OCCLUSION-BASED DEPTH ORDERING ON MONOCULAR IMAGES WITH BINARY PARTITION TREE

Guillem Palou and Philippe Salembier

Technical University of Catalonia (UPC), Dept. of Signal Theory and Communications, Barcelona, SPAIN

ABSTRACT

This paper proposes a system to relate objects in an image using occlusion cues and arrange them according to depth. The system does not rely on any a priori knowledge of the scene structure and focuses on detecting specific points, such as T-junctions, to infer the depth relationships between objects in the scene. The system makes extensive use of the Binary Partition Tree (BPT) as the segmentation tool jointly with a new approach for T-junction estimation. Following a bottom-up strategy, regions (initially individual pixels) are iteratively merged until only one region is left. At each merging step, the system estimates the probability of observing a T-junction which is a cue of occlusion when three regions meet. When the BPT is constructed and the pruning is performed, this information is used for depth ordering. Although the proposed system only relies on one low-level depth cue and does not involve any learning process, it shows similar performances than the state of the art.

Index Terms— Binary partition tree, occlusion cues, monocular depth, T-junction estimation.

1. INTRODUCTION

Humans are known for their ability to recognize objects and scenes in a large variety of situations. Nowadays, computers are far from human vision performances, but high efforts are put in this research area. Usually vision systems rely on the information obtained from multiple viewpoints to estimate disparity to infer depth. Only a few approaches try to infer depth from the observation of a single image. These approaches focus on what is called monocular depth perception. Two main approaches may be distinguished for monocular depth perception. The learning-based ones and the ones that operate over the image structure looking for low level cues. In the former class, [1] and [2] oversegment the image and gather for each region color, texture, vertical and horizontal features to use them within a Markov Random Field (MRF) framework to estimate the depth. These approaches are learning-based because the MRF has been trained with ground truth data. The main drawback of learning-based approaches is that their use is limited to the kind of images they have been trained for. The latter type of systems, where [3] can be included, uses no training but focuses on the detection of relative depth cues such as occlusion or convexity to order the objects in the scene. Note that occlusion does not permit to infer absolute depth as learning-based approaches may offer, but it is more generic as it assumes nothing about the type of scene. T-junction points are known to be strong occlusion cues. It is unlikely that a robust system for depth perception can be defined only using T-junction detection. However, in this paper, we are interested in studying the limitation and weakness of such a system. Nevertheless, surprisingly we will show that the results obtained with only a low-level occlusion cue are comparable with the state of the art systems that use learning and, therefore, higher level information.

The approach in [3] consists of an estimation of T-junctions, followed by an image segmentation and a final depth ordering stage. Its performances are quite good but some of the difficult T-junctions are wrongly estimated in the initial step. The system presented here has two fundamental differences with [3]: it combines the first and second stages of [3] into a single step to improve the robustness of the T-junction estimation and proposes a new iterative procedure to determine the final depth order. The final depth order image is obtained from the minimization of a specific cost function. The system described in Section 2 consists of two basic steps. The first one is the combined BPT construction and T-junction estimation, presented in Section 2.1. In second and last part of the system, described in Section 2.2, the T-junction candidate points are selected and the depth ordering is estimated. Results are presented in Section 3 along with the conclusions and some insights about possible future work.

2. PROPOSED APPROACH

In contrast to [3], the proposed system integrates the BPT construction and the T-junction estimation in a single process. The BPT [4] is constructed from an iterative region merging process where the two most similar and neighboring regions are merged at each step. To define the similarity measure between regions, a region model has to be defined. Color information is represented in the $CIE\ Lab$ color space. Regions are modeled with 3 histograms (one per color channel) of 64 bins. The histograms of initial pixels is estimated as in [3] using self-similarity: each pixel pdf is computed as a

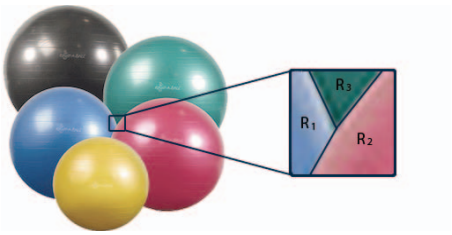

Fig. 1. T-junction example. Locally, region R_2 is the one forming the largest angle, appearing to be over R_1 and R_3.

weighted contribution of its neighbors, depending on neighborhoods similarities. The distance between regions relies on color, area, contour and depth information.
The color distance between regions R_1 and R_2, $d^{1,2}_{color}$, is computed using the Earth Mover's Distance (EMD), weighted by the area of the smallest region. For contour, the measure $d^{1,2}_{contour}$ is the increase of perimeter of the new region with respect to the biggest one [4]. This measure is used only when both regions exceed an area threshold of 50 pixels so that their shape is meaningful. The intermediate distance (discarding depth information) is then

$$d^{1,2}_{BPT} = \alpha \times d^{1,2}_{color} + (1-\alpha) \times d^{1,2}_{contour} \qquad (1)$$

With $\alpha = 0.07$. Once $d^{1,2}_{BPT}$ is available, depth information is introduced using T-junction candidate points. Candidates are the points where three regions meet. If R_1 and R_2 share a common neighbor R_3, at least a T-junction candidate n is present at the contact point. In a neighborhood of n, there is one region $R_i \in (R_1, R_2, R_3)$ which occludes the other two. For example in Figure 1, R_2 occludes R_1 and R_3.

For each candidate n, the probability $p_{i,n}$ that the top region is R_i is computed as described in Section 2.1. As several T-junction candidates may give information about the relative depth of R_1 and R_2, the probability that R_1 is on top of R_2 is evaluated by:

$$p_1 = \left(1 - \prod_{n}^{N_1} (1 - p_{1,n})\right) \prod_{n}^{N_2} (1 - p_{2,n}) \qquad (2)$$

Where N_1 and N_2 are the number of T-junctions indicating R_1 and R_2 is the top region respectively. p_2, the probability that R_2 is the top region, is computed similarly. Note that since each $p_{i,n}$ is independent from the others, $p_2 \neq 1 - p_1$. The final distance between two regions is $d^{1,2} = \frac{d^{1,2}_{BPT}}{(1-|p_1-p_2|)}$. Having defined the region distance, a BPT is constructed assuming that individual pixels form the initial regions. During the construction, the similarity measure increases the distance of regions that do not belong to the same depth plane. Thus, the tree is expected to be partially depth-structured.

2.1. T-junction Candidates Estimation

In this section, we assume that we are analyzing a candidate local configuration n_o in which R_1 may be on the top of R_2, that is, we want to estimate the value p_{1,n_o} of equation (2). To simplify the notation, we call p this value of p_{1,n_o}. To estimate the confidence value p of a T-junction, color difference, angle structure and boundary curvature confidence are evaluated at each candidate point within a centered circular window ($R = 10$), except for the angle. Color contributes to differentiate between contrasted regions. Angle helps to infer the depth relationship and curvature detects if the junction has clearly defined boundaries. As they are independent features, $p = p_{color} \times p_{angle} \times p_{curve}$.

2.1.1. Color

The color confidence is the result of several measures. In a local neighborhood of the candidate point, the three regions are modeled with their mean color vector and their covariance matrix. Pixels neighboring the boundaries are neglected because, their color value is a combination of various regions and they may introduce a bias in the estimation. For each pair of regions $(i, j) = (1, 2), (1, 3)$ and $(2, 3)$ statistical and perceptual measures are first computed. The former, $c^{i,j}_s$, is computed as a two sample Hotelling's T^2 test and the latter, $c^{i,j}_p$, is the euclidean distance between region mean colors. Experiments showed that these measures were prone to false alarms at edges, so another distance was added. To penalize high statistical distance dissimilarities, the minimum statistical distance, $c^{min}_s = \min\left(c^{1,2}_s, c^{1,3}_s, c^{2,3}_s\right)$, is used to obtain $c_r = c^{min}_s - \left(\frac{1}{c^{1,2}_s} + \frac{1}{c^{1,3}_s} + \frac{1}{c^{2,3}_s}\right)^{-1}$.
Reliable candidates are expected to have large color distances. A total of 7 measures (3 for $c^{i,j}_s$, 3 for $c^{i,j}_p$ and c_r) are obtained for each candidate point. The confidence for each measure is known as p-value. It is defined as the probability that a candidate has equal or less distance than the observed measure [5]. The p-value is computed assuming a Rayleigh distribution for the measures. The final color distance p_{color} is obtained from the product of the 7 p-values.

2.1.2. Angle

Angle plays an important role in a T-junction as the region forming the largest angle is assumed to be the one lying on top (see Figure 1). To calculate the angle feature, the three boundaries that meet at a candidate point are used. Information at the junction center is considered to be unclear, so a small nucleus of radius 4 is omitted from the analysis. First the average orientation of each branch is computed. Then, angles θ_i, between pairs of orientation are estimated and the junction angle characteristic is evaluated. Considering the angles, ideal shaped T-junctions have a maximum angle of π and two minimum angles of $\frac{\pi}{2}$. Two measures, $\Delta\theta_{max}$ and

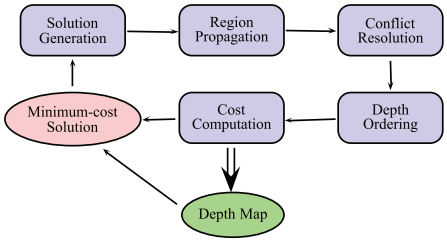

Fig. 2. Block diagram of the minimization loop

$\Delta\theta_{min}$, are defined as the absolute difference of the maximum and minimum angles with π and $\frac{\pi}{2}$ respectively. To obtain the confidence value, $\Delta\theta_{min}$ and $\Delta\theta_{max}$ are considered to follow Rayleigh distributions. Under the hypothesis that the T-junction is ideally shaped, the corresponding p-values are found. The final angle distance p_{angle} is the product of both values.

2.1.3. Curvature

Although curvature is not as important as color and angle, it serves to measure the branches' straightness. If boundaries are highly curved, the point may not be perceived as a junction. The curvature of the boundaries is measured using the level sets theory. Each region is isolated creating a binary image of the local window. Then, the mean absolute value of the curvature is computed at the region boundary points as in [6]. The 3 measures are assumed to be Rayleigh distributed. Under the hypothesis that an ideal candidate point has three straight branches, the final curvature distance p_{curve} is the product of the corresponding p-values of the three measures.

2.2. Selection and Depth Ordering

If all the T-junction candidates are used to generate the final depth order, two problems may arise. First, points corresponding to false alarms may be used. Second, since occlusion cues only give information on the relative depth order, the depth of each region is assumed to be constant. As a result, depth order conflicts may appear when several T-junctions give contradictory order information for the same pair of regions. In this case, if all T-junctions are considered, depth ordering is impossible. To converge to a solution some of the candidates should be discarded. The problem selecting the best set of T-junctions can be solved by the iterative minimization of a cost function, which can be performed by the scheme described by the block diagram of Figure 2.

To define the cost function three concepts are needed. First, since the candidate points are characterized by their confidence value, true T-junction are expected to have high p value. Second, it is expected that in real images, the number of true T-junctions is rather low. Therefore, the resulting depth order image is also expected to have few regions and depth planes. Third, regions are expected to have at least one depth relationship with their neighbors, represented in a Depth Order Graph (DOG). From these concepts, the cost function is defined as:

$$C(R) = \sum_{i \in R} \frac{c_i}{c_{max}} + \gamma_N \times N + \gamma_u \times U \qquad (3)$$

R is the set of rejected candidates. The individual cost function of a junction i is defined as $c_i = p_i \times \min(n_1^i, n_2^i, n_3^i)$. n_x^i, $x \in (1, 2, 3)$ are the region areas on the local window. $\gamma_u = 2$ is a penalization factor for isolated nodes U in the DOG, and N is the number of 4-connected components in the depth order image. γ_N is set to the half of the geometric mean of all the normalized T-junctions costs $\frac{c_i}{c_{max}}$.

The minimization loop is used to minimize the cost defined by eq. (3). Previously, an initial T-junction selection is performed by thresholding the T-junction probability value. At each iteration, the first stage of the loop generates a solution from the final T-junctions candidates, selecting the points randomly. The probability to select a candidate depends directly on its p value. The second step selects nodes on the BPT forming a partition preserving the selected T-junction points, allowing the construction of the DOG from T-junction depth relationships. Conflicts between T-junction information can be identified in the DOG as cycles. These cycles can be removed sequentially by rejecting the T-junctions of lowest probability values. After that, the depth ordering is performed by depth-labeling each region with its partial order in the DOG. Finally, the depth order image is generated. The cost defined by eq. (3) is calculated and the loop is restarted until stabilization, choosing the minimum cost solution as the final depth order image.

3. RESULTS AND CONCLUSIONS

Figure 3 presents some results of the algorithm and compares them with the state of the art solutions. The chosen images were selected from a variety of locations such as database [7] or downloaded from the Internet. They gather both indoor and outdoor environments. The original images are shown on the top row. The final set of T-junctions selected after the minimization loop are indicated in the second row. The three last rows present the depth map. Bright values correspond to regions that are close to the camera. Comparing the results of our system (last row) and the results from [2] (third row), one can see that occlusion cues are a good feature for relative depth ordering but, in some cases, it is insufficient to classify depth planes. Results of [1] (fourth row) show that surface orientation greatly increases the perception of depth. As mentioned in the introduction, learning-based strategies offer absolute depth information, which is unavailable considering only occlusion cues.

Several conclusions can be extracted from the results of Figure 3. First, the proposed system performance may compete

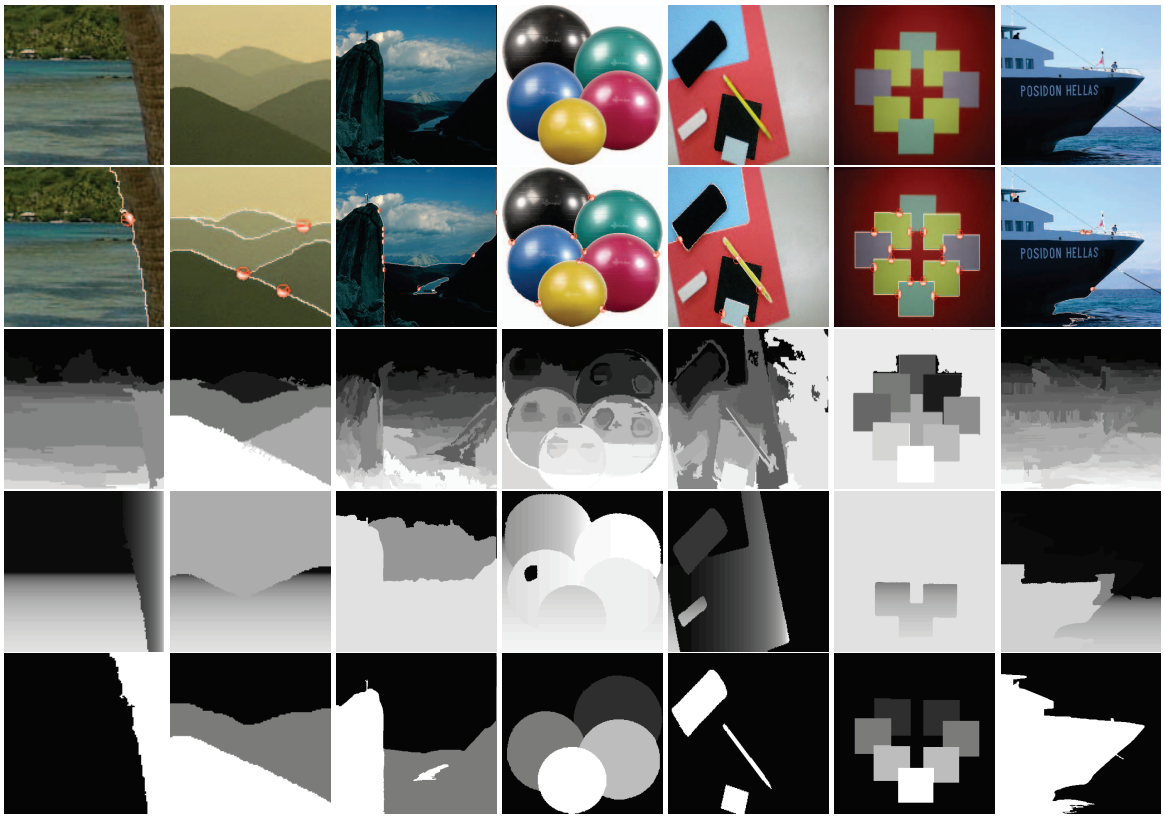

Fig. 3. Results of the algorithm compared with [2] and [1]. From top to bottom, in rows: original image, final image partition with marked T-junctions, results from [2], [1] and results of the proposed system. Bright (dark) values correspond to regions close to (far from) the camera.

with the current state of the art techniques and although [1, 2] infer absolute depth, relative depth relations are usually correctly estimated by our scheme. Second, our approach defines much more precisely the object boundaries, thanks to the use of the joint T-junction estimation and BPT construction. Third, due to the training process, algorithms from [2, 1] perform worse when the images correspond to indoor scenes. Furthermore, our system can also be used for figure/ground segregation. Nevertheless, it is possible to see that in some images not all objects/regions are indeed correctly ordered (images 2 and 5) due to the lack of T-junctions between them. Taking into account the last observation, future efforts will focus on improving the system to include other depth cues such as convexity to solve more general cases.

4. REFERENCES

[1] D. Hoiem, A. N. Stein, A.A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," in *IEEE Int. Conf. on Computer Vision ICCV 2007*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[2] A. Saxena, A. Ng, and S. Chung, "Learning depth from single monocular images," *Neural Information Processing systems (NIPS)*, vol. 18, 2005.

[3] M. Dimiccoli and P. Salembier, "Hierarchical region-based representation for segmentation and filtering with depth in single images," in *Int. Conf. on Image Processing ICIP 2009*, Cairo, Egypt, 2009, pp. 3497–3500.

[4] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. on Image Processing*, vol. 17, no. 11, pp. 2201 –2216, nov. 2008.

[5] K. V. Mardia, J. M. Bibby, and J. T. Kent, *Multivariate analysis*, Academic Press, London; New York, 1979.

[6] F. Guichard and J.M. Morel, "Image analysis and PDEs," *Institute for Pure and Applied Mathematics GBM Tutorial*, 2001.

[7] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, July 2001, vol. 2, pp. 416–423.