Displets: Resolving Stereo Ambiguities using Object Knowledge

Fatma Güney, Andreas Geiger
MPI Tübingen

Stereo techniques have witnessed tremendous progress over the last decades, yet some aspects of the problem still remain challenging today. Striking examples are reflecting and textureless surfaces which cannot easily be recovered using traditional local regularizers. In this paper, we therefore propose to regularize over larger distances using object-category specific disparity proposals (displets) which we sample using inverse graphics techniques based on a sparse disparity estimate and a semantic segmentation of the image. The proposed displets encode the fact that objects of certain categories are not arbitrarily shaped but typically exhibit regular structures. We integrate them as non-local regularizer for the challenging object class 'car' into a superpixel based CRF framework and demonstrate its benefits on the KITTI stereo evaluation. At time of submission, our approach ranks first across all KITTI stereo leaderboards.

Introduction: In this paper, we investigate the utility of mid-level processes such as object recognition and semantic segmentation for the stereo matching task. In particular, we focus our attention on the reconstruction of well-defined objects for which the data term is weak and current methods perform poorly, such as cars. Due to their textureless, reflective and semi-transparent nature, those object categories represent a major challenge for current state-of-the-art algorithms, as illustrated in Fig. 1 (top). In contrast, as humans we are able to effortlessly extract information about the geometry of cars even from a single image thanks to our object knowledge and shape representation. Inspired by this fact, we introduce object knowledge for well-constrained object categories into a slanted-plane MRF and estimate a dense disparity map. We leverage semantic information and inverse graphics to sample a set of plausible object disparity maps given an initial semi-dense disparity estimate. We encourage the presence of these 2.5D shape samples (or *displets*) in our MRF formulation depending on how much their geometry and semantic class agrees with the observation. Intuitively, displets can be thought of as a representative finite subset of the infinitely large set of possible disparity maps for a certain semantic category conditioned on the image. For example, car displets should cover the most likely 3D car configurations and shapes given the two input images.

Model: We assume that the image can be decomposed into a set of planar superpixels $\mathcal{S}$ and each superpixel $i \in \mathcal{S}$ is associated with a random variable $\mathbf{n}_i$ describing a plane in 3D. $\mathcal{D}$ denotes the set of displets in the image and each displet $k \in \mathcal{D}$ is associated with its class label c_k, a fitness value, and a set of superpixels $\mathcal{S}_k \subseteq \mathcal{S}$ on which it is defined. An additional random variable $d_k \in \{0,1\}$, which can be interpreted as auxiliary variable in a high-order CRF, denotes the presence ($d_k = 1$) or absence ($d_k = 0$) of the displet in the scene. Furthermore, we assume that we have access to a rough semantic segmentation of the image. Our goal is to jointly infer all superpixel plane parameters $\mathbf{n}_i$ as well as the presence or absence of all displets d_k in the scene. We specify our CRF in terms of an energy function

$$E(\mathbf{n},\mathbf{d}) = \sum_{i\in\mathcal{S}} \varphi_i^{\mathcal{S}}(\mathbf{n}_i) + \sum_{i\sim j} \psi_{ij}^{\mathcal{S}}(\mathbf{n}_i,\mathbf{n}_j) + \sum_{k\in\mathcal{D}} \varphi_k^{\mathcal{D}}(d_k) + \sum_{k\in\mathcal{D}} \sum_{i\in\mathcal{S}_k} \psi_{ki}^{\mathcal{D}}(d_k,\mathbf{n}_i)$$

where $\mathbf{n} = \{\mathbf{n}_i | i \in \mathcal{S}\}$ and $\mathbf{d} = \{d_k | k \in \mathcal{D}\}$ and $i \sim j$ denotes the set of adjacent superpixels in $\mathcal{S}$. In addition to the data term $\varphi_i^{\mathcal{S}}(\mathbf{n}_i)$ and pairwise constraints $\psi_{ij}^{\mathcal{S}}(\mathbf{n}_i,\mathbf{n}_j)$, we introduce long-range interactions into our model using displets: The unary potential $\varphi_k^{\mathcal{D}}(d_k)$ encourages image regions with semantic class label c_k to be explained by a displet of the corresponding class. Furthermore, $\psi_{ki}^{\mathcal{D}}(d_k,\mathbf{n}_i)$ ensures that the displet and the associated superpixels in the image are consistent.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

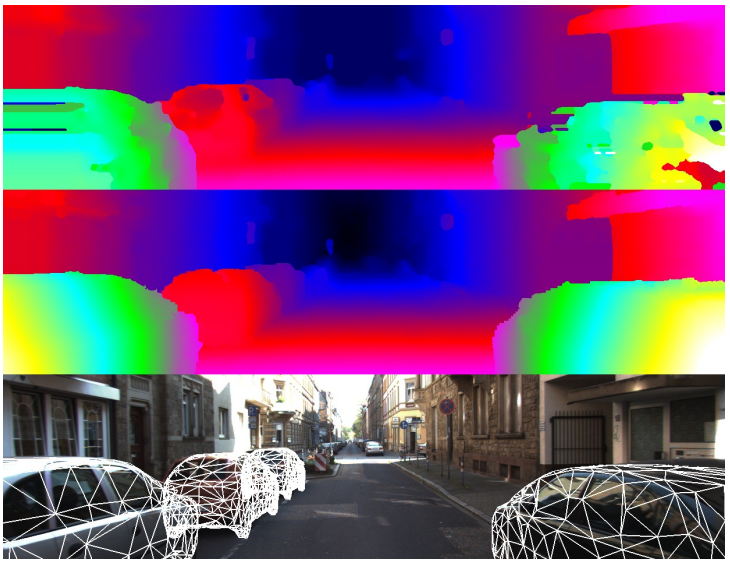

Figure 1: **Resolving Stereo Matching Ambiguities:** Current stereo methods often fail at reflecting or texturelesssurfaces (top, [5]). By using object knowledge, we encourage disparities to agree with plausible surfaces (center). This improves results both quantitatively and qualitatively while simultaneously recovering the 3D geometry of the objects in the scene (bottom).

Results: Our experiments indicate that the proposed framework is able to resolve stereo ambiguities on challenging stereo pairs from the KITTI benchmark [1] as illustrated in Fig. 1 (center). At the same time our method is able to extract 3D object representations which are consistent with the estimated disparity map and may serve as input to higher-level reasoning, see Fig. 1 (bottom) for an illustration. Table 1 and Table 2 show quantitative results on the KITTI stereo benchmark at the time of submission using the default error threshold of 3 pixels. The numbers represent outliers (in %) and average disparity errors (in pixels). Methods marked with an asterisk are scene flow methods which use two or more stereo image pairs as input.

Rank	Method	Out-Noc	Out-All	Avg-Noc	Avg-All
1	Our Method	**2.47 %**	**3.27 %**	**0.7 px**	**0.9 px**
2	MC-CNN [5]	2.61 %	3.84 %	0.8 px	1.0 px
3	SPS-StFl* [4]	2.83 %	3.64 %	0.8 px	**0.9 px**
4	VC-SF* [2]	3.05 %	3.31 %	0.8 px	0.8 px
5	SPS-St [4]	3.39 %	4.41 %	0.9 px	1.0 px
6	PCBP-SS [3]	3.40 %	4.72 %	0.8 px	1.0 px

Table 1: **All Regions**

Rank	Method	Out-Noc	Out-All	Avg-Noc	Avg-All
1	Our Method	**8.40 %**	**9.89 %**	**1.9 px**	**2.3 px**
2	VC-SF* [2]	11.58 %	12.29 %	2.7 px	2.8 px
3	PCBP-SS [3]	14.26 %	18.33 %	2.4 px	3.9 px
4	SPS-StFl* [4]	14.74 %	18.00 %	2.9 px	3.6 px
⋮	⋮	⋮	⋮	⋮	
11	MC-CNN [5]	18.45 %	21.96 %	3.5 px	4.3 px

Table 2: **Reflective Regions**

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[2] Christoph Vogel, Stefan Roth, and Konrad Schindler. View-consistent 3D scene flow estimation over multiple frames. In *ECCV*, 2014.

[3] K. Yamaguchi, D. McAllester, and R. Urtasun. Robust monocular epipolar flow estimation. In *CVPR*, 2013.

[4] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014.

[5] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network, 2014.