicted that a given disparity assignment to a pixel was reliable, we set the cost of all other disparities for the pixel to a constant value c_{GCP}, leaving the cost for the selected disparity unchanged. Using cross-validation as above, it was determined that the most effective value for the cost of disparities that have not been selected was $c_{GCP} = 2$. This allowed the MRF to override the GCPs, at a higher cost, and was more effective than setting these costs to infinity. The cost of all disparities of non-GCPs remained unchanged in the [-1, 1] range of negated NCC.

Stereo pair	GCP Selection	Accuracy	Density
Plastic	NCC	84.0	50.3
	LRC	91.2	48.5
	LRD	91.4	16.0
	RF	99.2	25.2
Midd1	NCC	87.1	64.5
	LRC	90.2	65.8
	LRD	88.9	25.9
	RF	98.5	47.1
Average	NCC	94.0	89.8
	LRC	98.0	81.2
	LRD	98.2	43.4
	RF	99.7	73.4

Table 2. Accuracy and density of GCPs over non-occluded pixels. Our method (RF) is compared against three baselines: the matching score (NCC), LRC and LRD. GCPs were chosen if NCC$>$ 0.5, LRC$=$ 1, LRD $>$ 100 or RF $>$ 0.7, respectively. All thresholds were learned via cross-validation on the final disparity maps after global optimization. Shown are results on: Plastic, on which RF achieves its minimum density, by far; Midd1 on which RF achieves its lowest accuracy; and averages on all 27 stereo pairs.

7. Globally Optimized Disparity Maps using GCPs

The random forest, comprising 50 trees, was trained using three-fold cross validation as described in Section 5. The MRF minimizes an energy function with the data and smoothness terms denoted by E_{data} and E_{smooth}, respectively. The former is equal to the negated NCC values modified according to the previous paragraph. The latter follows a simple Potts model with edge weights modulated by the strength of the intensity edges between neighboring pixels. We used the implementation of Komodakis [11] and partially adopted the settings of Wang and Yang [25] and defined the smoothness energy of the disparity map D as:

$$E_{smooth}(D) = \lambda \sum_{p \in I_L} \sum_{q \in N_4(p)} \omega_{pq}[d_p \neq d_q], \qquad (3)$$

where p is a pixel in the left image I_L with disparity d_p, q is a pixel in p's neighborhood with disparity d_q, λ is a parameter and the edge weights are defined as:

$$\omega_{pq} = max\{e^{-\frac{\Delta c_{pq}}{\gamma_c}}, 0.0003\}, \qquad (4)$$

with Δc_{pq} the Eulidean distance of the RGB values of p and q, and γ_c equal to 3.6. The data term is set as described at the end of the previous section. These settings are constant regardless of how the GCPs were chosen.

Figure 4 presents the relative error rates of the final disparity maps after MRF optimization using our method compared to four baselines: a basic MRF optimizer without GCPs, as well as MRFs with GCPs selected as the pixels with the highest NCC, LRC or LRD values. Absolute error rates can be seen in Table 3. The values for c_{GCP} and λ and the threshold for each method were determined by cross validation. Our results show significant improvements in accuracy compared to all baseline methods. Sensitivity to the parameters was low in general. Changing the RF prediction threshold from 0.7 to 0.6 results in an average error rate of 7.396% instead of 7.394%. Representative disparity maps are shown in Fig. 5.

GCP type	None	NCC	LRC	LRD	RF
Average error	9.84	9.95	10.28	8.69	7.39

Table 3. Error rates of the final disparity maps after MRF optimization. Our method (RF) is superior to a basic MRF without GCPs and MRFs with GCPs determined according to various criteria.

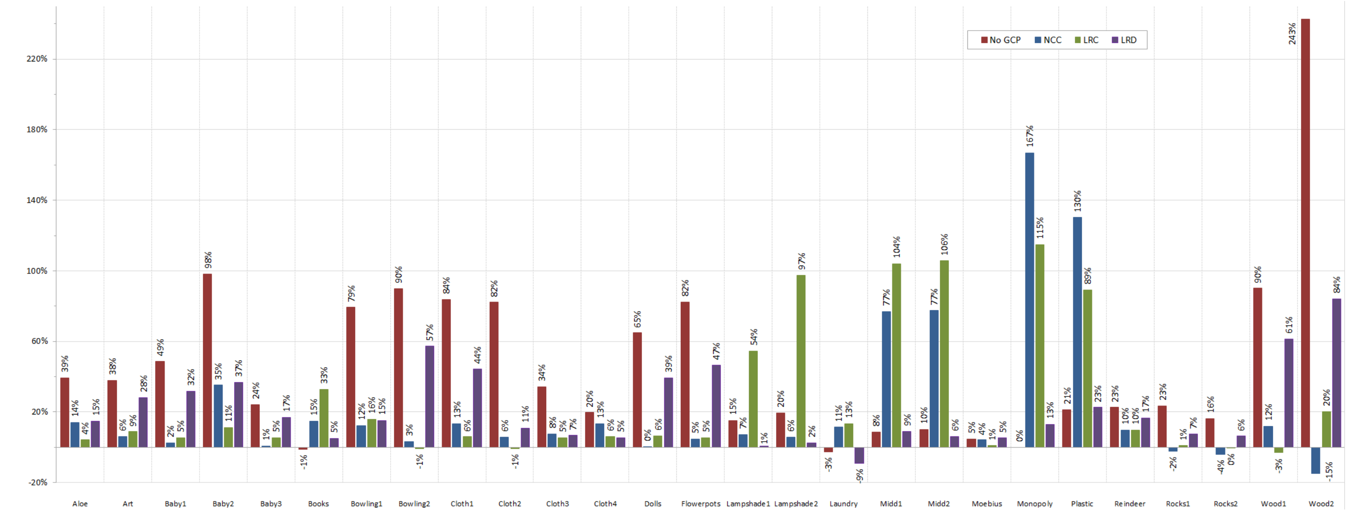


Figure 4. Relative difference of error rates between our method and the baselines after MRF optimization. The first bar for example, represents $(\varepsilon_{none} - \varepsilon_{\text{RF}})/\varepsilon_{\text{RF}}$, which is the increase in error rate between an MRF without GCPs and one with GCPs selected according to RF on Aloe. The difference is a 39% increase. Four bars corresponding to no GCPs and GCPs selected using NCC, LRC and LRD are shown in red, blue, green and magenta respectively.

On the 2005 Middlebury benchmark (Art, Books, Dolls, Laundry, Moebius, Reindeer), our method achieved an error rate of 10.41%. Other results include those of Hirschmüller and Scharstein [8] who report error rates of 8.13% using SGM and 10.88% using graph cuts, Weinman et al. [26] 16.05%, Li et al. [15] 14.36%, Alahari et al. [1] 13.34%, and Pal et al. [18] 18.22%. It should be noted that, unlike [8] who optimized the choice of cost function, we initialize our algorithm using NCC in small windows.

8. Conclusions

We have presented a supervised learning approach that is able to classify and rank stereo matches according to the likelihood of being correct. Experiments on standard data with ground truth demonstrate 91.6% classification accuracy, as well as ranking accuracy that is much closer to being optimal than any single confidence measure in isolation. We have also presented a stereo algorithm that builds upon the aforementioned capabilities and global optimization techniques to improve disparity estimation accuracy. To our knowledge, these are the first results that show that disparity maps can be improved using confidence. Being able to achieve the right balance between density and accuracy of the GCPs and their use as soft constraints are important factors in the overall accuracy of our final disparity maps. Only 9 out of 108 baseline disparity maps (4 methods on 27 stereo pairs) are more accurate than our MRF-optimized disparity maps. Moreover, there is only one publication [8] reporting higher accuracy than ours on a subset of the benchmark.

Acknowledgements This research has been supported in part by the National Science Foundation award #1217797 and by Google Inc. via a Google Research Award.

References

[1] K. Alahari, C. Russell, and P. Torr. Efficient piecewise learning for conditional random fields. In *CVPR*, pages 895–901, 2010. 7

[2] A. Bobick and S. Intille. Large occlusion stereo. *IJCV*, 33(3):1–20, 1999. 2

[3] L. Breiman. Random forests. *Machine Learning Journal*, 45:5–32, 2001. 2, 4

[4] J. Cruz, G. Pajares, J. Aranda, and J. Vindel. Stereo matching technique based on the perceptron criterion function. *Pattern Recognition Letters*, 16(9):933 – 944, 1995. 1, 2

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 2

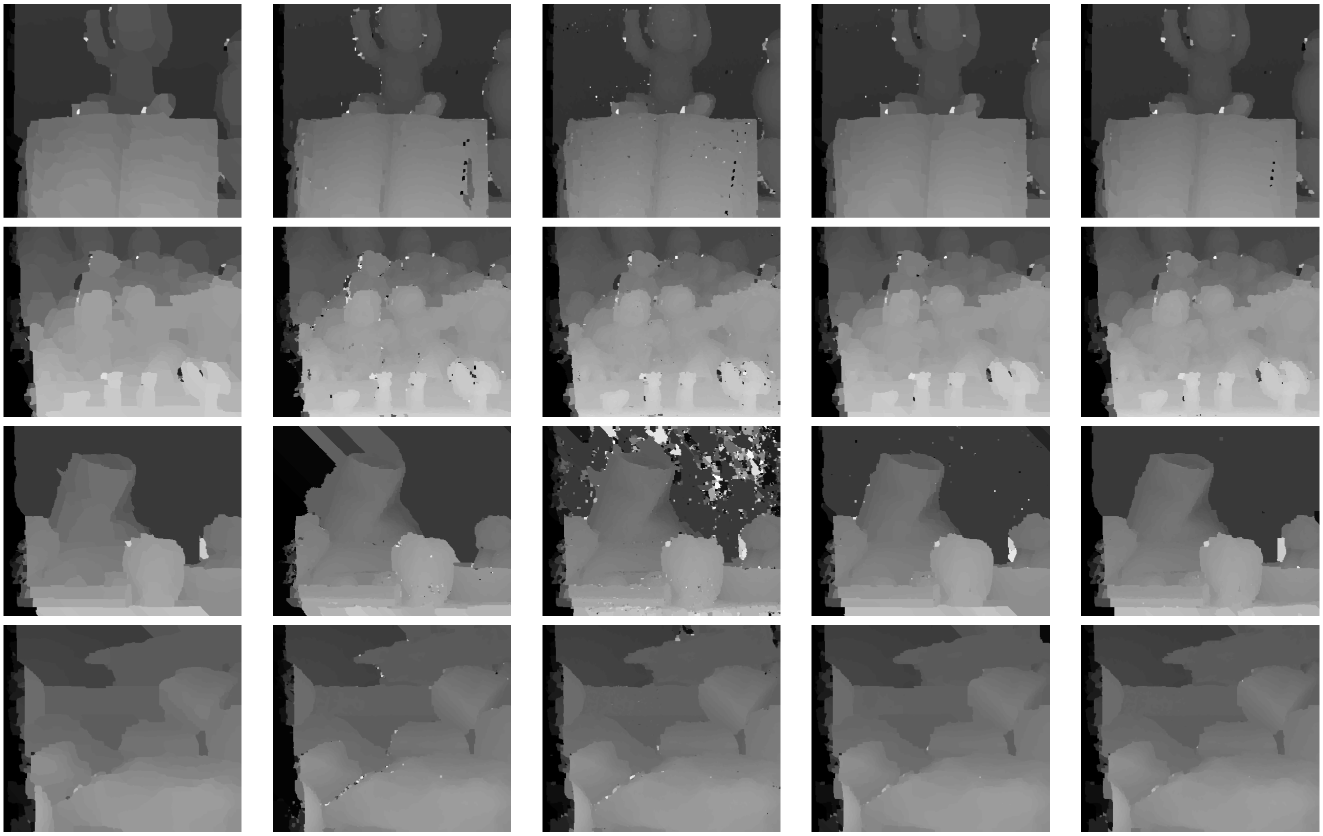
[6] R. Haeusler and R. Klette. Analysis of kitti data for stereo analysis with stereo confidence measures. In *ECCV Workshops*, pages II: 158–167, 2012. 1, 2

[7] R. Haeusler, R. Nair, and D. Kondermann. Ensemble learning for confidence measures in stereo vision. In *CVPR*, 2013. 1, 2, 4

[8] H. Hirschmüller and S. Gehrig. Stereo matching in the presence of sub-pixel calibration errors. In *CVPR*, pages 437–444, 2009. 7

[9] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 34(11):2121–2133, 2012. 1, 2, 3, 4, 5

[10] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *CVPR*, pages 1075–1082, 2005. 2



(a) Basic MRF (b) NCC GCPs (c) LRC GCPs (d) LRD GCPs (e) RF GCPs

Figure 5. Final disparity maps using an MRF without GCPs (leftmost column) and MRFs with GCPs determined according to NCC, LRC, LRD and the RF predictions (left to right). Results are shown for Baby2, Dolls, Midd1 and Rocks1 from [22]. Speckles are due to wrong GCPs that have affected their neighborhoods.

[11] N. Komodakis, G. Tziritas, and N. Paragios. Fast, approximately optimal solutions for single and dynamic MRFs. In *CVPR*, 2007. 6

[12] D. Kong and H. Tao. A method for learning matching errors for stereo computation. In *BMVC*, 2004. 1, 2

[13] D. Kong and H. Tao. Stereo matching via learning multiple experts behaviors. In *BMVC*, 2006. 1, 2

[14] M. Lew, T. Huang, and K. Wong. Learning and feature selection in stereo matching. *PAMI*, 16(9):869 –881, 1994. 2

[15] Y. Li and D. Huttenlocher. Learning for stereo vision using the structured support vector machine. In *CVPR*, 2008. 7

[16] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow. Learning a confidence measure for optical flow. *PAMI*, 35(5):1107–1120, 2012. 2, 5

[17] A. Motten, L. Claesen, and Y. Pan. Trinocular disparity processor using a hierarchic classification structure. In *IEEE/IFIP International Conference on VLSI and System-on-Chip*, 2012. 2

[18] C. Pal, J. Weinman, L. Tran, and D. Scharstein. On learning conditional random fields for stereo: Exploring model structures and approximate inference. *IJCV*, 99(3):319–337, 2012. 7

[19] D. Pfeiffer, S. Gehrig, and N. Schneider. Exploiting the power of stereo confidences. In *CVPR*, pages 297–304, 2013. 2

[20] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. Brostow. Capturing time-of-flight data with confidence. In *CVPR*, pages 945–952, 2011. 1, 2, 5

[21] N. Sabater, A. Almansa, and J. Morel. Meaningful matches in stereovision. *PAMI*, 34(5):930–942, 2012. 1, 2

[22] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *CVPR*, 2007. 1, 2, 4, 8

[23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2, 3

[24] X. Sun, X. Mei, S. Jiao, M. Zhou, and H. Wang. Stereo matching with reliable disparity propagation. In *3DIMPVT*, pages 132–139, 2011. 3

[25] L. Wang and R. Yang. Global stereo matching leveraged by sparse ground control points. In *CVPR*, pages 3033–3040, 2011. 3, 6

[26] J. J. Weinman, C. Pal, and D. Scharstein. Sparse message passing and efficiently learning random fields for stereo vision. Technical Report UM-CS-2007-054, University of Massachusetts Amherst, 2007. 7