Paper Review Summary:

Graphical Models for Stereo from Videos

Xiang Xiang

March 24, 2013

1 Introduction to Stereo Problem

Vision-based stereo has a solid support from the visual perception, is also a fundamental part of the broader 3D reconstruction problem, and is a classical problem in not only 3D computer vision [3, 10, 2] but also photogrammetry [3]. The key task of stereo is to estimate the depth from the camera to any point in a stationery object, or equivalently the disparity of the same object point in two different views. The disparity or depth is the key information for 3D reconstruction, display and visualization, as well as scene understanding or parsing, and robotic perception. Surely, it is natural to think about obtaining depth in a physical way as widely done in photogrammetry, rather than an algorithmic way. There are numerous range sensors in the market: active range scanner, structured light scanner, RGB-D sensors (essentially infrared cameras or formally thermographic cameras) such as Primesense Sensor embedded in the Microsoft Kinect with a projector, laser scanners such as Velodyne LIDAR Sensor used in Google Autonomous Cars, and so on. Once breaking the black-box, we will know that RGB-D cameras rely on either active stereo or time-of-fight sensing to generate depth *estimates* [4]. Nonetheless, they do have application limitations, such as the distance

to the object, precision, noisiness, field of view, restrictive environments, and so forth. For an example, range sensors are seldom embedded in endoscopic sensors for clinical uses. Hence, in this proposal we will focus on vision-based stereo and will not consider range sensors.

Typically, thinking about human vision, the binocular stereo or the formally called stereopsis is stereo's basic problem for estimating the disparity or parallax from two synchronized views with the perspective angle in a certain range, which is usually called 2.5D vision, namely not real 3D. Note that there are some constraints: (1) *synchronized*: no trouble if the object is static; (2) *angle in a certain range*: neither too close nor too faraway if the radius if fixed. Multiple synchronized views can be handled either in a straightforward way as multiple pairs of binocular stereo problems or jointly as a single stereo problem considering the cross-view spatial smoothness or formally photoconsistency, which can be achieved by the rectification or Plane Sweep (*i.e.*, a series of homographies) [11]. Furthermore, continuous overlapping views from video is an extension of multiple-views [5, 14]. In this case, the object or surface of interest is still unique, while the cross-view spatial smoothness becomes cross-frame temporal coherence. We can still assume that adjacent sequential views are synchronized, as long as the video frame rate is *reasonably* high and the angle constrain is satisfied. Moreover, stereo from non-overlapping views in videos is more like a structure-from-motion paradigm. Namely, then we are trying to solve a set of continuous stereo problems. Finally, as a lax intuitive illustration, the generated disparity map is similar to a color-coded segmentation map: pixels in the same region more or less have similar depth, while those in different region normally have different depth. Formally, in stereo, a disparity space image (DSI) is generated and analyzed. Notably, there are works on the monocular depth estimation without range sensors. Some use the de-focus cue or foreground-background segmentation as a prior, while some others employ supervised learning [8], as we will explore in this project.

For the fundamental two-view stereo matching problem, feature matching based epipolar geometric methods are basically solid [3, 10, 2]. There are a systematic series of multi-view

geometric theories to handle various types of view transformation and even the camera's radial distortion [3]. The key is to optimize some measures or statistics, such as minimizing the classic re-projection error or maximizing the photoconsistency normally defined as [11]

$$d(x,y,k) = argmax_d \quad C(x,y,d) = argmax_d \quad \sum_k \left(\tilde{I}(x,y,d,k) - I_r(x,y)\right)^2 \tag{1}$$

where d denotes the disparity, (x, y) is a pixel location, k denotes the camera ID, I_r denotes the intensities of a reference image relating to a chosen reference camera, and $\tilde{I}(x, y, d, k)$ denotes the intensities of a generalized disparity space volume (a group of DSI). The parameterization is normally achieved through a planar homology [3]. In this optimization framework, the routine is the classic Least Squares and its numerous improvements in robust statistics. Normally, there are only a few matched features, so this class of methods are also called sparse correspondence based methods [11]. However, they rely too much on robust and accurate feature matching, which is not always the case. In some applications such as endoscopic reconstruction, the object has distortions or formally nonrigid deformations even over a few frames. Then, the basic assumption of a stationery object does not strictly hold. Besides, illumination conditions can be poor and change a lot, which may result in that the object in some view's image is textureless.

In the recent decade, graph-based global optimization methods has break the domination of the above optimization framework (*e.g.*, maximizing photoconsistency), such as MRF-based Graph cuts [1]. Now, global optimization methods has dominated the top of the Middlebury stereo rankings, and provides the core methods among newly-established more challenging KITTI stereo benchmark's ranking. Graph cuts is a generic optimization framework incorporating problem formalization as well as optimizing the objective by Max-flow/Min-cut [1]. Similar with graph-based segmentation methods, graph-based stereo methods try to assign disparities by minimizing a global function considering both pixel

intensity (*i.e.*, data term) and neighboring links (*i.e.*, smoothness term). One merit of this framework is a certain flexibility to design the objective and constraint, according to the actual problem. This way can be heuristic, while the most commonly seen formulation is

$$E(\mathbf{x}, \mathbf{y}) = \mu \sum_{i} U(x_i, \mathbf{y}) + (1 - \mu) \sum_{i \sim j} V(x_i, x_j, \mathbf{y}) \tag{2}$$

where E is the total MRF energy, U is one data term, V is one smoothness term, μ specifies a weight, x_i is a random variable denoting the disparity of one pixel i, y_i is another random variable denoting the observed intensity of one pixel i, pixel i and pixel j are neighbors, $\mathbf{x}$ denotes a matrix or an array composed of all pixels' disparity, and $\mathbf{y}$ denotes a matrix or an array formed by all pixels' intensity. This optimization framework often produce better performances than epipolar geometric methods, because they can incorporate precise constraints from either defined or learned prior knowledge, and can give a better balance of region and boundary properties [6]. This superiority is not only verified in stereo but also image segmentation, for which Graph cuts dominates the mainstream methods (see Berkeley benchmark evaluations and more challenging PASCAL VOC Segmentation competitions). Since the data representation is in the pixel level, this group of methods are also called dense correspondence based methods [11].

Furthermore, it is natural to think about learning the parameters for the graph energy objective using probabilistic models [15, 9, 13, 7], instead of heuristically setting them. While its flexibility and capability to handle complex data such as video streams is appealing, the primary challenge lies in the lack of training data with ground-truth disparities. Either we try to design unsupervised/weakly-supervised learning methods by estimating disparities, or we have to prepare labeled training data. One drawback of supervised learning is that before the learner can be generalized to output labels for unseen testing data, training data with labels neend to be inputed. Then, when labels of training data are even not available,

researchers will argue: if I can estimate the labels for training data, why not directly doing it for testing data? Empirically, there do exist a trade-of between the cost of learning and direct estimation, both in complexity and performance. In general, it is welcome to pay off a little bit first and win more later. However, there is no easy conclusion till now. That's why we will perform an empirical analysis through this project.

Actually, Zhang and Seitz actually have provide an solution [15], by iterative parameter estimation from previous disparity predictions. Moreover, Scharstein and Pal extend the formulation 2 to a discriminative probabilistic model - CRF. Their learning based stereo has shown a potential usage in the view synthesis for 3D TV display [12].

2 Problem Formulation by Graphical Models

Unlike a classical application of probabilistic graphical models - clinical data parsing, stereo matching is not so intuitive. To encode the domain knowledge from 3D Vision, it is necessary to briefly review its formulation.

CRF is a discriminative extension of MRF, by conditionally normalizing the MRF energy E in Eqn. 2 over all possible values for each x_i and each pixel i as [7].

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{y}) = \frac{1}{Z(\mathbf{y})} \exp\big(-E(\mathbf{x}, \mathbf{y})\big) \tag{3}$$

where the normalizer

$$Z(\mathbf{y}) = \sum_{\mathbf{x}} \exp\big(-E(\mathbf{x}, \mathbf{y})\big) \tag{4}$$

In the above formulas, $E(\mathbf{x}, \mathbf{y})$ is the total graph energy defined in Eqn. 2, X_i is a discrete random variable taking on values x_i from a finite alphabet $\mathscr{X} = \{0, ..., (N-1)\}$, the concatenation of all random variables $\mathbf{X}$ takes on values denoted by $\mathbf{x}$, and $\mathbf{y}$ denotes the conditioning observation [7]. The key distinction between a CRF and a jointly defined MRF

is that the partition function of an MRF does not depend on the observation y and normalizes a joint distribution over the random variables X and a set of random variables Y defined for y. See an elaboration in [7].

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via Graph cuts. In *IEEE ICCV*, 1999.

[2] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *IEEE T-PAMI*, 25(8), 2003.

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[4] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In *International Symposium on Experimental Robotics*, 2010.

[5] R. Koch, M. Pollefeys, and L. V. Gool. Multi viewpoint stereo from uncalibrated video sequences. In *ECCV*, 1998.

[6] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via Graph cuts. In *ECCV*, 2002.

[7] C. J. Pal, J. J. Weinman, L. C. Tran, and D. Scharstein. On learning conditional random fields for stereo - exploring model structures and approximate inference. In *IJCV*, 2012.

[8] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*, 2005.

[9] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *IEEE CVPR*, 2007.

[10] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.

[11] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[12] L. C. Tran, C. J. Pal, and T. Q. Nguyen. View synthesis based on Conditional Random Fields and Graph cuts. In *IEEE ICIP*, 2010.

[13] J. J. Weinman, L. C. Tran, and C. J. Pal. Efficiently learning random fields for stereo vision with sparse message passing. In *ECCV*, 2008.

[14] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. In *ECCV*, 2012.

[15] L. Zhang and S. M. Seitz. Parameter estimation for mrf stereo. In *IEEE CVPR*, 2005.