HandyDepth: Example-based Stereoscopic Hand Depth Estimation using Eigen Leaf Node Features

Rilwan R. Basaru[1], Gregory G. Slabaugh[1], Christopher Child[1], Eduardo Alonso[1]
[1] Department of Computer Science, City University London, London, United Kingdom.
Remilekun.Basaru.1@city.ac.uk

Abstract - **This paper presents a data-driven method to estimate a high quality depth map of a hand from a stereoscopic camera input by introducing a novel regression framework. The method first computes disparity using a robust stereo matching technique. Then, it applies Random Forest (RF) to learn the mapping between the estimated, noisy disparity and actual depth given ground truth data. We introduce Eigen Leaf Node Features (ELNFs) that perform feature selection at the leaf node in each RF tree to identify features that are most discriminative for depth regression. Experimental results demonstrate the promise of the method to produce high quality depth images of a hand using an inexpensive stereo camera.**

Keywords **- Stereo Vision; Random Forest; ELNF; Hand; Depth**

I. INTRODUCTION

Depth estimation is a fundamental problem in computer vision and an essential input to many recent tracking and pose estimation techniques. While many depth estimation methods have been described in the literature, including active methods such as time-of-flight imaging and photometric stereo, this paper focuses on passive stereovision, which shows a number of advantages, namely: (a) its low power; (b) it does not dissipate energy into the scene; (c) it is inexpensive, and (d) it is suitable over a wide range of distances from the camera.

The aim of this paper is to compute a robust depth image of a hand using an inexpensive RGB video stereo camera, as shown in Fig. 1. The RGB images are matched to form a disparity image. However, disparity images are well known to have errors resulting from ambiguities inherent to stereo matching. To address this issue, we propose a novel, data-driven regressive Random Forest (RF) framework that learns the mapping between a noisy, lower quality disparity estimation to actual depth based on a ground truth dataset. When applied to a new disparity image, it corrects errors in the estimated disparity. As part of this regression framework, we propose a new regressive feature selection technique called Eigen Leaf Node Features (ELNFs) that factorizes for the posterior probability and regresses the depth using features that have been found to be highly discriminative. We demonstrate the ability of ELNF to more accurately estimate depth of hands compared to conventional random RF regression. Although this paper is focused on depth estimation, our ultimate goal is to enable hand pose estimation using passive stereovision in an egocentric application. The rest of the paper is structured as follows: the next section presents a general survey of related work in the field of depth estimation and sensing; in Section III, our data acquisition process using image registration is explained. Section IV describes our RF approach to mapping disparity images to depth; experiments are discussed in Section V; and the paper concludes in Section VI.

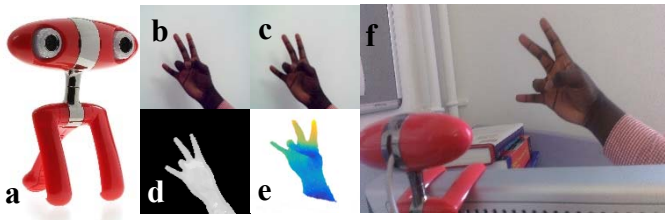

Figure 1: Using an inexpensive stereo camera (a), RGB images of the hand from two perspectives are captured (b, c), a disparity shift image is generated (d) and mapped to an improved depth image (e). The proposed technique can potentially use a stereo rig to estimate hand articulation and pose (f).

II. RELATED WORK

Recently, depth recovery from a single image has been proposed [2, 10], modeling the depth estimation as a Markov Random Field (MRF) solved using Convolutional Neural Networks (CNNs) [8]. While showing much promise, such methods lack the advantage of using stronger geometric features (like disparity) highly correlated with depth. In [5], a two-layered RF framework is used to establish the mapping between near infrared images of a scene consisting of articulated hand poses captured from modified RGB cameras. While this is a unique and relatively inexpensive technique, it suffers from ambient infrared radiation (e.g. when used in an outdoor scene). Also, it requires nontrivial hardware modifications.

Relatedly, Nyugen et al. [9] argue that, for problems with high dimensional data, the performance of RF can degrade as a result of randomization in both bagging samples and feature selection, and employ feature sampling at split nodes to identify informative features. Another recent approach that relates to our work is [6], where feature selection is done by optimizing for an appropriate weighting allocated to different features when building a decision tree. In the context of our work, we choose to focus on feature selection at a *leaf* node by introducing Eigen Leaf Node Features (ELNFs). Our approach yields results which greatly outperform conventional RF regression. A recent increase in interest in hand pose estimation has led to the advent of a number of techniques, particularly those working with data captured from active depth sensors or monocular cameras [3, 9]. However, less attention has been given to hand pose recovery based on stereoscopic images [7]. We contribute to this area by developing a framework that recovers the depth of a hand from stereo.

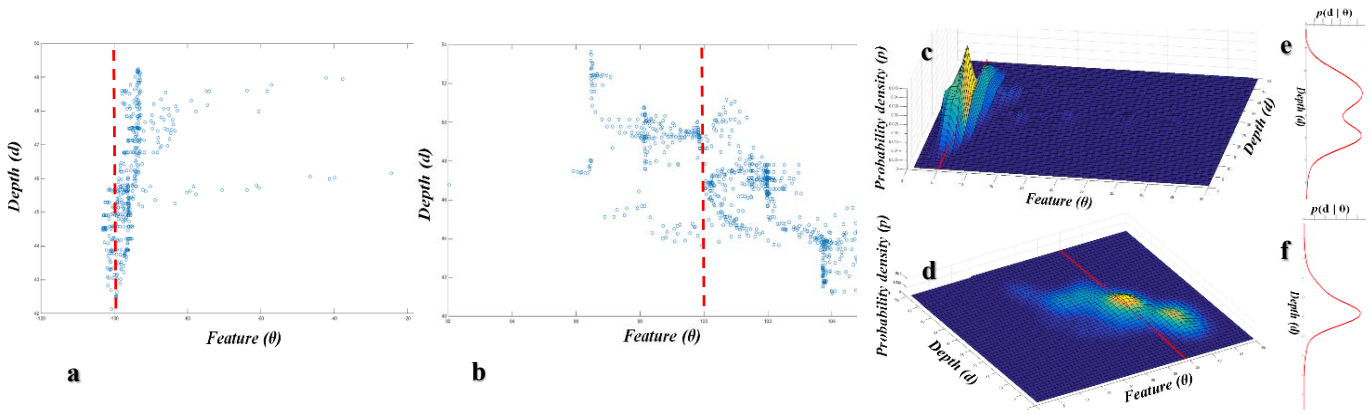

Figure 2: At a leaf node, a depth-feature distribution is established (a). Images a, c & e illustrate a poor feature-depth distribution (vertically orientated). In contrast, b, d & f illustrate a better feature-depth distribution (obliquely orientated) as factorizing yields a more confident posterior (f). ELNF is biased towards the obliquely orientated distribution.

III. DISPARITY AND DEPTH IMAGE DATABASE

The first stage in our framework involves building a database of disparity and depth images for different hand poses to train the regressive framework. Disparity is estimated from a stereo image pair using a robust stereo matching cost function (Quantized Census) [1]. Simultaneously, a depth image is acquired using a RGBD camera. The depth image is registered to the left image of the stereo pair, so for each disparity image, we have a well registered ground truth depth. This dataset is then used to train the regression framework, described in the next section. Once training is complete, the regressive framework can predict a depth image solely from the disparity image computed from stereo RGB input.

We generated a real dataset consisting of 1,000 instances of hand poses across 5 different subjects (200 from each) with a variety of different skin tones, genders and hand sizes. 700 of these were used in training while the remaining 300 were used in independent testing. In addition, a synthetic dataset was produced using OpenGL based computer generated hand poses, and rendered from the perspective of a synthetic stereo camera pair. The synthetic dataset consisted of 10,000 instances of different hand articulations.

IV. MAPPING DISPARITY TO GROUND TRUTH DEPTH

Our task is to establish the mapping between the computed disparity image $I_{disp}(\mathbf{x})$ and the actual depth $I_{depth}(\mathbf{x})$ at a pixel position $\mathbf{x}$ using the ground truth data. We model this mapping with a regressive RF [11]. At each split node in each tree, we employ a feature f_θ based on the difference in disparity between two points with random offset vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, similar to [3], but using a disparity image,

$$f_\theta(I, \boldsymbol{x}) = d_I\left(\boldsymbol{x} + \frac{\boldsymbol{u}}{d_M}\right) - d_I\left(\boldsymbol{x} + \frac{\boldsymbol{v}}{d_M}\right) \qquad (1)$$

where d_M is a normalizing factor representing the maximum disparity in the hand region. The offset vectors are stored at each split node so that they can be used for later prediction.

A. Random Forest

We grow N decision trees by recursively splitting the training data to reduce entropy. The distribution of the depth value (regression target) is modeled using the differential entropy as

$$E(S) = \log(\sigma_s), \qquad (2)$$

where σ_s is the standard deviation of the ground truth depth values of the pixels within a collection of samples S. Statistical analysis is carried out on the pixels that land at each leaf node, and, as a result, the distribution of the features computed against the actual depth established (Fig. 2a & 2d). Acknowledging that for a single pixel position an infinite amount of features based on Equation 1 could be generated, it would be impractical and redundant to use all these features. Hence a subset of these features is used to establish the relationship between features and ground truth depth. We propose ELNF to determine this subset of features, as described in the next subsection. For a determined subset, multivariate Kernel Density Estimation (KDE) is applied by convolving the features-ground truth depth distribution with a Gaussian kernel [4]. For a subset of N features, this yields a continuous (N+1)-D distribution of the feature(s) against the actual depth. Fig. 2c and 2d show the resulting distribution when N=1, i.e. the number of features used is one. In this setup, the frequency of this distribution is represented in the third dimension of the plot. The resulting continuous distribution is stored at the leaf node to be evaluated during testing (Fig. 2e & 2f).

At test time, each pixel, $\mathbf{x}$, whose depth is to be predicted, is passed through each of the trees in the forest. At each node the learned splitting function, $\mathbf{F}(f_\theta, \varphi)$ $\{L_n, R_n\}$ is evaluated, and, based on the feature, f_θ , and on the threshold, φ, the pixel sent to the left, L_n, or to the right node, R_n. This is repeated recursively until the leaf node is reached. At this point, the set of features (computed using learned vector pairs as described in the section below is used in factorizing for the posterior probability of depth, d, given the pixel's set of features, (as in Fig. 2e & 2f.)

$$p(d|\mathbf{x}) = \frac{1}{N}\sum p_t(d|\,\boldsymbol{\theta}) \quad (3)$$

This probability is aggregated across the ensemble of trees, t. Note again that in Fig. 2c & 2d, the number of features is just one, computed from a single learned vector pair. However, we found improved results using two features (i.e. N=2). Since the cost of the multi-dimensional KDE and the size of its resulting distribution increase exponentially, we have not attempted to use more than two features.

B. *Eigen Leaf Node Features (ELNF):*

An issue faced when factorizing for the posterior probability, $p(\mathrm{d}|\boldsymbol{\theta})$, is that the distribution might not show a strong correlation between the feature and the depth to be estimated. Consider Fig. 2b and 2d, they convey a strong negative correlation, hence factorizing for the posterior probability of the depth yields a small standard deviation, and, subsequently, more distinct predictions is achieved by maximum likelihood (Fig. 2f). In contrast, the distribution in Fig. 2a and 2c exhibits a weak correlation. As a result, the factorized posterior yields less distinct peak (Fig. 2e). As each pixel position at the leaf node has potentially infinite features, we would like to select those that are most discriminative for regression. The task, thus, is to ensure that feature(s) selected at the leaf node will yield a strong positive or negative correlation. To establish this, we exploit the principal eigenvector and the ratio of the two eigenvalues of the covariance matrix of the distribution, using what we call Eigen Leaf Node Features (ELNF). In this case we want to establish an obliquely orientated distribution (Fig. 2a) as opposed to a vertically orientated principal distribution (Fig. 2b). The ratio of the two eigenvalues represents how compact the distribution is in the principal direction relative to the perpendicular direction. Hence, at the leaf node we select the feature that minimizes the following cost function:

$$E\left(I_{depth}(\boldsymbol{x}), f_{\boldsymbol{\theta}}\right) = \alpha(|\Delta(\boldsymbol{v}_1)| - 1)^2 + (1-\alpha)\frac{\lambda_2}{\lambda_1} \quad (4)$$

where $\Delta(v_1)$ is the slope of the principal eigenvector $\boldsymbol{v}_1$ of the covariance matrix of the distribution of the actual depth, $I_{depth}(\mathbf{x})$, and the feature f_θ for a set of pixel points $\mathbf{x}$ in the leaf node; λ_1 and λ_2 are the two eigenvalues, the former being the principal eigenvalue; $\alpha \in [0, 1]$ is a weight providing a convex combination of the terms. More generally, when the number of features selected at the leaf node is more than one, we aim to maximize the dependency between all possible pairs of dimensions. Subsequently, we apply Eq. (4) to the distribution of all pairs of each feature with ground truth depth, i.e.

$$E(I_{depth}(\boldsymbol{x}), \boldsymbol{f_\theta}) = \sum_{n,p} C(n,p), \quad \forall n,p = 1,\dots,N+1 \mid n \neq p \quad (5)$$

as a generalized cost, where

$$C(n,p) = \alpha\left(\left|\Delta(\boldsymbol{v}_{1,n,p})\right| - 1\right)^2 + (1-\alpha)\frac{\lambda_{2,n,p}}{\lambda_{1,n,p}} \quad (6)$$

where $\Delta(\boldsymbol{v}_{1,n,p})$ is the slope of the principal eigenvector, $\boldsymbol{v}_{1,n,p}$, that corresponds to the distribution of the n^{th} and p^{th} columns of a data matrix, $\boldsymbol{D}$.

$$\boldsymbol{D} = [\boldsymbol{d_x}|\boldsymbol{f_\theta^1}, \boldsymbol{f_\theta^2}, .., \boldsymbol{f_\theta^N}] \quad (7)$$

Here, $\boldsymbol{D}$ is the resulting matrix when the ground truth depth vector $\boldsymbol{d_x}$ (consisting of the depth of all pixels at the leaf node) is concatenated with the features matrix (consisting of the features values computed at these pixel locations). In our implementation, α was set to 0.7. Distributions that minimize the cost function, E in Eqs. (4) and (5) are those for which the principal orientation has a slope closer to 1 or -1, and greater compactness along the principal orientation.

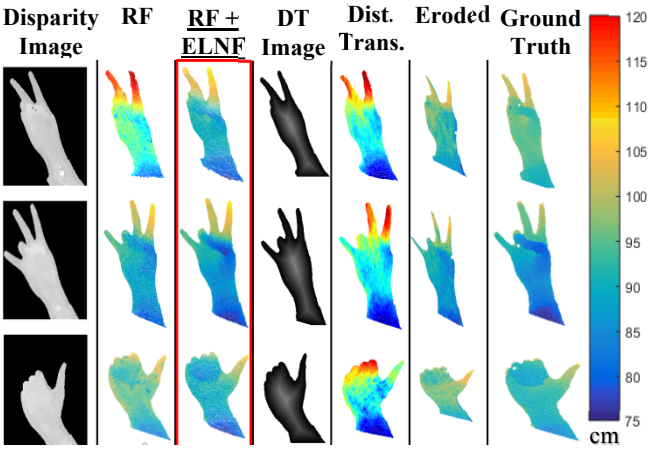

Figure 3: Qualitative Results using real captured poses. The 1st and 4th columns are input to the Distance Transform-based Prediction and to the Disparity based Prediction respectively. The 2nd and 3rd columns are predicted depth from our proposed framework with and without determining ELNF respectively, while the 5th and 6th columns show predicted depth based on distance transformed images and eroded disparity images.

V. Experimental Results

Experiments were carried out with the aim of exploring (i) the significance of ELNF, and (ii) the usefulness of disparity to predict depth.

A. *Evaluating the significance of ELNF*

We evaluated our ELNF approach by comparing it to a conventional regression forest. A qualitative comparison is illustrated in the 2nd and 3rd columns in Fig. 3. A substantial improvement in the predicted depth can be seen based on how well the algorithm matches depth to the ground truth (last column). In all cases, ELNF predicts a better hand shape compared to RF. For example, the digits are more discernible. ELNF is equipped to select the features that better predict the ground truth depth. Quantitative results are presented in Fig. 4, and measure the average absolute difference between the actual depth and the predicted depth across all hand region pixels. At low tree depth, the significance of ELNF is very apparent, as the entropy at the leaf node is high, and ELNF implicitly reduces entropy when the distribution is factorized. This superiority of ELNF is still maintained even at high tree

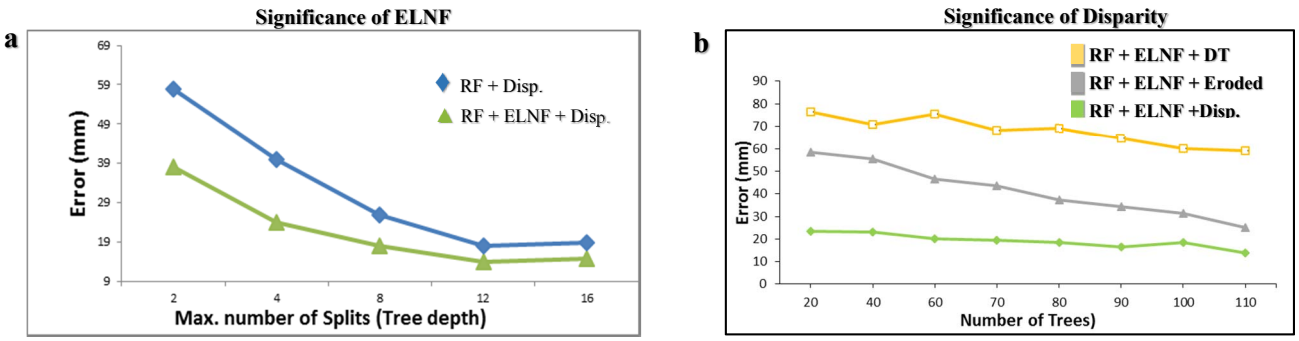

Figure 4: Quantitative results showing error in depth prediction at different tree depth and number of trees. Regressive RF is compared to conventional RF, (a) and evaluated in its ability to predicting depth from a distance transform (DT) input and from a disparity that its segmentation had being eroded (b).

depth, which inherently has less entropy. At a depth of 12, ELNF still produces a 33.834% reduction in error compared to RF (12.631mm compared to 19.090mm). In both cases, the error begins increases again at even deeper tree depth (16) due to over fitting.

B. Disparity input compared to distance transform input

One of our early concerns was that the successful prediction of depth was mainly dependent on the contour of the hand and not the disparity image. Two experiments were carried out to explore this. First, we trained and tested our framework with an image generated solely from a hand region segmentation of a single view image, as opposed to a disparity image. We used a distance transform (DT) of the hand region segmented image (4th column in Fig. 3). The predicted depth image using the DT is shown in the 5th column of Fig. 3. Notice that the entire shape of each finger is not discernable from the contour –for instance, in the distal end of the ring finger in the second row. Quantitative results in Fig. 4b show that the average error in using disparity is 19.047mm in comparison to 77.89mm when DT is used. This clearly illustrates the significance of depth information from the disparity image. The distance transformed input and disparity input are affected similarly as the depth of the tree increases and the trees become more specialized.

To investigate to what extent our framework depends on the segmentation pre-step, we tested our method on instances of eroded segmentation of the disparity images, which remove shape information so depth prediction is based more on disparity. Qualitatively, the results from the eroded disparity look promising (Fig. 3, 6th column), as one can still discern from part of the bent finger. Therefore, we infer the method is not highly dependent on the hand contour.

VI. CONCLUSION

In this paper an innovative regression RF technique for upgrading disparity information to depth for images of the hand is presented. More specifically, we have proposed ELNF, a new way to identify features more suitable for regression. ELNF regression is applicable beyond the context of this paper. We have demonstrated the use of a relatively inexpensive stereo camera to generate a high quality depth image of the hand. Quantitative and qualitative analysis convey promising results in terms of retrieving high quality depth from a stereo setup.

The proposed technique relies on a robust hand segmentation procedure [12]. Future research will aim at eliminating this step from the system. Furthermore, we also will explore the use of the recovered depth to predict hand pose and articulation.

REFERENCES

[1] R. Basaru, E. Alonso, C. Child, and G. Slabaugh, "Quantized Census for Stereoscopic Image Matching, " Proc. of the 3DV Conference: Workshop, DSMA, 2014, pp. 22-29,

[2] A. Saxena, S.H. Chung, and A.Y. Ng, "Learning depth from single monocular images," Proc. Adv. Neural Inf. Process. Syst. , 2005, pp.1-8

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-Time Human pose recognition in parts from single depth images," Proc. CVPR, 2011, pp. 1297-1304.,

[4] T. Duong, "Kernel density estimation and kernel discriminant analysis for multivariate data in R," Journal of Statistical Software: Vol. 21, Issue 7, pp 1-16, 2007.

[5] S. Fanello, C. Keskin, S. Izadi, P. Kohli, D. Kim, D. Sweeney, A. Criminisi., J. Shotton, S.B. Kang, and T. Paek, "Learning to be a Depth Camera for Close-Range Human Capture and Interaction," ACM Transactions on Graphics (TOG). Vol. 33, pp. pp 1-11, 2014,

[6] Z. Xu, G.Huang, K. Q.Weinberger, and A. X.Zheng, "Gradient boosted feature selection," Proc. 20th ACM ICKDD, 2014, pp. 522-531.

[7] R. Grzeszczuk, G. Bradski, M. H. Chu, and J-Y. Bouguet ,"Stereo based gesture recognition invariant to 3d pose and lighting.," Proc. CVPR, 2000, pp. 826-833.

[8] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network,". Proc. Adv. Neural Inf. Process. Syst, 2014, pp.2366-2374

[9] T. T. Nguyen, H. Zhao, J. Z. Huang, T. T. Nguyen, and M. J. Li, "A New Feature Sampling Method in Random Forests for Predicting High-Dimensional Data," Proc. Advances in Knowledge Discovery and Data Mining,, 2015, pp. 459 – 470.

[10] A. Saxena, S. H. Chung, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," IEEE Trans. Pattern Anal. Mach. Intell, Vol. 31, pp. 824-840, 2009.

[11] A. Criminisi, and J. Shotton, "Decision Forests for Computer Vision and Medical Image Analysis." Berlin: Springer, 2013

[12] M.M. Hasan., and P. K. Mishra , "Superior Skin Color Model using Multiple of Gaussian Mixture Model," British Journal of Science, 6(1), pp. 1-14, 2012.

.