arXiv:1601.03890v1 [cs.CV] 15 Jan 2016

Stereo Matching by Joint Global and Local Energy Minimization

Hongyang Xue, Deng Cai

Abstract

In [7], Mozerov et al. propose to perform stereo matching as a two-step energy minimization problem. They formulate cost filtering as a local energy minimization model, and solve the fully connected MRF model and the locally connected MRF model sequentially.

In this paper we intend to combine the two steps of energy minimization in order to improve stereo matching results. We propose to jointly solve the fully connected and locally connected models, taking both their advantages into account. The joint model is solved by mean field approximations and the experiments show that the locally connected potential makes up for the loss caused by fast approximations. While remaining efficient, our joint model outperforms the two-step energy minimization approach on the Middlebury stereo benchmark.

1 Introduction

Stereo matching is one of the fundamental approaches to acquire depth information. It is important in many fields such as autonomous driving, robotics, intermediate view generation, and 3D scene reconstruction. Stereo matching can be divided into two categories: cost filtering methods and energy minimization methods. Stereo matching is actually a discrete labeling problem where we need to find a solution which

- minimizes the label costs
- is spatially smooth

It is straightforward and popular to utilize a Conditional(Markov) Random Field Model. By this formulation, label costs are encoded in unary potentials and spatially smooth solution is enforced by pairwise potentials. The total energy can be minimized using approaches such as graph cut or loopy belief propagation. However, the global energy minimization approach suffer from a drawback that they are often very slow. Fast approximations usually introduce large errors, making the solutions less-global optimal [8].

Cost filtering approaches are usually fast and tractable compared with energy minimization schemes. They are based on the assumption that all pixels in a matching neighborhood have similar disparities, assuming that same color possesses the same disparity. It works reasonably well when addressing occlusions. However, this assumption limits the range on which information can be propagated due to its fixed spatial range. When facing occlusion problems, the

spatial extend of the bilateral filters are enlarged, causing outliers due to violation of the disparity smoothness in the color space. On the contrary, energy minimization approaches produces errors especially in the occluded regions of scenes.

Mozerov et al. [7] showed that bilateral filtering can be written as an energy minimization problem on a fully connected CRF. Densely connected CRFs are rare in stereo matching problems, because they are usually considered intractable due to unsolvable computational complexity. However, under the situation of Gaussian edge potentials, an efficient solution to fully connected CRFs was proposed in [5] with accelerated message passing by high-dimensional filtering. The fast solution to fully connected CRFs has been proved to be effective and efficient in semantic image segmentation tasks.

1.1 Motivation

Our work is inspired by and based on the two-step energy minimization approach [7]. In [7], Mozerov et al. showed that the cost filtering method can be rewritten as an energy minimization problem. They proposed an idea by unifying cost filtering and energy minimization methods together into a single two-step energy minimization algorithm to compensate the drawbacks of both approaches. In their unified approach, the energy min-marginals of the fully connected model(FCM) are calculated first, and then these values are used as an input cost for the second step of energy minimization with locally connected model(LCM).

Stereo matching approaches usually contain two parts: the matching cost computation and the stereo method. In the two-step energy minimization framework, the authors expect that the output of the FCM minimization problem yields a more robust unary potential for the LCM. There are many approaches focusing on improving the unary cost for matching, including methods based on complex handcrafted features and convolutional neural networks [13]. We propose a stereo approach which can utilize all kinds of stereo matching costs with little modification to the whole algorithm.

There are different algorithms for solving fully connected and locally connected CRFs. For the LCM, the TRW-S [4] algorithm is the most frequently used. For the FCM, there're early methods such as belief propagation [3] and graph cuts [1]. Lately, as mentioned above, an efficient algorithm based upon mean field inference was proposed in [5].

1.2 Contribution

The fully and the locally connected models have their advantages and weaknesses respectively. The fully connected models have loss in quality due to fast approximations. Thus the solutions of the fully connected models are *biased*. In addition, they do not scale well to high-resolution images. The locally connected one has solutions that are smooth limited in a small neighborhood. It lacks global information and is somewhat *short-sighted*. It is apparent and reasonable that a stereo approach need to consider both the models simultaneously rather than solving them sequentially. Thus we propose to jointly solve the FCM and LCM model, maximizing their cooperation.

Our contributions are two-fold: First, we propose to jointly minimize the global energy and local energy terms so that their drawbacks can be inhibited mutually. Experiments show that out joint model outperforms the original two-step model. Second, we follow the fast mean field algorithm in [5] for solving the joint model. To our knowledge, their algorithm has not been applied in stereo matching. The global energy term obtains good stereo results while the local term inhibits the errors introduced by fast approximation solutions. Experiments show that the local energy term do penalize the loss caused by approximate inference.

The joint model will be explained in Section 2. The fully connected part will be explained in Section 2.1, and the locally connected part will be explained in Section 2.2. In Section 3, the mean field inference solution to our joint model will be derived and explained. In section 4 we will discuss post-processing techniques. In section 5 we will display experiments and results.

2 Problem Definition

Given a pair of rectified images, we wish to compute the disparity d for each pixel in the left image. Disparity refers to the difference in horizontal location of an object in the left and right image. Assume there are M disparity levels, for a pair of images, disparity of each pixel takes finite discrete values from $\mathcal{D} = \{0, 1, \cdots, M-1\}$.

The stereo matching problem is usually formulated as a random field $\mathbf{D}$ defined over a set of variables $\{D_1, \cdots, D_N\}$ and the domain of each variable is $\mathcal{D}$. The Gibbs energy of a labeling $\mathbf{d} \in \mathcal{D}^N$ is

$$E(\mathbf{d}) = \sum_{i \in \mathcal{V}} \psi_u(d_i) + \sum_{(i,j) \in \mathcal{E}} \Psi_p(d_i, d_j) \tag{1}$$

where the set $\mathcal{V}$ corresponds to pixels and set $(i, j) \in \mathcal{E}$ to edges of an image graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$; d_i denotes the disparity of pixel i and belongs to the discrete set of disparities $\mathcal{D}$. ψ_u is the unary term which takes the conventional penalty cost; $\Psi_p(i, j)$ is the pairwise potential which encodes the interactions between pixels (i, j).

The choice of unary potential is a key in the process of stereo matching. There are many kinds of costs, from simple per-pixel matching dissimilarity measure and non-parametric transforms [12], [6] with a support region, to costs calculated by convolutional neural networks [13]. In [7], the authors argue that cost with a support region is redundant for their fully connected model. However, in our joint model, we select this kind of cost. In their approach, the result of FCM only provides a better unary cost for the LCM, so the unary cost with a support region is redundant for FCM. In our model, both the FCM and LCM utilizes the same unary cost. As a result, we apply the cost by gradient image and Census transform which is used in semi-global matching [2] for simplicity and sufficiency. Notice that it is convenient to employ more sophisticated and representative costs to our model.

In our model, the pairwise potential consists of two parts: the fully connected pairwise $\psi_p(d_i, d_j)$ and the locally connected ones $\tilde{\psi}_p(d_i, d_j)$. We can write the full model as:

$$E(\mathbf{d}) = \sum_i \psi_u(d_i) + \sum_i \sum_{j>i} \psi_p(d_i, d_j) + \sum_i \sum_{j \in \mathcal{N}(i)} \tilde{\psi}_p(d_i, d_j) \tag{2}$$

The problem is to find $\mathbf{d}$ that minimizes the above energy. Following conventional formulation, we transform the energy into Gibbs distribution, where Z is the partition function and $\mathbf{X}$ is the observation of the image pair.

$$P(\mathbf{d}|\mathbf{X}) = \frac{1}{Z} \mathrm{e}^{-E(\mathbf{d})} \tag{3}$$

In the other word, we need to do MAP(maximum a priori) inference to the Gibbs distribution $P(\mathbf{d}|\mathbf{X})$ given images $\mathbf{X}$. For convenience we will omit $\mathbf{X}$ in the remaining part of this paper.

$$\arg\max_{\mathbf{d}} P(\mathbf{d}) \tag{4}$$

2.1 Fully Connected Pairwise Potential

The fully connected pairwise potential in our joint model is

$$\psi_p(d_i, d_j) = \omega \cdot \mu(d_i, d_j) \cdot \exp(-\frac{|x_i - x_j|^2}{2\sigma_x^2} - \frac{|f_i - f_j|^2}{2\sigma_f^2}) \tag{5}$$

where $\mu(d_i, d_j)$ is the label compatibility function and we adopt the widely used Potts model, x_i is the position of pixel i and f_i is the feature of pixel i. We take x_i the coordinate of the pixel and f_i the color vector of the pixel.

This term is a Potts model bilateral filter and called Gaussian edge potential in [5]. It takes into account the interactions of every pair of pixels in the image. The Potts model implies that when two pixels have the same disparity value, their pairwise potential will not be considered. The bilateral kernel enforces smoothness in both color(appearance) and spatial domain. The appearance term is inspired from the observation that pixels with similar colors tend to own the same disparity value. The spatial kernel confines the influence of the appearance kernel within some spatial range.

We denote the bilateral filter as $k(i, j)$, that is,

$$k(i, j) = \exp(-\frac{|x_i - x_j|^2}{2\sigma_x^2} - \frac{|f_i - f_j|^2}{2\sigma_f^2}) \tag{6}$$

In semantic image segmentation [5], the pairwise potential contains a Gaussian kernel and a bilateral kernel. Noting that different from the bilateral kernel, the Gaussian kernel only enforces spatial smoothness which is also addressed in the locally connected potential in our model. To reduce redundancy we omit the fully connected Gaussian pairwise energy.

2.2 Locally Connected Pairwise Potential

The locally connected pairwise potential (7) is the same as the one in [7]. It only takes into account a small neighborhood of each pixel and penalizes disparity unsmoothness in a small region. It enforces local smoothness of the disparity map:

$$\tilde{\psi}_p = \tilde{\omega} \cdot \tilde{w}(i,j) \cdot \tilde{\phi}(d_i, d_j) \tag{7}$$

following [1] and [7], the function $\tilde{w}(i,j)$ divides the set of edges into two or more subsets corresponding to a high or low image gradient:

$$\tilde{w}(i,j) = \begin{cases} \lambda_1 & |\mathbf{I}_i - \mathbf{I}_j| < \mu_1 \\ \lambda_2 & \mu_1 \leq |\mathbf{I}_i - \mathbf{I}_j| < \mu_2 \\ \lambda_3 & |\mathbf{I}_i - \mathbf{I}_j| \geq \mu_2 \end{cases} \tag{8}$$

where

$$|\mathbf{I}_i - \mathbf{I}_j| = \sum_{c \in \{r,g,b\}} |I_i^c - I_j^c|$$

The weights in (8) implies strong dependency between nodes when the colors are similar. The smoothness multiplier $\tilde{\phi}(d_i, d_j)$ is chosen as:

$$\tilde{\phi}(d_i, d_j) = \begin{cases} 0 & |d_i - d_j| = 0 \\ \beta & |d_i - d_j| = 1 \\ 1 & |d_i - d_j| > 1 \end{cases} \tag{9}$$

As pointed out in [7], a nonzero value of $|d_i - d_j|$ is usually considered as the discontinuity of the disparity map and this is not true in general because a value 1 for $|d_i - d_j|$ could be caused by discretization errors. Thus the case $|d_i - d_j| = 1$ is specifically picked out and assigned a small penalization β. We also tried to pick out the case $|d_i - d_j| = 2$ and experiments show that adding more pieces to the multiplier function is useless.

3 Inference

Our model (2) contains a fully connected energy term which is intractable to find an exact solution. In [5], a fully connected CRF is applied for semantic image segmentation where krahenbuhl et al. employ mean field approximation with high-dimensional filtering for the inference. Following their idea, we also employ mean field approximation to inference our joint model.

Mean Field Inference with high-dimensional filtering has been proved in [5] to be an efficient solver for dense connected CRF models. The basic idea of Mean Field Approximation is that instead of computing the exact distribution $P(\mathbf{D})$, we approximate the original distribution with a factored distribution $Q(\mathbf{D})$ that minimizes the KL-divergence $\mathbf{D}(Q||P)$. The factored distribution $Q(\mathbf{D})$ is the product of distribution on individual pixels:

$$Q(\mathbf{D}) = \prod_i Q_i(D_i) \tag{10}$$

When minimizing the KL-divergence, the mean field approximation yields an iterative update equation:

Algorithm 1 Mean Field in our joint model

Initialize $Q_i(d_i) \leftarrow \frac{1}{Z_i}\exp(-\psi_u(d_i))$
while not converged **do**

1. $\tilde{Q}_i(l) \leftarrow \sum_{j\neq i} k(i,j)Q_j(l)$, $\tilde{P}_i(l) \leftarrow \sum_{j\in\mathcal{N}(i)} \tilde{w}(i,j)Q_j(l)$
2. $\hat{Q}_i(d_i) \leftarrow \sum_{l\in\mathcal{D}} \omega \cdot \mu(d_i,l)\tilde{Q}_i(l)$, $\hat{P}_i(d_i) = \sum_{l\in\mathcal{D}} \tilde{\omega} \cdot \tilde{\phi}(d_i,l)\tilde{P}_i(l)$
3. $Q_i(d_i) = \exp(-\psi_u(d_i) - \hat{Q}_i(d_i) - \hat{P}_i(d_i))$
4. normalize $Q_i(d_i)$

end while

$$\begin{aligned} Q_i(d_i) \leftarrow \exp\{\psi_u(d_i) - &\sum_{j\neq i}\sum_{d_j\in \mathrm{Val}(D_j)} \psi_p(d_i,d_j)Q_j(d_j) \\ &- \sum_{j\in\mathcal{N}_i}\sum_{d_j\in \mathrm{Val}(D_j)} \tilde{\psi}_p(d_i,d_j)Q_j(d_j)\} \end{aligned} \tag{11}$$

Based on this iterative update equation, we extend the inference algorithm in [5] which has meassage passing, compatibility transform and local update steps to solve our joint model(see Algorithm 1)

After the marginal of the factored distribution is computed on each pixel, we employ the winner-take-all strategy to decide the disparity of each pixel.

$$d_i = \arg\max_d Q_i(d) \tag{12}$$

3.1 Time Complexity Analysis

One iteration of the algorithm has time complexity $\mathcal{O}(MN)$ where N is number of pixels and M is the total number of disparity levels. It is obvious that the second(Compatibility transform), third(Local update) and fourth step in the iteration of Algorithm 1 has complexity $\mathcal{O}(MN)$. The first step(Message passing) seems to have complexity quadratic in N. Luckily in [5] Krahenbuhl et al. propose to employ a trick of fast high-dimensional filtering which accelerates the message passing step to $\mathcal{O}(MN)$.

4 Post-processing

Post-processing steps are indispensable for obtaining good stereo results. All sate-of-the-art algorithms in stereo use post-processing techniques. Some post processing approaches are fast, while the others can be quite time-consuming. We utilize several post-processing techniques. First we follow the post processing methods in [7] and take only the first step: Left-to-right disparity map cross-checking(LRC). Experimental results show that our approach outperforms the two-step algorithm even with only this single step of post-processing.

Besides the LRC step, we also employ several other techniques to post-process our stereo results. We describe the different post processing we apply in this section including occlusion filling, median filter and slanted plane

smoothing(SPS). And we will compare and analyze the impact of them in the experiments.

4.1 Disparity Map Occlusion Filling

We fill the occluded pixels through a simple strategy. First of all, we pick out all the inconsistent pixels using LRC. Rather than correcting them with the smaller value [7], we mark the inconsistent pixels as invalid. Then we fill them with the smallest disparity of the nearest valid pixels.

4.2 Weighted Median Filter

As is mentioned in [7], weighted median filter is a robust extension of the bilateral filter and widely used in stereo matching processing. Mozerov et al. [7] propose to apply weighted median filter to the whole disparity map which is quite time-consuming. We propose to employ weighted median filtering only to the invalid pixels. After the invalid pixels are filled with occlusion filling, we apply weighted median filter to the disparity map.

4.3 Slanted Plane Smoothing

Slanted plane smoothing(SPS) is proposed by Yamaguchi et al. [11] as a fast approach to dense depth estimate. Their idea is to jointly solve segmentation, occlusion labeling and stereo estimation based on a semi-dense depth map. Though our joint model outputs a dense depth map, we adopt their idea to post process our results. Experiments show that SPS is very effective and efficient.

5 Experiments and Results

In this section we evaluate our joint energy minimization approach on the Middlebury stereo benchmark v3. All the evaluations are run and compared under the metric of **avgErr** with nocc. We start by comparing the locally connected pairwise potential with the fully connected ones. By analyzing the results we show that the joint model outperform every single model, and the LCM and the FCM improves the stereo results of each other.

Then we compare the results of our joint model to the two-step energy minimization approach [7] and show that our joint model outperforms theirs.

Next we analyze and compare the effects of the post-processing techniques we employ. Furthermore, we compare our approach with other approaches on the Middlebury Benchmark v3 in both time and error.

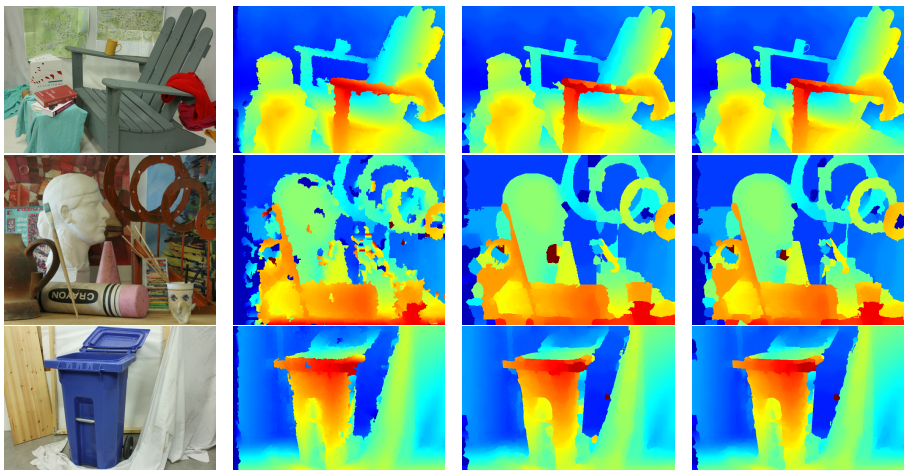
Our algorithm run on quarter-size images of the Middlebury dataset and they are up-sampled by the official evaluation toolkit.

5.1 FCM and LCM Analysis

Here we compare the results of FCM, LCM and FCM+LCM(the joint model, we will also use JEM for abbreviation). The result is shown in Table 1 and Figure 1. In this section, we only adopt the post-processing strategy of LRC(left-to-right cross checking).

Model	LCM	FCM	JEM(FCM+LCM)
Avg	7.32	6.58	**6.34**
Adiron	2.64	2.32	**1.93**
ArtL	7.56	5.40	**4.92**
Jadepl	19.1	16.8	**15.5**
Motor	4.06	3.46	**3.26**
MotorE	4.19	3.77	**3.51**
Piano	5.69	**4.52**	4.62
PianoL	17.1	**13.0**	13.2
Pipes	6.66	5.61	**5.33**
Playrm	7.99	**5.69**	5.79
Playt	**8.95**	16.6	14.2
PlaytP	3.89	4.05	**3.36**
Recyc	3.00	2.75	**2.47**
Shelves	13.7	**11.2**	12.4
Teddy	4.21	**3.32**	4.16
Vintage	**13.3**	14.2	14.9

Table 1: Compasion of LCM, FCM and JEM(LCM+FCM) models in avgErr. The table shows the error of the results of all 15 data sets and their average.



(a) Left (b) LCM (c) FCM (d) JEM

Figure 1: Results of LCM, FCM and JEM on Middlebury 3.0 [9] benchmark (Adirondack, ArtL, Recycle).

In average, out joint model outperforms the single FCM or LCM. We can also see from the table that on most data sets, our joint model works better than single FCM or LCM model. Mozerov et al. [7] employ the two-step energy minimization approach to improve stereo results in occluded regions. But in our evaluation we only consider non-occuluded regions which is the default of Middlebury benchmark v3, and as a result, we can see that the results of FCM is rather close to JEM compared with LCM. However, in most cases, e.g. Adiron, ArtL, PlaytP, JEM works better than FCM. We can see that LCM effects like a regularization term for the approximation inference of the random field.

5.2 Joint Model vs. Two-step Model

Now we compare out joint model JEM to the previous two-step energy minimization approach(TSGO) in [7]. For our JEM model, we only apply one step post-processing: LRC. The TSGO has several steps of post processing. The result is shown in Table 2.

Algorithm	TSGO	JEM
Avg.	7.07	**6.34**
Adiron	2.02	**1.93**
ArtL	**3.07**	4.92
Jadepl	32.5	**15.5**
Motor	**3.12**	3.26
MotorE	**2.94**	3.51
Piano	**2.96**	4.62
PianoL	16.1	**13.2**
Pipes	**4.90**	5.33
Playrm	**4.02**	5.79
Playt	18.7	**14.2**
PlaytP	**2.20**	3.36
Recyc	**2.33**	2.47
Shelves	**8.34**	12.4
Teddy	**2.46**	4.16
Vintage	**12.6**	14.9

Table 2: Comparison of TSGO with our joint model JEM in avgErr. Our JEM has only one simple post-processing step: LRC. We can see JEM is better than TSGO in average and especially for data set Jadeplant.

Even with only a single simple step of post-processing, our model outperforms TSGO in average. The result shows that the joint model has advantages over the splitted two-step model. Especially for the data set Jadeplant, TSGO has average error 32.5 while our JEM achieves 15.5 which is surprisingly only half of it.

After we add more post-processing steps, our results are further improved(see Table 4).

5.3 Post-processing Analysis

In this section, we compare the results of different post-processing techniques. First we utilize left-to-right cross checking(LRC), and then we apply slanted plane smoothing(SPS). In the end we employ occlusion filling(OF) with weighted median filter(WMF) in addition to the previous SPS. The result is shown in Table 3.

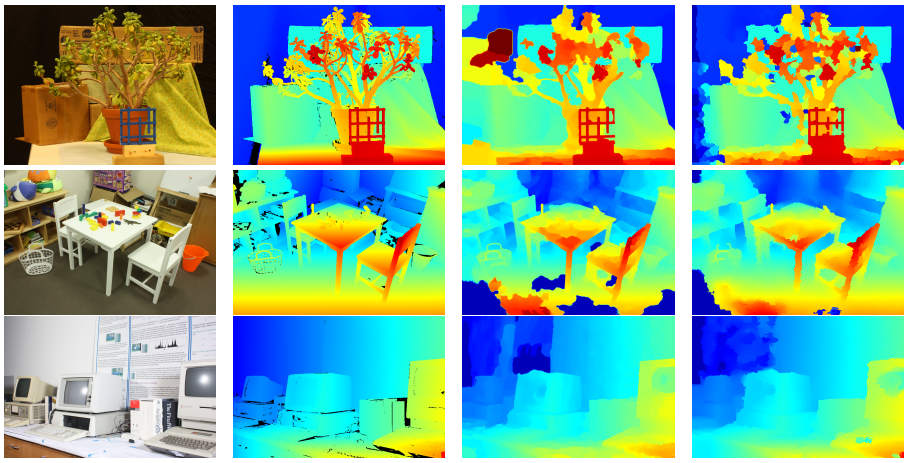
Algorithm	LRC	SPS	WMF + OF + SPS
Avg	6.34	5.52	**5.33**
Adiron	**1.93**	1.96	1.96
ArtL	4.92	4.90	**4.32**
Jadepl	15.5	16.4	**15.5**
Motor	3.26	3.28	**3.23**
MotorE	3.51	**3.25**	3.26
Piano	4.62	3.60	**3.57**
PianoL	13.2	12.4	**12.1**
Pipes	5.33	5.33	**5.14**
Playrm	5.79	5.69	**5.69**
Playt	14.2	6.73	**6.71**
PlaytP	3.36	2.62	**2.56**
Recyc	2.47	2.40	**2.34**
Shelves	12.4	9.49	**9.28**
Teddy	4.16	3.39	**3.17**
Vintage	14.9	9.40	**9.40**

Table 3: Comparison of different post-processing techniques in avgErr. LRC refers to left-to-right cross checking. OF refers to occlusion filling. WMF refers to weighted median filter and SPS mean slanted plane smoothing.

From Table 3 we can see that in average WMF + OF + SPS effects the best. This is of no surprise because more post processing usually lead to better results. Surprisingly we see little improvement on several data sets when adding WMF and OF to SPS, and even worse result on MotorE. We observe that the Middlebury stereo dataset mainly contains images of indoor scenes where walls and planar objects occupy the majority of the scenes. Thus stereo methods which rely on planar elements have natural advantages over the others. As a result, the SPS can play a key role in the results and it improves the original stereo results a lot.

5.4 More Comparisons

In this section, we compare our approach to the other approaches on the table of Middlebury benchmark v3. We compare our results with the others in both aspects of running time and accuracy.



(a) Left (b) GT (c) TSGO (d) JEM(Ours)

Figure 2: Results on the Middlebury 3.0 [9] benchmark with ground truth (Jadeplant, Playtable, Vintage).

5.4.1 Error

We compare our results with several other methods(most of them are recently proposed). Table 4 lists the results of several approaches on the Middlebury

Alg.	MC-CNN-art [13]	MeshStereo [14]	JEM	SGM [2]	TSGO	LPS [10]
Avg	3.82	4.38	5.33	5.82	7.07	7.58
Adiron	0.76	1.55	1.96	4.62	2.02	2.15
ArtL	2.55	2.93	4.32	2.34	3.07	12.6
Jadepl	16.3	16.2	15.5	13.2	32.5	15.4
Motor	1.27	2.83	3.23	2.97	3.12	1.45
MotorE	1.27	3.42	3.26	2.49	2.94	27.6
Piano	1.83	2.71	3.57	3.24	2.96	3.40
PianoL	5.07	4.94	12.1	5.18	16.1	20.1
Pipes	2.29	6.03	5.14	5.35	4.90	3.63
Playrm	2.27	4.57	5.69	4.40	4.02	4.60
Playt	3.11	3.82	6.71	23.9	18.7	3.24
PlaytP	3.03	2.02	2.56	3.61	2.20	3.15
Recyc	2.48	1.65	2.34	4.08	2.33	2.44
Shelves	4.41	10.7	9.28	8.83	8.34	9.51
Teddy	1.12	1.09	3.17	1.52	2.46	1.83
Vintage	14.8	4.64	9.40	16.2	12.6	4.76

Table 4: Comparison with other approaches in avgErr. Notice first that JEM outperforms TSGO especially on Jadeplant, Playtable and Vintage (see Figure 2). Our result are also better than that of MC-CNN-art on Jadeplant and Vintage.

benchmark v3. First of all, we can see our result outperform TSGO on many data sets and especially on Jadeplant, Playtable and Vintage. On the other data sets the results are rather close. We also observe that different approaches have advantages on different data sets. Even though LPS has the worst average result, it shows the best effect on Vintage. Our JEM also outperforms MC-CNN-art on Jadeplant and Vintage, and outperforms MeshStereo on several data sets including Jadeplant and Shelves.

5.4.2 Runtime Analysis

Alg.	MC-CNN-art	MeshStereo	JEM	SGM	TSGO	LPS
Runtime	106	62.0	20.8	52.3	53.8	9.35

Table 5: Comparison with other approaches in time(in seconds). We can see that the CNN based method is extremely slow. MeshStereo and TSGO are also slower than JEM. Only LPS is faster than JEM.

Comparing our approach with others in running time (see Table 5), we can see our approach is efficient compared with high-accuracy methods(MC-CNN-art and MeshStereo). JEM is also efficient when compared with approaches that have close error(SGM, TSGO). Though it takes a lot more time than LPS, it is more accurate than LPS.

6 Conclusion

We propose a joint energy minimization approach(JEM) based on the two-step energy minimization method(TSGO) [7]. The joint energy term consists of two parts: a global pairwise potential and a local one. JEM enables the effect of both models simultaneously and shows superior results to TSGO. In order to solve the joint model efficiently, we employ mean field approximation [5] to inference the posterior probability of our model. The mean field inference to the joint model is much more efficient than separate solutions to each models. As a result, JEM is faster and more effective than TSGO which is shown by experiments on the Middlebury benchmark v3. JEM is also comparable to several the state-of-the-art approaches in both computational time and estimation error.

References

[1] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[2] Heiko Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 807–814. IEEE, 2005.

[3] Alexander T Ihler, John Iii, and Alan S Willsky. Loopy belief propagation: Convergence and effects of message errors. In *Journal of Machine Learning Research*, pages 905–936, 2005.

[4] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, 2006.

[5] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *arXiv preprint arXiv:1210.5644*, 2012.

[6] John P Lewis. Fast template matching. In *Vision interface*, volume 95, pages 15–19, 1995.

[7] Mikhail G Mozerov and Joost van de Weijer. Accurate stereo matching by two-step energy minimization. *Image Processing, IEEE Transactions on*, 24(3):1153–1163, 2015.

[8] Christoph Rhemann, Asmaa Hosni, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3017–3024. IEEE, 2011.

[9] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[10] Sudipta N Sinha, Daniel Scharstein, and Richard Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1582–1589. IEEE, 2014.

[11] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Computer Vision–ECCV 2014*, pages 756–771. Springer, 2014.

[12] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer VisionECCV'94*, pages 151–158. Springer, 1994.

[13] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *arXiv preprint arXiv:1510.05970*, 2015.

[14] Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2057–2065, 2015.