Stereo Matching via Learning Multiple Experts Behaviors

Dan Kong and Hai Tao
 Department of Computer Engineering
 UC, Santa Cruz, CA 95064
 {kongdan,tao}@soe.ucsc.edu

Abstract

Window-based matching such as normalized cross-correlation (NCC) can reliably estimate depth even when the constant brightness assumption is violated in stereo due to imaging noise or different camera gains. However, fixed window methods tend to have poor performance at depth discontinuities and in low-texture regions. In this paper, we describe a novel learning-based algorithm, for stereo matching. The algorithm is based on the observation that the matching behavior of each expert is determined by the image texture and the underlying scene structure. In the proposed approach, the behaviors of multiple experts are first learned from ground truth using a simple histogram-based method and the likelihood under each expert is then combined probabilistically into a global MAP-MRF depth estimation framework. Since the resultant likelihood is a function of both stereo image and scene depth in a large neighboring area, we present an iterative Metropolis-Hastings algorithm for the MAP estimation that alternates between predicting expert behaviors and updating the disparity map. The experimental results show that our algorithm is comparable with state-of-the-art methods when the stereo images have identical intensity level but outperforms them when the intensities vary.

1 Introduction

Stereo matching has been extensively studied in the computer vision community for more than three decades. Main challenges in stereo computation stem from the ambiguous or erroneous correspondences caused by the aperture effect, repetitive textures, occlusion, and scene appearance changes. Due to the intrinsic ill-posed property of stereo vision, methods have been proposed to solve these problems by improving the local matching function or by applying sophisticated optimization schemes to minimize energy functions in a global sense. A detailed taxonomy of stereo algorithms can be found in [8].

The main problem of local method is to determine the optimal size, shape, and weight distribution of aggregation support for each pixel inside the support window. An ideal support region should be large enough in textureless regions but be small across the depth boundary. Several stereo algorithms in this category include the adaptive windows matching [5][11], variable window methods [10] and non-linear diffusion [7].

Figure 1: (a) Tsukuba left image. (b) Synthetical alteration of the Tsukuba right image. (c) Depth map computed using Belief Propagation. (d) Depth computed using 9×9 correlation window.

Figure 2: (a)(b) Toy stereo image pair from CMU VASC image database. (c) Depth computed using graph cut (d) Depth computed using 9×9 correlation window

Most global methods model the stereo problem using Bayesian models. There are two main categories of global method: dynamic programming-based or Markov Random Field (MRF)-based. Dynamic programming methods [4][2] are based on the assumption of "uniqueness constraint" and "ordering constraint" and the matching problem is formulated as a path-finding problem in the Disparity Space Image (DSI). Markov Random Field (MRF) theory provides a effective way for modeling contextual constraints. Two recent successful stereo algorithms, graph cut [3] and belief propagation [9], are both based on MRF model.

Inspired by recent research on learning Markov random field (MRF) priors [6], we proposed a method in this paper to learn the matching behaviors of local methods and combine them probabilistically in a global optimization framework.

2 Motivation

Most state-of-the-art stereo methods compute the likelihood using single pixel cost based on the assumption that the corresponding pixels in the two images should have identical intensity. However, this assumption is violated when intensity changes. To show this, we use the Tsukuba image pair from [1] and slightly change the intensity of the right image. Then, we apply the belief propagation algorithm to compute depth and the result is shown in Fig. 1. As a comparison, we also compute the depth map by using 9×9 normalized cross correlation (NCC). It can be seen from Figure 1, NCC gives much better results in this case. Another example is shown in Fig. 2 to compare graph cut and NCC. The stereo pair is taken from CMU VASC image database and the original left and right images have different intensity levels. Again, NCC outperforms graph cut.

Window-based methods such as the normalized cross correlation (NCC) aggregate support in local image regions and are robust against intensity changes. However, window-based methods suffer from the well-known limitations of poor performance at depth discontinuities and in low-texture regions. On the other hand, single pixel matching can be used to produce accurate depth boundaries but fails when the image intensity levels vary

in the two views. The question is how to overcome both problems so that we can deal with the situation of intensity changes while at the same time preserve the depth discontinuities and produce accurate results in textureless regions.

In this paper, we propose a novel learning-based approach to learn the matching behavior of local methods and integrate the learned knowledge into a global probabilistic framework to estimate the depth. We consider normalized cross-correlation methods with different window sizes and matching centers in this paper due to its robustness against image intensity changes. Each NCC is called an expert in the algorithm. In the current work, we limit the expert shape to be rectangular window with 4 scales (3×3 , 5×5 , 7×7 and 9×9) and 9 matching centers. Therefore, there are total of 36 experts and each expert makes local decision based on which disparity level the maximum matching score is obtained (winner-takes-all).

Figure 3: (a) probability of depth error as a function of distance to the nearest foreground object. (b) probability of depth error as a function of texture strength. (c) Probability of estimating true depth for 36 experts on the textured and textureless foreground (d) Probability of estimating true depth for 36 experts at depth discontinuity regions

The proposed method is based on the observation that the matching behavior of a particular expert, such as the centered 7×7 normalized cross-correlation, is determined by both the image texture and the underline scene structure. Using the fattening effect as an example, this refers to the phenomenon that foreground objects appear to be bigger in the depth map. Whether this type of error occurs is determined by how far a pixel is from a close-by foreground object, how large the matching window is, and how strong the background texture is. In other words, the behavior of an expert is a function of the texture and the scene structure. This observation is also true for other matching errors caused by low texture strength, aperture effect, or occlusions. To illustrate this observation, we plot the disparity error probability as a function of the distance to nearest foreground object

and the texture strength using ground truth depth map, as shown in Fig. 3(a)(b). Both the distance to foreground objects and the texture strength are quantized to discrete levels. We can see from Fig. 3 that the disparity error probability is high when the pixel is very close to a foreground object or the texture strength is very low Fig. We also observed that some experts have better matching performance than others depending on the texture and scene configuration. As an illustration, we again use the Tsukuba image and apply the 36 experts to estimate the depth. We compute the probability of estimating true depth for all the experts in three regions: textured foreground, textureless foreground, and textured background close to the foreground. The results are plotted in Fig. 3(c)(d). It can be seen that the behaviors of the experts are very similar for the first two regions in Fig. 3(c): high probability of computing correct depth in textured foreground and low probability in textureless area. However, for the third region in Fig. 3(d), some experts are more likely to estimate the true depth than others because they use left-sided or right-sided correlation window to avoid including the foreground pixels into the matching cost aggregation.

3 Representing and Learning Expert Behaviors

3.1 Representation

To learn the behaviors of experts, we need to describe them first. For standard normalized cross-correlation matching, we observe that besides the correct matching results, there are fattening effects for the background pixels close to the depth boundary and matching ambiguities in low-texture regions. In the approach, we define the behaviors of the expert as whether the window-based matching leads to the true depth $\mathbf{T}$, the nearby foreground depth $\mathbf{FG}$, or other wrong depth $\mathbf{B}$ and we use an indicator variable l to denote this value at each pixel, i.e., $l \in \{\mathbf{T}, \mathbf{FG}, \mathbf{B}\}$. As discussed, the expert behavior is a function of both image textures and the scene structures. We represent this function as conditional distribution, which is denoted as $P_k(l_i|D, I)$. Here, I is the reference image, D is the true depth and k is the expert index. Representing the expert behaviors is equivalent to learn this conditional distribution. Once this distribution is available, it can be used to interpret the raw matching scores given the disparity estimation. To do that, we define a new weighted likelihood function using the learned distribution as weights. More specifically, suppose the correlation score under expert k is $C_{i,j}^k$ for each pixel i at disparity level j . Conditioned on the behavior indicator l_i , the combined new likelihood for expert k becomes:

$$\begin{aligned} P_k(C_i^k|D, I) &= P(C_i^k|l_i = \mathbf{T}, D, I)P_k(l_i = \mathbf{T}|D, I) + \\ &P(C_i^k|l_i = \mathbf{FG}, D, I)P_k(l_i = \mathbf{FG}|D, I) + \\ &P(C_i^k|l_i = \mathbf{B}, D, I)P_k(l_i = \mathbf{B}|D, I) \end{aligned} \quad (1)$$

where $P(C_i^k|l, D, I)$ is the conditional probability of C_i^k , the correlation score vector at position i . $P(C_i^k|l, D, I)$ is approximated in this paper as:

$$P(C_i^k|l_i = \mathbf{T}, D, I) \approx \frac{\exp(-C_{i,d_i}^k)}{\sum_{d'} \exp(-C_{i,d'}^k)} = \alpha_1 \quad (2)$$

$$P(C_i^k|l_i = \mathbf{FG}, D, I) \approx \frac{\exp(-C_{i,fg(i,D)}^k)}{\sum_{d'} \exp(-C_{i,d'}^k)} = \alpha_2 \quad (3)$$

$$P(C_i^k | l_i = \mathbf{FG}, D, I) \approx 1 - \alpha_1 - \alpha_2 \quad (4)$$

Where C_{i,d_i}^k is the correlation score at pixel i when the depth estimation d_i is correct. $fg(i, D)$ is a depth extrapolation function that returns the depth of the nearest foreground object for pixel i under current depth estimation D . $C_{i,fg(i,D)}^k$ is the matching score at pixel i given the depth $fg(i, D)$. α_1 and α_2 are the normalized likelihood for C_{i,d_i}^k and $C_{i,fg(i,D)}^k$ respectively. The normalization is necessary especially in flat area, where the raw correlation score is usually very high for all the disparity level. Finally, $P_k(l_i = \mathbf{T} | D, I)$, $P_k(l_i = \mathbf{FG} | D, I)$ and $P_k(l_i = \mathbf{B} | D, I)$ are the weighting probabilities that can be retrieved from the learned distributions. When we have multiple experts, we could combine them by taking the product of the likelihood of individual expert defined in Eq.(1):

$$P(C_i | D, I) = \prod_k P_k(C_i^k | D, I) \quad (5)$$

The conditional distribution $P_k(l_i | D, I)$ is high dimensional because the matching scores and therefore the variables l_i is affected by neighboring image pixels and 3D scene structures. For example, for a 9×9 matching window, the fattening effect can reach at least 4 to 5 pixels away from a depth boundary. Therefore l_i can be affected by scene structure in a 9×9 window around it. When the size of the matching window increases, this influence region also grows. Direct learning of this high dimensional distribution is impractical. In this paper, we approximate this distribution by extracting a small set of local image and structure attributes and replacing the original image I and depth map D with these attributes. The image and structure attributes should be descriptive in the sense that all the factors that influence the window-based matching are included. Meanwhile, the attribute set should also be compact enough to avoid overfitting. Currently, we choose the following attributes as the conditional variables for l_i . They are horizontal texture strength $\{T_s^i\}_{i=1}^3$ at three scales: 3×3 , 7×7 and 11×11 , the distance F_d and the orientation F_o of the displacement vector to the nearest foreground object. Using these attributes, the conditional distribution $P_k(l_i | D, I)$ can be approximated as:

$$P_k(l_i | D, I) \approx P_k(l_i | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i}) \quad (6)$$

3.2 Learning

Learning $P_k(l_i | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i})$ is not trivial since we do not know if the true distributions are Gaussian, Poisson or multimodal. In our approach, we use histograms as the basis for probabilistic representation. Histograms are flexible non-parametric methods and have the advantages that probabilities can be retrieved by a table look-up. Estimation of a histogram is also simple: we count how often each attribute value occurs in the training data. To learn the conditional probability $P_k(l_i | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i})$, we quantize the texture and structure attributes to discrete levels and represent the joint probability distributions of all the attributes as multi-dimensional histograms. We use the data set from [1] as examples and each training instance consists of stereo image pair (I', J') , the depth map D_k^t computed from (I', J') using NCC and the ground truth depth map D^t . The learning procedure is as follows:

- Step 1: For each pixel i , compare the ground truth depth map D^t with D_k^t to determine if the matching leads to correct depth, nearby foreground depth or an outlier, the comparison result is l_i .

- Step 2: Compute texture attributes $\{T_s^i\}_{i=1}^3$ from I' and scene structure attributes $F_{d,i}, F_{o,i}$ from ground truth depth map D' and quantize them to finite levels.
- Step 3: Depending on l_i , increase one of the following histograms by one

$$P_k(l_i = \mathbf{T} | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i})$$

$$P_k(l_i = \mathbf{FG} | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i})$$

$$P_k(l_i = \mathbf{B} | T_{s,i}^1, T_{s,i}^2, T_{s,i}^3, F_{d,i}, F_{o,i})$$

- Step 4: normalize each histogram so that the sum of the bins equal to one to make it a probability distribution

4 A Probabilistic Stereo Model

4.1 Stereo as a MAP-MRF problem

We formulate the stereo matching as a MAP problem: given an image pair I and J , we seek to estimate D by maximizing the joint posterior

$$P(D|I, J) = \frac{P(I, J|D)P(D)}{P(I, J)} \propto P(I, J|D)P(D) \quad (7)$$

Now we need to define the likelihood $P(I, J|D)$ and prior $P(D)$. Most global methods define $P(I, J|D)$ using single pixel dissimilarity, which have problems for noisy images and in flat area. Our likelihood function, however, fuses the raw NCC matching score from all the expert probabilistically based on the learned behaviors. Given the image observations (I, J) , we compute the NCC matching cost for each expert from (I, J) and denote the raw NCC matching cost as $C = \{C^k, k = 1, \dots, K\}$. Our likelihood function is conditioned on both the structure D and the texture I :

$$P(C|I, D) = \prod_i P(C_i|I, D) \quad (8)$$

Where $P(C_i|I, D)$ is defined in Eq. 3.1

The prior term encodes smoothness of disparity field D . Let $N(i)$ be the neighbors of the pixel i and $S = \{i, j | i < j, j \in N(i)\}$ be the set of all adjacent pixel pairs. The Markov property asserts that the conditional probability of a site in the field depends only on its neighboring sites. By assuming homogeneous MRFs, the Gibbs distribution for the prior can be written as

$$P(D, O) \propto \prod_i \exp(-V_c(d_i)) \propto \prod_{i \in S} \prod_{j \in S} \exp(-\rho_d(d_i, d_j)) \quad (9)$$

To achieve discontinuity-preserving property, line process is integrated into the formulation using the same robust function used in [9]. More specifically,

$$\rho_d(d_i, d_j) = -\ln((1 - e_p) \exp(\frac{-|d_i - d_j|}{\sigma_p}) + e_p) \quad (10)$$

Figure 4: Graphical model for stereo (a): Traditional MRF model. (b): MRF model in this paper.

By combining the likelihood (4.1) and prior (3.1), the depth map D can be estimated in a maximum a posterior (MAP) way as:

$$D^* = \arg \max_D P(D|I, J) \propto \arg \max_D P(C|I, D)P(D) \quad (11)$$

Eq. (11) is similar to the MRF formulation previously proposed in [3] and [9]. However, it should be noticed that the likelihood function needs to be computed using neighboring depth values. This is equivalent to estimating a higher order MRF depicted in Fig. 4.

4.2 Optimization

It is difficult to optimize a high-order MRF here due to the likelihood function. We apply the Metropolis-Hastings sampling algorithm, which is an example of Markov Chain Monte Carlo (MCMC) method in the optimization. Using Metropolis-Hastings algorithm, a proposal move from current depth solution D to a new solution D' is accepted with the probability

$$P(D \rightarrow D') = \min\{1, [\frac{P(D'|I, J)q(D' \rightarrow D)}{P(D|I, J)q(D \rightarrow D')}]^{1/T}\} \quad (12)$$

Where $P(D|I, J)$ and $P(D'|I, J)$ are the posterior probabilities of two depth solutions and $q(D \rightarrow D')$ and $q(D' \rightarrow D)$ are the proposal probabilities. The temperature T in Eq. (12) controls the speed of the cooling process and decreases according to $T^{(t)} = \kappa T^{(t-1)}$, where κ is a constant between 0.8 and 0.99. Before starting the algorithm, we keep a set of possible depth labels for each pixel i using the winner-take-all output of all the experts. We denote it as $\eta(i)$. For each pixel i , we define the proposal move either takes a value in $\eta(i)$ or changes the depth to one of its neighboring pixels $N(d_i)$. In other words, $D' \in \{N(d_i), \eta(i)\}$. To evaluate the acceptance probability, we need to be able to compute the posterior probability ratio $P(D'|I, J)/P(D|I, J)$ and proposal probability ratio $q(D' \rightarrow D)/q(D \rightarrow D')$. By changing the depth of a single pixel, the likelihood values of pixels in a large neighborhood are affected and need to be recomputed. The smoothness term is also changed due to the flip of the pixel. Thus, the computation of the posterior probability ratio can be simplified because only the ratio between the likelihood values and the priors are needed in Eq. 12. To compute this ratio, the likelihood values of the affected pixels in a neighborhood around current pixel i are computed for configurations

Table 1: Performance comparisons using NCC matching cost

Algorithm	Learning			BP			GC		
	all	untex.	disc.	all	untex.	disc.	all	untex.	disc.
Tsukuba	2.03	0.77	11.75	3.62	3.83	15.17	4.91	4.36	20.79
Sawtooth	1.34	0.90	5.88	4.82	9.73	12.13	3.94	1.89	19.86
Venus	2.00	2.26	16.80	12.72	2.60	18.5	2.79	3.00	25.69
Map	0.75	1.67	6.70	0.76	2.60	18.5	1.33	2.14	12.7

D and D' . More specifically, the posterior ratio is calculated as

$$\frac{P(D'|I, J)}{P(D|I, J)} = \frac{P(D')P(C|D', I)}{P(D)P(C|D, I)} = \exp(-(V_c(D'_i) - V_c(D_i))) \prod_{j \in A_i} \frac{P(C_j|D', I)}{P(C_j|D, I)} \quad (13)$$

where A_i defines the set of pixels whose likelihood values are affected by the depth change at pixel i .

5 Experimental Results

We first test the proposed algorithm using new Middlebury benchmark images [1]. A leave-one-out strategy is adopted in the training and testing process to avoid using the same images in training and testing. Fig. 5 shows depth maps estimated using our method and the corresponding ground truth depth maps. It can be observed that the fattening effect around the foreground objects such as the lamp has been suppressed in our results and correct depth is computed in the textureless areas. Next, we compare our method with belief propagation (BP) and graph cut (GC). It should be noted that most state-of-the-art stereo methods like BP and GC use single pixel matching cost in their likelihood functions, which result in lower error rates for Middlebury benchmarks. However, in real applications, the quality of stereo image pairs are not as good as Middlebury images due to intensity changes, as mentioned in section 2. In this case, computing the likelihood using window-based matching will be more reliable. The proposed method, starting from a NCC depth map with errors, can gradually suppress the those matching errors using the learned behaviors of NCC. To validate this, we run BP and GC algorithms with the likelihood function computed using 7×7 correlation window. The results for "Tsukuba" and "Venus" are shown in Fig. 6 and the quantitative comparisons are shown in Table 1. We can see that the proposed method has the lowest error rates, especially in the discontinuity regions. Finally, we apply the proposed algorithm to a face stereo image pair with one of them darker than the other and compare the results with NCC-based BP and GC. As shown in Fig. 7, our method has better results than BP and GC.

Figure 5: Results on the new testbed (a) Tsukuba (b) Venus (c) Teddy (d) Cones

Figure 6: Comparisons of the disparity maps for the "Tsukuba" and "Venus" images using 7×7 NCC matching cost as the likelihood. (a) Our method. (b) Belief propagation. (c) Graph cut.

6 Conclusion

In this paper, we present a learning-based method for stereo matching. Our approach combines the advantages of both local methods and global methods. Preliminary results show that this approach is very promising. However, several issues need to be further studied. Among them, the first question is what are the other texture and structure attributes that can be used to learn the behaviors of the experts. Several promising features include thin structures, depth gaps, surface curvatures and orientations. In addition, how to select distinctive features from a feature pool such that the expert behaviors can be best discriminated under different image/scene configuration would also be an interesting research topic. Finally, we are interested in improving current algorithm by using more efficient optimization method.

Figure 7: Comparisons of the disparity maps for the "face" stereo pair. (a) Left image (b) Right Image (c) Initial depth from 7×7 correlation window. (d) Belief propagation result. (e) Graph cut result. (f) Our result.

References

- [1] <http://www.middlebury.edu/stereo>.
- [2] P. N. Belhumeur. A bayesian-approach to binocular stereopsis. *IJCV*, 19(3):237–260, August 1996.
- [3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.
- [4] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs. A maximum-likelihood stereo algorithm. *CVIU*, 63(3):542–567, 1996.
- [5] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: theory and experiment. *PAMI*, 16(9):920–932, September 1994.
- [6] S. Roth and M. J. Black. On the spatial statistics of optical flow. In *ICCV 05*, pages 42–49, 2005.
- [7] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *IJCV*, 28(2):155–174, July 1998.
- [8] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.
- [9] J. Sun, H. Y. Shum, and N. N. Zheng. Stereo matching using belief propagation. In *ECCV02*, pages II: 510–524, 2002.
- [10] O. Veksler. Fast variable window for stereo correspondence using integral images. In *CVPR03*, pages I: 556–561, 2003.
- [11] K. J. Yoon and I. S. Kweon. Locally adaptive support-weight approach for visual correspondence search. In *CVPR05*, pages II: 924–931, 2005.