FROM LOCAL OCCLUSION CUES TO GLOBAL MONOCULAR DEPTH ESTIMATION

Guillem Palou, Philippe Salembier

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

ABSTRACT

In this paper, we propose a system to obtain a depth ordered segmentation of a single image based on low level cues. The algorithm first constructs a hierarchical, region-based image representation of the image using a Binary Partition Tree (BPT). During the building process, T-junction depth cues are detected, along with high convex boundaries. When the BPT is built, a suitable segmentation is found and a global depth ordering is found using a probabilistic framework.

Results are compared with state of the art depth ordering and figure/ground labeling systems. The advantage of the proposed approach compared to systems based on a training procedure is the lack of assumptions about the scene content. Moreover, it is shown that the system outperforms previously low-level cue based systems, while offering similar results to a priori trained figure/ground labeling algorithms.

Index Terms— binary partition tree, depth estimation, occlusion, T-junctions, convexity

1. INTRODUCTION

Perceiving depth is a fairly easy task for humans. In most situations, people estimate disparity from two images, one for each eye. However, even if only one point of view is available, good scene interpretations can be done barely at a glance. It is believed that a priori scene knowledge helps to infer the final scene structure. Regardless, humans are able to derive the depth successfully in unknown situations. This capacity is also due to a low level vision process, where depth cues (features indicating depth relationships) are detected.

State of the art algorithms working on monocular depth perception rely mostly on a priori trained systems [1, 2]. Although these approaches may be suitable for common situations, they cannot handle all possible types of scenes. Another approach to arrive at a plausible depth estimation is to imitate the human vision system by detecting low level depth cues.

To this purpose, the region based approach in [3] proposes three steps: detection of T-junction points, BPT construction and depth ordering. The system presented here has two fundamental differences with [3]: it combines the first and second stages of [3] into a single step and proposes a new iterative probabilistic framework to determine the final depth order. The proposed system is an extension of [4] where less depth cues were considered to determine depth.

The organization of the paper is as follows: first, in Section 2, the estimation of local depth cues is explained. The procedure to go from local depth cues to global depth ordering is outlined in Section 3. The last Section 4 presents the results compared with state of the art systems on monocular depth ordering and figure/ground segregation.

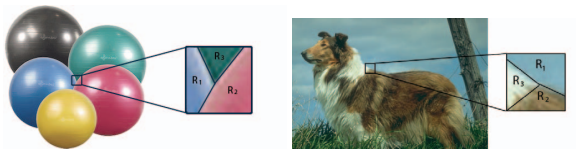


Fig. 1. The local depth gradient of T-junctions cannot be determined using only local depth cues. At the left, the red ball occludes the green and blue balls, producing a *normal* T-junction. The right image shows an *inverted* T-junction. Due texture variations, the sky appears to be locally in front of the dog.

2. ESTIMATION OF LOCAL DEPTH CUES

The proposed system adopts a very similar approach to [4] to estimate depth cues. This process is jointly performed with the construction of a BPT [5]. The BPT is a bottom-up approach and at each iteration merges the two most similar regions according to a defined distance using color, area, shape and depth features. Still, two major changes are introduced in the estimation of local depth cues. First, instead of considering only one kind of T-junctions [6], *normal* and *inverted* types are introduced, allowing different occlusion relations. Second, it is known that, typically, humans relate convex boundaries to foreground regions [7]. Hence, in addition to occluding points, convexity is also estimated as an occlusion cue between adjacent regions.

2.1. T-junction Estimation

To estimate T-junctions, all points i in the image are assigned a T-junction confidence value $0 \leq p_i \leq 1$. This confidence indicates the probability that the point i is indeed a true T-junction. p_i is estimated at each BPT merging, where all points where three regions meet are considered plausible candidates to indicate occlusion.

The process is the same as the one exposed in [4] but a brief explanation follows. Three features are examined locally at each point: color, angle and curvature. Since good T-junctions should have good color contrast, with straight and T-shaped branches, the three measured features are compared with the ideal values: high contrast, perfect angles and no curved branches. The result of this comparison is a confidence value for each feature. Finally, to obtain the overall T-junction confidence for a point i, color p_c, angle p_a and curvature p_κ confidences are combined to obtain: $p_i = p_c \times p_a \times p_\kappa$.

2.2. Depth gradient at T-junctions

Previous work on T-junctions [3, 4] imposed a strong depth configuration for these cues: the region forming the largest angle (top region) is lying closer to the viewer. However, experience shows that

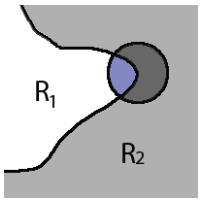


Fig. 2. Normally, convex shapes present less area in small neighborhoods. Convex regions, as R_1 here, are perceived as foreground while R_2 is perceived as background.

T-junctions may also indicate the opposite depth relation. Since, locally, all kinds of junctions are similar, deciding whether T-junctions are *normal* or *inverted* should be done by looking other than local features, Fig. 1.

The depth relation created by a T-junction proved to be very uncertain. Normally, if an object is really occluding other objects in the background, more than one T-junction is likely to be formed in the image, and all these T-junctions may have the same region/object as the occluding region. On the other way, it is possible to detect a T-junction even though no real occlusion relation exists. False detections often occur due to color or texture variations. As an initial guess, prior to global reasoning, all T-junction are considered *normal*, indicating true occlusion relationships.

This initial guess, being low confident, is allowed to change when estimating the global depth ordering of the scene. That is, in some circumstances, the depth gradient of a T-junction is changed if there are many other occlusion relations indicating the opposite depth relationship.

2.3. Convexity Estimation

Convexity depth cues are defined locally at region boundaries. A region R_1 is convex with respect to R_2 if, on average, the curvature vector on the common boundary is pointing towards R_1. If R_1 appears to be convex, it is perceptually seen as the foreground region (and thus, closer to the viewer). To determine convexity, the curvature vector should be computed but, instead, a faster strategy is proposed. Generally, when examining boundary pixels, if R_1 presents less area than R_2 in a local neighborhood, R_1 may be seen as convex, see Fig. 2. Convexity cues are also characterized by a confidence value which can be computed using:

$$\zeta_c(R_1, R_2) = \frac{\lambda}{L} \sum_{(x,y)\in\Gamma} \alpha(x,y) \tag{1}$$

With $\alpha(x,y) = 1$ if the area of R_1 is greater than the area of R_2 in a local window $\Omega(x,y)$, $\alpha(x,y) = -1$ otherwise. The weighting parameter λ is chosen to be the average of the normalized gradient magnitude along the boundary. L is the number of points where the measure $\alpha(x,y)$ is calculated. The overall convexity confidence of a boundary is:

$$\zeta(R_1, R_2) = 1 - \exp\left(-\frac{1}{\gamma_c} \times |\zeta_c(R_1, R_2)|\right) \tag{2}$$

With $\gamma_c = \frac{1}{12}$ determined experimentally. If the result $\zeta_c(R_1, R_2)$ is positive, R_1 is considered to be convex and, therefore, on top of R_2. The converse indicates that R_2 is on top of R_1. To make the measure as scale invariant as possible, $\Omega(x,y)$ is chosen to be a circular window with a radius of 5% of the contour length. Points lying near junctions, image borders and other regions are discarded for the measures. Contours having small lengths ($L < 100$ pixels) are considered to be non-significant for convexity cues.

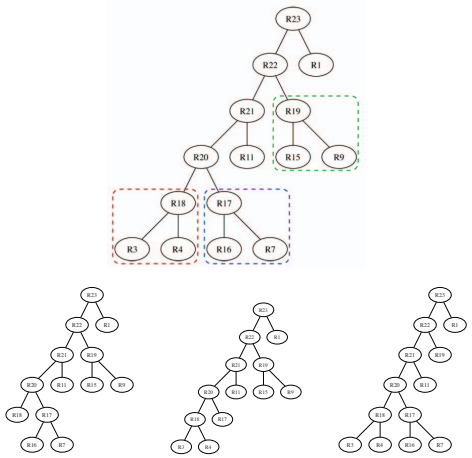


Fig. 3. Three allowed prunings of a given BPT. For each pruning, the framed leave nodes are merged to its parent, reducing the number of BPT nodes by one. All other possible prunings in this BPT reduce the number of leaves of the tree by more than one. The results of the prunings (red, blue and green) are shown a the bottom (left,center and left respectively).

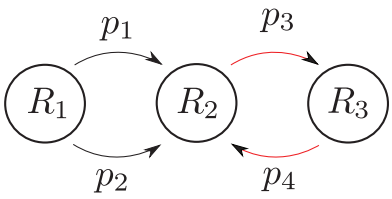
3. GLOBAL DEPTH ORDERING

After the BPT construction, only local occlusion relations have been detected. The goal now is to find a depth ordered partition D from the previously constructed BPT. To arrive at a global depth ordering, the local depth cues should propagate their depth information to relate all the regions in the image. It is possible, however, that two (sets of) local cues contradict each other, telling that two regions are at the top of each other at the same time. If that is the case, no depth ordering is possible and a conflict arises. To solve such cases, the problem of arriving at a consistent solution is stated as finding D that minimizes a cost function.

To define such a function, three concepts are proposed. First, the algorithm should try to use as many non-conflicting cues as possible. Second, since humans can decompose an image in few depth planes, D should be simple, with few regions. Third, no region should be isolated, i.e. all regions should be related with the others by at least one occlusion relation. With these concepts, the following cost function is stated:

$$C(D) = \sum_{i\in R} p_i + \gamma_N \times N + \gamma_u \times U \tag{3}$$

Where R is the set of rejected depth cues (T-junctions and convexity relations) due to depth conflicts. p_i can act as the confidence of a T-junction or as the confidence of a convexity relation between two region boundaries. U is the number of isolated regions, that is, regions which do not have any depth relationship with any other in the final depth partition. Finally, N stands for the number of regions composing the final depth image. $\gamma_u \approx 2$ is set high enough to discourage isolated regions. $\gamma_N = \max(0.1, p_{min})$ is the penalization



ρ_{ij}	R_1	R_2	R_3
R_1	-	$p_1 + p_2 - p_1p_2$	$(p_1 + p_2 - p_1p_2)p_3$
R_2	0	-	p_3
R_3	0	p_4	-

Fig. 4. Example of a DOG. Nodes and edges represent regions and depth cues respectively. The table shows the PoP for each pair of nodes. Red edges form a cycle.

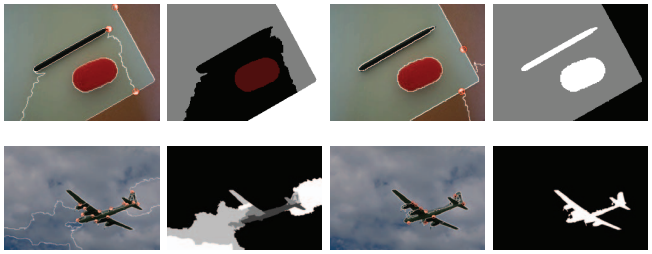


Fig. 5. Two first columns: results from [4]. Two last columns: proposed system results. In the original images, boundaries are overlaid and the junctions are marked in red. The closer region in junctions is filled with white if the T-junction is *normal*, or with black if it is *inverted*. In the depth maps, white regions are close to the viewer, black are further away. The red regions have no related depth cues with their neighbors.

factor for scenes with many regions, where p_{min} is the minimum confidence value of all the detected depth cues.

The minimization begins with a tree B_0 and iteratively seeks for simpler trees greedily, Fig. 3. B_0 is obtained by pruning the tree at nodes where the most confident T-junctions appear. The confidence p_i of these T-junctions satisfy $p_i > 0.1$ and $p_i > 0.2p_{max}$, where p_{max} is the maximum T-junction confidence. To obtain B_0, the BPT is pruned at the regions forming these T-junctions.

At each iteration t, for each tree B_t, a set of K feasible solutions B_t^k, $k = 1..K$, are generated by considering all the possible prunings that reduce the number of leaves in B_t by exactly one. In the example of Fig. 3, three such prunings are possible. Since the leaves of each pruned BPT define a partition, the depth ordered partition D_t^k is obtained for each B_t^k. With all D_t^k available, the next tree B_{t+1} is the tree corresponding to the partition of minimum cost $D_{t+1} = \arg\min_{D_t^k} \left(C(D_t^0), C(D_t^1) \ldots, C(D_t^K)\right)$. The pruning process is applied successively, obtaining at each iteration B_t and D_t, $t = 1..T$. At the final iteration T, the tree has only one leaf and cannot be further pruned. The final depth ordered partition is $D_{min} = \arg\min_{D_t} \left(C(D_0), C(D_1) \ldots, C(D_T)\right)$.

As can be seen in the previous minimization procedure, a depth ordered partition has to be generated from each pruned tree B. To this end, local depth cues should propagate their depth information through regions. Since conflicts may appear, a probabilistic scheme is proposed.

3.1. Probabilistic Framework for Depth Ordering

Since the initially computed cues are merely local, a global reasoning should be done to arrive at a consistent solution for the whole image. To do so, a Depth Order Graph (DOG) is constructed. Nodes in the graph represent regions of the partition extracted from the leaves of the BPT. The depth relations are represented in the DOG by directed weighted edges, going from the foreground region to the background one. There is exactly one edge going from region R_1 to R_2 if there is a depth cue i (T-junction or convexity) telling that R_1 is in front of R_2. The weight of this edge is the cue confidence, p_i.

To order the regions according to depth, the DOG should be acyclic (with no conflicts). To achieve such a graph structure, the DOG can be seen as a network of reliable links [8]. Each edge in the DOG associated with a cue i is reliable with probability p_i. A region R_j is reachable from R_i if there exists at least one directed path that goes from the former region to the latter. The probability of existence of this path ρ_{ij} is defined as reliability in [8], and referred in this work as probability of precedence (PoP) due to its nature. That is, the PoP ρ_{ij} is the probability of a region R_i to be foreground with respect to R_j.

The probability ρ_{ij} can be calculated exactly by the inclusion-exclusion principle [8], Fig. 4. Nevertheless, its computation cost encourages to find approximate solutions. Since the exact value of ρ_{ij} is not the ultimate goal of the conflict resolution step, an upper bound proved to give reasonable results. The computation is performed using a modification of the classic Floyd-Warshall algorithm [9]. When all ρ_{ij} are computed, the existence of a conflict (or a cycle) in the DOG is straightforward. If for some R_i, R_j, both $\rho_{ij}, \rho_{ji} \geq 0$ a cycle is present. In such case, some depth cues should be modified/discarded.

The proposed approach aims to break low-confident depth relations. Assuming $\rho_{ij} < \rho_{ji}$, the modified cue is the one corresponding to the edge with lowest confidence that goes from R_i to R_j. Two different cases appear. First, if the edge represents a convexity depth cue, the cue is discarded and the corresponding edge removed. Second, if the edge nature comes from a T-junction, a slightly different approach is performed. Since the occlusion relation is not clear, the edge is first reverted, thus changing a *normal* T-junction to an *inverted* one. If it still creates a conflict, it is discarded.

This process is repeated until no cycles in the DOG are found. When an acyclic graph is available, the depth order of each region can be computed using a topological partial sort to obtain the depth ordered partition D.

4. RESULTS AND CONCLUSIONS

4.1. Improvements of the System

Two different sources of improvements can be found in the proposed system with respect to the work in [4]. First, the introduction of more depth cues (convexity) may help the system to find depth relations where T-junctions are not present. For example, in the first row of Fig. 5, T-junctions are not able to relate the pen and the box with their background. Nevertheless, convexity cues help to determine the correct depth relations of the objects with their background.

The second source of improvement is due to the global reasoning. A good example of such behavior is shown in the second row of Fig 5. While there are many conflicting cues (there are many T-

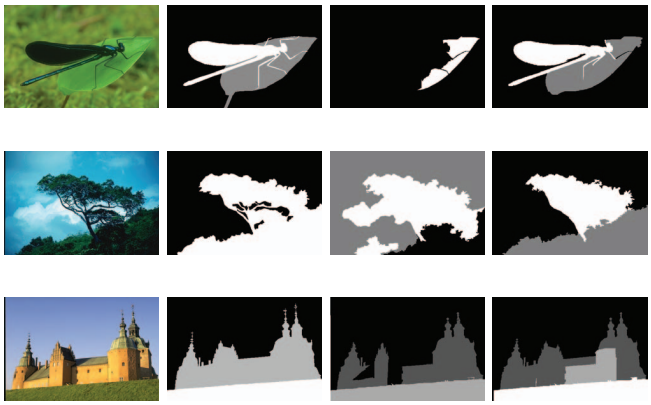


Fig. 6. From left to right. Original image, possible depth ordering from a ground truth segmentation, results of the system in [4] and results of the proposed system.

junctions telling that the sky is the front region), a global reasoning achieves a correct depth interpretation by *inverting* all the conflicting junctions.

As a more qualitative comparisons, Fig. 6 and 7 show the comparison with the previous systems [4] and [3]. It can be seen that the systems perform similarly but due to the inclusion of more depth cues, the proposed shows more consistent results than [4]. Moreover, the integration of T-junction estimation and segmentation may lead to slight better segmentations, as can be seen qualitatively in Fig. 7.

4.2. Comparison with the State of the Art

Qualitatively comparison can be performed against other low level cue based algorithms, such as [3]. Comparisons with learning-based approaches were already presented in [4], and differences on the algorithm behavior were clearly stated.

Results are also compared with figure/ground labeling algorithms, [10, 11] using the dataset [12]. Figure/ground labels are assigned at occlusion boundaries, where two depth planes meet. The closer side is assigned the figure label, while the further side is considered the ground. The comparison is performed using automatic segmentations, although the proposed system can be modified to work also with segmentations generated by humans.

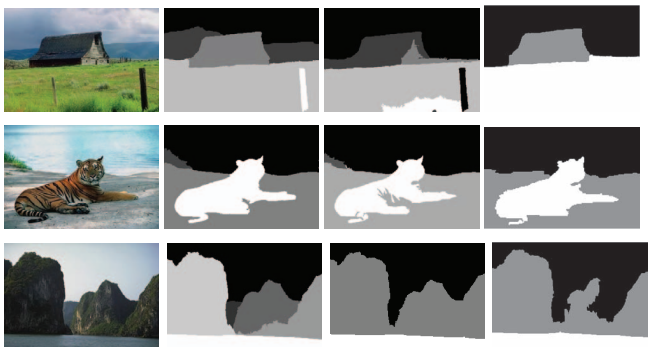


Fig. 7. From left to right. Original image, possible depth ordering, results of the proposed system and results of the region based approach in [3]

Algorithm	Proposed	[10]	[11]
Results	71.3%	68.9%	69.1%

Table 1. Percentage of correct assigned figure/ground labeling on automatically generated contour points.

To evaluate the ground truth labels on automatic generated contours, detected contour pixels are matched to a ground truth contour point having the same orientation. Human marked and automatic occlusion boundaries may not coincide, so unmatched pixels are ignored. Results in Table 1 count the number of correct matches.

Similar performance with both systems is observed even though [11] makes use of more depth cues. Moreover, both [10, 11] rely on a priori training and they do not provide region information, only labels at detected edges. The proposed approach, however, provides a full ordered depth partition which can be a good starting point for other segmentation or even 3D visualization applications.

5. REFERENCES

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

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from an image," *International Journal of Computer Vision*, pp. 328–346, 2011.

[3] M. Dimiccoli, *Monocular Depth Estimation for Image Segmentation and Filtering*, Ph.D. thesis, Universitat Politecnica de Catalunya, 2009.

[4] G. Palou and P. Salembier, "Occlusion-based depth ordering on monocular images with binary partition tree," in *IEEE ICASSP'11*, Prague, Czech Rep., May 2011.

[5] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp. 561 –576, apr 2000.

[6] J. McDermott, "Psychophysics with junctions in real images," *Perception*, vol. 33, no. 9, pp. 1101–1127, 2004.

[7] C. C Fowlkes, D. R Martin, and J. Malik, "Local figure-ground cues are valid for natural images.," *J Vis*, vol. 7, no. 8, pp. 2, 2007.

[8] R. Terruggia, *Reliability Analysis of Probabilistic Networks*, Ph.D. thesis, Universita degli Studi di Torino, 2010.

[9] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to Algorithms*, The MIT Press, 2nd revised edition edition, Sept. 2001.

[10] X. Ren, C. Fowlkes, and J. Malik, "Figure/ground assignment in natural images," in *ECCV*. 2006, pp. 614–627, Springer.

[11] I. Leichter and M. Lindenbaum, "Boundary ownership by lifting to 2.1d," in *IEEE ICCV'09*, 2009, pp. 9 –16.

[12] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," Tech. Rep. UCB/EECS-2010-17, EECS Department, University of California, Berkeley, Feb 2010.