A REGULARIZATION FRAMEWORK FOR STEREO MATCHING USING IGMRF PRIOR AND SPARSENESS LEARNED FROM AUTOENCODER.

Sonam Nahar

The LNMIIT, Jaipur (India)

Manjunath V. Joshi

DA-IICT, Gandhinagar (India)

ABSTRACT

In this work, we propose to use an Inhomogeneous Gaussian Markov Random Field (IGMRF) and sparsity based priors in a regularization framework in order to estimate the dense disparity map. The IGMRF prior captures the smoothness as well as preserves sharp discontinuities and the sparsity prior captures the sparseness in the disparity map. We present a sparse autoencoder based approach for learning and inferring the sparse representation of disparities. An iterative two phase algorithm is proposed to solve our energy minimization problem. Experimental results on the standard datasets demonstrate the effectiveness of the proposed approach.

Index Terms**—** Stereo, IGMRF, Sparsity, Autoencoder.

1. INTRODUCTION

The main goal of stereo vision is to find the disparity between corresponding pixels [1]. Since it is an ill-posed problem, regularization can be used to restrict the solution space. Many of the state of the art disparity estimation techniques are based on the MRF regularization [1, 2, 3, 4]. Other regularization based methods such as mumford shah regularization [5], second order smoothness prior [6], ground control points [7] and learned conditional random field (CRF) [8] have also been used. Many of these techniques use single or a set of global parameters and they may not capture local variation among disparities. We need a prior which considers the spatial variation among the disparities locally as well as computationally less taxing. This motivates us to use IGMRF as one of the prior which was first proposed in [9]. IGMRF can handle smooth as well as sharp changes in disparity since the local variation among disparities are captured by these parameters. IGMRF prior for disparity estimation has been used earlier in [10]. However, the method do not use the sparsity prior for regularization.

Although IGMRF captures the smoothness with discontinuities, it fails to capture added feature such as sparseness. In general, disparity maps are redundant due to the smoothness and one can represent it in a domain in which it is sparse. Learning an efficient sparse representation from a large set of examples rather than using a fixed set of bases such as discrete cosine transform (DCT) and discrete wavelet transform (DWT) have achieved better performance in solving inverse problems [11, 12]. This motivates us to learn the sparse representation of disparities from a large database of true disparities and use a sparsity prior that complements the IGMRF prior. In [13], authors proposed a two layer graphical model for inferring the disparities and learning their sparse representation using a nonstationary sparse coding method. This method is complex and computationally intensive. In [14], the disparities are reconstructed from few disparity measurements based on the principle of compressive sensing but in this case accuracy of estimation depends on the reliable support points.

In this paper, we propose to use sparsity prior and IGMRF prior in an energy minimization framework. We present a novel method for learning and inferring the sparseness of disparities using a sparse autoencoder. A sparse autoencoder reconstructs the input data with its sparse representation [15] and have been used in denoising [16] and inpainting [17]. Sparse autoencoders are efficient when compared to dictionary learning methods. The advantage of our approach is that sparsity can be better learned due to the use of large amount of true disparity maps. Note that this is offline operation and do not add to computational complexity. Finally, we propose a two phase, iterative algorithm for estimating the dense disparity map. To start with, we need an initial estimate of disparity map which is obtained using the local method where the *absolute intensity differences* (AD) with truncation, aggregated over a fixed window is used as matching cost. In order to reduce computation time, we propose to optimize this cost by graph cuts instead of the classic *winner take all* (WTA) optimization. We apply the post processing operations such as left-right consistency check, interpolation, median filtering [1] in order to obtain the better initial estimate for faster convergence while regularizing.

2. PROPOSED APPROACH

2.1. Problem Formulation

Dense stereo matching problem is commonly formulated in a global energy minimization framework. Typically the following energy function is used [1]:

$$E(d) = E_{data}(d) + E_{prior}(d). \tag{1}$$

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

For a given disparity map $d \in \mathcal{R}^{N \times N}$, we use the following data term [18]:

$$E_{data}(d) = \sum_{(x,y)} \min(|I_L(x,y) - I_R(x + d(x,y), y)|, T). \tag{2}$$

$I_L(x,y)$ and $I_R(x + d(x,y), y)$, represents the left and right image pixel intensities having $d(x,y)$ as disparity at (x,y) location. Here T is the truncation constant on pixel wise matching cost. We consider disparity search from left to right as well as from right to left directions and hence relax the traditional ordering constraint used in disparity estimation. $E_{prior}(d)$ represents prior terms which is the sum of IGMRF and sparsity priors, and it is given as:

$$E_{prior}(d) = E_{IGMRF}(d) + \gamma E_{sparse}(d), \tag{3}$$

where $E_{IGMRF}(d)$ and $E_{sparse}(d)$ represent the IGMRF and sparsity priors, respectively. γ controls the weightage of the $E_{sparse}(d)$.

2.2. IGMRF Model of Disparity

In practice, stereo images consist of various textures, sharp discontinuities as well as smooth areas making the disparity inhomogeneous. Hence, the use of IGMRF prior in this work, enforces the smoothness in disparity map with discontinuity preservation. For modeling IGMRF, $E_{IGMRF}(d)$ is chosen as the square of finite difference approximation to the first order differentiation of disparities. Considering the differentiation in horizontal and vertical directions at each pixel location, we can write $E_{IGMRF}(d)$ as defined in [9]:

$$\begin{aligned} E_{IGMRF}(d) &= \sum_{(x,y)} b^X_{(x,y)}(d(x-1,y) - d(x,y))^2 \\ &\quad + b^Y_{(x,y)}(d(x,y-1) - d(x,y))^2. \end{aligned} \tag{4}$$

Here b^X and b^Y are the spatially adaptive IGMRF parameters in horizontal and vertical directions, respectively. The IGMRF parameters at each pixel location (x,y) are given by [9] as:

$$b^X_{(x,y)} = \frac{1}{\max(4(d(x-1,y) - d(x,y))^2, 4)}. \tag{5}$$

$$b^Y_{(x,y)} = \frac{1}{\max(4(d(x,y) - d(x,y-1))^2, 4)}. \tag{6}$$

To start the regularization process we use the estimated initial disparity map to compute these parameters which are then used to estimate the disparity map. These parameters and the disparity map are refined in two phases alternatively and iteratively.

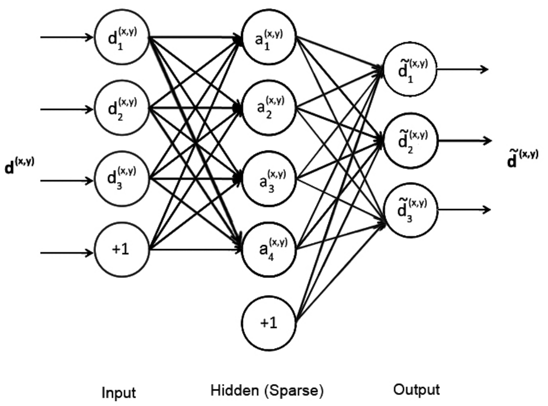

Fig. 1: A sparse autoencoder with $n = 3$ and $K = 4$. Here +1 represents a bias unit.

2.3. Sparse Model of Disparity

In this work, we present a new method for learning and inferring the sparse representation of disparities using sparse autoencoder, which is then used to define the sparsity prior. An autoencoder is an artificial neural network which sets the desired output same as the input and has only one hidden layer. In reality, finding the sparse representation of a disparity map is computationally expensive and therefore a better choice would be to find the sparse representation of disparity patches of small size individually and average the resultant sparse patches at the end in order to get complete sparse representation of disparity map.

Let the input to an autoencoder be a disparity patch $d^{(x,y)} \in \mathcal{R}^n$, centered at location (x,y) in a disparity map d [1], its corresponding hidden representation at hidden layer be $a^{(x,y)} \in \mathcal{R}^K$ and the reconstructed output be $\tilde{d}^{(x,y)} \in \mathcal{R}^n$. The autoencoder has weights (W, U, r, s), where $W \in \mathcal{R}^{n \times K}$ is the encoder weight matrix between the input layer and hidden layer, $U \in \mathcal{R}^{K \times n}$ is the decoder weight matrix between the hidden layer and output layer, and $r \in \mathcal{R}^K$ and $s \in \mathcal{R}^n$ are the bias weight vectors for hidden and output layer, respectively. For a fixed set of weights (W, U, r, s), $a^{(x,y)}$ and $\tilde{d}^{(x,y)}$ can be computed as follows:

$$a^{(x,y)} = f(W^T d^{(x,y)} + r), \tag{7}$$

$$\tilde{d}^{(x,y)} = f(U^T a^{(x,y)} + s), \tag{8}$$

where f is an activation function and we use sigmoid for this.

An autoencoder is called as sparse autoencoder when the sparsity constraint is imposed on its hidden layer. It learns an overcomplete sparse representation of data when the number of hidden units K are greater than the number of input units n, $K > n$. An example of sparse autoencoder is shown in Figure 1.

[1] Note that $d(x,y)$ is the disparity at location (x,y) and $d_{(x,y)}$ is the disparity patch extracted at location (x,y) in d.

Let $a_j^{(x,y)}$ be the activation of hidden unit j. A sparsity constraint on the activations of hidden units are imposed by forcing them to be inactive most of the time. A unit is active when its activation value is close to one and inactive when it is close to zero. We define ρ, as a global sparsity parameter for all hidden units, typically a small value close to zero. Let $\hat{\rho}_j$ be the average activation of hidden unit j (averaged over training set of disparities). Then the sparsity constraint for each j^{th} hidden unit is enforced by a penalty term which penalizes $\hat{\rho}_j$ deviating significantly from ρ as:

$$\sum_{j=1}^{K} KL(\rho||\hat{\rho}_j) = \sum_{j=1}^{K} \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}, \quad (9)$$

where $KL(\rho||\hat{\rho}_j)$ is the Kullback Leilbler (KL) divergence.

Given a large training set of ground truth disparity patches $\mathcal{G}$=$\{d^{(1)}, d^{(2)}, \ldots, d^{(m)}\}$, with each patch $d^{(i)} \in \mathcal{R}^n$. We train a sparse autoencoder in order to learn the weights (W, U, r, s). To do this we minimize the following energy function formed using (7), (8) and (9):

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}||d^{(i)} - f(U^T(f(W^T d^{(i)} + r)) + s)||_2^2 \\ + \frac{\lambda}{2}(\sum_{i=1}^{n}\sum_{j=1}^{K}(W_{ij})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n}(U_{ij})^2) \\ + \beta \sum_{j=1}^{K} KL(\rho||\hat{\rho}_j), \quad (10) \end{aligned}$$

Here the parameters λ control the overfitting and β is the weightage of the sparsity penalty term. We minimize (10) as a function of W, U, r, s using back propagation algorithm [19].

Once the autoencoder is trained, (W, U, r, s) are used in defining our sparsity prior as follows:

$$E_{sparse}(d) = \sum_{(x,y)} \left|\left| d^{(x,y)} - f(U^T a^{(x,y)} + s) \right|\right|_2^2. \quad (11)$$

$E_{sparse}(d)$ measures how well the disparity patches agree with their sparse representations.

2.4. Dense Disparity Estimation

Using (2), (4) and (11) defined for $E_{data}(d)$, $E_{IGMRF}(d)$ and $E_{sparse}(d)$ terms, respectively, we can rewrite our final energy function defined in (1) as (12). We minimize this non convex energy function using graph cuts optimization using α-β swap moves [2]. Our algorithm proceeds with the use of initial estimate of disparity map, and iterates and alternates between two phases until convergence as given in Algorithm 1.

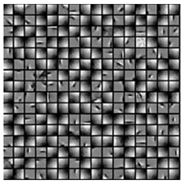

Fig. 2: Encoder weights. Training and testing root mean square error between input and output is 0.27 and 1.8.

$$\begin{aligned} E(d) = \sum_{(x,y)} \min(|I_L(x,y) - I_R(x + d(x,y), y)|, T) \\ + b^X_{(x,y)}(d(x-1,y) - d(x,y))^2 \\ + b^Y_{(x,y)}(d(x,y-1) - d(x,y))^2 \\ + \gamma \sum_{(x,y)} \left|\left| d^{(x,y)} - f(U^T a^{(x,y)} + s) \right|\right|_2^2. \quad (12) \end{aligned}$$

Algorithm 1: Proposed Algorithm

```
Input: Stereo image pair $I_L$ and $I_R$, a set of ground
       truth disparity patches
       $\mathcal{G}$=$\{d^{(1)}, d^{(2)}, \ldots, d^{(m)}\}$.
Train a sparse autoencoder using $\mathcal{G}$ by minimizing (10)
and obtain weights $(W, U, r, s)$;
Obtain an initial disparity map $d_0$;
Initialization: $d = d_0$;
repeat
    Phase 1:With $d$ being fixed, infer the sparse vector
    $a^{(x,y)}$ for each disparity patch $d^{(x,y)}$ in $d$ using (7).
    Compute IGMRF parameters $b^X_{(x,y)}$ and $b^Y_{(x,y)}$
    using (5) and (6), at each pixel location;
    Phase 2: With $\{a^{(x,y)}\}$, $\{b^X_{(x,y)}, b^Y_{(x,y)}\}$ fixed as
    obtained in phase 1, minimize the (12) for $d$ using
    graph cuts;
until convergence;
```

3. EXPERIMENTAL RESULTS

We conducted several experiments and evaluated the results on the Middlebury stereo benchmark images [20]. Note that we used gray scale stereo images, one can also use color images for better performance. The sparse autoencoder was trained using a set of $m = 2 \times 10^5$ disparity patches extracted from the 2014 Middlebury training ground truth set [20]. The size of each disparity patch was chosen as 8×8, i.e., $n = 64$. For training, the parameters were chosen as: $K = 256$, $\lambda = 10^{-4}$, $\beta = 0.1$ and $\rho = 0.01$. The parameter γ in (12) was initially set at 10^{-4} and exponentially increased at each iteration from 10^{-4} to 10^{-1}. All the parameters were empirically chosen for obtaining the best performance. The same parameters were used in all the experiments and our

Method	**Venus**	**Teddy**	**Cones**
Initial	3.47	19.65	16.43
IGMRF	2.56	16.38	13.33
IGMRF-DCT	2.11	15.4	12.1
Proposed	**0.22**	**10.7**	**9.64**

Table 1: Evaluation results on the Middlebury datasets [20] in terms of bad matching pixels computed over whole image. Comparisons are made considering different cases: (1^{st} row): Initial estimate. (2^{nd} row): Using IGMRF prior only. (3^{rd} row): IGMRF and sparsity prior with DCT. (4^{th} row): Proposed Method.

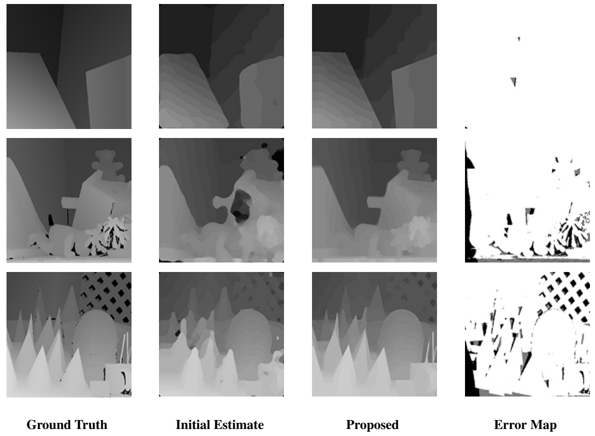

Fig. 3: Results for the datasets of [20], *Venus* (1^{st} row), *Teddy* (2^{nd} row) and *Cones* (3^{rd} row).

algorithm converged within 5 to 10 iterations for the large number of stereo pairs. We tested our algorithm on a Core i7-3632QM CPU @2.20 GHz and 8.00 GB RAM. The trained sparse autoencoder's encoder weights W are shown in Figure 2.

In order to evaluate the performance of our approach quantitatively, we used the percentage of bad matching pixels with a disparity error tolerance of 1 as reported in [1] and compute this error over the whole image as well as in non occluded regions only. Performance was measured under different scenarios. The disparity map was estimated using the IGMRF prior only, IGMRF and sparsity priors with sparseness obtained using DCT, and IGMRF and sparsity priors with sparseness obtained using sparse autoencoder. The results in Table 1 show that the performance using the proposed method is better when compared to the other two experimented cases. We can see that incorporation of sparsity prior in addition to IGMRF prior significantly improves the performance. Figure 3 shows the disparity maps and corresponding error maps computed using our proposed method. The error map shows the regions where our method differs from the ground truth (black and gray regions are errors in occluded and non occluded regions, respectively and white indicates no error). We can see that our method achieves greater accuracy in discontinuous as well as non occluded regions. We mention here that although in problem formulation our method does not consider occlusions, it works well in these areas as well.

Method	**Venus**		**Teddy**		**Cones**	
	all	*nonocc*	*all*	*nonocc*	*all*	*nonocc*
GC [2]	3.44	1.79	25.0	16.5	18.2	7.70
MultiGC [3]	3.13	2.79	17.6	12.0	11.8	4.89
BP [4]	0.45	0.13	8.30	3.53	8.78	2.90
Mumford [5]	0.76	0.28	14.3	9.34	9.91	4.14
2OP [6]	0.49	0.24	15.4	10.9	10.8	5.42
GCP [7]	0.53	0.16	11.5	6.44	9.49	3.59
CRF [8]	1.3	-	11.1	-	10.8	-
CS [14]	0.68	0.31	13.30	7.88	9.79	3.97
Sparse [13]	-	-	11.98	-	8.14	-
Proposed	**0.22**	**0.15**	**10.7**	**4.65**	**9.64**	**3.55**

Table 2: Comparison with state of the art regularization based disparity estimation methods in terms of bad matching pixels (all and non occluded). Here '-' indicates the result not reported.

Adiron.	Motorcyc.	Piano	Pipes	Playr.	Recyc.
39.4	41.6	43.5	39.6	52.8	52.0

Table 3: Evaluation of proposed method on Middlebury 2014 stereo datasets in non occluded regions.

We compare the performance of our method with other state of the art regularization based stereo methods which is shown in Table 2. One can see that our algorithm performs better for all the three datasets in non occluded regions and gives the least bad matching error for the *Venus* stereo pair. As far as occlusion is concerned, it gives superior performance except that method proposed in [4] because the method of [4] handles the occlusions explicitly and use belief propagation for minimization of their energy function. These results reflect the effectiveness of IGMRF and learned sparseness using sparse autoencoder. We also evaluate our method on the latest Middlebury 2014 stereo datasets [20] as shown in Table 3. On the Middlebury website [20], the version 3 of Middlebury evaluation page gives the comparison of latest best performing stereo methods experimented on these data sets. Our method is comparable to those methods listed on that evaluation page.

4. CONCLUSION

We have proposed to use an IGMRF and sparsity priors in a regularization framework for dense disparity estimation. A sparse autoencoder based method is proposed for learning the sparseness. Disparity map is estimated using a two phase, iterative algorithm and our results show the effectiveness of the proposed approach.

5. 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] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in *Computer Vision and Pattern Recognition, IEEE Conference on*, June 2007, pp. 1–8.

[9] 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.

[10] S. Nahar and M.V. Joshi, “A learning based approach for dense stereo matching with IGMRF prior,” in *Computer Vision, Pattern Recognition, Image Processing and Graphics, IEEE National Conference on*, Dec 2013, pp. 1–4.

[11] M. Aharon, M. Elad, and A. Bruckstein, “K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[12] J. Mairal, M. Elad, and G. Sapiro, “Sparse representation for color image restoration,” *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 53–69, January 2008.

[13] I. Tosic, B.A. Olshausen, and B.J. Culpepper, “Learning sparse representations of depth,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 5, pp. 941–952, September 2011.

[14] S. Hawe, M. Kleinsteuber, and K. Diepold, “Dense disparity maps from sparse disparity measurements,” in *Computer Vision, IEEE International Conference on*, November 2011, pp. 2126–2133.

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

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, December 2010.

[17] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” in *Advances in Neural Information Processing Systems 25*, pp. 350–358. 2012.

[18] 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.

[19] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, USA, 1997.

[20] D. Scharstein, R. Szeliski, and R. Zabih, “Middlebury stereo,” .