QUALITATIVE EVALUATION OF FEATURE DESCRIPTORS FOR MONOCULAR DEPTH ESTIMATION

Dimce Kostadinov
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia

Zoran Ivanovski
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia

ABSTRACT

The paper presents results from the research that is focused on the quality evaluation of the descriptors used for 3D depth map estimation form single image. A research was performed to evaluate a feature descriptor that employs all or combination of the known monocular depth cues. An investigation and analysis was applied for monocular depth cues that can be extracted using image processing techniques.

Key words: machine learning, computer vision, single image 3D depth estimation, monocular depth cues usefulness

I. INTRODUCTION

Humans appear to be extremely good at judging depth from single monocular images. They use various monocular cues to infer the 3D structure of the scene. Some of the cues are local properties of the image, such as texture variations and gradients, color, haze, defocus, etc. [1]. However, local image cues alone are usually insufficient to infer the 3D structure. The ability of humans to "integrate information" over space, i.e., to understand the relation between different parts of the image, is crucial to understanding the 3D structure.

Motivated by the biological and physiological studies of human visual system, a research was performed to evaluate a feature descriptor that employs all or combination of the known monocular depth cues. The evaluation was focused on monocular depth cues that can be extracted using image processing techniques. Some monocular depth cues are successfully used in paper [1], hence, we considered this paper potentially significant. The main concern in our research is the qualitative evaluation of the extracted depth cues and the feature descriptor usefulness, its reliability and consistency and the accuracy of the depth maps estimated using this descriptor.

In [1] A. Saxena, Sung H. Chung, and A. Y. Ng considered the problem of monocular depth estimation as machine learning problem. They showed that it is possible to give fine dense depth map for single monocular image using supervised learning approach and Markov Random Field (MRF) model. They were able to estimate 3D depth maps for monocular images of rural and urban outdoor environment, which were to some point both quantitatively accurate as well as visually pleasing.

Quality evaluation for the monocular cue based descriptor used in the algorithm proposed by [1] is difficult task. This is related to the fact that for a given test image, in general case, there is no reference or exact precise information about the depth to compare with. The depth information is available only for limited set of images specially recorded to be used for the testing. The size of the test set is too small to ensure relevant statistical evaluation. Even if big enough test set was available, which would be useful in error analyses, ablative analyses and give insight into the problem and may point out which monocular cues are best suited for this problem, statistically more accurate predictions may give worst visually distinguishable object depths. Hence, we have to have in mind visually pleasing vs. correct performance compromise burden. At present the judgment for quality performance is observer subjective, and we consider visual inspection by multiple observers as a usable performance evaluation method.

In the next section 2, 3, 4 is presented the approach used in [1] when 3 depth cues are used and an explanation is given how the depth cues are incorporated into one feature descriptor for single image depth estimator modeling.

The last 2 sections are devoted to experiments and quality evaluation of the proposed monocular cue based feature descriptor in [1].

II. THE APPROACH AND ALGORITHM DESCRIPTION

The algorithm in [1] consists of feature extraction phase and depth inferring phase. In the feature extraction phase a set of cues are extracted from the image using selected image processing techniques. The feature extraction is applied on block level and a feature vector is extracted for every image block (patch). Texture intensity, edge directions and haze are extracted as cues to infer depth, since they appear different at different depths.

The depth inferring phase relays on learned relationship between extracted features and the actual depth of the scene. Supervised learning approach is used to learn the mapping between patch features and the depth. In order to perform learning, a training set is obtained, consisting of pairs of images and corresponding depth maps.

Inferring depth from the cues of an individual patch is not reliable; most of these monocular cues contain "contextual information," in the sense that they represent global properties of an image. Due to ambiguities like these, the overall organization of the image should be considered in order to determine depths. This is done by utilizing the cues of neighboring patches and using MRF model to enforce smoothness constraints.

III. FEATURE EXTRACTION PHASE

Two types of features are used: absolute depth features, used to estimate the absolute depth at a particular patch location, and relative depth features, which are used to estimate relative

depths (the difference in depth between two patches in horizontal and vertical direction).
Many objects' texture will look different at different distances from the viewer. Texture gradients, which capture the distribution of the direction of edges, also help to indicate depth. Haze is another depth cue, and is caused by atmospheric light scattering. So the following three types of local cues are used: texture variations, texture gradients, and haze. For each patch of the image the following computations are performed:

1. Law filter bank is applied on the intensity channel to extract texture energy:

Figure 1: Law filter bank.

2. Low-pass filter is applied on the two color channels C_b and C_r to capture haze:

Figure 2: Low-pass filter for haze capturing.

3. Navatia Babu filters oriented at six angles (0,30,60,90,120,150 degrees) are applied on the intensity channel to extract edge directions:

Figure 3: Navatia Babu filters for extracting edge directions.

The absolute and squared filter outputs are summed over all pixels within the patch. There are 17 filters in total; hence the initial feature vector is of dimension 34. Each element of the vector is calculated using following equation:

$$E_i(n) = \sum_{(x,y)\in patch(i)} \left| I(x,y) F_n(x,y) \right|^k, k = 1,2 \qquad (1)$$

In the above equation $E_i(n)$ is the sum of "absolute energy" (k=1) and the sum of "squared energy" (k=2), respectively, for particular patch i in the image $I(x,y)$. $F_n(x,y)$ is the kernel function of the filter n=1,2,...17.
An attempt to incorporate or capture contextual information is made by using image features extracted at multiple scales (image resolutions). To capture additional global features (e.g. occlusion relationships), the features used to predict the depth of a particular patch are computed from that patch, as well as from the four neighboring patches. This is repeated at each of the three scales, so that the feature vector for a particular patch includes features of its immediate neighbors, its far neighbors (at a larger scale), and it's very far neighbors (at the largest scale), as shown in Fig. 4.
Many objects (such as trees and buildings) found in outdoor images show vertical structures and are usually spread in large vertical area in the image. In order to better capture the depth structure of these objects, the image is divided in columns, and every column is divided in four patches referred to as column patches. The features of the column patches are added to the features of patches that lie in that particular column. The structure of the absolute feature vector: central patch, the four cardinal neighboring patches at each resolution and the column features is illustrated in Fig. 4.

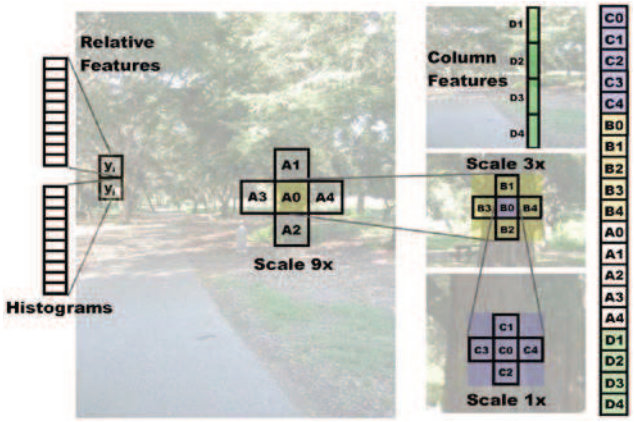

Figure 4: The absolute depth feature vector for a patch includes features from its immediate neighbors and its more distant neighbors (at larger scales). The relative depth features for each patch utilize histograms of the filter outputs.

Final feature vector has 646 elements: 34 features for each patch X (3 resolutions X 5 patches for each resolution + 4 column patches). Relative feature vectors are used to estimate the difference of the depths of two adjacent patch locations. Relative depth features are calculated as differences between the histograms of filter outputs computed from two neighboring patches. Each histogram has 10 bins. There are 17 filter outputs; hence, the dimension of the relative feature vector is 170. The structure of the relative feature vector is illustrated in Fig. 4.

IV. The Depth Inferring Phase

In the depth inferring phase the actual depth at particular patch locations is estimated using feature descriptors extracted from the image. The depth estimation is performed using probabilistic model based on discriminatively-trained Markov Random Field (MRF). The model incorporates multiscale local and global image features, and models both depths at individual points, as well as the relation between depths at different points.
In [1] two probabilistic models were proposed: Gaussian probability model and Laplace probability model.

A. The Gaussian Probability Model

The Gaussian probability model for depth inferring is given with the following equation:

$$P(d \mid X;\theta,\sigma) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{M}\frac{\left(d_i - x_i^T\theta_r\right)^2}{2\sigma_{1r}^2} - \sum_{s=1}^{3}\sum_{i=1}^{M}\sum_{j\in N_s(i)}\frac{\left(d_i(s)-d_j(s)\right)^2}{2\sigma_{2rs}^2}\right) \quad (2)$$

In (2), Z is the normalization constant for the model, M is the total number of patches in the image (at the lowest scale); x_i and d_i are the absolute depth feature vector and the depth for patch i, respectively, and θ and σ are parameters of the model. The depth at a higher scale is constrained to be the average of the depths at lower scales.

In equitation (2), because the considered image is assumed to be recorded using horizontally mounted camera, different rows of the image have different statistical properties and different parameters (θ_r, σ_{1r}, σ_{2rs}) are used for each row in the image. The first term in equation (2) models depth as a function of multi-scale features of a single patch i. The second term in equitation (2) places a soft constraint on the depths to be smooth. The model parameter θ_r is found solving last squares problem [1].

The log-likelihood is quadratic in d_i, hence the MAP (maximum a posteriori) estimate can be found in closed form. In practice, though, usually iterative procedure is used to control the smoothing of the estimated depths in each iteration.

B. The Laplace Probability Model

The Laplace probability model for depth inferring is given with the following equation:

$$P(d \mid X;\theta,\sigma) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{M}\frac{\left|d_i - x_i^T\theta_r\right|}{2\lambda_{1r}} - \sum_{s=1}^{3}\sum_{i=1}^{M}\sum_{j\in N_s(i)}\frac{\left|d_i(s)-d_j(s)\right|}{2\lambda_{2rs}}\right) \quad (3)$$

In this model the parameters are the same as in equation 2, except for the variance terms. Here, λ_{1r} and λ_{2rs} are the Laplacian spread parameters. Maximum-likelihood parameter estimation for the Laplacian model is not tractable. By analogy to the Gaussian case, approximation is found by solving a linear system of equations $Xd\theta \approx$ that minimizes L_1, instead of L_2 norm. Robust multi-linear regression is one of the techniques that can be applied.

Given a test image, MAP inference for the depths d_i is tractable. Solution is found by maximizing the model in terms of d_i and using linear programming.

V. Experimental Results and Analysis

This part consists of the results from different experiments preformed with the algorithm proposed in [1]. In order to test the performance of the algorithm in its original form, an implementation code was obtained from Stanford University. Unfortunately, the code was a test version, it had lot of errors and it was not complete. It was fixed, completed and tested by the DIPteam. The difference from what is published in [1] and current implementation is that the absolute depth feature vector is of dimensionality 816. It includes filter outputs on power 4 (fourth moments) and the column feature patch is not divided into 4 pieces. Relative feature vector, which has 170 elements, and gives estimation how different the depths of two adjacent patch locations can be, is replaced with two relative feature vectors with dimensionality 170 thus using separate difference measurement in two different directions. All of those changes were made by the authors in the test code, but are not documented in the paper [1].

The test results are shown in log scale to better visually distinguish the values of the predicted depths at particular location. The color map is shown in Fig. 5.

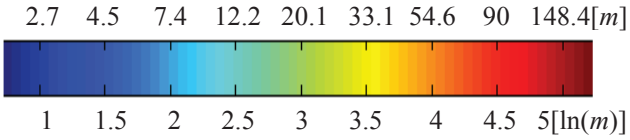

Figure 5: The color map in logarithm meter units ln(m);

A. The Probabilistic Model

A collection of 300 images, with 90% of them having unstructured (rural) outdoor environment content (including forests, trees, etc.), was obtained from Stanford University. In the test 3/4 of the image collection were used for training and the rest was used for testing. These results were useful in determining the qualitative difference between the Gauss and Laplace model in predicting the depth maps.

In Fig. 5.1 are shown some of the results produced by the algorithm using Gaussian model. In Fig. 5.2 are shown some of the results produced by the algorithm using Laplace model. It seems that Laplace model in some cases gives better predictions in terms of more sharp depth maps and is more robust to outliers. In other cases it is worst then Gaussian model. Despite the fact that it is difficult to perform error analysis, it can be concluded that the Gaussian model has more consistent predictions, so in what follows the Gaussian model is used.

In Fig. 5.1 are shown some of the results produced by the algorithm using Gaussian model.

B. Quality Evaluation of the Feature Descriptor

In order to see which of the extracted image features have the biggest contribution to algorithm accuracy, and are most important and successful in capturing the relation depth feature, the absolute feature vector was disassembled into few absolute feature vectors and for each one of them a separate training and testing was performed. The following partial absolute feature vectors were tested: extracted only texture cues, extracted only haze and the distribution of edge orientation cues, summary statistic for only 1 resolution scale, summary statistic for only 2 resolution scales, only absolute and squared summary statistic and full absolute feature vector for images scaled down 5 times. Same training and test sets were used for all experiments. The training set is a collection of pairs - urban outdoor images/ground truth depth maps, with 95% of them having urban outdoor content. Part of the training pairs (140) is obtained from Stanford University.

Another 400 urban images were recorded from Skopje and 150 of them were used in the training set. For those images depth maps were estimated using Google Earth, with average error rate of 10%. This error rate is acceptable in comparison to the depth maps made in Stanford. The depth maps in Stanford set were obtained using laser scanning. Often the laser scans have errors due to reflections and missing laser scans and due to the noise in the motor system. Also, the depth maps are not perfectly aligned with the images, and have an alignment error of about 2 patches. The depth maps in the complete training set have a maximum range of 110m.
A selection was made on the images taken in Skopje and 70 images different from the ones used in the training set were chosen for testing. Some of the results are shown in Figs. 6.1-8.2.
Edge distribution and texture variations contribute separately to the estimation of the 3D structure. Following the results in Figs. 6.1 and 6.2, it seems that captured texture variations (Fig. 6.1) are more essential for 3D rezoning, in this algorithm, then the edge distribution (Fig. 6.2). This is because texture variations are more reach with information about 3D depth cues. When only depth cues extracted at two or one resolution scales are used (Fig. 7.1, 7.2), the ability to distinguishing the precise local 3D depth which is related with the overall 3D structure is low, and the resulting predicted depth map is coarse.

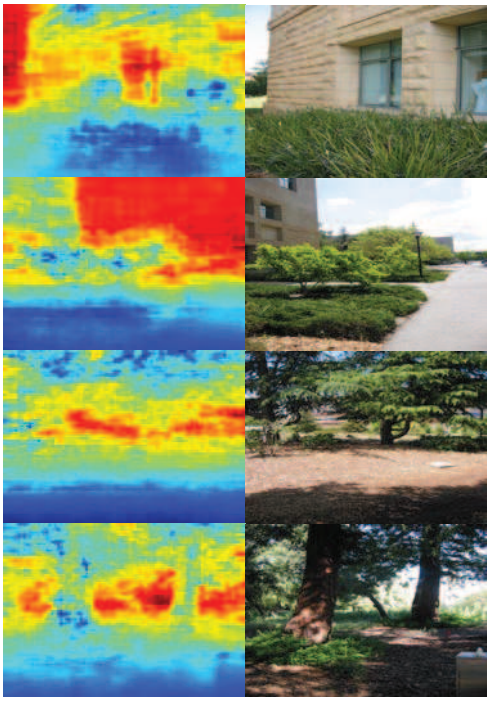

Figure 5.1: Predicted absolute depth map (Gaussian model) (column 1); Original image (column 2).

In Fig. 5.2 are shown some of the results produced by the algorithm using Laplace model.

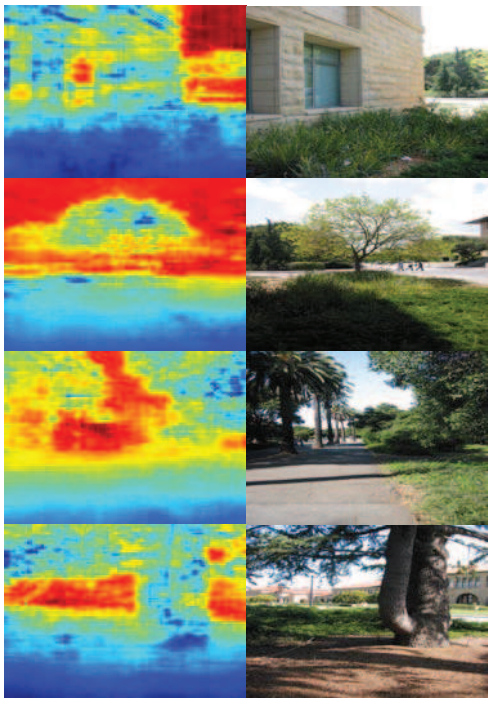

Figure 5.2: Predicted absolute depth map (Laplace model) (column 1); Original image (column 2).

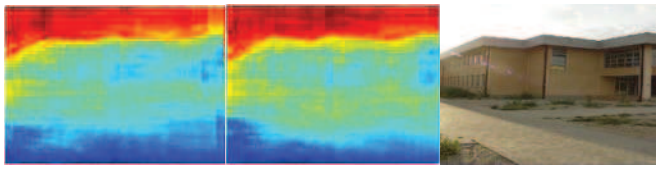

Figure 6.1: Estimated absolute depth map using texture based features only (column 1), complete feature vector (column 2); Original image (column 3).

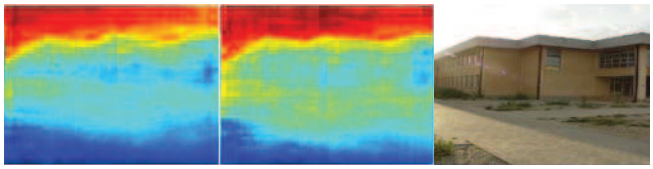

Figure 6.2: Estimated absolute depth map using the haze and the distribution of edge orientation features only (column 1), complete feature vector (column 2); Original image (column 3).

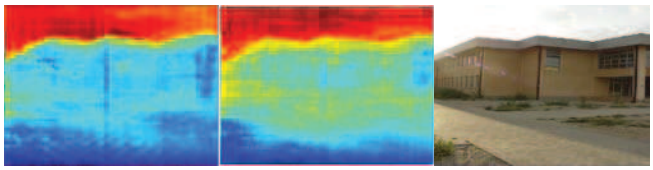

Figure 7.1: Estimated absolute depth map using only 2 resolution scales (column 1), complete feature vector (column 2); Original image (column 3).

The results shown in Fig. 8.1 reveal that absolute and square summary statistics in insufficient, the accuracy in predicting the depth map is low. As can be seen in Fig. 8.2, downscaling of the image results in loses of image information about the image details. When the algorithm reasons about the depth on downscaled image, the overall depth structure is obtained, however the accuracy in estimating 3D details in the depth map is low.

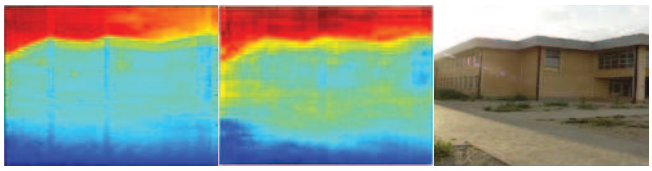

Figure 7.2: Estimated absolute depth map using only 1 resolution scale (column 1), complete feature vector (column 2); Original image (column 3).

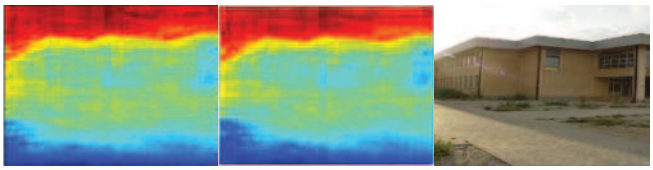

Figure 8.1: Estimated absolute depth map using only absolute and squared summary statistics (column 1), complete feature vector (column 2); Original image (column 3).

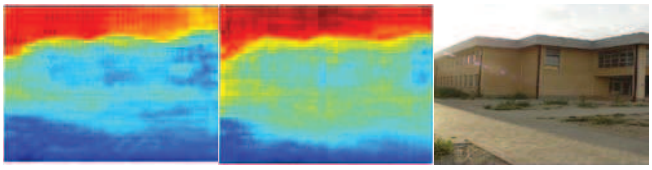

Figure 8.2: Estimated absolute depth map for images scaled down 5 times (column 1), original size (column 2); Original image (column 3).

Tests were also performed to see the influence of the size of the training set on the accuracy of inferring. Results of the preformed test are shown in Fig. 9.

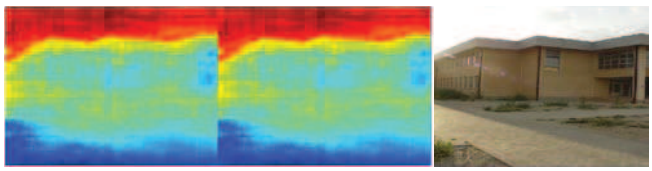

Figure 9: Gaussian model: predicted absolute depth map using training set of 160 images obtained from Stanford (column 1), and training set of 290 images (column 2); Original image (column 3).

In general, the test showed that by increasing the number of training images the inferring accuracy for test image with strongly similar content with the content of the training set is increased; however the improvement is not dramatic.

The ability to reason about variety of images with different contents is not increased.

In this algorithm depth is inferred for single image patch, and the patches are considered as basic units for image representation. The image is divided into number of rectangular patches, and with this kind of representation the information regarding the boundaries of the real image objects is lost. Hence, in the process of depth map inferring it is difficult to differentiate if a particular variation in the depth is due to noise in the image or simply due to depth discontinuities.

VI. Conclusions

It can be concluded, in general, that all components of the feature vector are important for the accuracy of the algorithm. Also, the feature vector is nicely fitted to the size of the image. However, due to the division of the image into number of rectangular patches, and with that loss of information regarding the boundaries of the real image objects, the estimation of the depth map is sensitive to noise in the image.

The algorithm mainly focuses on the "learning" part. However, the exact relation between the extracted depth cues and the depth is not explicitly resolved, nor the effectiveness of the used depth cue extracting filters for that manner. In that context, we expect that a research dealing with these issues could be of significant importance in both ways, for computational effectiveness and accuracy improvement.

VII. References

[1] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng "Learning Depth from Single Monocular Images" ***NIPS 18***, 2005

[1.1] Ashutosh Saxena, Jamie Schulte, Andrew Y. Ng. "Depth Estimation using Monocular and Stereo Cues" ***IJCAI***, 2007

[3] Ashutosh Saxena, Sung H. Chung, Andrew Y. Ng. "3-D Depth Reconstruction from a Single Still Image" ***(IJCV)***, Aug 2007

[4] A. Torralba, A. Oliva., "Depth estimation from image structure," PAMI, 24(9): 1-13, 2002

[5] V. Nedovic, A. W.M. Smeulders, A. Redertand J.M. Geusebroek, "Depth Information by Stage Classification," ICCV, 2007

[6] K. Boulanger, K. Bouatouch and S. Pattanaik "ATIP: A Tool for 3D Navigation inside a Single Image with Automatic Camera Calibration"