DENSE STEREO MATCHING USING KERNEL MAXIMUM LIKELIHOOD ESTIMATION

A. Jagmohan, M. Singh, N. Ahuja

University of Illinois at Urbana-Champaign, IL 61801, USA
{jagmohan, msingh, n-ahuja}@uiuc.edu

Abstract

There has been much interest, recently, in the use of Bayesian formulations for solving image correspondence problems. For the two-view stereo matching problem, typical Bayesian formulations model the disparity prior as a pairwise Markov random field (MRF). Approximate inference algorithms for MRFs, such as graph cuts or belief propagation, treat the stereo matching problem as a labelling problem yielding discrete valued disparity estimates. In this paper, we propose a novel robust Bayesian formulation based on the recently proposed kernel maximum likelihood (KML) estimation framework. The proposed formulation uses probability density kernels to infer the posterior probability distribution of the disparity values. We present an efficient iterative algorithm, which uses a variational approach to form a KML estimate from the inferred distribution. The proposed algorithm yields continuous-valued disparity estimates, and is provably convergent. The proposed approach is validated on standard stereo pairs, with known sub-pixel disparity ground-truth data.

1. INTRODUCTION

The dense stereo matching problem is to determine pixel pairs corresponding to common scene points, given two stereo images under a known camera configuration. The pixel pair correspondences are usually represented in the form of a disparity or depth map. Stereo matching finds applications in several vision tasks such as image-based rendering, and 3-D scene reconstruction. The key issues complicating the task of stereo matching include the presence of imaging noise, specularities, occlusions, and textureless regions.

Recently, Bayesian formulations for establishing image correspondences have attracted much interest [1–5], due to the ease with which prior knowledge can be incorporated, and because modeling assumptions explicitly emerge in such formulations. Let $\mathcal{I} = \{I_L, I_R\}$ denote the given stereo image pair, where $I_L, I_R \in \mathbb{Z}^{N_1 \times N_2}$ denote the left and right images respectively. The aim, in stereo matching, is to find a disparity map $D = \{d_i\}_{i \in \{1,\ldots N_1\} \times \{1,\ldots,N_2\}}$, which establishes $\{\mathcal{I}_i\} = \{(I_L(i), I_R(i + d_i))\}$ as the desired correspondences. In practice, the epipolar constraint is typically used to constrain d_i to be one-dimensional. Bayesian approaches formulate this problem as one of inferring the posterior probability distribution $P(D|\mathcal{I}) = P(\mathcal{I}|D)P(D)$. An estimate of the true disparity map can be subsequently obtained from the inferred distribution, using, for example, maximum a-posteriori (MAP) or minimum mean-squared error (MMSE) estimation.

Typical Bayesian formulations model the conditional distribution $P(\mathcal{I}|D)$ as a product of independent marginal distributions $P(\mathcal{I}_i|d_i)$, where the marginal distributions are selected in accordance with standard imaging models (such as the Lambertian model). The disparity prior $P(D)$ is usually modeled as a pairwise Markov random field (MRF) to enforce spatial smoothness of the disparity map [3, 5]. While performing exact inference on MRFs is computationally intractable, approximate inference algorithms based on the use of graph cuts [4, 5], and belief propagation [3], have been shown to yield good performance for the two-view stereo matching problem [6].

These algorithms treat the stereo matching problem as a labelling problem, with the disparity estimates constrained to take values from a discrete set, for e.g., $d_i \in \mathbb{Z}$. The use of such discretized disparity maps in applications such as image-based rendering causes artifacts in the synthesized views [6]. While the fidelity of the computed maps can be enhanced by considering larger discrete sets, the computational complexity scales up quickly—the order of complexity for the belief propagation algorithm, for example, is $O(L^2)$ where L is the number of disparity levels [3].

In this paper we propose a novel Bayesian formulation for the two-view stereo problem. The presented formulation is based on the recently proposed kernel maximum likelihood (KML) estimation framework [7]. The proposed formulation uses probability density kernels to infer the posterior distribution $P(D|\mathcal{I})$. We present an efficient variational approach for finding the KML disparity estimate. The key features of the proposed framework are as follows. Firstly, it does not require the disparity estimates to take values from

The support of the Office of Naval Research under grant N00014-03-1-0107, the National Science Foundation under grant ECS02-25523, and the Computational Science and Engineering Department, UIUC is gratefully acknowledged.

a discrete set. Secondly, it is an exact inferential framework and the proposed variational solution is provably convergent. Finally, the underlying kernel density estimation framework provides robustness to occlusions and outliers. We illustrate the efficacy of the proposed algorithm by comparing its performance to that of a Potts model [8] based formulation, which uses belief propagation for inference.

2. KERNEL MAXIMUM LIKELIHOOD STEREO MATCHING

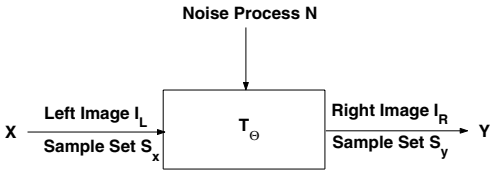

Fig. 1. The underlying model assumed by the KML framework. The left (reference) image is characterized by sample set S_X, the right image is characterized by sample set S_Y. The sample sets are related by a parametric transformation T_Θ, and a noise process $\mathcal{N}$.

We present a novel stereo matching formulation based on our recently proposed kernel maximum likelihood (KML) registration framework (cf. Chapter 7, [7]). The KML approach is a maximum likelihood approach which uses probability kernels for estimating priors for an image from the data (as opposed to a preselected pdf model).

We now explain the proposed formulation using the abstraction shown in Fig. 1. The left stereo image I_L undergoes a transformation T_Θ and a random distortion, represented by the process $\mathcal{N}$, to yield the right stereo image I_R. T_Θ represents the change in the spatial locations of the pixels as well as their intensities. The stereo matching problem is solved by estimation of the parameters Θ, characterizing this change. This task is complicated by the fact that $\mathcal{N}$ is difficult to model, as it includes the combined effects of occlusion (say, $\mathcal{N}_o$) and additive imaging noise $\mathcal{N}_1$.

The KML framework forms non-parametric probability density estimates using kernel density estimation [9]. This approach has the following advantages: (1) Since the estimated image pdf is produced by the aggregate of all image pixels, the effect of a small subset of pixels being occluded (i.e. the effect of $\mathcal{N}_0$) is mitigated. (2) The effect of additive noise $\mathcal{N}_1$ is easily incorporated in the definition of the density kernels. (3) The approach provides us with an estimate of a probability distribution for the input image—this is crucial as real-world objects cannot always be modeled through known parametric density models. (4) It is convenient to add other priors to the resulting Bayesian formulation and variational frameworks are easy to formulate for exact Bayesian inference.

As shown in Fig. 1, we represent I_L and I_R using random variables X and Y respectively, where X, Y are defined over the joint range space and domain space of the respective images. The given stereo pair is treated as independently drawn samples from the distributions, $f_X(x)$ and $f_Y(y)$ of X and Y respectively. The corresponding sample sets are denoted by $\mathcal{S}_X = \{x_i\}_{i=1}^m$ and $\mathcal{S}_Y = \{y_j\}_{j=1}^n$. We further use $\mathcal{S}_D = \{1, \ldots, N_1\} \times \{1, \ldots, N_2\}$, and $\mathcal{S}_R \subset \mathbb{Z}$ to denote the 2-D spatial domain space and the intensity range space of the images, respectively. Then, $x_i = [i; I_l(i)]$, with spatial location $i \in S_D$, and intensity $I_l(i) \in \mathcal{S}_R$. The elements $y_j \in S_Y$ are similarly defined for the right image. For stereo matching, the two sample sets may be considered to be related through a spatially varying transform T_{Θ_i}, which may be defined in terms of the disparity map $D = \{d_i\}$, $d_i \in \mathbb{R}$. We index each sample by its spatial location, and use the Lambertian assumption to define

$$T_{\Theta_i}(x_i) = T_{\Theta_i}([i; I_l(i)]) = [i + [d_i, 0]; I_l(i)] \quad (1)$$

and write the kernel estimate for the conditional density as,

$$P_{\text{KML}}(I|D) \propto \prod_{i=1}^{m} \sum_{j=1}^{n} K(H^{-1}(T_{\Theta_i}(x_i) - y_j)) \quad (2)$$

We model the prior $P(D)$ as a pairwise MRF, as is conventionally done in Bayesian formulations [3, 5], and use a robust difference function $\rho(\cdot)$ for disparity differences,

$$P(D) \propto \prod_{(i,j),\|i-j\|\leq 1} e^{\rho(|d_i - d_j|)} \quad (3)$$

We seek a disparity field $\widehat{D}$ for the left image, which maximizes the posterior likelihood $P(D|I)$, i.e., $\widehat{D} = \text{argmax}_D$ $(\log P(I|D) + \log P(D))$. This can be written as

$$\widehat{D} \quad = \quad \underset{D}{\text{argmax}}(E_{data}(D) + E_{smooth}(D)) \quad (4)$$

Equation (4) explicitly represents the log-posterior distribution as a sum of a data term and a smoothness term. The data term models the requirement that the disparity estimate be consistent with the observed data $\mathcal{I}$, and the smoothness term models the requirement that the disparity estimate be spatially smooth. From (2) and (3), the KML estimate is,

$$\begin{aligned}\widehat{D}_{\text{KML}} &= \underset{D}{\text{argmax}}(\log P_{\text{KML}}(I|D) + \log P(D)) \\ &= \underset{D}{\text{argmax}}(\sum_{i=1}^{m} \log \sum_{j=1}^{n} K(H^{-1}(T_{\Theta_i}(x_i) - y_j)) + \lambda \times \\ &\sum_{\|i-j\|\leq 1} \rho(|d_i - d_j|)) = \underset{D}{\text{argmax}}(P_1(D) + P_2(D)) \quad (5)\end{aligned}$$

(a) (b)

Fig. 2. (a) Left image of *venus* stereo pair. (b) Left image of *map* stereo pair.

3. VARIATIONAL OPTIMIZATION ALGORITHM

The maximization required to be performed in (5) is for a non-convex, nonlinear function defined over a $N_1 \times N_2$ dimensional state-space. We now present an iterative algorithm for performing this optimization based on the following variational principle. In each iteration, a lower bound to the function in (5) is found for a given value of D. The next value of D is then found by maximizing this lower bound.

Consider, first, the smoothness prior term $P_2(D)$ in (5). For a convex function $\rho(x)$, $\rho(x_2) - \rho(x_1) \geq \rho\prime(x_1)(x_2 - x_1)$. Thus, taking $\rho(x) = \alpha \exp(-x^2/h_d^2)$ yields, for some $d_{i,0}, d_{j,0} \in \mathbb{R}$, $\rho(|d_i - d_j|) \geq \rho(|d_{i,0} - d_{j,0}|) + (\alpha/h_d^2) \times \exp(-(d_{i,0} - d_{j,0})^2/h_d^2)((d_{i,0} - d_{j,0})^2 - (d_i - d_j)^2)$. Thus,

$$P_2(D) \geq P_2(D_0) + \lambda \sum_{i=1}^{m} \sum_{j,\|j-i\|\leq 1} v_{ij} \left(k_{ij} - \frac{(d_i - d_j)^2}{h_d^2} \right) \tag{6}$$

where $v_{ij} = \alpha e^{\frac{-(d_{i,0}-d_{j,0})^2}{h_d^2}}$, $k_{ij} = \frac{(d_{i,0}-d_{j,0})^2}{h_d^2}$, and $D_0 = \{d_{i,0}\}$ is an initial disparity map.

Consider, next, the data term $P_1(D)$ in (5). We make use of the concavity of the $\log(\cdot)$ function to find a lower bound for this term. For the case where the kernel is exponential, we get $\log \frac{\sum_{j=1}^{n} K(H^{-1}(T_\Theta(x) - y_j))}{\sum_{j=1}^{n} K(H^{-1}(T_{\Theta_0}(x) - y_j))} \geq \sum_{j=1}^{n} p_j (c_j - \left\| H^{-1}(T_\Theta(x) - y_j) \right\|^2)$ where $c_j = \left\| H^{-1}(T_{\Theta_0}(x) - y_j) \right\|^2$. Finally, for a diagonal bandwidth matrix H, with intensity bandwidth h_I^2, and spatial bandwidths $h_{w_1}^2$, $h_{w_2}^2$, by summing over the samples in S_X we get

$$P_1(D) \geq P_1(D_0) + \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \times \left(c_{ij} - \frac{(I_l(i) - I_r(j))^2}{h_I^2} - \frac{(i + d_i - j)^2}{h_{w_1}^2} \right) \tag{7}$$

where $w_{ij} = \frac{K(H^{-1}(T_{\Theta_0}(x_i) - y_j))}{\sum_{k=1}^{n} K(H^{-1}(T_{\Theta_0}(x_i) - y_k))}$, and c_{ij} are both functions of D_0. Defining $g(D) \doteq P_1(D) + P_2(D)$ and combining (6) and (7) yields

$$g(D) \geq g(D_0) + A(D_0) - \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} \left(\frac{(i + d_i - j)^2}{h_{w_1}^2} \right) - \lambda \sum_{i=1}^{m} \sum_{j,\|j-i\|\leq 1} v_{ij} \left(\frac{(d_i - d_j)^2}{h_d^2} \right) \tag{8}$$

where $A(D_0)$ is defined appropriately.

Two observations can be made from (8). Firstly, both the values and the derivatives of the lower bound and the function $g(D)$ are identical at $D = D_0$. Secondly, the lower bound is easy to maximize. It requires solving the following sparse, linear system of $N_1 \times N_2$ equations in $\{d_i\}_{i=1}^m$

$$\sum_{j=1}^{n} \frac{w_{ij}}{h_{w_1}^2} (i + d_i - j) + \lambda \sum_{j,\|j-i\|\leq 1} v_{ij} \left(\frac{(d_i - d_j)}{h_d^2} \right) = 0 \tag{9}$$

where the weights w_{ij}, v_{ij} are functions of D_0.

To summarize, the iterative algorithm proposed to compute $\widehat{D}_{\text{KML}}$ is as follows: (1) Choose an initial disparity map D_0. (2) Find $\widehat{D}_{\text{KML}}$ by solving the linear system in (9). (3) Stop if $\|D_0 - D_{\text{KML}}\| \leq \epsilon$, else set $D_0 = \widehat{D}_{\text{KML}}$ and goto Step 2. In practice, the bandwidth parameters in the objective function in 5 are scheduled so as to provide a gradual reduction in scale, over the course of the iterations.

The iterative algorithm described above can be shown to converge to a local maximum of the cost function given in (5). The proof is omitted due to lack of space, but the interested reader is referred to the similar proof in ([7], Section 7.3.1). This algorithm provides an efficient computational method to solve the MRF in (5) and yields continuous-valued disparity estimates. In the next section, we present results for the performance of the proposed formulation.

4. RESULTS

To evaluate the performance of the proposed formulation, we use the rectified gray-scale *venus* and *map* stereo pairs, with known sub-pixel disparity ground-truth data [6]. Fig. 2 shows the left image of each stereo pair. Denoting the estimated disparity map by $D^{KML} = \{d_i^{KML}\}$, and the ground-truth by $D^{GT} = \{d_i^{GT}\}$, we use the following metrics to quantitatively evaluate algorithm performance: (1) B, the fraction of pixels for which $|d_i^{GT} - d_i^{KML}| > 1$, (2) $B_{\overline{O}}$, the fraction of pixels in non-occluded regions, for which $|d_i^{GT} - d_i^{KML}| > 1$, (3) $M_{\overline{O}}$, the disparity mean square error (DMSE) for non-occluded pixels, and (4) $M_{\overline{OC}}$, the DMSE for non-occluded pixels which additionally satisfy $|d_i^{GT} - d_i^{KML}| \leq 1$. We compare proposed algorithm's performance to that of a Potts model [8] based MRF formulation, which uses belief propagation for inference.

The results for the *venus* stereo pair are shown in Figures 3(a)-(c). Fig. 3(a) shows the true disparity map, Fig.

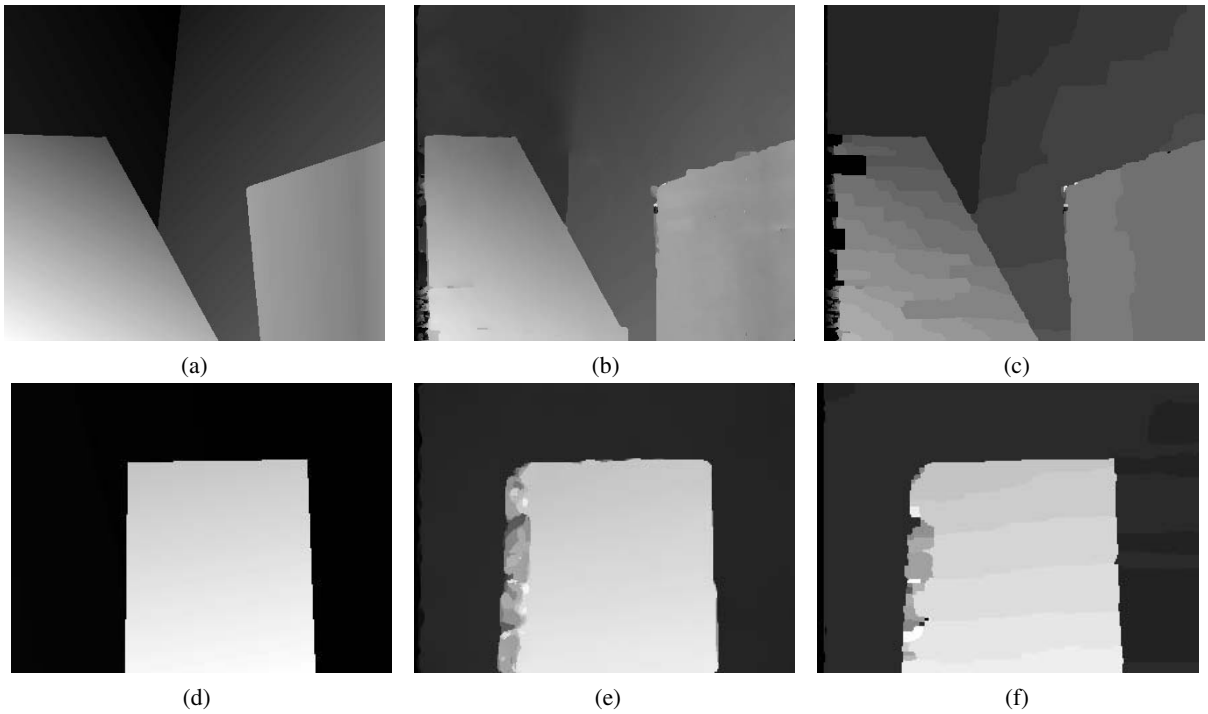

Fig. 3. Disparity maps for *venus* (top row) and *map* (bottom row) stereo pairs. (a), (d) Ground-truth. (b), (e) Proposed algorithm. (c), (f) Belief propagation.

3(b) shows the disparity map produced by the proposed algorithm, and Fig. 3(c) shows the disparity map produced by the belief propagation algorithm. Median filtering was used to eliminate noise in the estimated disparity map. As can be seen, unlike the belief propagation algorithm, the proposed algorithm produces a smooth, non-discretized disparity map. Figures 3(d)-(f) show similar results for the *map* stereo pair.

Table 1 quantitatively compares the performance of the two algorithms. For the *venus* stereo pair, the metrics B and $B_{\overline{O}}$, which do not penalize discrete-valued estimates, are comparable for the two algorithms. The DMSE metrics $M_{\overline{O}}$ and $M_{\overline{O}C}$ are significantly lower for the proposed algorithm. For the simple disparity configuration of the *map* stereo pair $M_{\overline{O}}$ is comparable for the two algorithms, but $M_{\overline{O}C}$ is significantly lower for the proposed algorithm.

Finally, we note that, unlike the Potts model based formulation, the proposed formulation does not incorporate gradient based cues in the smoothness prior. We anticipate that incorporating these cues will further improve the performance of the proposed algorithm.

Table 1. Quantitative comparison.

	$B(\%)$	$B_{\overline{O}}(\%)$	$M_{\overline{O}}$	$M_{\overline{O}C}$
Venus Proposed	5.94	2.64	0.29	0.04
Venus Belief Prop	6.19	2.95	1.42	0.13
Map Proposed	6.89	0.36	0.70	0.03
Map Belief Prop	6.63	0.20	0.66	0.16

5. REFERENCES

[1] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs, "A maximum likelihood stereo algorithm," *Comp. Vision Image Understanding*, vol. 63, pp. 543–567, 1996.

[2] W. Woo and A. Ortega, "Stereo image compression with disparity compensation using the mrf model," in *Proc. visual Commu. and image processing*, 1996, pp. 28–41.

[3] J. Sun, N. Zheng, and H. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, 2003.

[4] S. Birchfield and C. Tomasi, "Multiway cut for stereo and motion with slanted surfaces," in *Proc. IEEE Int. Conf. Comp. Vision*, 1999, pp. 489–495.

[5] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[6] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Comp. Vision*, vol. 47, pp. 7–42, 2002.

[7] M. K. Singh, "Image Segmentation and Robust Estimation Using Parzen Windows," 2003, PhD Thesis, Univ. of Illinois.

[8] O. Veksler, "Efficient graph-based energy minimization methods in computer vision," 1999, PhD Thesis, Cornell Univ.

[9] R.O. Duda and P.E. Hart, *Pattern Recognition and Scene Analysis*, John Wiley, 1973.