MAP Disparity Estimation using Hidden Markov Trees

Eric T. Psota Jędrzej Kowalczuk Mateusz Mittek Lance C. Pérez
Department of Electrical & Computer Engineering
University of Nebraska-Lincoln
{epsota, jkowalczuk, mmittek, lperez}@unl.edu

Abstract

A new method is introduced for stereo matching that operates on minimum spanning trees (MSTs) generated from the images. Disparity maps are represented as a collection of hidden states on MSTs, and each MST is modeled as a hidden Markov tree. An efficient recursive message-passing scheme designed to operate on hidden Markov trees, known as the upward-downward algorithm, is used to compute the maximum a posteriori (MAP) disparity estimate at each pixel. The messages processed by the upward-downward algorithm involve two types of probabilities: the probability of a pixel having a particular disparity given a set of per-pixel matching costs, and the probability of a disparity transition between a pair of connected pixels given their similarity. The distributions of these probabilities are modeled from a collection of images with ground truth disparities. Performance evaluation using the Middlebury stereo benchmark version 3 demonstrates that the proposed method ranks second and third in terms of overall accuracy when evaluated on the training and test image sets, respectively.

1. Introduction

Stereo matching, used to extract the positional offsets relating corresponding points in two views of a scene, remains one of the most studied problems in computer vision. These offsets, otherwise known as disparities, can be easily converted to per-pixel depths representing the three-dimensional properties of the scene. Depth information is critical in applications such as robotic navigation, augmented reality, image-based rendering, advanced human-computer interfaces, and visual metrology.

As an inverse computer vision problem that aims to recover depth from two-dimensional projections of the scene, stereo matching must overcome a variety of challenges resulting from both scene properties and the image acquisition process. The challenges associated with scene properties include reflections, occlusions, and repeating patterns; while those arising from the image acquisition process include the effects of quantization, noise, and image blur. Furthermore, since increasing the product of the image dimensions and the number of disparity hypotheses causes a proportional increase in the complexity of matching, devising efficient algorithms is an ongoing challenge that has become even more difficult due to increasing video and image resolutions.

The process of stereo matching generally includes the following four steps: computation of per-pixel matching costs, spatial aggregation of costs, disparity assignment, and refinement of disparities. Scharstein and Szeliski [20] recognized that stereo matching algorithms can be classified as either local or global, depending on how the cost aggregation and disparity assignment steps are performed. Local algorithms choose disparities that minimize the dissimilarity between small regions in the images, while global algorithms attempt to minimize an energy function that explicitly enforces pixel similarity and smoothness over the entire image. While region-based methods allow for efficient matching, their inherent locality makes them unsuitable for large textureless regions. Global algorithms are capable of overcoming these limitations, however they often rely on computationally complex optimization schemes. To reduce computational complexity, recent methods have exploited tree structures in order to accelerate the process of cost aggregation [9, 25, 3]. Rather than operating on traditional four-connected image grids, like early global algorithms, these methods reduce the computational complexity by efficiently passing messages between nodes in a tree.

Here, a global stereo matching method is proposed that performs cost aggregation through message passing within the minimum spanning tree (MST). Unlike traditional graph-based approaches that define and minimize global energy functions, the proposed method performs maximum *a posteriori* (MAP) disparity estimation using statistics that are generated from a wide range of ground truth data. Specifically, the proposed method uses the recently released Middlebury stereo benchmark version 3 image set [18]. This set includes 30 high-resolution image pairs and ground truth disparity maps, making it possible

to generate robust statistical models relating the observed image intensities and the corresponding disparities.

2. Background

Several different algorithms have been introduced that perform cost aggregation by allowing neighboring pixels to share information either iteratively or in sequence. These include dynamic programming (DP) [16], graph cuts [4], belief propagation [22], tree-reweighted message passing [12], and semi-global matching [9]. While some of these methods operate on conventional four-connected image grids, others have adopted tree structures to reduce the complexity of aggregation.

Szeliski *et al.* [23] compared energy minimization methods that operate on images represented as four-connected grids of pixels. Tree-reweighted message passing and graph cuts were shown to be most effective in arriving at solutions that approach the global minimum of the energy functional at the expense of high computational complexity. Similar solutions can be obtained more efficiently using belief propagation, however, this method is not guaranteed to converge and, in most cases, does not converge on cyclic graphs such as the four-connected image grids.

To improve the efficiency of cost aggregation and guarantee convergence, trees spanning the image space can be used instead of four-connected grids. The original method based on DP, for example, operates on horizontal scanlines in the image [16]. Since scanlines are processed independent of one another, unwanted streaking artifacts appear in the resulting disparity maps. Hirschmüller [9] recognized the limitations of such an approach and introduced the semi-global matching (SGM) method that considers eight or more lines converging at each pixel location to perform cost aggregation. SGM, which can be thought of as a multi-directional variant of dynamic programming, significantly reduces the streaking effect. An alternative proposed by Bleyer and Gelautz, termed simple-tree DP [3], achieves efficient global aggregation and addresses the streaking issue by performing four passes of DP through every pixel. Effectively, aggregation occurs on trees composed of horizontal scanlines that are connected vertically. Both SGM and simple-tree DP have been shown to be effective cost aggregation/disparity selection methods, both in terms of computational efficiency and accuracy of matching.

In [25], Yang introduced a stereo matching method that operates on minimum spanning trees constructed from the input images. The minimum spanning tree, which is obtained through successive removal of high-cost edges from the four-connected image grid, covers the full extent of the image and provides a unique path between every pair of pixels in the image. Instead of exchanging disparity evidence among the pixels in the standard four-connected grid, the information is propagated sequentially along paths in the minimum spanning tree. Yang's method is essentially a global variant of adaptive support-weight cost aggregation [27], where the connectivity of pixels within the MST and their color similarity determine the effective scope of aggregation. Note that minimum spanning trees were previously used in the context of stereo matching by Veksler [24] to orchestrate dynamic programming in a way that enforces inter-scanline consistency of disparities.

Stereo matching is frequently interpreted as a labeling problem, where every pixel in the image is assigned a disparity label. Proven maximum likelihood (ML) and maximum *a posteriori* (MAP) estimators exist that solve labeling problems through message passing on trees. Such algorithms have been extensively studied in the field of coding theory [17]. Yang's method fits this framework by performing message passing on trees, however, it is limited by the fact that aggregation is performed independently for each slice of the cost volume corresponding to a specific disparity. As such, this method tends to flatten objects in the disparity map. SGM addresses this problem by considering disparity transitions while aggregating cost, which results in improved performance on slanted surfaces. In many ways, SGM's rule of operation closely resembles the Viterbi algorithm [8], a well-known ML estimator.

This paper introduces a stereo matching method that performs MAP disparity estimation through message passing on the minimum spanning tree. It is assumed that the MST can be modeled as a hidden Markov tree (HMT) that governs the relationships between the disparities assigned to neighboring pixels. This assumption makes it is possible to efficiently compute disparities on the MST using the upward-downward algorithm [6]. In addition, the proposed method does not rely on heuristic formulations of matching cost or disparity smoothness constraints. Scharstein and Pal [19] have previously demonstrated that stereo matching performance can be significantly improved by learning the matching parameters from ground truth data. Similarly, the proposed method relies on probability models estimated from ground truth disparity data using statistical analysis and machine learning techniques.

3. MAP Disparity Estimation on the MST

The proposed method introduces a novel tree-based approach to stereo matching that uses message passing on the MST in order to estimate the MAP disparity likelihood of every pixel. Specifically, message passing is performed using the upward-downward algorithm, a known MAP estimator designed to operate efficiently on trees. In the context of stereo matching, the application of the upward-downward algorithm requires a tree structure that covers the entire space of the image (here, the MST) and probabilistic models that both govern the interactions between pixels and capture the relationship between disparities and pixel

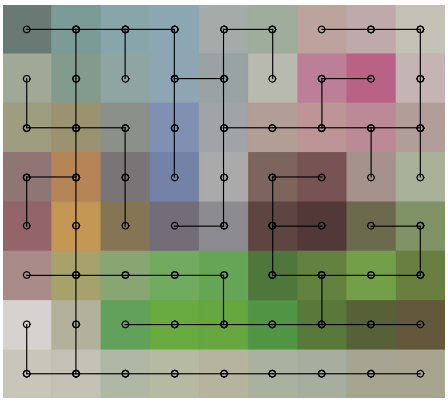


Figure 1: An example of a minimum spanning tree generated from an image.

intensities. The construction of the MST, the message passing scheme, and the associated probabilistic models are discussed in the following sections.

3.1. Minimum Spanning Trees from Images

Prior to aggregation, minimum spanning trees are constructed from the input images to facilitate the sharing of disparity evidence between pixels. An assumption is made that neighboring pixels with similar color are more likely to have similar disparity, and thus an algorithm stands to benefit from sharing disparity evidence along edges in the tree. Fig. 1 shows an example of a minimum spanning tree that connects all of the pixels in a 9×8 image. Notice that short paths through the tree tend to connect neighboring pixels with similar color. In contrast, long paths tend to connect pixels with significantly different color.

In order to find the minimum spanning tree, costs must first be assigned to each edge linking neighboring pixels in the four-connected image grid. Since changes in disparity often coincide with noticeable changes in image intensity [27], it is natural to choose costs that are proportional to the intensity difference between neighboring pixels. Here, the cost of each edge is assigned according to the distance between pixels as measured by the sum of absolute intensity differences. Once costs have been assigned, Kruskal's algorithm [14] is applied to iteratively remove high-cost edges from the grid until the minimum spanning tree is obtained. The representation of the MST used by the proposed method is a set of child/parent pixel pairs, ordered in a way that all paths from the leaf nodes up to the root node can be traversed in a single scan through the pairs.

3.2. Upward-Downward Algorithm

The MSTs representing the images are assumed to have the properties of a hidden Markov tree (HMT), i.e., the disparity of each pixel is conditionally independent of the disparities of all other pixels given the disparities of its immediate neighbors in the tree. Hidden Markov trees are defined by 1) a set of connections between nodes that have hidden states, and 2) a set of observations associated with the state of each node. In the context of stereo matching, the hidden state is the disparity d_n of pixel n, and the observation is the vector of costs $\mathbf{c}_n$ of choosing all possible disparity values. Fig. 2 illustrates an example of such an HMT.

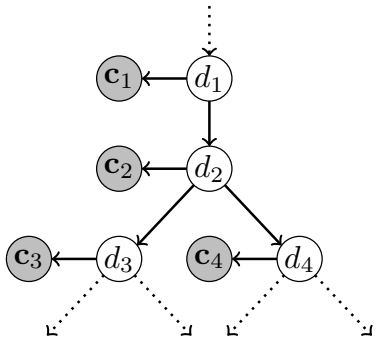


Figure 2: Subsection of a hidden Markov tree (HMT) model where each state node, denoted by d_n, is associated with an observation node, denoted by $\mathbf{c}_n$. In this illustration, parent nodes are placed above, and connected to, their child nodes. For example, the children of d_2 are $c(d_2) = \{d_3, d_4\}$ and the parent of d_2 is $p(d_2) = d_1$.

The HMT model allows efficient calculation of the maximum a posteriori (MAP) disparity estimate

$$\hat{d}_n = \underset{d_n}{\operatorname{argmax}}\, P(\mathbf{C}|d_n)P(d_n)$$

of each pixel n, where $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ denotes the set of all observed matching costs throughout the entire image, $P(\mathbf{C}|d_n)$ is the likelihood of observing the cost $\mathbf{C}$ given disparity d_n, and $P(d_n)$ is the prior probability of disparity d_n. The structure and properties of the HMT can be exploited to calculate the large multivariate distribution $P(\mathbf{C}|d_n)P(d_n)$ efficiently using the upward-downward algorithm [6].

The probability distribution that is computed using the upward-downward algorithm can be reformulated as

$$\begin{aligned} P(\mathbf{C}|d_n)P(d_n) &= P(d_n, \mathbf{C}) \\ &= P(\mathbf{C}_n|d_n)P(d_n, \mathbf{C}_{\setminus n}) \end{aligned} \quad (1)$$

using conditional independence relationships assumed by the HMT model, where $\mathbf{C}_n$ denotes the collection of costs that belong to the subtree rooted at node d_n, and $\mathbf{C}_{\setminus n}$ denotes the collection of all costs excluding those in $\mathbf{C}_n$. In Fig. 2, $\mathbf{C}_2$ includes $\mathbf{c}_3$, $\mathbf{c}_4$, and all costs associated with children of d_3 and d_4. The cost $\mathbf{C}_{\setminus 3}$ includes $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_4$, and all other costs not connected to children of d_3.

The calculation of (1) using the upward-downward algorithm is decomposed into two stages, where β messages are first passed up from the leaf nodes to the root node (the upward stage) and then α messages are passed down from the

root node to the leaf nodes (the downward stage). After initializing $P(\mathbf{C}_n|d_n)$ to $P(\mathbf{c}_n|d_n)$ at the leaf nodes, messages in the upward stage are computed recursively using

$$\overbrace{P(\mathbf{C}_n|d_{p(n)})}^{\beta_n^{p(n)}(d_{p(n)})} = \sum_{d_n=d_{\min}}^{d_{\max}} \overbrace{P(\mathbf{C}_n|d_n)}^{\beta_n(d_n)} P(d_n|d_{p(n)}) \quad (2)$$

$$\overbrace{P(\mathbf{C}_n|d_n)}^{\beta_n(d_n)} = \left[\prod_{v=c(n)} \overbrace{P(\mathbf{C}_v|d_{p(v)})}^{\beta_v^{p(v)}(d_{p(v)})} \right] P(\mathbf{c}_n|d_n) \quad (3)$$

where $c(n)$ denotes the set of children of node n, $p(n)$ denotes the parent of node n, and $[d_{\min}, d_{\max}]$ represents the range of disparities. The messages in the downward stage are then computed using the recursion

$$\overbrace{P(d_n, \mathbf{C}_{\setminus n})}^{\alpha_n(d_n)} = \sum_{d_{p(n)}=d_{\min}}^{d_{\max}} P(d_n|d_{p(n)}) \frac{\overbrace{P(\mathbf{C}_{p(n)}|d_{p(n)})}^{\beta_{p(n)}(d_{p(n)})}}{\underbrace{P(\mathbf{C}_n|d_{p(n)})}_{\beta_n^{p(n)}(d_{p(n)})}} \overbrace{P(d_{p(n)}, \mathbf{C}_{\setminus p(n)})}^{\alpha_{p(n)}(d_{p(n)})}. \quad (4)$$

In order to evaluate (2)-(4), it is necessary to estimate the disparity transition probabilities $P(d_n|d_m)$ between all neighboring pixels n and m and the disparity likelihoods $P(\mathbf{c}_n|d_n)$ for all pixels n prior to applying the recursion. These probabilities depend on the stereo image capture configuration, which is affected by factors such as resolution, exposure, illumination, baseline, and structural properties of the scene. For the remainder of this work, transition probabilities and disparity likelihoods are estimated from half-resolution training images and ground truth disparity maps provided by the Middlebury stereo benchmark version 3. After being rounded to integer values, each available ground truth disparity in the training set is used to estimate transition probabilities and disparity likelihoods.

3.2.1 Transition Probabilities

The majority of stereo matching methods that incorporate local message passing abstract the disparity transition probability into an explicit smoothness term in the energy function. The smoothness term typically takes the form of a pair-wise penalty function that discourages significant discontinuities in the disparity assignment of neighboring pixels. The linear truncated model and variants of the Potts model, all of which increase the penalty values with increasing disparity transitions, are commonly used in stereo matching [3, 9]. To allow for disparity discontinuities around object edges, both models saturate the penalty value once a certain disparity transition threshold is met. Edge-preserving properties of the smoothness term can otherwise

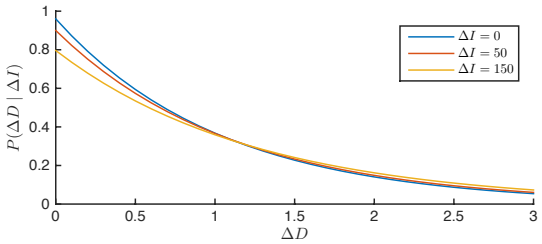


Figure 3: Transition probabilities given intensity difference.

be achieved by making the penalty function dependent on the intensity difference between connected pixels.

Consider two neighboring pixels labeled n and m in the HMT with observed intensities I_n and I_m. The proposed method aims to directly model the probability of a disparity transition $\Delta D = |d_n - d_m|$ conditioned on the pair-wise pixel intensity change $\Delta I = |I_n - I_m|$. The underlying probability distribution was estimated by first constructing a bivariate histogram of the disparity transitions ΔD and intensity differences ΔI extracted using training images and ground truth disparity maps. Conditional probability distributions were calculated from the bivariate histogram by fixing the values of ΔI and normalizing by the number of observed pixels with that particular intensity difference.

Continuous transition probabilities, illustrated for intensity changes of ΔI=0, 50, and 150 in Fig. 3, are observed to follow the exponential distribution where the decay rates of the exponentials vary with the intensity change. This is not surprising, since only a small number of pixels in most images belong to depth edges. The majority of pixels reside inside object boundaries and experience little to no change in disparity when compared to their neighbors.

Probabilities of selected disparity transitions as functions of intensity differences are shown in Fig. 4. This illustrates the effect that increasing intensity differences have on the probability of disparity transitions. For example, as the intensity difference between neighboring pixels increases from 0 to 150, the probability that the disparity stays the same (i.e., $\Delta D = 0$) drops from 0.95 to 0.73, while the probability that the disparity changes by two, i.e., $\Delta D = 2$, increases from 0.0 to 0.02. The red lines in Fig. 4 represent linear models fitted to the observed probabilities (plotted in blue). The transition probabilities obtained for $\Delta D > 4$ have been observed to experience very little change, and are collectively represented by the model shown in the last row of Fig. 4. These models allow for the transition probability $P(d_n|d_m)$ to be computed for all pixels n and m based on the intensity differences between the pixels.

3.2.2 Likelihoods

As previously indicated, the upward-downward algorithm requires an estimate of the likelihood $P(\mathbf{c}_n|d_n)$ of observ-

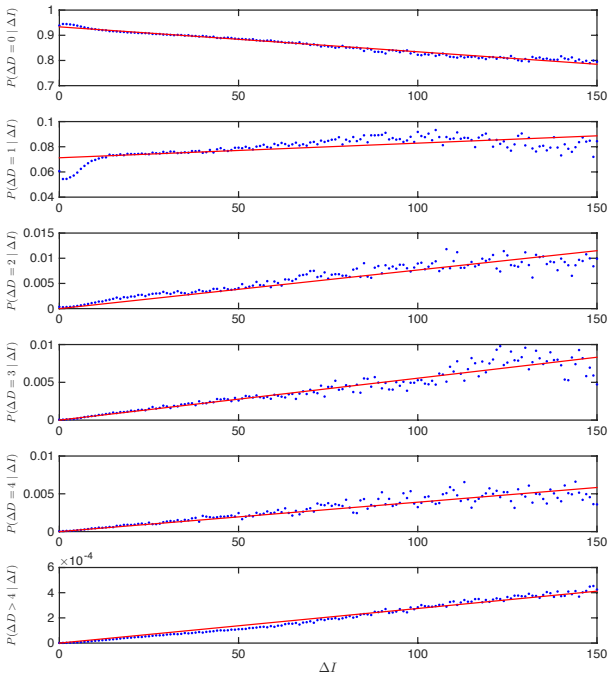


Figure 4: The probability of disparity transitions $\Delta D = 0$, 1, 2, 3, 4, and $\Delta D > 4$ as a function of intensity difference. The blue dots represent probabilities extracted from ground truth disparity maps and fitted lines are drawn in red.

ing a vector of costs $\mathbf{c}_n$, given a disparity d_n. In the context of stereo matching, $\mathbf{c}_n$ is a metric that quantifies the visual similarity between a pixel n and each of its candidate matches in the other image. Selection of the appropriate cost metric is crucial to the performance of stereo matching, and has been studied extensively. A thorough examination of cost metrics was presented by Hirschmüller *et al.* [10], who examined a variety of parametric and non-parametric cost metrics, and concluded that the census transform [28] provides the best overall accuracy, especially in the presence of illumination and exposure changes.

Our experiments indicate that the accuracy of matching can be further improved by combining the 9×7 census metric with the absolute difference of horizontal gradients around the pixels of interest. This is likely due to the fact that census imposes a hard threshold, which is easily dominated by noise in areas of the images that lack texture. In such a scenario, the real-valued absolute difference of gradients contributes to the disambiguation of matching. In the proposed method, the census metrics $\mathbf{c}_n^c$ and gradient metrics $\mathbf{c}_n^g$ evaluated for all disparity hypotheses are vertically concatenated to form the vector of costs $\mathbf{c}_n$.

While many existing stereo matching methods avoid estimating the likelihood $P(\mathbf{c}_n|d_n)$ and instead use the values of $\mathbf{c}_n$ directly within a cost minimization framework, the proposed method takes a statistical approach to estimate the likelihood from ground truth data. Estimation is performed using multinomial logistic regression under the following assumptions: 1) the disparity d_n can be expressed as a categorically distributed variable dependent on $\mathbf{c}_n$, 2) the probability of d_n can be calculated from a linear combination of the observed costs $\mathbf{c}_n$, and 3) this linear combination can be obtained by applying a training procedure to sample data. Precisely, multinomial logistic regression estimates the conditional probability $P(d_n|\mathbf{c}_n)$. Under the assumption of equiprobable disparities, this conditional probability is proportional to the sought likelihood $P(\mathbf{c}_n|d_n)$.

Using the softmax model [11], the formulation of the conditional probability provided by multinomial logistic regression is given by

$$P(d_n = k|\boldsymbol{c}_n) = \frac{\exp\left(\boldsymbol{c}_n^{\mathsf{T}}\boldsymbol{\theta_k}\right)}{\sum_{i=d_{\min}}^{d_{\max}} \exp\left(\boldsymbol{c}_n^{\mathsf{T}}\boldsymbol{\theta_i}\right)}, \tag{5}$$

where d_n is the true disparity of the n-th pixel, k is a candidate disparity, and $\boldsymbol{\theta_k}$ is the vector of the linear coefficients corresponding to the k-th disparity candidate.

The training set of the Middlebury stereo benchmark was used to fit the set of coefficients $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\boldsymbol{d}_{\min}}, \ldots, \boldsymbol{\theta}_{\boldsymbol{d}_{\max}}]$. The resulting elements along the diagonals corresponding to both census and gradient metrics are the dominant terms in the $\boldsymbol{\theta}$ matrix, whereas the values of elements outside of the diagonals are smaller by more than an order of magnitude. Since only the diagonal elements of $\boldsymbol{\theta}$ are significant, the matrix can be approximated by setting the off-diagonal terms to zero. To avoid favoring any particular disparity over another, the elements along the diagonals can be simply represented as constants γ^c and γ^g. This allows for simplified computation of equation (5) using the approximation

$$P(d_n = k|\boldsymbol{c}_n) \approx \frac{\exp\left(\gamma^c c_{n,k}^c + \gamma^g c_{n,k}^g\right)}{\sum_{i=d_{\min}}^{d_{\max}} \exp\left(\gamma^c c_{n,i}^c + \gamma^g c_{n,i}^g\right)}, \tag{6}$$

where $c_{n,k}^c$ is the k-th element of the census vector $\boldsymbol{c}_n^g$, and $c_{n,k}^g$ is the k-th element of the gradient vector $\boldsymbol{c}_n^g$. Within the proposed method, the coefficient values are $\gamma^c = -0.014$ and $\gamma^g = -0.289$.

3.3. Disparity Selection and Refinement

Once disparity likelihoods have been computed using the upward-downward algorithm, disparity maps for both images are obtained using the winner-take-all (WTA) disparity selection criteria by simply selecting the most likely disparities. Cross-checking [20] is then used to identify inconsistent pixels, i.e., pixels whose disparities do not satisfy a

Table 1: Results on the Middlebury training set evaluating the percentage of errors > 2 of non-occluded pixels.

Method, Res.	Avg. **bad 2.0**	1 Adiron	1 ArtL	1 Jadepl	1 Motor	1 MotorE	1 Piano	0.5 PianoL	1 Pipes	0.5 Playrm	0.5 Playt	1 PlaytP	1 Recyc	0.5 Shelvs	1 Teddy	0.5 Vintge
MeshStereo [1], H	**15.1** 1	**7.14** 1	9.55 4	23.0 9	9.42 4	10.1 5	**15.4** 1	**24.9** 1	12.8 8	23.7 4	23.4 2	**13.3** 1	13.5 4	39.9 2	7.21 4	**22.6** 1
TMAP, H	16.2 2	10.9 3	11.4 7	21.7 6	8.00 3	**8.01** 1	16.0 2	27.6 3	**8.85** 1	22.5 3	35.2 8	15.1 3	**12.4** 1	44.4 5	6.93 3	35.9 4
IDR [13], H	17.0 3	12.0 5	11.4 6	18.7 3	10.4 5	8.53 2	16.6 4	26.0 2	11.8 6	22.4 2	49.7 16	13.4 2	12.9 3	49.0 11	6.47 2	34.0 3
LCU [2], Q	17.3 4	10.9 2	12.6 9	22.0 8	11.3 7	12.1 10	20.9 11	32.2 9	12.1 7	23.9 5	24.6 4	16.5 6	12.5 2	40.3 3	**5.14** 1	38.8 7
SGM [9], H	17.8 5	15.3 7	8.87 3	**18.1** 1	10.9 6	8.90 3	16.4 3	29.1 6	11.5 5	**21.7** 1	52.5 18	15.8 4	14.6 5	46.4 7	7.47 6	39.3 8
PFS [5], F	19.9 6	20.9 13	7.96 2	18.5 2	11.6 8	9.87 4	19.9 8	28.9 5	14.2 11	28.2 8	49.5 15	17.2 7	16.5 6	55.4 20	7.88 8	45.7 15

Table 2: Results on the Middlebury test set evaluating the percentage of errors > 2 of non-occluded pixels.

Method, Res.	Avg. **bad 2.0**	0.5 Austr	1 AustrP	1 Bicyc2	1 Class	0.5 ClassE	1 Compu	1 Crusa	1 CrusaP	1 Djemb	0.5 DjembL	0.5 Hoops	1 Livgrm	1 Nkuba	1 Plants	0.5 Stairs
MeshStereo [1], H	**13.4** 1	**5.90** 1	4.88 4	10.8 8	12.9 5	**10.6** 1	13.6 2	12.2 3	**9.01** 1	5.39 3	27.4 2	23.5 2	**17.7** 1	21.0 5	15.4 5	20.9 2
LCU [2], Q	17.0 2	24.7 5	7.59 9	11.6 9	11.9 3	27.9 2	14.0 3	19.3 4	15.8 6	8.10 13	36.1 8	29.1 6	21.3 4	**18.4** 1	14.1 2	23.8 4
TMAP, H	17.1 3	20.2 4	4.94 5	8.13 4	12.8 4	30.0 3	14.1 5	27.9 9	20.4 10	**5.09** 1	31.5 6	**23.1** 1	20.9 3	19.0 2	18.8 9	**18.0** 1
IDR [13], H	18.4 4	37.5 11	**4.08** 1	7.49 2	23.3 12	40.6 6	15.7 12	24.5 5	11.3 5	5.46 5	33.1 7	26.0 3	21.5 5	21.7 6	15.3 4	21.2 3
SGM [9], H	18.7 5	40.3 12	4.54 3	8.03 3	22.9 11	40.5 5	14.6 8	24.7 6	10.1 3	5.40 4	29.6 4	28.5 5	23.9 7	20.0 3	14.2 3	30.9 8
LPS [21], H	19.4 6	6.14 2	5.34 6	9.24 5	**7.53** 1	96.0 20	15.0 10	**9.61** 1	9.40 2	5.18 2	92.4 22	27.4 4	24.3 8	23.0 9	**10.0** 1	25.6 6

bijective mapping. These pixels, which most often coincide with occlusions, should not be allowed to affect aggregation. This is achieved by forcing the disparity likelihoods $[P(\mathbf{c}_n|d_n = d_{\text{min}}), \ldots, P(\mathbf{c}_n|d_n = d_{\text{max}})]^{\mathsf{T}}$ at inconsistent pixels to be uniformly distributed, and reapplying the upward-downward algorithm. Finally, median filtering is performed on the resulting disparity maps in order to eliminate isolated mismatches.

3.4. Computational Complexity

Letting N be the number of pixels and D be the number of disparity hypotheses, the planarity of the four-connected image grid makes it possible to compute the MST in $\mathcal{O}(N)$ time [15], whereas the complexity of the upward-downward algorithm used for aggregation and refinement is $\mathcal{O}(ND^2)$. By explicitly considering only a fixed number of disparity transition probabilities as suggested in section 3.2.1, one can avoid full matrix multiplication in (2) and (4) and instead evaluate a sum of shifted vectors, reducing the complexity to $\mathcal{O}(ND)$. Similar complexity has previously been reported by Felzenszwalb and Huttenlocher [7] for their linear-time message passing method.

4. Results

The proposed method was evaluated using both the training and test image sets of the Middlebury stereo benchmark version 3. Results are given in Tables 1 and 2 for the evaluation of error rates using the default metric, which measures the percentage of non-occluded disparity errors that are greater than two with respect to the original resolution. The average error rate of the proposed method, denoted by TMAP (Tree-based Maximum A Posteriori disparity estimation), ranks 3rd and 2nd lowest in the test and training sets, respectively. It also achieves the lowest error rates among all methods for six of the image pairs. The similar performance on training and test sets in terms of both rank and average error rate (16.2% and 17.1%) suggest that the parameters used to model disparity likelihood and transition probabilities are not specific to the training set. The proposed method was implemented on a MacBook Pro computer with a 2.7 GHz Intel Core i7 CPU and 16 GB memory. The average runtime per image for the 30 half-resolution images in the testing and training sets is 6.8 seconds. Message passing using (2)-(4) accounted for 6.1 seconds, while MSTs and input likelihoods accounted for $<$0.7 seconds.

To illustrate the performance of the proposed method, select disparity maps and error images are given in Fig. 5. Results on the `MotorE` images demonstrate that TMAP allows for robust matching in the presence of exposure variations. This is due to the use of the census and gradient metrics, both of which are invariant to shifts in intensity. The `Pipes` images test the method's ability to handle depth discontinuities and narrow foreground objects. While the transition probabilities strongly favor continuity of disparities, the method manages to accurately capture the foreground objects. This is because the minimum spanning tree conforms to the foreground objects, requiring very few disparity discontinuities to occur between neighbors in the tree. The `Playt` images illustrate a case of imperfect rectification, where significant vertical disparities exist between

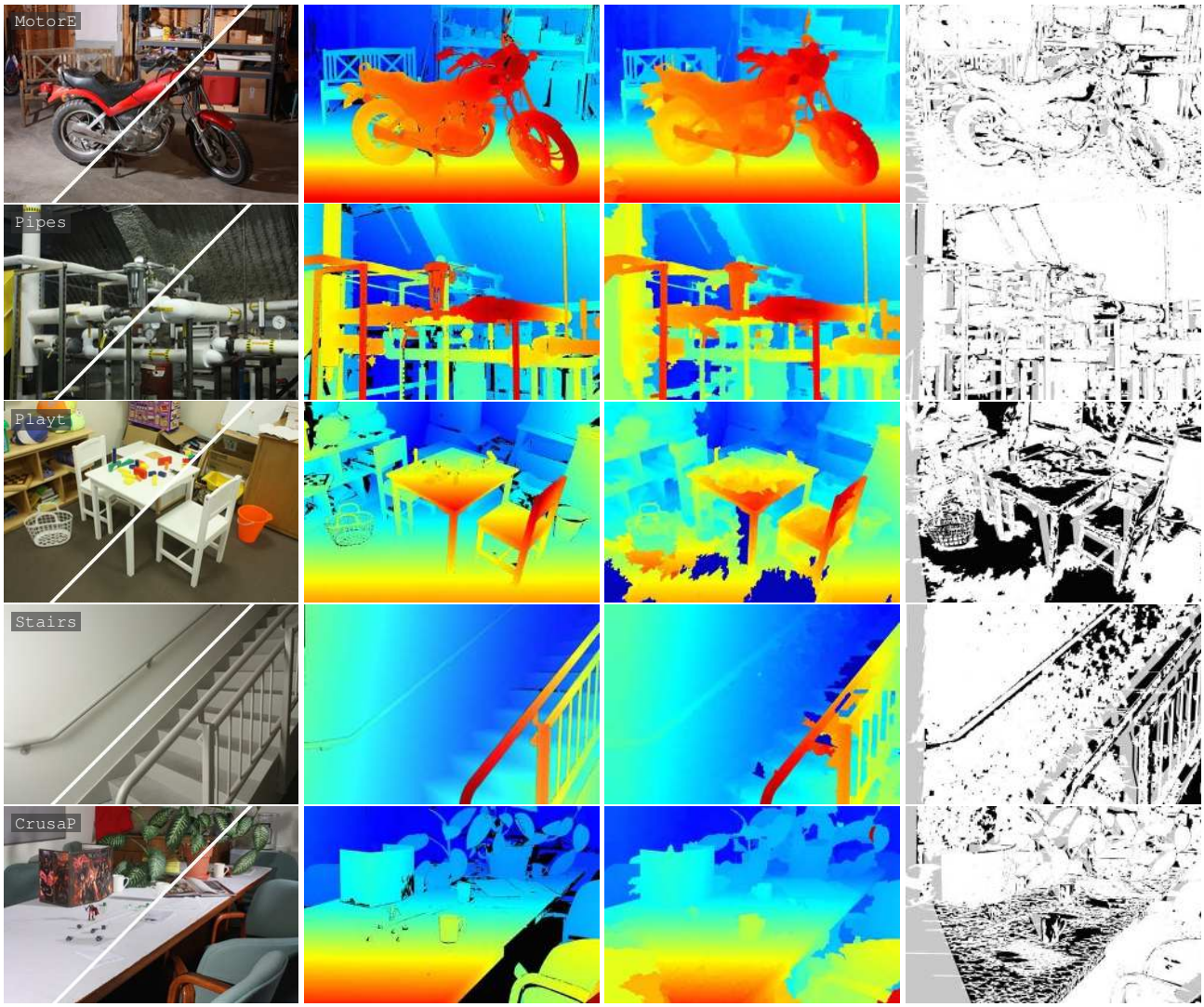


Figure 5: Disparity maps and the corresponding error images obtained for selected datasets of the Middlebury benchmark (version 3). Errors in occluded regions are shown in gray, while the errors in non-occluded areas are shown in black.

corresponding pixels. The way that the census cost metric is computed makes it inherently sensitive to vertical offsets, leading to unreliable matching on imperfectly rectified images. The two images from the test set where the proposed method produces the lowest and highest error rates relative to other methods are `Stairs` and `CrusaP`, respectively. While both of these image sets contain weakly textured surfaces, the proposed method reproduces the smooth surface in the `Stairs` image more accurately due to its gradual disparity gradient.

A comparison of error rates achieved using various tree-based cost aggregation methods is given in Fig. 6. In order to isolate cost aggregation from the other stages of stereo matching, disparity refinement and post processing were disabled, and the same census/gradient cost metric described in section 3.2.2 was used by all methods. The cost aggregation of TMAP outperforms all other methods for 12 out of 15 training images, while achieving the second lowest error rate for the remaining three images. Both the MST filtering [25] and the recursive bilateral filtering (RecurB) [26] perform aggregation independently for each disparity hypothesis. By doing so, they are making an implicit assumption that surfaces in the scene are fronto-parallel, resulting in increased error rates for images that contain slanted surfaces. In contrast, the aggregation schemes of semi-global matching [9] and simple tree matching [3] explicitly penalize disparity transitions between neighboring pixels in the tree. However, the pre-defined structure of the trees over

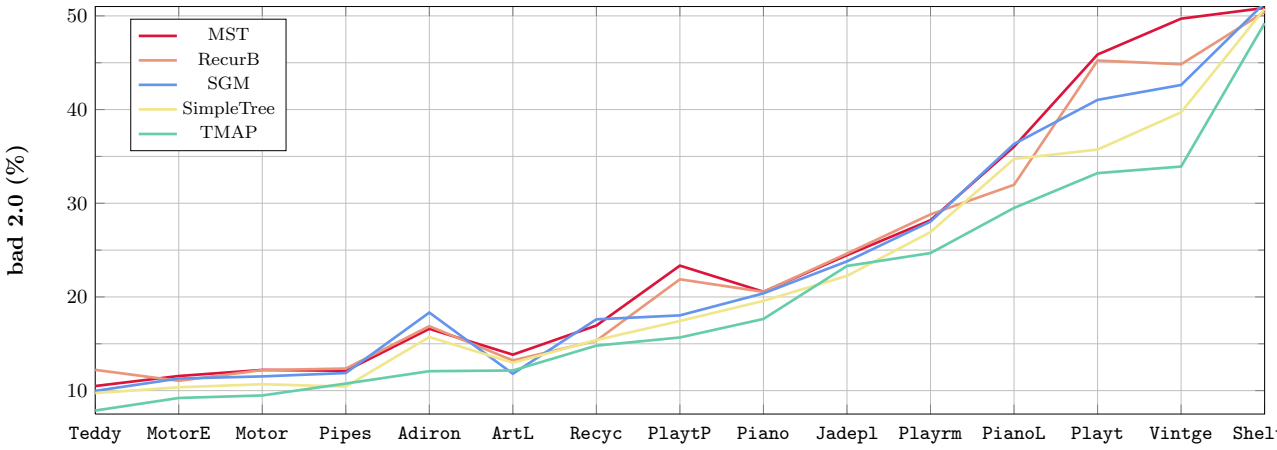
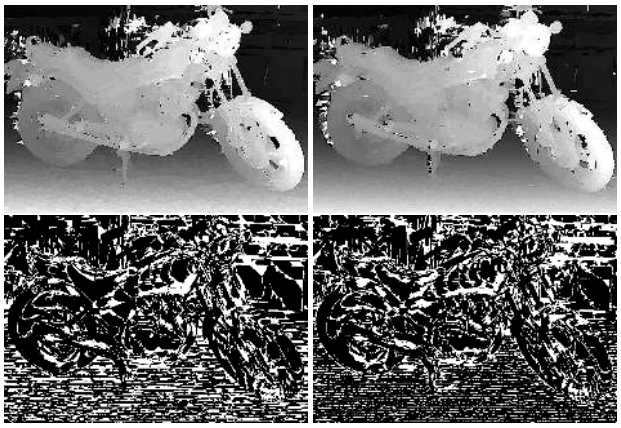


Figure 6: Accuracy of matching using the proposed (TMAP) and related aggregation methods (winner-take-all with no post-processing). To isolate the performance of aggregation, each method uses the same weighted combination of census/gradient cost derived in section 3.2.2. Related aggregation methods include MST filtering (MST) [25], semi-global (SGM) [9], recursive bilateral filtering (RecurB) [26], and simple trees (SimpleTree) [3].



(a) MST (b) **TMAP**

Figure 7: A comparison of results obtained using the cost aggregation schemes of the MST filtering method [25] and the proposed method (TMAP). In the bottom images, the white areas indicate absolute disparity errors ≥ 1.

which these methods aggregate does not take advantage of visual principles of grouping [27]. The minimum spanning trees used by TMAP, on the other hand, enforce visual grouping of shapes while still allowing explicit penalization of disparity transitions.

An illustration of the proposed method's ability to recover disparities along a slanted surface is provided in Fig. 7. Here, TMAP is compared to MST filtering, since both methods perform aggregation on minimum spanning trees. Recall that MST filtering does not consider disparity transitions during aggregation and, as such, produces a rough approximation of the surface corresponding to the ground plane in the scene. Conversely, the proposed method evaluates the probabilities of all possible disparity transitions. The transition probabilities estimated from the available ground truth data strongly favor gradual transition in disparity, as previously discussed in section 3.2.1. As a result, the proposed method accurately handles disparity estimation along slanted surfaces.

5. Conclusion

This paper presents a method for performing maximum a posteriori (MAP) disparity estimation on minimum spanning trees that facilitate the efficient exchange of disparity evidence across the entire image. The proposed method uses an implementation of the upward-downward algorithm to aggregate costs through message passing between nodes in the minimum spanning tree. The messages, which represent disparity likelihoods and probabilities of disparity transitions, are derived from statistical relationships between matching costs and true disparities.

When evaluated using the Middlebury stereo benchmark version 3, the proposed method is among the top performers. While the parameters of the method are learned from the training set, results on the test set demonstrate that these parameters are not specific to the training set, and that the method is capable of performing well under a variety of challenging conditions. The aggregation technique used by the proposed method has also been demonstrated to outperform existing tree-based aggregation techniques. When compared to another method that aggregates using MSTs, it is also shown that the proposed method allows for more accurate disparity estimation along slanted surfaces without requiring higher-order smoothness terms. It is noteworthy that these results are achieved without requiring contextual knowledge of the scene or surface fitting.

References

[1] Anonymous. A global stereo model with mesh alignment regularization for view interpolation. Listed on the Middlebury stereo benchmark, 2015. 6

[2] Anonymous. Using local cues to improve dense stereo matching. Listed on the Middlebury stereo benchmark, 2015. 6

[3] M. Bleyer and M. Gelautz. Simple but effective tree structures for dynamic programming-based stereo matching. In *In VISAPP*, pages 415–422, 2008. 1, 2, 4, 7, 8

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov. 2001. 2

[5] C. Cigla and A. A. Alatan. Information permeability for stereo matching. *Signal Processing: Image Communication*, 28(9):1072–1088, Oct. 2013. 6

[6] J.-B. Durand, P. Goncalves, and Y. Guedon. Computational methods for hidden Markov tree models-an application to wavelet trees. *IEEE Transactions on Signal Processing*, 52(9):2551–2560, Sept. 2004. 2, 3

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision*, 70(1):41–54, May 2006. 6

[8] J. Forney, G.D. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, Mar. 1973. 2

[9] H. Hirschmuller. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, Feb. 2008. 1, 2, 4, 6, 7, 8

[10] H. Hirschmuller and D. Scharstein. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, Sept. 2009. 5

[11] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 3rd edition, 2013. 5

[12] V. Kolmogorov. Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, Oct. 2006. 2

[13] J. Kowalczuk, E. Psota, and L. Perez. Real-Time Stereo Matching on CUDA Using an Iterative Refinement Method for Adaptive Support-Weight Correspondences. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1):94–104, Jan. 2013. 6

[14] J. B. Kruskal, Jr. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, Feb. 1956. 3

[15] T. Matsui. The minimum spanning tree problem on a planar graph. *Discrete Applied Mathematics*, 58(1):91–94, Mar. 1995. 6

[16] Y. Ohta and T. Kanade. Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(2):139–154, Mar. 1985. 2

[17] E. Psota and L. Perez. LDPC decoding and code design on extrinsic trees. In *IEEE International Symposium on Information Theory, 2009. ISIT 2009*, pages 2161–2165, June 2009. 2

[18] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In X. Jiang, J. Hornegger, and R. Koch, editors, *Pattern Recognition*, Lecture Notes in Computer Science, pages 31–42. Springer International Publishing, Jan. 2014. 1

[19] D. Scharstein and C. Pal. Learning Conditional Random Fields for Stereo. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, pages 1–8, June 2007. 2

[20] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, Apr. 2002. 1, 5

[21] S. Sinha, D. Scharstein, and R. Szeliski. Efficient High-Resolution Stereo Matching Using Local Plane Sweeps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1582–1589, June 2014. 6

[22] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, July 2003. 2

[23] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, June 2008. 2

[24] O. Veksler. Stereo correspondence by dynamic programming on a tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, pages 384–390 vol. 2, June 2005. 2

[25] Q. Yang. A non-local cost aggregation method for stereo matching. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1402–1409, June 2012. 1, 2, 7, 8

[26] Q. Yang. Recursive Bilateral Filtering. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, number 7572 in Lecture Notes in Computer Science, pages 399–413. Springer Berlin Heidelberg, 2012. 7, 8

[27] K.-J. Yoon and I.-S. Kweon. Locally adaptive support-weight approach for visual correspondence search. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, pages 924–931 vol. 2, June 2005. 2, 3, 8

[28] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In J.-O. Eklundh, editor, *Computer Vision — ECCV '94*, number 801 in Lecture Notes in Computer Science, pages 151–158. Springer Berlin Heidelberg, 1994. 5