Depth Estimation in Monocular Images: Quantitative versus Qualitative Approaches

Ralph Ewerth[1], Alexander Balz[2], Jan Gehlhaar[2], Krzysztof Dembczyński[3], Eyke Hüllermeier[4]

[1]Jena University of Applied Sciences, [2]University of Marburg, [3]Poznań University of Technology, [4]University of Paderborn
E-Mail: ralph.ewerth@fh-jena.de, {balz, gehlhaar}@informatik.uni-marburg.de, kdembczynski@cs.put.poznan.pl, eyke@upb.de

Introduction

Depth estimation in images and video sequences is an important prerequisite for a number of computer vision tasks and scene understanding. However, estimating the camera distance, i.e., depth information, of objects in monocular images is a difficult problem. Human visual perception of depth in single images relies on some monocular depth clues. Different approaches have been suggested to tackle the problem of automatic depth estimation in single images using regression, for example, by using conditional random fields (e.g. [8, 9]) or deep learning methods (e.g. [7]). Such approaches estimate the camera distance of scene parts and objects in a quantitative manner.

A natural question arising in this context concerns the suitability of the information contained in an image for training a quantitative regression model. More generally, one may wonder about the limits of an approach of that kind. Classical methods for three-dimensional image acquisition require additional information, for example obtained by a second camera or generated by structured light projection. There is a number of approaches of this kind, but they are not suitable for depth estimation in single images.

In this extended abstract, we discuss an alternative approach that is based on learning-to-rank methods from the field of preference learning [4]. The main idea is to qualitatively estimate the relative order of objects or image regions, respectively. This approach is motivated by the assumption that information extracted from a single image does not provide enough information about the correct (quantitative) camera distance. Moreover, what is typically needed in practice is indeed the relative order of scene objects, not the precise distances. Seen from this point of

view, a regression method is actually solving an unnecessarily difficult problem or, stated differently, tackles the actual problem only indirectly by solving a more difficult problem first. Finally, supervision becomes simpler for ranking methods, because qualitative comparisons are easier to obtain than numerical depth values, especially if humans are involved in the labelling of images.

In the presented approach, monocular depth features are needed as well. We suggest corresponding descriptors for some monocular depth criteria that are aimed at capturing information about the relative order of two regions in an image. These descriptors are then employed to compare and rank image regions using a ranking method. Rankboost is used in our first experiments. This approach is shown to be capable of qualitatively estimating depth in a single image. Based on these preliminary results, the advantages and disadvantages of a ranking-based approach are discussed with respect to several aspects: possibility of generating training data for different applications, space and runtime complexity, accuracy of estimation, and possible use cases.

Feature Extraction and Depth Estimation by Rankboost

In a first step, the image is partitioned into rectangular regions of equal size. Then, features are extracted for each region according to the modeled monocular clues. The following six monocular depth clues are used in the proposed system: relative height, atmospheric perspective, linear perspective, texture gradients (contours), and two kinds of color similarity.

Since the relative depth information is only weakly correlated with the monocular depth descriptors, an integration model is required that combines them in an optimal manner. As we tackle depth estimation as a learning-to-rank problem, a function is sought that orders image regions according to their camera distance, i.e. scene depth. Preference learning methods are designed for tasks of this kind [4]. More specifically, the problem of ranking image regions can be formalized as a so-called *subset ranking* or *object ranking* task, to which the Rankboost algorithm is applicable [5]. In general, boosting algorithms are meta learners that combine so-called weak learners in an ensemble. Each weak learner is expected to yield an error rate slightly better than random guessing. The decision of the ensemble is finally realized in the form of a linear combination of the weak learners.

Depth estimation in single images can now be tackled using Rankboost as follows. Each training image is considered as an exemplary object ranking, where the objects are image regions ranked according to their scene depth based on the given features. Each image region can be described via monocular depth features that

are used to infer a strong depth (ranking) function using Rankboost's training algorithm. The concept of combining weak rankings into a strong ranking function fits well to the depth estimation scenario, since the monocular depth clues itself provide only rather weak information about scene depth. A ranking of image regions based on a single monocular depth feature represents a weak ranking. The final depth map is learned by combining the best weak rankers, namely those with the minimum error in the corresponding training round.

Preliminary Experimental Results

We present some preliminary experimental results for depth estimation using a learning-to-rank approach. The publicly available data set provided by Saxena et al. [9] is used in these experiments. It consists of 534 outdoor images in total and is divided into a training set of 400 images and a test set of 134 images. The original image resolution is 1704×2272, the resolution of the related ground truth depth maps is 305×55. Each image is separated into image regions, and monocular depth descriptors are extracted for each region. The monocular depth clues are represented by a six-dimensional feature vector. This feature vector is used to generate a ranking-based depth map for an image that represents the ordering of image regions according to their scene depth.

We used the Rankboost implementation of Dang's Ranklib library [1]. In the Rankboost tests, 300 training rounds and 10 thresholds were used by default. The quality of generated depth maps is evaluated in terms of the error rate, which is the percentage of incorrectly ordered pairs of regions in a depth map representation of an image. The original resolution of depth maps has been scaled from 305×55 to 61×55 in order to have a more balanced ratio of width and height compared with the original images, for both the entire depth map and its regions.

The presented approach achieves an error rate of 23.4% on the test set. Some example results are shown in Figure 1. These results are promising and confirm that a qualitative estimation of relative depth in single images is in principle possible, although the accuracy of the best regression methods has not yet been reached. For example, we have evaluated Liu et al.'s approach [8], which achieves an error rate of 17.7%. It needs to be mentioned, however, that this approach uses additional information by predicting semantic categories for image regions to infer depth information.

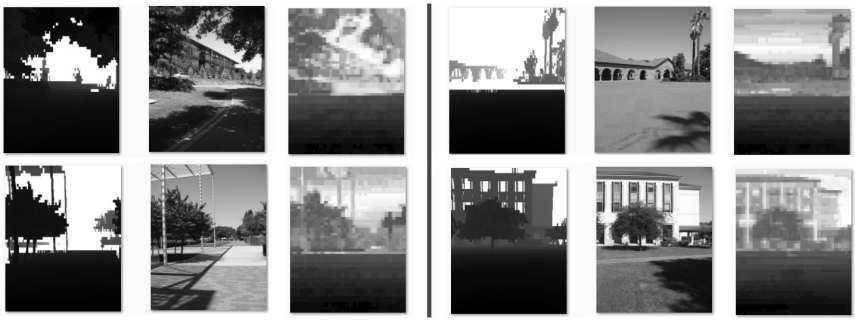

Figure 1: Some example results presented in two colums: ground truth depth map (left), original image (middle), and depth map created by Rankboost (right).

Conclusion: Comparing Regression and Ranking Methods

One of the main advantages of qualitative depth estimation is a broader range of application domains thanks to a simplification of data collection. For a qualitative approach, training data does not need to be acquired by a 3D-camera system but can easily be generated using manual labeling with an appropriate software tool. Hence, depth estimation becomes possible in domains for which no depth data has been generated before. One possible application domain lies in the field of visual concept detection in images or video retrieval [2].

In addition, the presented approach using Rankboost is efficient. Its run-time complexity is linear in the number of regions and training rounds, whereas, for example, solving the linear program for inference in Saxena et al.'s approach [9] has polynomial runtime complexity.

By now, the main disadvantage of the presented approach is that alternative methods still achieve a better accuracy. Hence, the main direction for future work is to model better monocular depth descriptors that capture depth information more accurately. Besides, further ranking methods ought to be investigated as alternatives to Rankboost.

References

[1] V. Dang. “Ranklib – A Library for Ranking Algorithms (Version 1.1)”. http://people.cs.umass.edu/ṽdang/ranklib.html 2010.

[2] R. Ewerth, R., M. Schwalb, and B. Freisleben, B. “Using Depth Features to Retrieve Monocular Video Shots”. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*. Amsterdam, The Netherlands. 210–217. 2007.

[3] P.F. Felzenszwalb and D.P. Huttenlocher. “Efficient Graph-Based Image Segmentation”. In: *International Journal of Computer Vision*. Volume 59. Issue 2. pp. 167–181. Kluwer Academic Publishers. 2004.

[4] J. Fürnkranz and E. Hüllermeier. “Preference Learning: An Introduction”. In: *Preference Learning* 4.1. pp. 1-17. Springer-Verlag. 2010.

[5] Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. “An Efficient Boosting Algorithm for Combining Preferences”. In: *Journal of Machine Learning Research*. Vol. 4, December. pp. 933–969. MIT Press. 2003.

[6] E. B. Goldstein. “Sensation and perception”. Wadsworth Cengage Learning. 2009.

[7] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. “Depth and Surface Normal Estimation from Monocular Images Using Regression on Deep Features and Hierarchical CRFs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA. pp. 1119–1127. 2015.

[8] B. Liu, S. Gould, and D. Koller. “Single Image Depth Estimation from Predicted Semantic Labels”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos, CA, USA. pp. 1253–1260. 2010.

[9] A. Saxena, M. Sun, A.Y. Ng. “Make3D: Learning 3D Scene Structure from a Single Still Image”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 31, No. 5. pp. 824–840. 2009.