Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching
Aristotle Spyropoulos, Nikos Komodakis, Philippos Mordohai

To cite this version:
Aristotle Spyropoulos, Nikos Komodakis, Philippos Mordohai. Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching. Computer Vision and Pattern Recognition 2014, Jun 2014, Columbus, Ohio, United States. 2014, ⟨10.1109/CVPR.2014.210⟩. ⟨hal-01246487⟩
Learning to Detect Ground Control Points
for Improving the Accuracy of Stereo Matching

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

Abstract

While machine learning has been instrumental to the on-going 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 \cite{4, 12, 13} and they have not gained much traction. Very recently, Haeusler et al. \cite{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 \cite{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 \cite{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-
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, 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 states 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.

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 states 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 [12] 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, \( d_{\text{min}} \) and \( d_{\text{max}} \), 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.

\[
f_{\text{AML}}(x_L, y) = \frac{1}{\sum_{x_R} e^{-\frac{c(x_L, x_R, y) - c_1(x_L, y)}{2\sigma_{\text{AML}}^2}}}
\]

**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.
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 $5 \times 5$ 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. Haesler 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.

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 $5 \times 5$ 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. $\sigma_{AML}$ 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 $A_{opt} = \int_{0}^{1} \frac{d_m - (1-\varepsilon)}{d_m} dd_m = \varepsilon + (1-\varepsilon)ln(1-\varepsilon)$, where $\varepsilon$ 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.

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.

### Table 2. Accuracy and density of GCPs over non-occluded pixels.

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

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.

### Table 3. Error rates of the final disparity maps after MRF optimization.

<table>
<thead>
<tr>
<th>GCP type</th>
<th>None</th>
<th>NCC</th>
<th>LRC</th>
<th>LRD</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average error</td>
<td>9.84</td>
<td>9.95</td>
<td>10.28</td>
<td>8.69</td>
<td>7.39</td>
</tr>
</tbody>
</table>

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 \((\varepsilon_{\text{none}} - \varepsilon_{\text{RF}})/\varepsilon_{\text{RF}}\), 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.


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

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.