Low-level Vision by Consensus in a Spatial Hierarchy of Regions

Ayan Chakrabarti[1], Ying Xiong[2], Steven J. Gortler[2], Todd Zickler[2]
[1]TTI-Chicago. [2]Harvard University.

We introduce a multi-scale framework for low-level vision, where the goal is estimating physical scene values from image data—such as depth from stereo image pairs. The framework uses a dense, overlapping set of image regions at multiple scales and a "local model," such as a slanted-plane model for stereo disparity, that is expected to be valid piecewise across the visual field. Estimation is cast as optimization over a dichotomous mixture of variables, simultaneously determining which regions are inliers with respect to the local model (binary variables) and the correct co-ordinates in the local model space for each inlying region (continuous variables). When the regions are organized into a multi-scale hierarchy, optimization can occur in an efficient and parallel architecture, where distributed computational units iteratively perform calculations and share information through sparse connections between parents and children. The framework performs well on a standard benchmark for binocular stereo, and it produces a distributional scene representation that is appropriate for combining with higher-level reasoning and other low-level cues.

The proposed framework consists of three main components: (1) the global scene map to be estimated, which is a function $Z(n) \in \mathbb{R}^d$ on the two-dimensional image plane, with $n = (x, y)$ indexing discrete spatial locations; (2) a dense set P of overlapping regions $p \in P$ within the image plane, each one a collection of locations n; and (3) a local model that is expected to apply piecewise across most of the scene. The local model is defined in terms of a matrix-valued function $U(n) \in \mathbb{R}^{d \times M}$, and requires that scene values within any inlying region p satisfy $Z(n) = U(n)\theta_p, \ \forall n \in p$, for some $\theta_p \in \mathbb{R}^M$.

Then, estimation requires determining: a) which regions are inliers with respect to the local model, indicated by a binary variable $I_p \in \{0, 1\}$; and b) for all inlying regions, values of the per-region variables θ_p. This is cast as a minimization over $\{I_p, \theta_p\}$ of the following *consensus objective*:

$$L(\{I_p, \theta_p\}_{p \in P}) = \sum_{p: I_p = 0} \tau_p + \sum_{p: I_p = 1} D_p(\theta_p) + \lambda \sum_n |J_n| \ \mathrm{Var}\left[\{U(n)\theta_p\}_{p \in J_n}\right].$$

The first term simply applies a cost τ_p for declaring region p an outlier, and the second scores local variables θ_p in inlying regions using the *data cost* $D_p(\cdot)$, typically measuring the ability of the scene values $U(n)\theta_p$, $n \in p$ to explain the relevant image data. The final term enforces consistency among all inlying regions, by penalizing the variance (weighted by a scalar factor λ) between scene estimates at each pixel n from all p in the set $J_n = \{p : I_p = 1, p \ni n\}$ of inlying regions that contain n.

Compared to traditional approaches based on Markov random fields (MRFs), the consensus framework reasons in a much larger variable space, and more critically, with orders of magnitude more links between variables. This is because it enforces simultaneous consistency between the thousands of regions that overlap any single pixel. Despite this complexity, two properties make estimation not only feasible, but efficient. First, since the dense region-set embodies an over-complete scene representation—with many more internal variables than values in the output scene map—good solutions can often be reached by a simple alternating algorithm similar to expectation-maximization. Second, we show analytically that when the regions are organized hierarchically by scale (*e.g.* Fig. 1 (c)), each region only needs to sum information from its parents and children (Fig. 1 (d)). This leads to a significant reduction in computation because the hierarchical connections constitute only a minuscule fraction of the total links that exist in the consensus objective.

As shown in Fig. 1 (d), the architecture that implements the estimation algorithm is composed of a large network of computational units, one for each region. Regardless of its region's scale, each unit carries out identical operations at each iteration, and these operations happen in parallel at each scale. By sharing information through sparse connections between parents and children, the units collaborate to produce a consistent scene representation over the image plane. From an implementation perspective, this structure allows estimation to be trivially parallelized across multiple cores, as well as across single instruction multiple data (SIMD) channels. Experiments on the binocular stereo problem show that the consensus framework achieves greater accuracy on the KITTI benchmark than comparable state-of-the-art variational and MRF approaches.

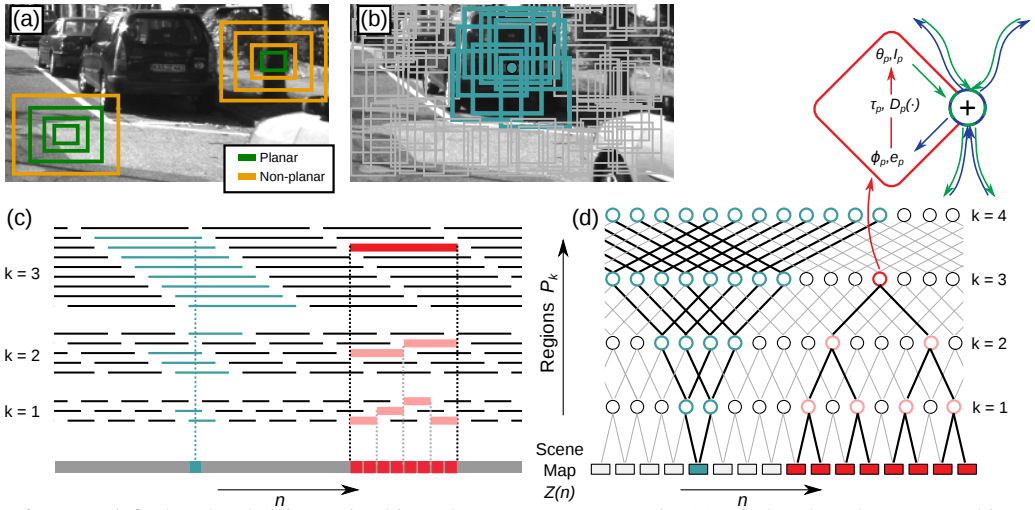

Figure 1: Consensus framework for low-level vision, using binocular stereo as an example. (a) Window-based stereo matching with a slanted-plane model reduces ambiguity, but it requires guessing the correct window shapes and sizes throughout the scene. Consensus addresses this by explicitly considering all regions at all locations (depicted as a 2D cartoon (b) and in 1D organized by scale (c)). It reasons simultaneously about which regions are inliers to the slanted-plane model and the correct slanted plane for each inlying region. The regional slanted planes must agree where they overlap, and in the objective this implies high-order factors that link the variables of thousands of regions that overlap each pixel (blue in (b) and (c)). When regions are organized hierarchically (red/pink in (b)), optimization becomes parallel and efficient. (d) The result is a distributed architecture, with computational units that iteratively perform the same set of computations and share information sparsely between parents and children. The framework can be applied to a variety of low-level tasks using a variety of regional models.

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