Direction Matters: Depth Estimation with a Surface Normal Classifier

Christian Häne, L'ubor Ladický , Marc Pollefeys
Department of Computer Science
ETH Zürich, Switzerland
{christian.haene, lubor.ladicky, marc.pollefeys}@inf.ethz.ch

Abstract

In this work we make use of recent advances in data driven classification to improve standard approaches for binocular stereo matching and single view depth estimation. Surface normal direction estimation has become feasible and shown to work reliably on state of the art benchmark datasets. Information about the surface orientation contributes crucial information about the scene geometry in cases where standard approaches struggle. We describe, how the responses of such a classifier can be included in global stereo matching approaches. One of the strengths of our approach is, that we can use the classifier responses for a whole set of directions and let the final optimization decide about the surface orientation. This is important in cases where based on the classifier, multiple different surface orientations seem likely. We evaluate our method on two challenging real-world datasets for the two proposed applications. For the binocular stereo matching we use road scene imagery taken from a car and for the single view depth estimation we use images taken in indoor environments.

1. Introduction

The problem of finding a dense disparity map from a stereo rectified image pair is well studied in the computer vision literature. Despite that, in real-world situations, where images contain noise and reflections, it is still a hard problem. The main driving force of the published techniques is to compare the similarity of image patches at different depths. For the rectified binocular case this is usually defined in terms of a displacement, called disparity, along the image scan lines. Often images contain texture-less areas, such as walls in indoor environments. Matching image patches will not lead to a confident estimate for the depth in this case, as many different disparities lead to low matching costs. Also overexposed spots on images , which frequently occur in real life images makes matching image patches infeasible. Another failure cases of standard approaches are ambiguities in the input data. For example when matching the often slightly reflective ground in indoor imagery it happens that the reflection on the ground is matched better than the often less textured floor.

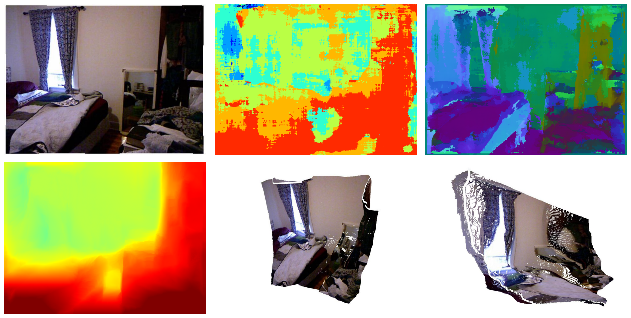


Figure 1. Overview of our method. Top Row: The input to our method is depicted in the top row. On a single input image (left) two classifiers are evaluated, single view depth estimation (middle) and surface normal directions (right). Bottom Row: On the bottom the obtained depth map by our surface normal direction based regularization is shown (left) together with two renderings of the obtained dense point cloud (middle and right).

To tackle all these difficulties, global optimization algorithms have been applied to this problem. It is common to not only include matching costs based on a dissimilarity measures into the energy, but also use image edge information and priors such that planar surfaces are preferred (for example [35]).

Recently, single view depth estimation has also started to be a topic in the computer vision literature. For this problem, strong assumptions, such that there are vertical objects standing on the ground, or data driven machine learning approaches that do not assume a special layout, have been utilized. The advances in machine learning approaches have also lead to classifiers that are able to estimate surface orientation based on a single image. We argue that information about the surface orientation that is extracted from the input image gives additional important cues about the geometry exactly in these cases where standard algorithms struggle. Therefore we propose a global optimization approach that allows us to combine responses of a surface normal direction classifier with matching scores for binocular stereo or scores of a classifier for the single view depth estimation problem. This automatically addresses the problems with standard approaches in stereo matching. In homogeneous area, such as walls, or on the reflective ground the surface normal directions can often be estimated reliably and hence constrain the depth estimation problem to the desired solution. An important feature of our method is that it is not restricted to use a single surface normal direction per pixel but allows the inclusion of the scores from multiple directions, which is important when the classifier is not able to reliably decide on a specific direction. An example result of our method for the single view depth estimation problem is depicted in Figure 1.

We evaluate our approach on two challenging real world benchmark datasets. Qualitative and quantitative improvements over a baseline regularization on the same matching scores but without using the information of the normal direction classifier are reported.

1.1. Related Work

Extracting depth maps out of a potentially noisy matching cost volume has been well studied over the last two decades. Traditionally, the problem is posed as a rectified binocular stereo matching problem. In this setting patch dissimilarity measures (matching scores) are evaluated for a range of disparity values. An overview of such stereo matching methods can be found in [28]. Due to areas in the images that are hard to match, such as texture-less and reflective surfaces, extracting the depth as the best matching disparity leads to unsatisfactory noisy solutions. The key to obtaining smooth surfaces is formulating the problem as an optimization problem with a unary data term (unary potential) based on the matching costs and a regularization term that penalizes spatial changes of disparity. There are approaches considering each scan line [2, 23], methods using dynamic programming over tree structures [3, 34] or methods taking into account multiple paths through the image simultaneously [12].

Formulating the problem as a Markov Random Field (MRF) enables the use of algorithms that find the solution with one global optimization pass [17, 6]. In general such formulations are NP-hard, but for convex priors, using a graph-cut through the cost volume that segments the volume into before and after the surface attains a globally optimal solution [26, 15]. This volume segmentation approach can also be formalized as a continuous cut through the volume [25]; here the regularization is based on the total variation (TV).

Apart from the matching costs, the input images also contain other valuable information that can be used. Depth discontinuities often correspond to image edges and this cue has been used frequently. One important example is [37], in which the anisotropic TV [5] is used to align the image edges and depth discontinuities.

In dense multi-view reconstruction, surface normals can contribute important information. Out of multiple views, a semi-dense oriented point cloud can be extracted [7]. The normal information encoded in these point clouds can be integrated into volumetric surface reconstruction, in order to improve the quality of the extracted 3D model [16]. Another example where surface orientations help to improve dense surface reconstruction is presented in [11]. In this work the 3D reconstruction and semantic labels are estimated jointly using a volumetric approach, thereby each transition between semantic classes gets a different prior on the normal directions. These priors help to reconstruct surfaces which are scarcely observed in the depth maps.

Recently, it has been demonstrated that dense surface normals can also be estimated based on a single image using data driven classification [13, 20]. We propose to use the responses of such a classifier to prefer surface directions, with a good classification score. Our approach is not limited to binocular stereo, as depth estimation from a single image has also become feasible [19]. Unary potentials origination from such a classifier can also be used as input to our method to get smooth depth maps out of a single image.

In the remainder of the paper we will first introduce the convex optimization framework, which we use to extract the depth maps. Afterwards we explain our normal direction based regularization followed by an experimental evaluation on two challenging real-world datasets.

2. Formulation

In this section, we explain the formulation which we are using to extract a regularized depth or disparity map out of a matching cost volume, that for example originates from binocular stereo matching or depth classification based on a single image. Posing this problem as an energy minimization over the 2D image grid poses the main difficulty that the energy is generally non-convex because of the highly non-convex data cost. By lifting the problem to a 3D volume it has been shown that globally optimal solutions can be achieved [37].

More formally, the goal is to assign to each pixel (r, s) from a rectangular domain $\mathcal{I} = \mathcal{W} \times \mathcal{H}$ a label $\ell_{(r,s)} \in \mathcal{L} = \{0, \ldots, L\}$. Instead of assigning labels to pixels directly an indicator variable $u_{(r,s,t)} \in [0, 1]$ for each $(r, s, t) \in \Omega = \mathcal{I} \times \mathcal{L}$ is introduced. Using the definition

$$u_{(r,s,t)} = \begin{cases} 0 & \text{if } \ell_{(r,s)} < t \\ 1 & \text{else}, \end{cases} \tag{1}$$

the problem of assigning a label to each pixel is transformed to finding the surface through Ω that segments the volume into an area in front of and behind of the assigned depth. Adding regularization and constraints on the boundary allow us to state the label assignment problem as a convex minimization problem [37], which can be solved globally optimally.

$$E(u) = \sum_{r,s,t} \left\{ \rho_{(r,s,t)} | \left(\nabla_t u\right)_{(r,s,t)} | + \phi_{(r,s,t)} \left(\nabla u\right)_{(r,s,t)} \right\}$$
$$\text{s.t. } u_{(r,s,0)} = 0 \quad u_{(r,s,L)} = 1 \quad \forall (r,s) \tag{2}$$

The values $\rho_{(r,s,t)}$ are the data costs or also called unary potential, for assigning label t to pixel (r, s), they for example originate from binocular stereo matching. With the symbol ∇_t we denote the derivative along the label dimension t, and ∇ denotes full 3 component gradient. In both cases we use a forward difference discretization. The regularizer $\phi_{(r,s,t)}$ can be any convex positively 1-homogeneous function. This term allows for an anisotropic penalization of the surface area of the cut surface. The main novelty of our algorithm is the use of a normal direction classifier to define the anisotropic regularization term. The boundary constraints on u enforce that there is a cut through the volume.

In the remainder of the manuscript we will use $\mathbf{r}$ and (r, s, t) interchangeably as position index within the volume.

2.1. Normal Classifier Based Regularization Term

The input to the optimization is not limited to unary data costs $\rho_{\mathbf{r}}$. Also the regularization term $\phi_{\mathbf{r}}$ can be dependent on the input data. An important cue for a faithful surface reconstruction is its orientation. Recent advances in data driven classification show that classifying surface normals based on a single image is feasible [20]. For dense stereo matching, surfaces with little texture and/or surfaces seen on a very slanted angle pose problems. In such cases the surface normals can often be estimated reliably and hence contribute crucial information to the optimization problem. Important examples are the road surface in automotive applications or the ground and walls in an indoor environment. In the following we will introduce our proposed approach to including the scores of a surface normal classifier into the above formulation Eq 2.

The classifier outputs a score $\kappa_(r, s, n)$, for each pixel (r, s) of an image, for a discrete set of surface normals n. In order to use this information given by the classifier in the optimization, the cut surface is penalized anisotropically, based on the classifier responses. By this approach surfaces that are aligned with directions having good scores κ, will be preferred by the regularization. A requirement to the regularization term $\phi_{\mathbf{r}}$ is that it is a convex positively 1-homogeneous function. Fulfilling these conditions directly can be difficult. However, this can be tackled by not directly defining the function in the primal but using the notion of the so-called Wulff shape, which defines the smoothness in a primal-dual form [5].

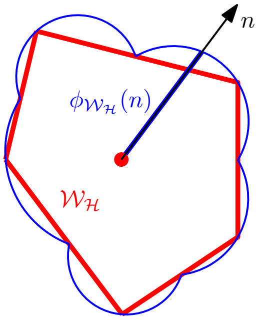

n
$\phi_{\mathcal{W}_{\mathcal{H}}}(n)$
$\mathcal{W}_{\mathcal{H}}$

Figure 2. The red line indicates the outline of the Wulff shape $\mathcal{W}_{\mathcal{H}}$. The distance of the blue line to the origin in direction n, defines the value of the function $\phi_{\mathcal{W}_{\mathcal{H}}}(n)$ and hence the cost of a surface with normal direction n.

$$\phi_{\mathcal{W}}(\nabla u) = \max_{p \in \mathcal{W}} p^T \nabla u, \tag{3}$$

where $\mathcal{W}$ is a convex, closed and bounded set that contains the origin, the so-called Wulff shape. Doing this does not restrict the space of the possible regularization terms as any convex, positively 1-homogeneous function can be defined in terms of a convex shape. With this reformulation of the regularization term the problem of specifying a function has been transformed into specifying a convex shape.

We use a recent idea to form the Wulff shape as intersection of half spaces [10], which allows for a good convex approximation of the input scores κ of the surface normals. Every convex shape can be approximated as the intersection of half spaces. Assume we have a discrete set $\mathcal{H}$ of half spaces with outward pointing normals $n \in \mathcal{S}^2 \subset \mathbb{S}^2$ containing the origin and the distance of the halfspace boundaries to the origin are denoted by d^n. $\mathcal{S}^2$ denotes a discrete subset of the 3-dimensional unit length vectors $\mathbb{S}^2$. The convex shape obtained as an intersection of the halfspaces $\mathcal{H}$ is denoted as discrete Wulff shape $\mathcal{W}_{\mathcal{H}}$. Using the definition that a halfspace $h \in \mathcal{H}$ is active if it shares a boundary with $\mathcal{W}_{\mathcal{H}}$, it follows that for each active half space

$$\phi_{\mathcal{W}_{\mathcal{H}}}(n) = \max_{p \in \mathcal{W}_{\mathcal{H}}} p^T n = d^n. \tag{4}$$

This means setting $d^n_{\mathbf{r}} = \kappa_{(r,s,n)}$ penalizes the directions of the active halfspaces according to the classifier and smoothly interpolates in between. An illustration of this behaviour is given in Fig 2. As the active halfspaces in general correspond to the most likely surface orientations this convex approximation of the scores κ will be most accurate for the best scoring directions.

Before we can plug the regularizer into the formulation, the input normal directions of the classifier that are given in the standard Euclidean space need to be transformed into the space of the volume Ω, in which the cut surface is computed.

2.2. Transforming the Normal Directions

In order to use the normal direction classifier scores $\kappa_{(r,s,n)}$ in the volume Ω we need to derive the mapping of the normal directions n from the standard Euclidean space to Ω. Although this mapping depends on the actual application and is different for single view depth estimation and binocular stereo, the general recipe to derive it is the same. First the transformation of a vector from one space into the other is derived. This is done by taking the Jacobian of the mapping of a point. We call this transformation matrix M. The transformation of the normal directions is then given as $N = \left(M^{-1}\right)^T$ [32, Appendix C].

Binocular Stereo: In binocular stereo matching the points (x, y, z) get mapped to (r, s, t) by

$$\begin{pmatrix} r \\ s \\ t \end{pmatrix} = \begin{pmatrix} f_x \frac{x}{z} + c_x \\ f_y \frac{y}{z} + c_y \\ f_x \frac{b}{z} \end{pmatrix}, \tag{5}$$

where f_x and f_y are the focal length in x and y direction in pixels and b denotes the baseline of the stereo rig. The depth labels t correspond to disparities. We finally get

$$N = \begin{pmatrix} \frac{z}{f_x} & 0 & 0 \\ 0 & \frac{z}{f_y} & 0 \\ -\frac{xz}{bf_x} & -\frac{zy}{bf_x} & -\frac{z^2}{bf_x} \end{pmatrix} \tag{6}$$

Single View Depth: It has been pointed out that by scaling an image and checking if in the scaled image a patch corresponds to a chosen cannonical depth z_0, single view depth estimation becomes feasible [19]. This means, if a pixel is classified having depth z it should have depth z/α if the image is scaled by α. We derive that a point (x, y, z) gets mapped to (r, s, t) by

$$\begin{pmatrix} r \\ s \\ t \end{pmatrix} = \begin{pmatrix} f_x \frac{x}{z} + c_x \\ f_y \frac{y}{z} + c_y \\ \log_\alpha(z_0) - \log_\alpha(z) \end{pmatrix}, \tag{7}$$

where f_x and f_y are the focal length in x and y direction in pixels and the canonical depth z_0 anchors the logarithmic single view depth labels. The transformation of the normal direction is then given as

$$N = \begin{pmatrix} \frac{z}{f_x} & 0 & 0 \\ 0 & \frac{z}{f_y} & 0 \\ -x\ln(\alpha) & -y\ln(\alpha) & -z\ln(\alpha) \end{pmatrix} \tag{8}$$

From the transformation matrices we can see that the transformed normals change along the viewing rays. Hence a different discrete Wulff shape $\mathcal{W}_{\mathcal{H}_\mathbf{r}}$ will be needed at each position $\mathbf{r}$ in the volume.

2.3. The Final Optimization Problem

Before we plug the normal direction based regularizer into the formulation we need to make two remarks:

- The classified normal directions n are based on clustering the training data. Therefore there is no regular sampling of all the possible directions. This can lead to long thin corners when intersecting the half spaces. To avoid overpenalizing these directions we limit the maximal cost of any direction by intersecting the discrete Wulff shape with the unit ball $\mathbb{B}^3$, meaning a sphere with radius 1 containing its interior. We are normalizing the scores of the classifier such that a cost of 1 corresponds to a very unlikely direction.

- Our formulation naturally allows the inclusion of image edge information to the regularization. This allows us to handle surfaces that connect depth discontinuities and as such cannot be handled by the normal direction classifier. However, often a depth discontinuity is present in the input image as an edge. In this case the cut surface normal should be aligned with the image gradient direction or the negative image gradient direction and the viewing ray should lie on the cut surface in the region of the discontinuity. An algorithm that prefers such an alignment in the formulation Eq. 2 has been presented in [37]. In our model we can nicely include this preference by adding the two normal directions (image gradient and negative image gradient) into the discrete Wulff shape with a score $\kappa = k_1 + k_2 e^{-\|\nabla I\|/k_3}$ based on the strength of the image gradient $\|\nabla I\|$ and the parameters k_1, k_2 and k_3.

The combined Wulff shape formed of the intersection of $\mathcal{W}_{\mathcal{H}_\mathbf{r}}$ and $\mathbb{B}^3$ can now be stated as:

$$\phi(\nabla u) = \max_{p \in \mathbb{B}^3 \cap \mathcal{W}_\mathcal{H}} \left\{ p^T(\nabla u) \right\}. \tag{9}$$

For the ease of notation we dropped the position index $\mathbf{r}$. Before we plug everything together and write the final optimization problem in its saddle point form for minimization with the first order primal-dual algorithm [24] we rewrite the above smoothness term to

$$\begin{aligned} \phi(\nabla u) = & \max \left\{ p^T \nabla u \right\} \\ \text{subject to} \quad & p = q, \quad p \in \mathcal{W}_\mathcal{H}, \quad q \in \mathbb{B}^3. \end{aligned} \tag{10}$$

This avoids a costly projection step to the intersection of a ball and a convex polytope. The final primal-dual saddle point problem can now be stated as

$$
\begin{aligned}
E(u,\eta,p,q)=&\sum_{\mathbf{r}}\left\{\rho_{\mathbf{r}}\left|\left(\nabla_t u\right)_{\mathbf{r}}\right|+p_{\mathbf{r}}^T\left(\nabla u\right)_{\mathbf{r}}+\eta_{\mathbf{r}}^T\left(p_{\mathbf{r}}-q_{\mathbf{r}}\right)\right\}\\
\text{subject to}\quad & u_{(r,s,0)}=0,\quad u_{(r,s,L)}=1,\quad \forall(r,s)\\
& p_{\mathbf{r}}\in\mathcal{W}_{\mathcal{H}_{\mathbf{r}}},\quad q_{\mathbf{r}}\in\mathbb{B}^3,\quad \forall\mathbf{r} \qquad (11)
\end{aligned}
$$

The Lagrange multiplier η enforces the equality constraint on p and q. This primal-dual saddle point energy can now be minimized with respect to u and η and maximized with respect to p and q [24]. The algorithm does gradient descent steps in the primal and gradient ascent steps in the dual followed by proximity steps. The required projection step to $\mathcal{W}_{\mathcal{H}_{\mathbf{r}}}$ can be done efficiently by the procedure given in [10].

3. Results

In our evaluation we demonstrate that our regularization improves quantitatively and qualitatively on binocular stereo matching and single view depth estimation. We start with some notes about our implementation and the used classifiers and then show the results for two applications, binocular stereo matching and single view depth estimation.

3.1. Implementation

For both of our applications we use a normal direction classifier trained on the respective training set, using the same method. The training normal maps are clustered into 40 clusters. The likelihoods of the normal directions are estimated using the boosting regression framework [20] by combining various contextual and superpixel-based cues. The contextual part of the feature vector consists of bag-of-words representations over a fixed random set of rectangles surrounding the pixel [31, 18], the superpixel-based part consists of bag-of-words representations over the superpixel, to which the pixel belongs to [20]. Bag-of-words representations are in both cases built using 4 dense features - texton [22], self-similarity [29], local quantized ternary patterns [14] and SIFT [21], each clustered into 512 visual words. Unsupervised superpixel segmentations are obtained using MeanShift [4], SLIC [1], GraphCut-segmentations [38] and normalized cuts [30]. The regressor is trained individually for 5 colour models - RGB, Luv, Lab, Opponent and Grayscale and the final likelihoods are averaged over 5 independent predictions. For further details we refer the reader to [20].

The output from the classifier for each pixel are the 40 scores for the cluster center's normal direction. We normalize the scores to the interval $[0, 1]$ and use them as an input for our normal based regularization.

As we pointed out earlier at each position $\mathbf{r}$ in the volume Ω we get a different transformation matrix N, that transforms from the standard Euclidean space into Ω. However when looking at the shape that is found as an intersection of the halfspaces $\mathcal{H}_{\mathbf{r}}$ we observe that the neighborhood structure only changes rarely when traversing along a ray. This means we do not need to save this information for all the positions t along a ray. We also observe that often there is a clear best normal or only a few well scoring normals. This lead to the decision to only use the 10 best scoring normal directions per pixel. Additionally, to the classifier output, our normal direction based regularizer also includes image edge information. We compute image gradients for each pixel (r, s) using forward differences and include the two directions aligned with the gradient and its negative to the Wulff shape, if there is a strong enough gradient. The score for these directions are chosen dependent on the image gradient magnitude. The costs $\rho_{\mathbf{r}}$ are application specific and will be described together with the results.

We use the first order primal-dual algorithm from [24] to minimize the energy function. The algorithm does gradient descent steps in the primal and gradient ascent in the dual, followed by proximity operations to project back to the feasible set. The proximity steps are either clamping operations or projections to Wulff shapes that can be derived in closed form. For the proximity step of the discrete Wulff shape we refer the reader to [10].

We compare all our results to a baseline approach using isotropic regularization of the cut surface. For this we simply set the variables $p_{\mathbf{r}} := q_{\mathbf{r}}$ and drop the constraint on $p_{\mathbf{r}}$ in the energy Eq 11.

3.2. Binocular Stereo

We evaluate our method on the KITTI benchmark dataset [8]. The dataset contains 195 rectified binocular stereo pairs for testing and 194 for training. They are taken using a stereo rig with 54cm baseline mounted forward facing on a car. The images are challenging due to the many reflections, cast shadows and overexposed areas in the images. Initially, the benchmark was on grayscale images only. Recently, colorized images were released, which we decided to use for maximal quality of the normal direction classification. The costs $\rho_{\mathbf{r}}$ are computed by evaluating standard image dissimilarity measures for a predefined disparity range. We use the average of two dissimilarity measures computed on the grayscale images which we extracted from the colorized images. For the first dissimilarity measure we compute the image gradients with a Sobel filter in r and s direction. The matching score is the average of the absolute differences of the individual components of the Sobel filtered images over a 5×5 pixel window, similar to [9]. The second one is the Hamming distance of the Census transformed images [36] over a 5×5 pixel window.

Examples of the disparity maps we get on the KITTI benchmark dataset for the baseline algorithm and our normal based regulariztion are depicted in Figure 3, together with the respective error images obtained from the bench-

Error	Out-Noc	Out-All	Avg-Noc	Avg-All
Baseline: Test set average				
3 pixels	7.90 %	9.65 %	1.4 px	1.8 px
Normal direction based: Test set average				
3 pixels	6.57 %	7.54 %	1.2 px	1.4 px
Baseline: Reflective regions				
3 pixels	24.92 %	29.13 %	5.6 px	8.0 px
Normal direction based: Reflective regions				
3 pixels	20.31 %	23.89 %	3.0 px	4.3 px

Table 1. Quantitative results on the KITTI benchmark for a disparity error threshold of 3 pixels

mark website. We observe that the normals mainly help to improve flat surfaces with little texture; this is visible in the error images as darker colored pixels on the ground for the normal based version with respect to the baseline. Also on building facades and walls we see an improvement using the normals. These areas often contain little or ambiguous texture which makes matching based on just a simple image dissimilarity measures very challenging, and hence information about the surface direction helps to better constrain the solution. Also on the left hand side of the images, where matching is not possible, we see an improvement using the surface normal classifier. Looking at the quantitative results of the benchmark given in Table 1, we also see a clear improvement using the normal direction based regularization term. This is especially the case in reflective areas.

3.3. Single View Depth Estimation

For the evaluation of the single view depth estimation task we use the challenging NYU indoor dataset [33]. Here the task is to estimate the depth based on just a single image. We trained a classifier with the method described in [19] using 725 images for training and 724 for evaluation. For the single view depth classifier the labels indicate how often an image has to be downscaled by a factor of $\alpha = 1.25$ to see a patch at the canonical distance of $6.9m$, chosen based on the depth range of the training set. For the regularization we linearly interpolated the classifier scores of the original 7 labels to a total of 46 labels. To quantitatively evaluate single view depth estimation, different error measures have been proposed. We believe that one of the most natural error measures is the absolute difference to the ground truth in terms of label distance. We use the following error measures in our evaluation:

- M1: Abs. difference of label: $|\log_{\alpha}(z_{\text{gt}}) - \log_{\alpha}(z_{\text{res}})|$
- M2: Abs. rel. diff. to the ground truth: $|z_{\text{gt}} - z_{\text{res}}|/z_{\text{gt}}$

In Figure 4 we show qualitative and quantitative results. It is apparent that the normal directions contribute valuable information about the surface direction which is not encoded in the single view classifier. This can be visually seen through details in the results such as table corners and pillows which are visible using our proposed regularization but not in the baseline. The quantitative results also show a clear improvement. As an average over the whole test dataset we managed to decrease the absolute difference of the labels, measure M1, from an initial value of 1.2548 to 1.1724 using our proposed regularization. For the measure M2 we observed a decrease from 0.2878 to 0.2728.

4. Conclusion

In this paper we present an approach to incorporate a surface normal direction classifier into the continuous cut formulation for extracting a depth map from unary potentials for different labels. The strength of our method is that we can use classifier scores for a whole range of normal directions and do not need to choose a normal direction estimate per pixel prior to the final optimization. We demonstrated the benefit of our formulation over a baseline without the normal direction classifier for two different tasks, namely binocular stereo matching and single view depth estimation. For both tasks we used publicly available, widely used benchmark datasets.

The results show that a surface normal direction classifier contributes valuable information to both tasks. This is in agreement with earlier works on single view depth estimation, where directions of surfaces play an important role in the formulation [27]. The significant improvement in reflective areas for binocular stereo matching suggests that also for binocular stereo matching in indoor environments, where the often slightly reflective ground poses problems, this approach could help. For areas such as the ground or even texture less walls state of the art semantic classification seems to work well. Therefore, in the future we want to investigate how other additional cues such as semantic information can be brought into stereo matching while still optimizing a single convex energy.

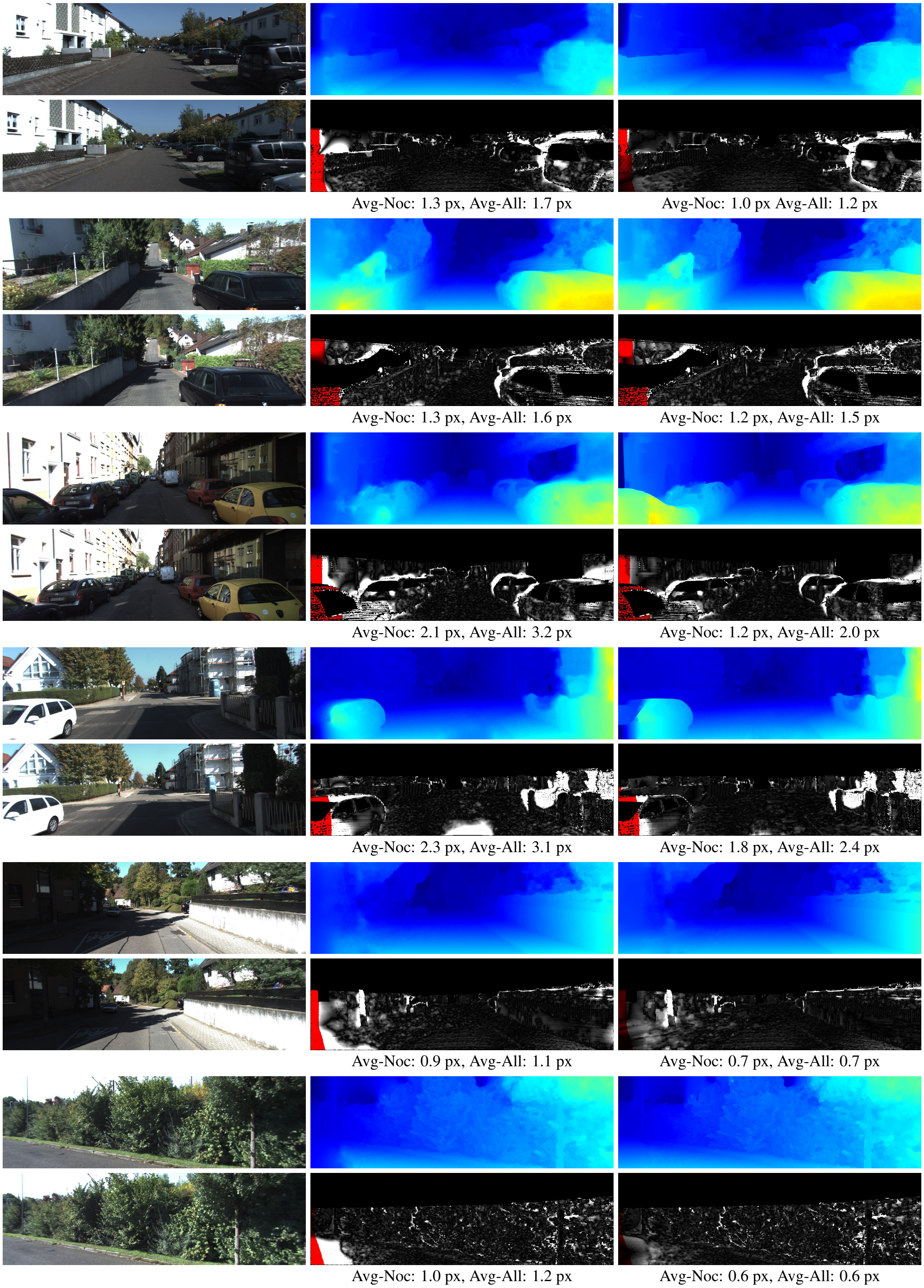
Acknowledgements: We acknowledge the support of the V-Charge grant #269916 under the European Community's Seventh Framework Programme (FP7/2007-2013).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2012. 5

[2] H. H. Baker and T. O. Binford. Depth from edge and intensity based stereo. In *International joint conference on Artificial intelligence*, 1981. 2

[3] M. Bleyer and M. Gelautz. Simple but effective tree structures for dynamic programming-based stereo matching. In



Avg-Noc: 1.3 px, Avg-All: 1.7 px

Avg-Noc: 1.0 px Avg-All: 1.2 px

Avg-Noc: 1.3 px, Avg-All: 1.6 px

Avg-Noc: 1.2 px, Avg-All: 1.5 px

Avg-Noc: 2.1 px, Avg-All: 3.2 px

Avg-Noc: 1.2 px, Avg-All: 2.0 px

Avg-Noc: 2.3 px, Avg-All: 3.1 px

Avg-Noc: 1.8 px, Avg-All: 2.4 px

Avg-Noc: 0.9 px, Avg-All: 1.1 px

Avg-Noc: 0.7 px, Avg-All: 0.7 px

Avg-Noc: 1.0 px, Avg-All: 1.2 px

Avg-Noc: 0.6 px, Avg-All: 0.6 px

Figure 3. Example results from the KITTI benchmark. First column input images, middle column baseline results with error image, right column normal based regularization result with error image. The average disparity error in pixels for the non-occluded areas and the whole image are indicated underneath the respective error image.

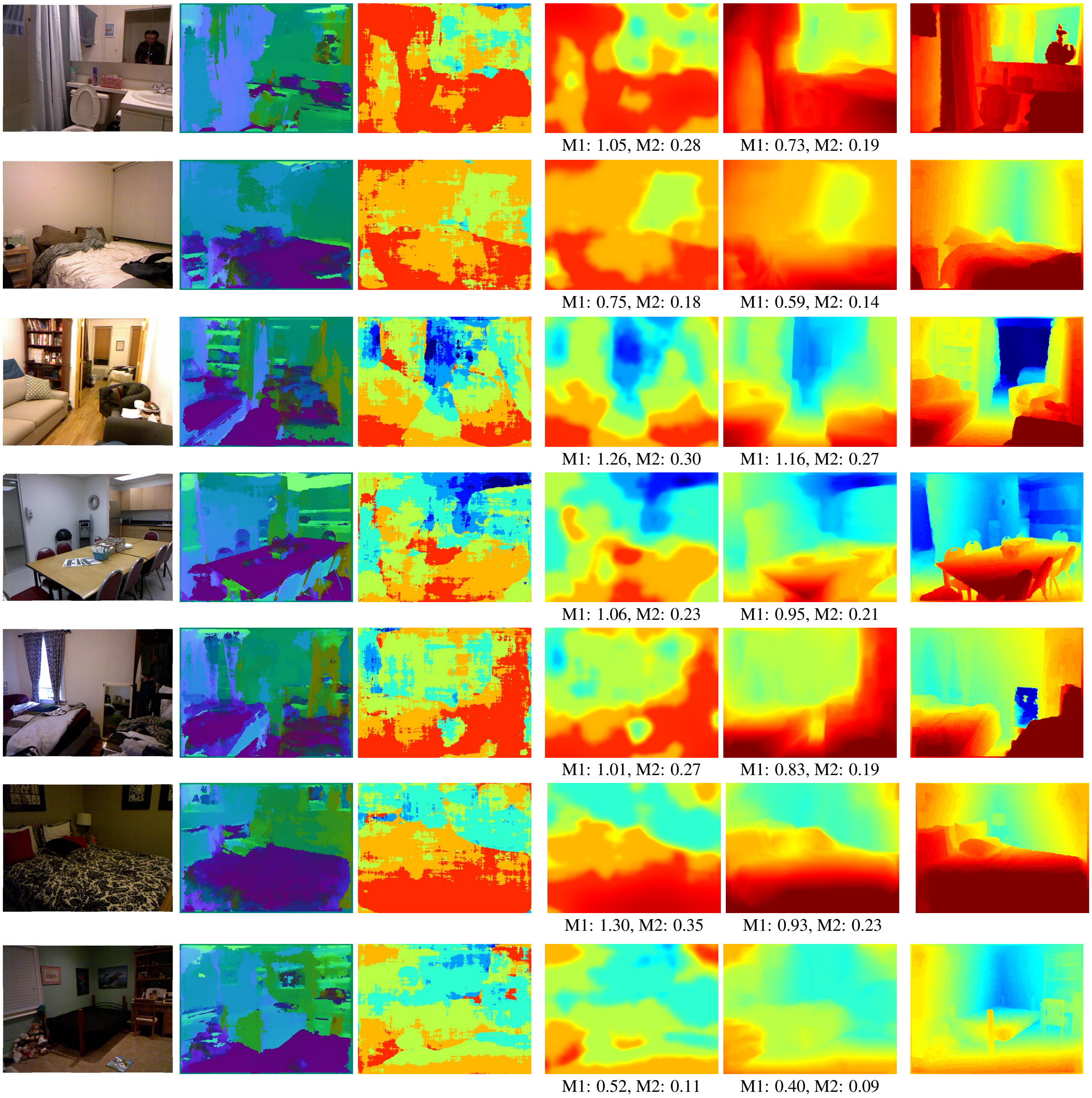


Figure 4. Qualitative and quantitative single view results. From left to right, input image, best cost normals, best cost single view depth, regularized without normals, our proposed regularization with normals, ground truth. M1 denotes the abs. difference of label error measure and M2 denotes the abs. rel. diff. to the ground truth error measure.

International Conference on Computer Vision Theory and Applications (VISAPP), 2008. 2

[4] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2002. 5

[5] S. Esedoglu and S. J. Osher. Decomposition of images by the anisotropic rudin-osher-fatemi model. *Communications on pure and applied mathematics*, 2004. 2, 3

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International journal of computer vision (IJCV)*, 2006. 2

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010. 2

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 5

[9] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision (ACCV)*, 2010. 5

[10] C. Häne, N. Savinov, and M. Pollefeys. Class specific 3d object shape priors using surface normals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3, 5

[11] C. Häne, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3d scene reconstruction and class segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[12] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision (IJCV)*, 2007. 2

[14] S. u. Hussain and B. Triggs. Visual recognition using local quantized patterns. In *European Conference on Computer Vision (ECCV)*, 2012. 5

[15] H. Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2003. 2

[16] K. Kolev, T. Pock, and D. Cremers. Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. In *European Conference on Computer Vision (ECCV)*, 2010. 2

[17] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Eighth IEEE International Conference on Computer Vision (ICCV)*, 2001. 2

[18] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. In *International Conference on Computer Vision (ICCV)*, 2009. 5

[19] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. In *Conference of Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 4, 6

[20] L. Ladicky, B. Zeisl, and M. Pollefeys. Discriminatively trained dense surface normal estimation. In *European Conference on Computer Vision (ECCV)*, 2014. 2, 3, 5

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004. 5

[22] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision (IJCV)*, 2001. 5

[23] Y. Ohta and T. Kanade. Stereo by intra-and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1985. 2

[24] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 4, 5

[25] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. A convex formulation of continuous multi-label problems. In *European Conference on Computer Vision (ECCV)*, 2008. 2

[26] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *International Conference on Computer Vision (ICCV)*, 1998. 2

[27] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2009. 6

[28] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision (IJCV)*, 2002. 2

[29] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 5

[30] J. Shi and J. Malik. Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2000. 5

[31] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European Conference on Computer Vision (ECCV)*, 2006. 5

[32] D. Shreiner. *OpenGL programming guide*. Addison-Wesley, seventh edition, 2010. 4

[33] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012. 6

[34] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[35] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *European Conference on Computer Vision (ECCV)*, 2012. 1

[36] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision (ECCV)*. 1994. 5

[37] C. Zach, M. Niethammer, and J.-M. Frahm. Continuous maximal flows and wulff shapes: Application to mrfs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 3, 4

[38] Y. Zhang, R. I. Hartley, J. Mashford, and S. Burn. Superpixels via pseudo-boolean optimization. In *International Conference on Computer Vision (ICCV)*, 2011. 5