Ensemble Learning for Confidence Measures in Stereo Vision

Ralf Haeusler
Computer Science Department
The University of Auckland
r.haeusler@aucklanduni.ac.nz

Rahul Nair, and Daniel Kondermann
Heidelberg Collaboratory for Image Processing
University of Heidelberg
http://hci.iwr.uni-heidelberg.de

Abstract

With the aim to improve accuracy of stereo confidence measures, we apply the random decision forest framework to a large set of diverse stereo confidence measures. Learning and testing sets were drawn from the recently introduced KITTI dataset, which currently poses higher challenges to stereo solvers than other benchmarks with ground truth for stereo evaluation.

We experiment with semi global matching stereo (SGM) and a census dataterm, which is the best performing real-time capable stereo method known to date.

On KITTI images, SGM still produces a significant amount of error. We obtain consistently improved area under curve values of sparsification measures in comparison to best performing single stereo confidence measures where numbers of stereo errors are large. More specifically, our method performs best in all but one out of 194 frames of the KITTI dataset.

1. Introduction

A vast amount of algorithms to solve the stereo problem have been proposed with the target to yield improved error statistics on popular benchmarking datasets. It is now well known that good rankings in benchmarks do not imply satisfying results for challenging image data. Recently, this issue has been approached through definition of a more challenging benchmark [9], and further improvements on performance of stereo solvers are anticipated. However, little attention has been paid to the question whether current solutions in increasingly challenging matching problems are actually reliable. This question becomes more important, with increasing degree of ill-conditioning in a matching task. We illustrate this for the stereo case: If, in a worst case scenario, one of the two cameras fails, dense matching results can be computed, but these are not reliable in any location.

Related areas of mismatches need to be detected. A common method is to match in both directions and evaluate the consistency. We illustrate that this method is not perfect but quite effective, by plotting consistency gaps over disparity errors, see Figure 1.

Applications where accurate stereo confidence measures are essential in raising reliability of computer vision include sparse [19] or dense [16] 3D scene reconstructions.

Proposals have been made in the literature on how matching reliability could be captured [5, 13]. However, each of the proposals incorporate certain weaknesses, that is, these may be suitable only for specific image data and fail in situations where discriminative power for particular matching errors is low. This has initiated attempts to combine several confidence measures with the aim of achieving superior accuracy in detection of bad matching estimates. Previous solutions [14, 17] were based on a very limited set of features capturing confidence and were tested only on data not presenting much challenge to stereo.

In this paper, we employ strong energy based confidence clues and use a larger and significantly more challenging stereo dataset introduced recently [9], where results compare much better to real-world scenarios than was the case with benchmarks proposed previously.

The paper is organized as follows: Section 2 provides a brief overview of related work. Section 3 details challenges in defining confidence for matching tasks, compiles some proposals for stereo confidence definition and introduces new confidence definitions used in this paper. Section 4 explains the machine learning framework used for confidence accuracy improvements. Section 5 describes experiments conducted. Sections 6 and 7 contain results and discussion. Section 8 concludes.

2. Related Work

Kong and Tao [14] proposed a stereo matcher, where distributions of labels for good, bad and foreground fattening affected disparities are estimated in a MAP-MRF framework based on horizontal texture and distances to closest foreground objects drawn from ground truth. Motten *et al.* [17] derived binary confidence labels by learning from a larger set of features amenable to hardware processing using decision trees and ANNs. Both approaches were evalu-

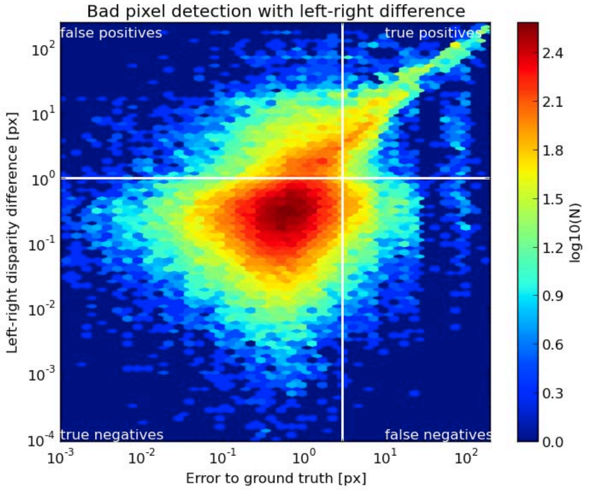

Figure 1. Error of SGM stereo result to ground truth plotted against left-right difference of corresponding points in disparity maps of both views. In challenging images, a significant proportion of bad pixels is not filtered (false negative). Likewise, many good estimates are removed (false positives).

ated only on well-behaved stereo data.

For optical flow, Gehrig and Scharwächter [8] used Gaussian mixtures to model a feature space composed of spatial and temporal flow variance, residual flow energy and structure tensor eigenvalues on small image patches. Mixture coefficients and diagonal covariances were estimated using supervised learning (fifteen classes, defined by intervals of flow end point error) in an expectation maximization framework. Multi-cue confidence was defined as classification outcome according to the highest class posterior. It was not clear whether the combination consistently outperformed single features.

Aodha *et al*. [2] used a similar set of features as Gehrig and Scharwächter [8] to estimate multi-feature flow confidence, but also included image gradient and distance transform on the Canny image of the estimated flow field. Additionally, these features were densely scale-space sampled with a rescaling factor of 0.8. Learning was performed using a random decision forest [6] framework. It was demonstrated, that the combination outperforms single features on many elements of a medium sized dataset of engineered and synthetic images.

Work of Aodha *et al*. makes the assumption that solutions of the flow problem are well defined in general. This is expressed in the idea that a confidence measure can successfully select best fitting results from multiple algorithms, ignoring the fact that flow is often undefined, *e.g*., in areas becoming occluded in subsequent frames.

Quality of results in combinaton-by-classification approaches entirely depends on the strength of contributing features and on the capability of the learning algorithm to deal with correlated variables.

Regarding above mentioned confidence features for optical flow [2], image gradients in conjunction with flow variance are likely to detect lowly textured areas in input images with high variance in flow. Indeed, such flow results are likely to be unreliable in our experience.

However, in stereo and motion alike, reasons for failure may not be restricted to low texturedness. Hence, using a more diversified set of confidence measures as contributing features is very likely to result in improved accuracy for good or bad pixel detection due to consideration for an increased number of possible reasons for algorithm failure.

So, in the following section, we discuss various stereo confidence measures proposed in the literature, and attempt to motivate a selection of most promising measures.

3. Confidence Measures for Stereo

Causes for errors in disparity estimation within a global stereo optimization framework can be based on inappropriate model assumptions, highly nonconvex energies causing multiple strong local minima or numerically instable global minima.

Confidence is commonly understood either as error prediction or only as a measure for uncertainty of results. Assuming error prediction worked out, we would know error magnitudes and could plug these into the stereo estimation model to improve stereo results directly.

However, we can only hope to gain knowlege about suitability of signals to provide good estimates of stereo disparities in most cases, *e.g*., in untextured or repetitive image regions. If this is the case, all we can do is to attempt prediction of potentially large matching errors.

In the absense of a strong theoretical foundation to account for properties of global energies in commonplace stereo aggregation schemes, many spatially local stereo confidence measures have been proposed [5, 13]. However, evaluation has been carried out for a local stereo matching algorithm and on a small dataset only. Also, these measures may be accurate only in specific matching situations. Below we briefly discuss the most prominent proposals for stereo confidence.

To clarify the intention behind defining confidence measures for matching, we would like to point out again, that confidence is not supposed to be a measure for potential disparity error magnitudes. Rather it should be a measure for the likelihood of an algorithm to fail due to high challenges of a specific matching situation. Failing means to exceed a certain error bound. For low confidence matching situations, no improved or specialist algorithm may exist for obtaining a solution. Good confidence measures detect areas that cannot be matched reliably.

In the following definitions, c refers to matching costs resulting from a *Semi global matching* (SGM) [11] aggregation scheme.

Curvature of a parabola fit to matching costs c for subpixel estimation at a pixel $\mathbf{p}$ is frequently considered to be a confidence measure. However, this curvature rarely provides accurate information about gross mismatches. It may only be useful to estimate variances of disparities, given that a match is known to be correct.

The peak ratio measure is widely used in descriptor matching to reject correspondences with close matching costs which are believed to be ambiguous. In the following, d_1 denominates the disparity with lowest associated cost $c(\mathbf{p}, d_1)$ and d_2 is a disparity where $c(\mathbf{p}, d_2)$ is a local minimum with second lowest cost at pixel $\mathbf{p}$. The peak ratio for a disparity at pixel $\mathbf{p}$ is then defined as

$$\Gamma_0(\mathbf{p}) = c(\mathbf{p}, d_1)/c(\mathbf{p}, d_2) \ .$$

Note that with some dataterms (*e.g*. census), image noise can propagate through aggregation and lead to large peak ratios even in reliable matches. This can be the case if $|d_1 - d_2|$ is very small.

Entropy of disparity costs for controlling a diffusion process in cost aggregation [20] attracted some attention as a potential confidence measure. Certainly, flat or noisy cost functions contain little information and are less likely to result in a good correspondence. For defining entropy, costs c need to be normalized into a probability distribution p:

$$\Gamma_1(\mathbf{p}) = -\sum_d p(d) \log p(d) \quad \text{with} \quad p(d) = \frac{e^{-c(\mathbf{p},d)}}{\sum_{d'} e^{-c(\mathbf{p},d')}}$$

Merrell *et al*. [16] propose another measure integrating costs for all disparity estimates. We coin it Perturbation measure due to its design target to capture the deviation of cost function c to an ideal function which is large at all locations except at the minimum d_1. The definition is

$$\Gamma_2(\mathbf{p}) = \sum_{d \neq d_1} e^{-\frac{(c(\mathbf{p},d_1) - c(\mathbf{p},d))^2}{s^2}}$$

We found careful scaling with parameter s crucial in avoiding numerical problems related to floating point accuracy.

Though not perfect, as illustrated in Fig. 1, consistency between left and right disparity is an established criterion for identification of mismatches and occlusions [11]. The definition requires disparity maps D^l and D^r of left and right image:

$$\Gamma_3(\mathbf{p}) = \left| D^l(\mathbf{p}) - D^r(\mathbf{p} - (d_1^l, 0)^T) \right|$$

Image gradient determines the ability of data terms to generate distinctive scores. In stereo, low texture along epipolar lines is critical. This motivates the definition of horizontal gradient as a confidence measure:

$$\Gamma_4(\mathbf{p}) = \| \nabla_x I^l(\mathbf{p}) \|$$

Note, however, that estimated depth edges often do not coincide with image gradients due to foreground fattening [18].

Disparity map variance, defined as

$$\Gamma_5(\mathbf{p}) = \| \nabla D_1^l(\mathbf{p}) \|$$

is usually a good indication of problematic correspondences as errors occur often on or near depth discontinuities. However, Γ_5 may be less suitable if used in conjunction with stereo algorithms that frequently locate discontinuities well. This may be the case in segmentation based stereo approaches.

A measure coined disparity ambiguity here is introduced to capture potential error magnitudes for the case of mismatches resulting from matching ambiguities (which may be detected by peak ratio Γ_0 defined above).

$$\Gamma_6(\mathbf{p}) = |D_1^l(\mathbf{p}) - D_2^l(\mathbf{p})|$$

Although not beneficial as a confidence measure itself, inclusion of disparity ambiguity into a learning framework is an attempt to separate small from large errors in image locations where the peak ratio may fail as explained above.

As an additional confidence measure, we use *Zero mean Sum of Absolute Differences* (ZSAD) matching costs between (left and right) image intensities I^l and I^r for the winning disparity d_1:

$$\Gamma_7(\mathbf{p}) = ZSAD\left(I^l(\mathbf{p}), I^r\left(\mathbf{p} - (d_1^l, 0)^T\right)\right)$$

Another proposal for confidence is what we call semi global energy: We compute the sum of data and smoothness term in a small neighborhood for each pixel, choosing a patch size of 25 × 25 and aggregate along emerging rays in eight directions r for these experiments. The feature is defined in analogy to the SGM objective energy, but with the winning disparity $d_1 = D_\mathbf{p}$ fixed:

$$\begin{aligned}\Gamma_8(\mathbf{p}) = \sum_r \sum_{\mathbf{q} \in r(\mathbf{p})} & c(\mathbf{q}, d_1) + b_1 t(|D_\mathbf{q} - D_{N(\mathbf{q})}| = 1) \\ & + b_2 t(|D_\mathbf{q} - D_{N(\mathbf{q})}| > 1)\end{aligned}$$

Here, $N_\mathbf{q}$ denotes the successor of $\mathbf{q}$ in the set of pixels $r(\mathbf{p})$ along ray r emerging from $\mathbf{p}$. b_1 and b_2 are distinct penalties for different magnitudes of disparity map gradient, and t is a decision function.

Feature Vector Setup

We define one feature vector $\mathbf{f}_7 \in \mathsf{R}^7$, containing only information derived from input images and computed disparity maps. It is defined pixel-wise as follows (we omit argument $\mathbf{p}$ and learning sample indices here):

$$\mathbf{f}_7 = (\Gamma_3^1, \Gamma_4^1, \Gamma_4^2, \Gamma_4^3, \Gamma_5^1, \Gamma_5^2, \Gamma_5^3)$$

Most features are included for three scales with a rescaling factor of two. The notation indicates this with superscripts. Features for lower scales are separately extracted from stereo computed on down-scaled images and not by downscaling of feature maps. Bi-linear interpolation was used for upscaling.

Vector $\mathbf{f}_7$ can be computed for arbitrary stereo results. Another feature vector, $\mathbf{f}_{23} \in \mathsf{R}^{23}$, in addition contains information of spatially aggregated costs, as captured by the features defined above. This feature vector is therefore defined only for stereo schemes with pixel-wise cost computations for each matching candidate. We define:

$$\begin{aligned}\mathbf{f}_{23} = (&\Gamma_0^1, \Gamma_0^2, \Gamma_0^3, \Gamma_1^1, \Gamma_1^2, \Gamma_1^3, \Gamma_2^1, \Gamma_2^2, \Gamma_2^3, \Gamma_3^1, \Gamma_4^1, \Gamma_4^2, \Gamma_4^3,\\ &\Gamma_5^1, \Gamma_5^2, \Gamma_5^3, \Gamma_6^1, \Gamma_6^2, \Gamma_6^3, \Gamma_7^1, \Gamma_7^2, \Gamma_7^3, \Gamma_8^1)\end{aligned}$$

4. Ensemble Learning for Confidence Measures

In the following, we explain the machine learning approach chosen for combining confidence measures. In particular, we motivate to formulate this as a standard classification problem over feature vectors defined in the previous section, that is, estimation of a mapping

$$R : F \mapsto \{-1, 1\} \ ,$$

where R maps to each sample $\mathbf{f} \in F$ one of two class labels, depending on stereo error bounds derived from ground truth. Separate models are created with $\mathbf{f}_7$ and $\mathbf{f}_{23}$ feature vectors, denoted here RDF_7 and RDF_{23}.

We choose a classification approach instead of regression, as confidence measures do not contain matching error magnitude information as explained previously.

Random tree ensembles [6], which have several amenable properties over other learning approaches, are used for this study. Advantages over other classification methods include robustness towards parametrization, low tendency of overfitting data and interpretation of feature relevance.

Each decision tree in the random forest partitions feature space recursively by greedily choosing a feature and a binary test thereupon, which minimizes an entropy based objective function. Once the resulting partition is pure or some other stopping criterion is met, class counts in this partition are recorded. This corresponds to a tree structure that can then be traversed during prediction time from root to the leaf containing the predicted density by performing the binary tests learned during training.

In random tree ensembles, T randomized decision trees are grown independently with two ways of introducing randomization:

- Bagging: Each tree only uses a random subset R of all available samples.
- Random subspace selection: For each space partitioning decision, only the best possible split of a random subset of all possible variables is considered.

During prediction, each tree casts a vote for a class density.

Random tree ensembles can also provide information on variable importance. Two different measures are used here for discussion [3]: The GINI importance measures the contribution of each variable to the decrease of the objective function, while the permutation importance calculates the decrease in accuracy on the out of bag samples after permuting the values of the feature.

5. Experiments

Stereo estimates are computed using semi global matching stereo [11] (penalties $b_1 = 20$, $b_2 = 100$) with a binary census data term on 7×7 matching windows. The choice of this algorithm is due to best overall performance on unconstrained image data in terms of stereo accuracy [12, 21] as well as computational costs low enough for on-line results in, *e.g.*, automotive applications [7]. We restrict our experiments to this powerful stereo algorithm, as we are not interested in stereo errors introduced through weak models. We intend to work only on genuinely hard matching problems.

We select training data from a few frames of KITTI with depth ground truth available, consisting of laser range finder measurements aggregated over five consecutive frames using ego-motion compensation [9]. In an effort to reduce adaptation to a specific matching problem domain, these frames are selected such that a variety of different challenges are posed to the stereo algorithm, including textureless areas, very large baseline, repetitive structures, transparencies and specular reflections. In particular, these frames are those containing following numbers in their filenames: $43, 71, 82, 87, 94, 120, 122, 180$. Samples of the above described feature vector are collected only in locations where data term values for stereo matching are available (that is, these are not set to be invalid) on all scales and for all disparity candidates. In practice, this excludes areas along image boundaries, in particular where occlusions are present near the left image border. The intention is to avoid biases in learning and classification due to non-uniform scaling of some of the used cost function based features in the presence of undefined matching cost values.

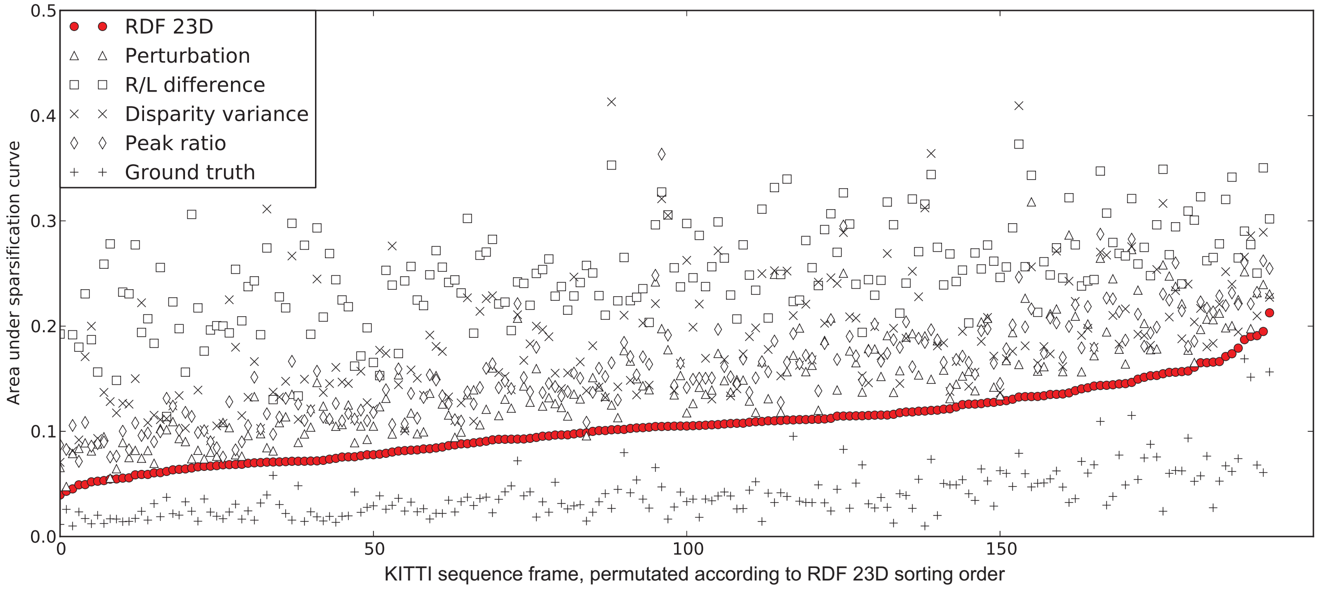

Figure 2. Area under curve measures of our result (red), in comparison to four confidence measures that usually perform best. Lower values are better. The proposed method outperforms other approaches consistently in all but one frame.

As confidence measures generally contain no information about error magnitudes, solving a regression problem for feature combination is not likely to yield the intended results. Therefore, we solve a classification task taking into account only error bounds as follows: Learning samples are categorized into two classes: good and bad disparities. The class boundary is defined by a threshold of 3 px between ground truth disparities and stereo estimates, in line with the default of the KITTI online evaluation. Presumably, higher accuracies of laser range finder measurements cannot be guaranteed.

Due to very high quality of stereo results on KITTI in general, these two classes are highly unbalanced, which may deterioate class model quality and result in unnecesary computational costs due to high data volumes. Therefore, we apply stratified sampling to balance training data. Learning was conducted using the machine learning module of the Vigra library [15]. Generalisation error is monitored within the random forest framework by computing out of bag errors for increasingly large stratified random subsets of the training set.

Decision forest parameters are chosen as follows: Number of trees: $T = 50$, Selection ratio: $R = 0.6$, Minimum sample size in each node to split: $M = 20$. Parameters T, R and M were tuned to achieve an optimum in computational cost and minimize the out of bag error of the random decision forest.

Combined confidence measures for $\mathbf{f}_7$ and $\mathbf{f}_{23}$ alike are defined as the posterior probability of the bad disparity class.

Confidence measures, including decision forest results, are compared using the sparsification strategy: Pixels in disparity maps are successively removed, in the order of descending confidence measure values, until the disparity map is empty. Stereo error measures are computed on remaining pixels in each iteration. If the area under the resulting curve (*AUC*) is smaller than for concurrent confidence measures, it indicates that this measure is more accurate. AUC values are normalized such that confidence measures discarding pixels randomly yield a value of 0.5.

6. Results

Area under the curve (AUC) values for the proposed RDF_{23} confidence measure indicate superior accuracy compared to best performing of all single confidence measures on 193 out of 194 frames on the KITTI dataset, see Fig. 2. Our result is slightly inferior only to the perturbation measure on KITTI Frame 30. The respective sparsification plot for this Frame is displayed in Figure 4. On few other frames (Frames 13,20 and 89), our method is just on a par with the best performing single measure in terms of AUC values.

In the presence of frequent gross stereo errors which are generally detected well by all features including the semi global energy feature proposed, the RDF_{23} results still show a slight improvement, see Fig. 5. Even if a single contributing confidence measure fails (see Fig. 6), results of RDF_{23} are not compromised.

Outstanding accuracy gains from RDF_{23} results are not achieved if the confidence feature set is reduced to such

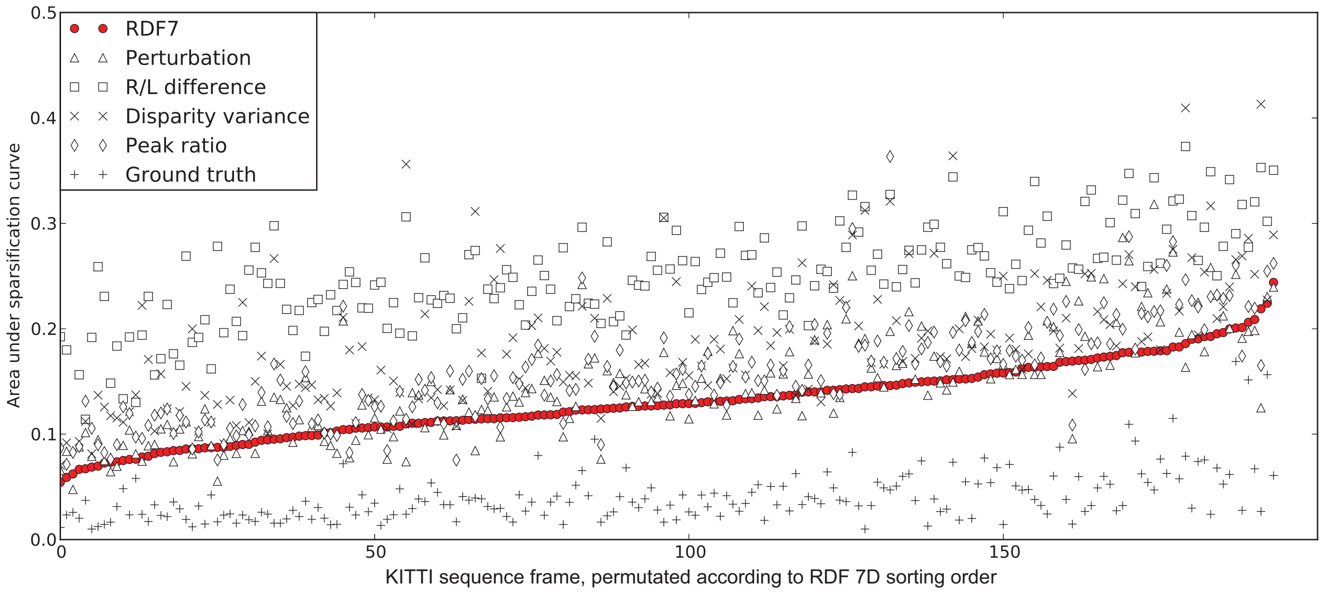

Figure 3. Area under curve measures of our result when the feature set is reduced to information from disparity maps and image intensities. Again, we compare to best performing single confidence measures. Lower values are better. Frequently, results from this reduced feature set are outperformed by single features. This demonstrates that energy based features as included in $\mathtt{f}_{23}$ are essential.

variables that can be obtained solely from disparity maps and image intensities, assuming the stereo algorithm be a black box (see Fig. 3). Still, results are above the average of single features.

In RDF_{23} estimation, disparity variance, perturbation, peak ratio and left-right difference have the largest contribution according to Gini importance in decision forest estimation (see Tab. 1).

In the reduced feature set $\mathbf{f}_7$, Gini importance is highest for the disparity variance variable as well (see Tab. 2).

Possible reductions of false positives and negatives of the proposed method in comparison to the standard consistency check method are illustrated in Fig. 7. A signifant reduction in both, false positives and false negatives, can be observed on depicted road surface and vehicles.

Out of bag errors do not decrease significantly when adding data beyond the choosen training set of size $2.2 \cdot 10^5$.

We provide a complete set of sparsification plots, i.e. plots for each KITTI frame [4].

7. Discussion

Class posteriors of $\mathbf{f}_7$ features yielding inferior results to those of $\mathbf{f}_{23}$ (compare Figures 2 and 3) samples is not surprising, as the main reason for stereo failure detectable by $\mathbf{f}_7$ is textureless areas with co-located disparity discontinuities, which are less frequent in KITTI data, as related objects (*e.g.* sky areas) are not covered by ground truth.

Samples from $\mathbf{f}_{23}$ better cover a wider range of potential matching problems, such as errors at depth discontinuities, despite the most important variable according to the Gini measure in both feature spaces being disparity variance.

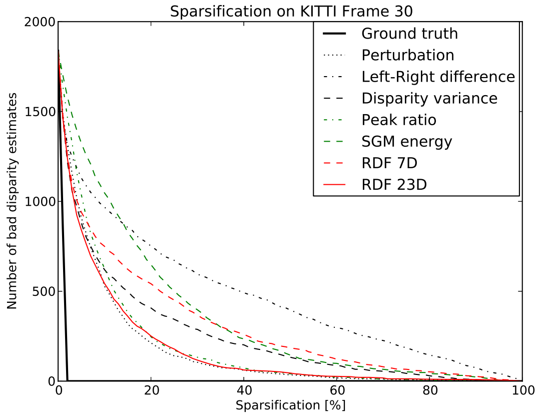

Figure 4. Sparsification plot for the worst result of our method on KITTI data on Frame 30. Note, however, that stereo estimates are almost perfect for this frame. So, this single negative result is little significant.

The perturbation measure attracting higher variable importance on a smaller scale suggests that confidence may be more appropriate to be looked upon at superpixel level.

In opposition to Aodha *et al.* [2], who apply a leave-one-out strategy for learning and testing, we use only a very small fraction of data for training. This is a closer match to applications in practice, where an extensive training dataset

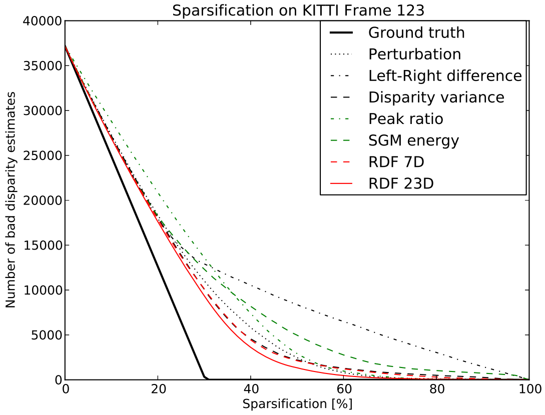

Figure 5. KITTI Frame 123, resulting in a significant amount of SGM stereo errors (approx. 30 percent), results in all confidence measures responding well. Our method still achieves an improvement on top of this.

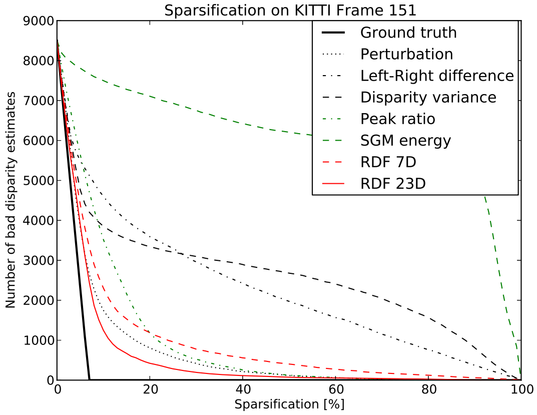

Figure 6. Though one of the contributing measures, SGM energy, fails on Frame 151, our method results in superior accuracy compared to all single measures over the entire sparsification range.

cannot be made available due to prohibitive costs or technical limitations. Still, failure of our method is extremely infrequent. For the only instance on KITTI Frame 30, error rates of the stereo algorithm are very low. In such a case, failure is of no relevance in practical applications.

Even if our class posteriors were only on par with the best single measure in each frame, this would be an advantage as each single measure may fail in some situations.

Undefined stereo values due to occluded regions cannot be handled separately in this study, as corresponding ground truth data is not yet made public in KITTI. However, this does not affect outcomes, as occlusions are simply considered to be a subclass of mismatches. Yet, separate evaluations, as done in stereo benchmarking, would be of interest.

Feature	Scale	Permutation	Gini
Disparity variance	1	0.099	4473
Perturbation	2	0.095	3469
Peak ratio	1	0.045	2548
Left right difference	1	0.046	1533
Perturbation	1	0.080	1474
Peak ratio	2	0.031	1424
Disparity variance	2	0.041	1375
Entropy	2	0.036	1234
ZSAD	1	0.028	1083
ZSAD	2	0.030	1029
Semi global energy	1	0.032	970
ZSAD	3	0.023	837
Entropy	2	0.029	809
Disparity variance	3	0.019	771
Entropy	1	0.023	587
Perturbation	3	0.028	560
Peak ratio	3	0.018	538
Gradient	3	0.011	502
Disparity ambiguity	3	0.006	496
Gradient	2	0.010	426
Disparity ambiguity	2	0.005	396
Gradient	1	0.007	345
Disparity ambiguity	1	0.005	317

Table 1. Variable importance in f_{23}

Feature	Scale	Permutation	Gini
Disparity variance	1	0.135	8095
Disparity variance	2	0.073	5303
Left right difference	1	0.071	4506
Disparity variance	3	0.028	2527
Gradient	3	0.035	1633
Gradient	2	0.047	1463
Gradient	1	0.032	1246

Table 2. Variable importance in f_7

8. Conclusion

We have demonstrated that learning a classifier on multivariate confidence measures is an appropriate approach to increase accuracy in stereo error detection if a suitable set of confidence features is selected. In particular, variance based features on image intensities and matching results as previously applied to the optical flow problem are insufficient for consistently outperforming contributing confidence measures in stereo analysis. This requires strong energy based features. Additionally, we confirm that scale space sampling of features is a crucial contributing factor for success. This suggests, that modeling of spatial dependencies may further improve results.

Apart from bias that may have been introduced due to

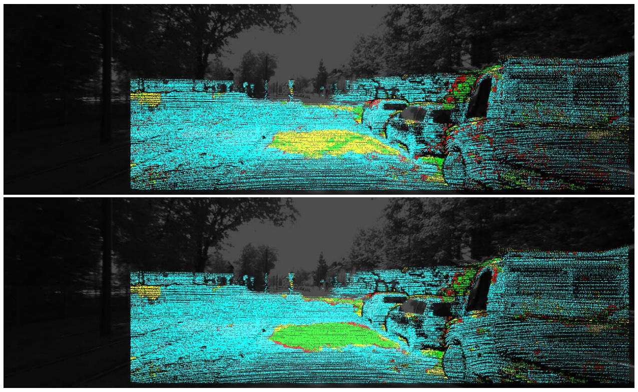

Figure 7. Visualization of true positives (green), false positives (red), true negatives (blue) and false negatives (yellow) according to the denominations given in the plot of Fig. 1 on KITTI Frame 112. Our result, based on 23 dimensional features (bottom), significantly reduces false positives and false negatives compared to left-right difference results (top).

flaws in the ground truth data [10] used here, advantages of the proposed method are larger where stereo is more challenging and hence produces more error prone results. Yet, to shed light on this, new challenges for stereo need to be defined (and come with ground truth), beyond what is present in KITTI data. These challenges could include dark scenes, harsh backlight or any kind of image degradation, including issues resulting from compromised recording equipment. This would help to shift attention to specific problems which need to be addressed before stereo vision systems can confidently be used in applicantions relevant to safety, such as driver assistance systems.

Acknowledgements

This work was supported by Prof. Bernd Jähne and Prof. Akira Nakamura. We thank Dr. Ullrich Köthe for technical support.

References

[1] *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*. IEEE, 2011. 8

[2] O. M. Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow. Learning a Confidence Measure for Optical Flow. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 2, 6

[3] K. Archer and R. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008. 4

[4] Authors. Supplemental material to CVPR 2013. Supplied as additional material `1685-Haeusler-sup.pdf`. 6

[5] J. Banks and P. I. Corke. Quantitative evaluation of matching methods and validity measures for stereo vision. *I. J. Robotic Res.*, 20(7):512–532, 2001. 1, 2

[6] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. 2, 4

[7] S. Gehrig and C. Rabe. Real-time Semi-Global Matching on the CPU. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 85 –92, june 2010. 4

[8] S. K. Gehrig and T. Scharwächter. A real-time multi-cue framework for determining optical flow confidence. In *ICCV Workshops* [1], pages 1978–1985. 2

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Computer Vision and Pattern Recognition (CVPR)*, Providence, USA, June 2012. 1, 4

[10] R. Haeusler and R. Klette. Analysis of KITTI Data for Stereo Analysis with Stereo Confidence Measures. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *ECCV Workshops (2)*, volume 7584 of *Lecture Notes in Computer Science*, pages 158–167. Springer, 2012. 8

[11] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008. 3, 4

[12] H. Hirschmüller and D. Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007. 4

[13] X. Hu and P. Mordohai. A quantitative evaluation of confidence measures for stereo vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 1, 2

[14] D. Kong and H. Tao. Stereo matching via learning multiple experts behaviors. In M. J. Chantler, R. B. Fisher, and E. Trucco, editors, *BMVC*, pages 97–106. British Machine Vision Association, 2006. 1

[15] U. Köthe. The VIGRA Computer Vision Library, Version 1.9.0, 2012. `http://hci.iwr.uni-heidelberg.de/vigra/`. 5

[16] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In *ICCV*, pages 1–8, 2007. 1, 3

[17] A. Motten, L. Claesen, and Y. Pan. Binary confidence evaluation for a stereo vision based depth field processor SoC. In *Pattern Recognition (ACPR), 2011 First Asian Conference on*, pages 456–460, 2011. 1

[18] M. Okutomi, Y. Katayama, and S. Oka. A simple stereo algorithm to recover precise object boundaries and smooth surfaces. *International Journal of Computer Vision*, 47(1-3):261–273, 2002. 3

[19] D. Pfeiffer, S. Gehrig, and N. Schneider. Exploiting the Power of Stereo Confidences. In *CVPR*. IEEE, 2013. 1

[20] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174, 1998. 3

[21] P. Steingrube, S. K. Gehrig, and U. Franke. Performance Evaluation of Stereo Algorithms for Automotive Applications. In M. Fritz, B. Schiele, and J. H. Piater, editors, *ICVS*, volume 5815 of *Lecture Notes in Computer Science*, pages 285–294. Springer, 2009. 4