2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Dense Disparity Estimation Based on Feature Matching and IGMRF Regularization

Sonam Nahar
The LNMIIT, Jaipur (India)

Manjunath V. Joshi
DA-IICT, Gandhinagar (India)

***Abstract*—In this paper, we propose a new approach for dense disparity estimation in a global energy minimization framework. We combine the feature matching cost defined using the learned hierarchical features of given left and right stereo images, with the pixel-based intensity matching cost to form the data term. The features are learned in an unsupervised way using the *deep deconvolutional network*. Our regularization term consists of an inhomogeneous Gaussian markov random field (IGMRF) prior that captures the smoothness as well as preserves sharp discontinuities in the disparity map. An iterative two phase algorithm is proposed to minimize the energy function in order to estimate the dense disparity map. In phase one, IGMRF parameters are computed, keeping the disparity map fixed, and in phase two, the disparity map is refined by minimizing the energy function using graph cuts, with other parameters fixed. Experimental results on the Middlebury stereo benchmarks demonstrate the effectiveness of the proposed approach.**

I. Introduction

Stereo vision aims to find the disparity between corresponding pixels, i.e. pixels resulting from the projection of the same 3D point onto the two image planes. However, estimation of disparities is an ill-posed problem due to depth discontinuities, photometric variation, lack of texture, occlusions etc. Global stereo methods tackle such problems by solving the dense disparity estimation problem in an energy minimization framework where the energy function represents a combination of data term and a regularization term [1].

The data term is generally defined by using the pixel based matching cost between the corresponding pixels in the left and right images [1]. Most of these matching costs are built on the brightness constancy assumption, and include absolute differences (AD), squared differences (SD), sampling insensitive absolute differences of Birchfield and Tomasi (BT), or truncated costs [1]. They rely on raw pixel values, and are less robust to illumination changes, view point variation, noise, occlusion etc. One may represent stereo images using a feature space where they are robust, distinct and transformation invariant. Feature based stereo methods use the features such as edges, gradients, corners, segments, or hand crafted features such as scale-invariant feature transform (SIFT) [2]–[5]. In [3], non overlapping segments of stereo images are used as features, and the dense stereo matching problem is cast as an energy minimization in segment domain instead of pixel domain where the disparity plane is assigned to each segment via belief propagation. However, the method assume that the disparities in a segment vary smoothly which is not true in practice due to the depth discontinuities. Also the solution here relies on the accuracy of segmentation which is itself a non trivial task. Authors in [6] find sparse correspondences by feature points and then the dense correspondences are obtained from these sparse matches using seed growing methods. In such approaches, the accuracy depends on the initial support points. In [7], the mutual information (MI) based feature matching is used in an MRF framework. However, matching with basic image features still results in ambiguities in correspondence search, especially for textureless areas and wide baseline stereo. Hence, to reduce these ambiguities, one needs to use more descriptive features. Recently in [4] and [5], dense stereo matching problem is solved in MRF regularization framework by matching the hand crafted features such as SIFT. The drawback of these approaches is that they are computationally expensive and require domain knowledge of the data.

In recent years, learning features from unlabeled data using unsupervised feature learning and deep learning approaches have achieved superior performance in solving many computer vision problems [8], which has also attracted the attention of stereo vision researchers. Nevertheless, little has been investigated to use learned features matching in an energy minimization framework for disparity estimation. Authors in [9] learn the features from a large number of image patches using K-singular value decomposition (K-SVD) dictionary learning approach. The limitation of their approach is that the features are learned from a set of image patches and do not consider the spatial correlation globally. Also, higher level features are not learned, instead they are estimated using a simple max pooling operation from the layer beneath. The higher layer correspondences are used to initialize the lower layer matching, and hence the accuracy depends on the higher layer matches only. They use the K-SVD method which is linear, whereas non-linear, unsupervised feature learning methods have shown superior performance in learning efficient representation of images at multiple layers [8], [10]–[13].

In this paper, we combine the feature matching cost defined using the learned hierarchical features of given stereo image pair, with the pixel-based intensity matching cost. To the best of our knowledge, this is the first work which combines the pixel based intensity and feature matching cost in an energy minimization framework for dense disparity estimation. These features are learned using the *deep deconvolutional architecture* [12]. Our deep deconvolutional network is trained over a large set of stereo images in an unsupervised way, which in turn results in a diverse set of filters. The learned

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

filters capture image information at different levels in the form of low-level edges, mid-level edge junctions and high-level object parts. Unlike deep convolutional neural networks, deep deconvolutional network is a top down approach where an input image is generated by summing a convolution of the feature maps with learned filters, and it does not spatially pool features at successive layers and hence preserves the mid level cues emerging from the data such as edge intersections, parallelism and symmetry. They scale well to complete images and hence learn the features for the entire input image instead of small size patches. This makes them to consider global contextual constraint while learning.

Since the disparity estimation is an ill-posed problem, one can make it better posed by incorporating a regularizing prior in the energy function. In general, disparities are piecewise smooth, thus making them inhomogeneous. This feature of disparities can be captured by the discontinuity preserving markov random field (MRF) based regularization prior as noted in [1], [14]–[16]. Other regularization based methods such as Mumford Shah regularization [17], second order smoothness prior [18], ground control points [19] etc. have also been used. Many of these techniques use single or a set of global MRF parameters which are either manually tuned or estimated. These global parameters may not adapt to the local structure of the disparity map and hence fail to better capture the spatial dependence among disparities. We need a prior that considers the spatial variation among disparities locally. This motivates us to use an IGMRF prior which was first proposed in [20] for solving the satellite image deblurring problem. IGMRF prior handles the smooth as well as sharp changes in disparity map. In our approach, we estimate the IGMRF parameters, and the same are used while regularizing. An iterative two phase algorithm is proposed to minimize the energy function in order to estimate the dense disparity map where the IGMRF parameters and the disparity map are refined, alternatively.

II. Problem Formulation

The dense stereo matching problem is formulated in an energy minimization framework. For a given rectified pair of stereo images, left image $I_L \in \mathbb{R}^{M\times N}$ and right image $I_R \in \mathbb{R}^{M\times N}$, our objective is to find a disparity map $d \in \mathbb{R}^{M\times N}$ that minimizes the following energy function:

$$E(d) = E_D(d) + E_P(d), \tag{1}$$

where the data term $E_D(d)$ measures how well the d to be estimated agrees with I_L and I_R of a scene. The prior term $E_P(d)$ measures how good it matches with the prior knowledge about the disparity map. For finding the correspondences, we consider search from left to right as well as from right to left and hence relax the traditional ordering constraint used in disparity estimation.

In our work, the data term is defined as a sum of the intensity and feature matching costs i.e.,

$$E_D(d) = E_I(d) + \gamma E_F(d), \tag{2}$$

where, γ controls the weightage of $E_F(d)$. For a given d, the intensity matching cost $E_I(d)$ measures the dissimilarity between the corresponding pixel intensities of I_L and I_R, while the feature matching cost $E_F(d)$ measures the dissimilarity between the corresponding learned features of I_L and I_R. In order to define $E_I(d)$, we use the robust and sampling insensitive measure (BT) proposed in [21]. At pixel location (x, y) having disparity $d(x, y)$, it is given as minimum absolute intensity difference between $I_L(x, y)$ and $I_R(x + d(x, y), y)$ in the real valued range of disparities along the epipolar line and can be written as:

$$E_I(d) = \sum_{(x,y)} \min((\min_{d(x,y)\pm\frac{1}{2}} |I_L(x,y) - I_R(x+d(x,y),y)|), \tau^I), \tag{3}$$

where, τ^I is the truncation threshold which is used to make intensity matching cost more robust against outliers. For defining the feature matching cost $E_F(d)$, we extract the features of stereo image pair at multiple layers of deep deconvolutional network, and the same is discussed in next section. In order to model d using its prior characteristics, we define the prior term $E_P(d)$ using IGMRF.

III. Deep Deconvolutional Network for Extracting Hierarchical Features and Defining $E_F(d)$

Deconvolutional network [12] is an unsupervised feature learning model that is based on the convolutional decomposition of images under sparsity constraint and generates sparse, overcomplete features. Stacking such network in a hierarchy results in a deep deconvolutional network. Layers of such network learn both the filters and features as done in blind deconvolution problem. We use the deep deconvolutional architecture as proposed in [12]. In order to explain how deep deconvolutional network extracts the hierarchical features, we first consider a deep deconvolutional network consisting of a single layer only. To train this network, a training set consisting of large number of stereo images $\mathcal{I}$=$\{I^1, \ldots, I^{m_s}\}$ are used where each image I^i is considered as an input to the network. Here m_s is the number of images in the training set $\mathcal{I}$, and we consider only left images of different scenes for training the network. Note that one may use right stereo images as well. Let P_1 be the number of $2D$ feature maps in a first layer. Considering the input at layer 0, we can write each image I^i as composed of P_0 channels $\{I^i_1, \ldots, I^i_{P_0}\}$. For example, if we consider a color image then we have P_0=3. Each channel c of input image I^i can be represented as a linear sum of P_1 feature maps s^i_p convolved with filters $f_{p,c}$ i.e.,

$$\sum_{p=1}^{P_1} s^i_p \oplus f_{p,c} = I^i_c, \tag{4}$$

where $\oplus$ represents the 2D convolution operator. Note that in this work, we use gray scale stereo images only and hence P_0=1. Eq. (4) represents an under-determined system since both the features and filters are unknown and hence to obtain a unique solution, a regularization term is also added that

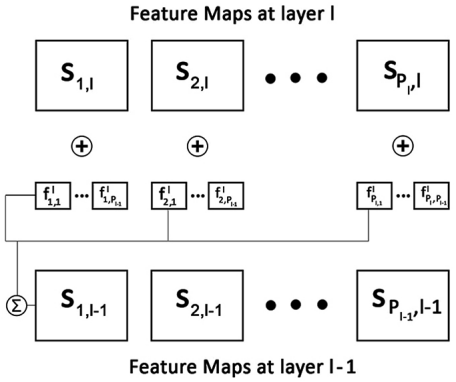

Fig. 1: A deep deconvolutional network consisting of single layer l.

encourages sparsity in the latent feature maps. This gives us an overall cost function for training a single layer deconvolutional network as:

$$C_1(\mathcal{I}) = \sum_{i=1}^{m_s} \left[\frac{\alpha}{2} \sum_{c=1}^{P_0} \left|\left| \sum_{p=1}^{P_1} s_p^i \oplus f_{p,c} - I_c^i \right|\right|_2^2 + \sum_{p=1}^{P_1} |s_p^i|^1 \right]. \quad (5)$$

Here, $|s_p^i|^1$ is the L_1-norm on the vectorized version of s_p^i. The relative weighting of the reconstruction error of each I^i and sparsity of their feature maps s_p^i is controlled by the parameter α. This network is learned by minimizing $C_1(\mathcal{I})$ with respect to s_p^i's and $f_{p,c}$'s when the input to network is $\mathcal{I}$. Note that the set of filters $f_{p,c}$ are the parameters of the network, common to all images in the training set while each image has its own set of feature maps s_p^i.

The single layer network described above can be stacked to form a deep deconvolutional network consisting of multiple layers. This hierarchy is achieved by considering the feature maps of layer $l-1$ as the input to layer l, $l > 0$. Let P_{l-1} and P_l are the number of feature maps at layer $l-1$ and l, respectively. The cost function for training the l^{th} layer of a deep deconvolutional network can be written as a generalization of Eq. (5) as:

$$C_l(\mathcal{I}) = \frac{\alpha}{2} \sum_{i=1}^{m_s} \sum_{c=1}^{P_{l-1}} \left|\left| \sum_{p=1}^{P_l} g_{p,c}^l (s_{p,l}^i \oplus f_{p,c}^l) - s_{c,l-1}^i \right|\right|_2^2 + \sum_{i=1}^{m_s} \sum_{p=1}^{P_l} |s_{p,l}^i|^1, \quad (6)$$

where $s_{c,l-1}^i$ and $s_{p,l}^i$ are the feature maps of image I^i at layer $l-1$ and l, respectively and thus it shows that layer l has as its input coming from P_{l-1} channels. $f_{p,c}^l$ are the filters at layer l and $g_{p,c}^l$ are the elements of a fixed binary matrix that determine the connectivity between the feature maps at successive layers, i.e. whether $s_{p,l}^i$ is connected to $s_{c,l-1}^i$ or not. For $l = 1$, we assume that $g_{p,c}^l$ is always 1, but for $l > 1$, we make this connectivity as sparse. Since $P_l > 1$, the model learns overcomplete sparse, feature maps. This structure is illustrated in Figure 1.

A deep deconvolutional network consisting of NL number of layers is trained upwards in a layer wise manner starting with the first layer ($l = 1$) where the inputs at layer l=0 are the training images $\mathcal{I}$. Each layer l is trained in order to learn a set of filters $f_{p,c}^l$ which is shared across all images in $\mathcal{I}$, and infer the set of feature maps $s_{p,l}^i$ of each image I^i in $\mathcal{I}$. To learn the filters, we alternately minimize $C_l(\mathcal{I})$ w.r.t. the filters and feature maps by keeping one of them constant while minimizing the other. We follow the optimization scheme as proposed in [12].

Once the deep deconvolutional network is trained, we can use it to infer the multi-layer features of a given left I_L and right I_R stereo images for which we want to estimate the dense disparity map. The network described above is top-down in nature i.e., given the latent feature maps, one can synthesize an image but there is no direct mechanism for inferring the feature maps of a given image without minimizing the cost function given in Eq. (6). Hence, once the network is learned/trained, we apply given I_L and I_R separately as input image to the trained deep deconvolutional network with the fixed set of learned filters and infer the feature maps $s_{p,l}^{I_L}$ and $s_{p,l}^{I_R}$ of I_L and I_R at layer l, respectively by minimizing the cost functions $C_l(I_L)$ and $C_l(I_R)$, respectively. Once, they are learned, we create a feature vector at each pixel location in I_L and I_R separately. In order to obtain the features of I_L at a layer l, we stack the P_l number of inferred feature maps $s_{p,l}^{I_L}$ and obtain a single feature map $Z_l^{I_L}$ where at each pixel location (x, y) in $Z_l^{I_L}$, we get a feature vector of dimension $P_l \times 1$. Similarly, using the same process we obtain the features of I_R. Thus, $Z_l^{I_L}$ and $Z_l^{I_R}$ represent the l^{th} layer features of I_L and I_R, respectively.

Once the multi-layer features of I_L and I_R are obtained, we can define our feature matching cost $E_F(d)$ as:

$$E_F(d) = \sum_{l=1}^{NL} \sum_{(x,y)} \min(|Z_l^{I_L}(x,y) - Z_l^{I_R}(x + d(x,y), y)|, \tau^F). \quad (7)$$

At each pixel location (x, y) having disparity $d(x, y)$, it measures the absolute distance between the feature vector $Z_l^{I_L}(x, y)$ and corresponding matched feature $Z_l^{I_R}(x + d(x, y), y)$. Here, τ^F is the truncation threshold which is used to make feature matching cost more robust against outliers.

IV. IGMRF Model for Disparity

Object distances from the camera i.e., depths are inversely proportional to disparities and hence are made up of various textures, sharp discontinuities as well as smooth areas making them inhomogeneous. This motivates us to use an IGMRF prior that can enforce the smoothness in disparity map as well as preserves the discontinuities. For modeling the d using IGMRF, $E_P(d)$ is chosen as the sum of squares of finite difference approximations to the first order derivatives of disparities. Considering the differentiation in both horizontal and vertical directions at each pixel location, one can write

$E_P(d)$ as [20]:

$$E_P(d) = \sum_{(x,y)} b^X_{(x,y)}(d(x-1,y)-d(x,y))^2 + b^Y_{(x,y)}(d(x,y-1)-d(x,y))^2. \quad (8)$$

Here b^X and b^Y are the spatially adaptive IGMRF parameters in horizontal and vertical directions, respectively. Using these estimated parameters in $E_P(d)$, one can obtain a solution which is less noisy in smooth areas and also preserve the depth inhomogeneity in other areas. By knowing the disparity map, the IGMRF parameters at a location (x, y) can be obtained as [20]:

$$b^X_{(x,y)} = \frac{1}{\max(4(d(x-1,y)-d(x,y))^2, 4)}. \quad (9)$$

$$b^Y_{(x,y)} = \frac{1}{\max(4(d(x,y)-d(x,y-1))^2, 4)}. \quad (10)$$

Since the true disparity map is unknown, and has to be estimated, we start the regularization process by using the estimated parameters of an initial estimate obtained using a suitable approach.

V. Dense Disparity Estimation

Our data term defined in Eq. (2) is formed by adding intensity and feature matching costs using Eqs. (3) and (7), respectively. Similarly, our prior term is defined by IGMRF prior using Eq. (8). Hence, the final energy function defined in Eq. (1) can be rewritten as given in Eq. (11). Note that although we do not consider the occlusions explicitly, they are handled by clipping matching costs using thresholds $\tau = \{\tau^I, \tau^F\}$. In order to minimize Eq. (11), our algorithm starts with an initial estimate of d, and it iterates and alternates between following two phases until convergence:

Phase 1:With d being fixed, compute IGMRF parameters $b^X_{(x,y)}$ and $b^Y_{(x,y)}$ using Eqs.(9) and (10), at each pixel location.

Phase 2: With $\{b^X_{(x,y)}, b^Y_{(x,y)}\}$ fixed as obtained in phase 1, minimize the Eq. (11) for d using graph cuts optimization based on α-β swap moves [22].

We use a classical local stereo method [1] for obtaining the initial disparity map in which the *absolute intensity differences* (AD) with truncation, aggregated over a fixed window is used as matching cost. Post processing operations such as left-right consistency check, interpolation and median filtering [1] are applied in order to obtain a better initial estimate for faster convergence while regularizing. However, any other suitable disparity estimation method can also be used in obtaining the initial estimate.

VI. Experimental Results

We demonstrate the efficacy of the proposed method by conducting various experiments using the Middlebury stereo benchmark images [23]. We first provide the details of parameters used in training the deep deconvolutional network. A 2-layer deep deconvolutional network i.e., NL=2 was trained over m_s=60 left stereo images obtained from the Middlebury 2005 and 2006 datasets [23]. We set the number of feature maps as P_1=9 and P_2=45 for layer 1 and 2, respectively. The feature maps at layer 1 were fully connected to the input in which each image has a single channel. In order to reduce the computations, 36 feature maps in layer 2 were connected to a pair of maps in layer 1, and remaining 9 were singly connected. In this way, we obtained 9 and $36 * 2 + 9 = 81$ filters at layer 1 and 2, respectively. The parameter α in Eq. (6) was set as 1 and the filters of size 7×7 were learned. These parameters were manually set as per the experimental settings done in [12] except that we used gray scale stereo images for training i.e., P_0=1. With these settings, our deconvolutional network was trained to obtain the set of filters. The learned filters at the first and second layers are shown in Figure 2, where the first layer learns Gabor like filters, and the filters in the second layer lead to mid-level features such as center-surround corners, T and angle-junctions, and curves.

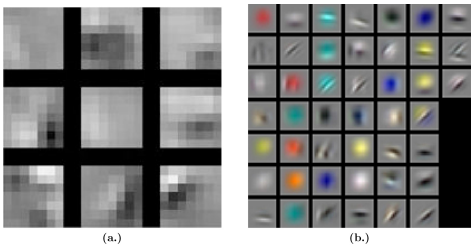

Fig. 2: Filters learned at first and second layers of deep deconvolutional network. (a.) Filters learned at first layer (9). (b.) Filters learned at second layer (81) where 36 filters in pair are shown in color and remaining 9 filters are shown as gray scale.

In order to estimate the dense disparity map, we experimented on the Middlebury stereo 2001 and 2003 datasets [23] which are different from the training datasets used earlier. Our algorithm converged within 5 iterations for all the stereo pairs used in our experiments. While minimizing Eq. (11), the data cost thresholds $\{\tau^I, \tau^F\}$ were set as 0.08 and 0.04, respectively. The parameter γ was chosen as 1 by hit and trial for the best performance. We used the same parameters for all the experiments and this demonstrates the robustness of our method. In this work, all the experiments were conducted on a computer with Core i7-3632QM, 2.20 GHz processor and 8.00 GB RAM.

In order to perform the quantitative evaluation, we use the percentage of bad matching pixels (*B%*) with a disparity error tolerance δ [1] as the error measure, and compute it over the entire image as well as in the non occluded regions. We first demonstrate the effectiveness of using the proposed data term by considering the energy functions with different data terms $E_D(d)$ but keeping the prior as IGMRF. The performance is compared using the $E_D(d)$ made up of traditional pixel based data terms such as AD and BT. We also consider

$$E(d) = \sum_{(x,y)} \min((\min_{d(x,y)\pm\frac{1}{2}} |I_L(x,y) - I_R(x+d(x,y),y)|), \tau^I) + \gamma \sum_{l=1}^{NL} \sum_{(x,y)} \min(|Z_l^{I_L}(x,y) - Z_l^{I_R}(x+d(x,y),y)|, \tau^F) + \sum_{(x,y)} (b^X_{(x,y)}(d(x-1,y) - d(x,y))^2 + b^Y_{(x,y)}(d(x,y-1) - d(x,y))^2). \quad (11)$$

$\mathbf{E_D(d)}$	**Venus**	**Teddy**	**Cones**
AD	1.90	16.49	12.14
BT	0.95	15.67	11.89
BT+gradient	0.89	14.9	11.32
Proposed	**0.40**	**11.41**	**9.64**

TABLE I: Performance evaluation in terms of % of bad matching pixels computed over the whole image with δ=1. Here, the optimization of energy function is carried out using different data terms $E_D(d)$ with IGMRF as prior term $E_P(d)$.

the use of BT+gradient as $E_D(d)$ for comparison where the BT is combined with gradient based feature matching. All the data terms were used with truncation on their costs. The results of these experiments are summarized in Table I. The results clearly show that the approach using proposed $E_D(d)$ outperforms those with traditional pixel based $E_D(d)$, illustrating the effectiveness of using the learning based $E_F(d)$ in our approach. In other words, when the proposed feature matching cost is combined with the pixel based intensity cost, the accuracy of disparity estimation is high. Since our deep deconvolutional network learns overcomplete, sparse features, the $E_F(d)$ when used as a data term in the IGMRF regularization framework, generates sparse disparity estimates. Hence in order to obtain the dense and accurate disparity estimates, the intensity matching cost is combined with the feature matching cost to form our data term.

We now demonstrate the performance of our approach by varying the number of layers in the feature matching cost $E_F(d)$. To do this we estimate the disparity maps by considering first NL=1 and then NL=2 in Eq. (11). Figure 3 shows that the performance improves when we use both layers for feature extraction. We also experimented using more than two layers i.e., $NL > 2$ but did not find significant improvement (see Figure 3). Based on this observation, we fixed NL=2 in all our experiments while comparing the performance with other approaches. Use of 2 layers indicates the advantage of deep deconvolutional network since it learns better features with limited number of layers.

In order to show the visual quality of our results we display the computed disparity maps and error maps for different stereo pairs, see Figure 4. One can see that the final disparity maps are piecewise smooth and visually plausible. The error maps shown in the last column of Figure 4 correspond to the difference between ground truth and estimated disparities (black and gray regions correspond to errors in occluded and non occluded regions, respectively and white indicates no error). We can see that the proposed method performs better in

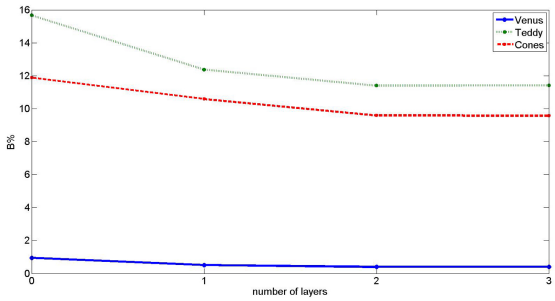

Fig. 3: Results in terms of % of bad matching pixels by varying the number of layers NL in $E_F(d)$. **Here,** NL=0 **means when** $E_F(d)$ **is absent in optimization of Eq. 11**

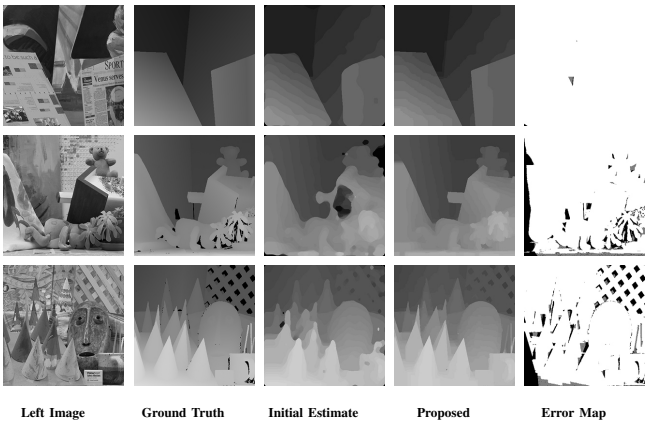

Fig. 4: Experimental results for the Middlebury stereo 2001 and 2003 datasets [23], *Venus* (1^{st} row), *Teddy* (2^{nd} row) and *Cones* (3^{rd} row)

occluded as well as in non occluded regions. This indicates the advantage of using IGMRF prior and proposed data term with truncation on their cost. Use of number of filters to capture sparse features in unsupervised way makes our method to better handle outliers and results in accurate disparity maps.

Now, we compare our results with the state of the art regularization and feature based global dense stereo methods. The comparison in terms of percentage of bad matching pixels is shown in Table II. We do not compare our method with global stereo methods based on hand crafted and learned features [4], [5], [9] since their results are not available for the Middlebury datasets. As seen from the Table II, our method performs best among all the other methods in non occluded regions and the overall performance is comparable to state of the art global stereo methods. In order to compare our method with latest best performing stereo methods, we experimented

Method	Venus		Teddy		Cones	
	all	*nonocc*	*all*	*nonocc*	*all*	*nonocc*
Initial	3.47	2.00	19.65	5.61	16.43	7.15
Proposed	0.40	**0.10**	11.41	**3.51**	9.64	**2.40**
AdapBP [3]	0.21	0.10	7.06	4.22	7.92	2.48
DblBP [16]	0.45	0.13	8.30	3.53	8.78	2.90
GCP [19]	0.53	0.16	11.5	6.44	9.49	3.59
TwoStep [6]	0.45	0.27	12.6	7.42	10.1	4.09
SGlob [7]	1.57	1.00	12.2	6.02	9.75	3.06
2OP [18]	0.49	0.24	15.4	10.9	10.8	5.42
MGC [15]	3.13	2.79	17.6	12.0	11.8	4.89
Mumfd [17]	0.76	0.28	14.3	9.34	9.91	4.14
GC [14]	3.44	1.79	25.0	16.5	18.2	7.70

TABLE II: Comparison with state of the art global dense stereo methods in terms of % of bad matching pixels over entire image as well as in non occluded regions with δ=1, experimented using Middlebury stereo 2001 and 2003 datasets [23] . First row shows the results of initial estimate used in our approach.

on the recently released Middlebury stereo 2014 datasets [23], and submitted our estimated disparity maps online to the server on Middlebury website [23] which in turn returned the overall evaluation and comparison chart. Our proposed method achieved a rank of 32. Though, the proposed method is not ranked among the top, the results indicate that it is comparable to other latest stereo methods.

VII. Conclusion

We have presented a new approach for dense disparity map estimation based on global energy minimization framework using the combination of intensity and the feature matching costs, and using IGMRF as prior. The feature matching cost is defined over the deep learned features of given stereo pair using deep deconvolutional network. An iterative two phase algorithm is proposed where the IGMRF parameters and disparity map are refined, alternatively. Experimental results validate the effectiveness of the proposed approach.

References

[1] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1/2/3, pp. 7–42, April-June 2002.

[2] N. Ayache and B. Faverjon, "Efficient registration of stereo images by matching graph descriptions of edge segments," *International Journal of Computer Vision*, pp. 107–131, 1987.

[3] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Pattern Recognition, IEEE International Conference on*, vol. 3, 2006, pp. 15–18.

[4] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 978–994, May 2011.

[5] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Computer Vision and Pattern Recognition, IEEE Conference on*, June 2013, pp. 2307–2314.

[6] Z. L. L. Wang and Z. Zhang, "Feature based stereo matching using two-step expansion," *Mathematical Problems in Engineering*, vol. 14, p. 14, December 2014.

[7] H. Hirschmller, "Stereo processing by semi-global matching and mutual information," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 328–341, February 2008.

[8] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[9] C. Zhang and C. Shen, "Unsupervised feature learning for dense correspondences across scenes," *CoRR*, vol. abs/1501.00642, 2015.

[10] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2." in *Neural Information Processing Systems*, 2007, pp. 873–880.

[11] G. E. Hinton and S. Osindero, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, p. 2006, 2006.

[12] M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition, IEEE Conference on*, June 2010, pp. 2528–2535.

[13] M. R. K. Jarrett, K. Kavukcuoglu and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Computer Vision, IEEE International Conference on*, 2009, pp. 2146–2153.

[14] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, November 2001.

[15] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Computer Vision, European Conference on*, 2002, pp. 82–96.

[16] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 492–504, March 2009.

[17] R. Ben-Ari and N. Sochen, "Stereo matching with mumford-shah regularization and occlusion handling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 11, pp. 2071–2084, November 2010.

[18] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, "Global stereo reconstruction under second order smoothness priors," in *Computer Vision and Pattern Recognition, IEEE Conference on*, June 2008, pp. 1–8.

[19] L. Wang and R. Yang, "Global stereo matching leveraged by sparse ground control points," in *Computer Vision and Pattern Recognition, IEEE Conference on*, June 2011, pp. 3033–3040.

[20] A. Jalobeanu, L. Blanc-Feraud, and J. Zerubia, "An adaptive gaussian model for satellite image deblurring," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 613–621, April 2004.

[21] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 4, pp. 401–406, April 1998.

[22] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts?" *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 2, pp. 147–159, February 2004.

[23] D. Scharstein, R. Szeliski, and R. Zabih, "Middlebury stereo." [Online]. Available: http://vision.middlebury.edu/stereo