From Occlusion to Global Depth Order, a Monocular Approach

Babak Rezaeirowshan, Coloma Ballester, and Gloria Haro(✉)

Department of Information and Communication Technologies,
Universitat Pompeu Fabra, Barcelona, Spain
babak.re.r@gmail.com, {coloma.ballester,gloria.haro}@upf.edu

Abstract. Estimating 3D structure of the scene from a single image remains a challenging problem in computer vision. This paper proposes a novel approach to obtain a global depth order of objects by incorporating monocular perceptual cues such as T-junctions and object boundary convexity, which are local indicators of occlusions, together with physical cues, namely ground contact points. The proposed combination of these local cues complement each other and creates a more thorough partial depth order relationship. The different partial orders are then robustly aggregated using a Markov random chain approximation to obtain the most plausible global depth order. Experiments show that the proposed method excels in comparison to state of the art methods.

Keywords: Monocular depth · Ordinal depth · Depth layering · Occlusion reasoning · Convexity · T-junctions · Boundary ownership · 2.1D

1 Introduction

Depth perception in humans enables a robust 3D vision even in the presence of a single view stimulus. Such a system is desirable in computer vision mainly due to its many applications and the abundance of monocular cameras. Human vision harnesses monocular cues to resolve inherent ambiguity caused by 3D to 2D projection in the image formation process and creates a sensible 3D perception. Monocular depth perception cues consist of dynamic cues and static cues. Dynamic cues, such as motion occlusion and motion parallax require multiple frames and motions in the scene as stimuli which are out of scope of this work. In this proposal, the focus is on static cues, namely, convexity and T-junctions; other cues in this category are perspective, relative dimensions, lighting and shadow.

While physiological aspects of these cues have been widely studied in the literature of psychophysics and vision, there is only a handful of research works that test these theories in a practical scenario using computer vision methods. Most of the work related to depth estimation in computer vision focuses on stereo disparity or motion parallax, both of which use triangulation to compute depth.

J. Braz et al. (Eds.): VISIGRAPP 2016, CCIS 693, pp. 575–592, 2017.
DOI: 10.1007/978-3-319-64870-5_28

While triangulation-based methods provide absolute depth, which is desirable in many applications, they require two or more views. Monocular static cues on the other hand, can be combined to create a depth perception in the absence of binocular and dynamic monocular cues or as a complement to improve existing depth perception in a much wider domain.

The goal of this paper is, given a single image from an uncalibrated camera and its decomposition in shapes (that are assumed to represent the projection of the 3D objects on the image plane; e.g., a segmentation), to create a globally consistent depth order of these shapes that constitute the image scene. For this purpose, occlusion cues between objects, namely T-junctions and convexities are used. Additionally, we use physical cues such as ground contact points. Following the underlying assumption for extracting depth from occlusion cues, we assume that the image is composed of objects that are fronto-parallel to the camera. This is also referred to as the *dead leaves model*, a term coined by Matheron [1], which constitutes a model for image formation where the image is made by objects falling on top of each others and partially occluding them. The reason for making such assumption is that in the presence of non fronto-parallel objects in the image, e.g. floor, occlusion does not translate to depth order (see Fig. 1).

Fig. 1. Dead leaves model (DLM) and correctness of convexity cues. The left image follows the DLM while the right one doesn't. Arrows indicate the occluding object suggested by convexity cues. Bright arrows indicate a correct depth order while dark arrows indicate a wrong one.

Given an image that satisfies the dead leaves model, the occlusion cues provide a depth order among neighbouring regions. However, we require a global order to establish a rough 3D model of the scene, which is understood here as obtaining a consistent global order from a number of partial orders, which may contain some discrepancies. This problem is in general referred to as rank aggregation and it has been dealt with in several fields of computer science [2,3]. This ordering problem appears whenever there are multiple operators providing partial orders with transitive relations. The goal is to use the transitivity to obtain a global robust order as consistent as possible with the partial orders. Transitivity between orders can be stated as the following property: if we have a partial order indicating $\mathrm{A} < \mathrm{B}$ and another one indicating $\mathrm{B} < \mathrm{C}$, thus we can infer the global order A < B < C. Our approach stems from the fact that transitivity of local orders can be utilized to obtain a global order using rank aggregation.

Our main contributions in this paper are (i) a depth ordering system based on monocular perceptual cues that allows reasoning without need for camera calibration, multiple frames, or motion, (ii) a novel general convexity cue detector that assigns a local depth order based on convexity and which is based on the convex hull of a shape, and (iii) the extraction of a global depth order by a robust integration of the partial orders.

2 Related Work

3D modeling has received a significant attention from the computer vision community, with studies focusing on various aspects of 3D perception. Due to the vastness of the literature in this field, we will focus on studies conducted on monocular static cues. Computational methods for depth extraction from a single image can be categorized into supervised methods and Gestalt-based methods. Alternatively, other approaches have been suggested in the literature that use human perception and vision as the basis from which to attempt to infer a computational model simulating the known processes of human vision. Our work falls in the latter category. Thus, we focus on the use of T-junctions and convexity cues for establishing a depth order. The role of T-junctions as a cue for recovering surface occlusion geometry was introduced by [4], and later stressed by [5,6]. Moreover, through the Gestalt school of thought in psychology, T-junctions were described as a basis of monocular depth perception by the work of Kanizsa [7]. Later on, more computational works demonstrated the capability of T-junctions for depth estimation; to the best of our knowledge, one of the first attempts at depth ordering methods using T-junctions was performed by [8]. Later on, an inspiring work of Nitzberg et al. [9,10] proposed the so-called 2.1D sketch through a joint segmentation and depth estimation model. More recently, studies have been conducted using energy minimization approaches which use either explicit [11,12] or implicit [13] junction detection algorithms.

In addition to T-junctions, convexity is considered to be one of the most dominant cues for figure-ground organization [7]. A computational model for utilizing convexity has been developed for figure-ground organization in the recent past [14]. Moreover, works on occlusion reasoning using Gestalt-based methods have used convexity as a complementary cue to T-junctions for a more robust relative depth estimation [15–17]. While it has been suggested that convexity affects human depth perception and is coded explicitly in the brain [18], the literature in computational models that use convexity is divided in this sense. In the works [15,17], convexity is explicitly detected and coded, while in [16] this is done implicitly. The proposed approach shares with [15–17] the use of convexity and T-junctions cues. In order to integrate the partial depth orders suggested by the monocular depth cues we use a graph-based approach. Previous works [17,19] also use a graph representation but need to reduce it to an acyclic graph and remove conflicts among different cues. In contrast, our work can directly handle conflicting transitive orders in the graph by using a rank-aggregation-based method [3], and obtain a globally consistent depth order. Here, transitive order is the order established by a path in the graph involving more than two nodes using the transitivity property mentioned in Sect. 1. A very recent work on depth layering using occlusion cues is the work of [20] where convexity, T-junctions and a ground contact cue is used to obtain a depth order of the image. An energy minimization scheme is used to find the correct depth order which makes their method more complex and time consuming than our proposed method. Moreover, they have to make more restrictive assumptions to obtain the correct ground contact cue which limits their method to a smaller domain. As the method proposed by [20] shows promising results and performs superior to other similar methods [17,21], it has been used as a benchmark for evaluation of our proposed method. A comparative evaluation using the experimental setup in [20] is presented in Sect. 4.

3 Proposed Method

We propose a method to extract a global depth order from a single image from an uncalibrated camera. The idea is motivated by studies showing human vision capability to integrate monocular depth cues to create a sensible depth perception. Given an input image, let us consider the set of its (segmented) shapes - the notion of shape used in this paper will be clarified in Sect. 3.1. Then, a global depth order can be obtained following the steps below:

1. Determine a local depth order between each pair of adjacent shapes by analysing the convexity of their common boundaries.
2. Detect T-junctions and use a multi-scale feature to determine a local depth order between the shapes that meet at each T-junction.
3. Establish a global depth order by rank aggregation of the previous partial local orders.
4. Refine the order using ground contact cue.

Each step of the proposed method is detailed in the following sections. Figure 2 illustrates the different steps of the algorithm.

3.1 Local Depth Cues Detection

Local depth cues are extracted to establish a local depth order between neighbouring objects. In this work, convexities (L-junctions) and T-junctions are used for this purpose. We use a segmentation of the image as an input to the cue detection mechanism. In order to compute a local depth order in a manner that follows the human perception based on psychophysics studies [7,18,22], T-junctions and convexities must be treated in a different manner. Thus, an explicit detection of such depth cues is required. In the following, we explain how we detect both kind of junctions.

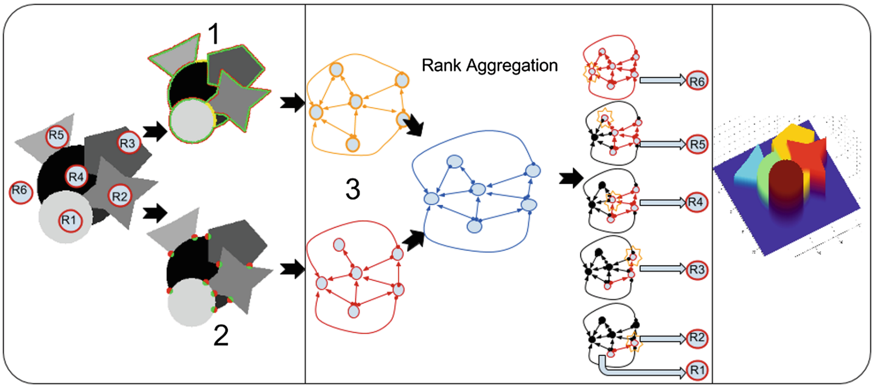

Fig. 2. Diagram of the proposed method. From left to right: The segmented image; detected convexities cues (above) and T-junctions cues (below) and the local depth order from each cue inferred by the local cues (green areas are in front of red ones, whereas yellow indicates an inconclusive cue); global depth order extraction by rank aggregation on a graph whose nodes represent the different shapes and the directed edges indicate local depth orders; final result with global depth order illustrated as a depth map, where warmer color values indicate closer objects to the camera. (Color figure online)

Convexity Cue. In this paper we propose a global convexity decision about each connected boundary between any two adjacent (segmented) regions in the image. The aim of this step is to determine which side of the boundary is the occluder and which side is the occluded, thus establishing a local depth order. Given the dead leaves model assumption, this cue can be used to infer the local depth order of the shapes that share a boundary.

To find the occluding region, we propose a method to determine which side of the boundary is closer to a convex shape. Figure 3 illustrates this process. Initially the segmented image is used to obtain the set of all the common boundaries

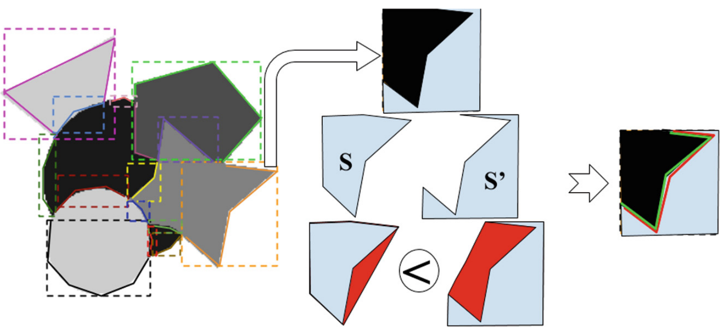

Fig. 3. Illustration of the main steps of the convexity cue detector and the estimated local order where green areas are estimated to be in front of red areas. (Color figure online)

between any two regions or objects in the image (Fig. 3, left image). For each connected common boundary, we consider its bounding box (shown in Fig. 3, middle-up). A connected common boundary divides the bounding box into two shapes (denoted by S and S' in Fig. 3). The shape whose area is closer to the area of its convex hull (i.e. smaller red area in Fig. 3, middle-down) is considered more convex and assigned as the occluder (S in the example of Fig. 3). On the other hand, the complement shape (S' in Fig. 3) is assigned as the occluded.

Let us notice that there is the possibility that a given boundary does not provide a conclusive depth cue. In other words, the convexity cue does not provide enough information to clarify which side is the occluder and which side is the occluded. This phenomenon appears, for instance, when the common boundary is either a straight line or a sinusoidal curve. To deal with such cases we define a criteria based on a threshold on our proposed global convexity measure of the connected boundary between two adjacent regions. This criteria is derived from the absolute difference between the convexity defect areas (red areas in Fig. 3) of the shapes (S and S'). If this value is not significant enough (i.e. it is lower than a prescribed threshold thr_{CX}) then these boundaries are considered inconclusive and will have no effect on the result. We define this threshold as $thr_{CX} = L \cdot \pi \cdot thr$, where L is the length of the boundary and thr is a tuning parameter that controls the sensitivity of the criteria and is independent of the length of the boundary. Examples of such inconclusive boundaries for different values of thr can be found in Fig. 4; namely, the figure displays examples for a smaller value of $thr = 0.0$ and a bigger value of $thr = 0.6$. In order to study the effect of this parameter, both on the local and global depth ordering, we present in Sect. 4 some experiments where the threshold thr is modified in the range of $[0.0, 0.6]$ with step size of 0.05.

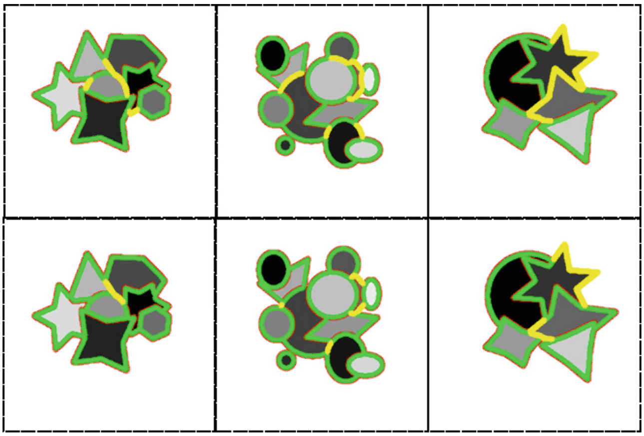

Fig. 4. Illustration of modifying the value of thr_{CX} through the parameter thr. Top row, $thr = 0.5$; bottom row, $thr = 0.15$. Decreasing thr leads to accepting more conclusive boundaries (less inconclusive boundaries in yellow). (Color figure online)

T-Junctions Cue. One of the pivotal depth cues used in this paper are T-junctions. T-junctions appear at the meeting points of three shapes boundaries and are related to occlusion configurations (see Fig. 2). Two of the three regions present in the T-junction are separated by the stem of the T; these two regions are perceived to be partly occluded by the region which presents a larger section or angle. The latter region is then in front of the other two. Moreover, the angle of each object forming the junction must satisfy some criteria to be classified as a T-junction.

In this paper, we compute T-junctions using the method in [23] where the authors gave a definition of T-junction which overcomes the difficulty of computing angles in a discrete image. They proposed an efficient algorithm which is mainly based on thresholding and computes junctions directly on the image without previous preprocessing or smoothing. The segmented image is used as an input to this method and the output is the locations of T-junctions.

The definition is based on the topographic map of an image $u : \Omega \in \mathbb{R}^2 \to \mathbb{R}$ (in our case, the segmented image), that is, the family of the connected components of the so-called, level sets of u, $[u \geq \lambda] := \{x \in \Omega : u(x) \geq \lambda\}$, and on its boundaries, the so-called level lines. Here, λ represents the gray level of the segmented image u. The set of level sets is invariant to monotonic non-decreasing illumination changes, a classical requirement in image processing and computer vision [24], and the level lines contain the boundaries of the parts of the physical objects projected on the image plane. In practice, the algorithm computes the T-junctions as all the pixels p where two level lines meet and such that the area of the connected component of each of the bi-level sets $[u \leq \alpha]$, $[\alpha < u < \beta]$, $[u \geq \beta]$, with $\alpha < \beta$, meeting at p is big enough.

After detecting the location of T-junctions, for establishing a local depth ordering one could use some angle or area of the regions meeting at the T-junction, both of which have been used in the literature [17,25]. Problems arise when certain configurations of the cue lead to an inaccurate computation. One of the problems is related to the scale at which the depth cue is obtained.

Noise in the image can also lead to incorrect cues, so one could use larger scales but they are less discriminative in depth. To avoid these issues, we stem from the work by [16] to create a reliable multi-scale measure to establish a local depth order (according to human vision) at the located cues. To this end, features are formulated using the curvature of the level lines of the distance function of each connected component in the segmented image at different scales. The features are computed for each scale s by adding the contribution from each connected component using the following formula:

$$E_s(x) = \sum_{c=1}^{nc} (e^{\beta_s |K_{c,s}(x)|^{\gamma_s}} - 1), \tag{1}$$

where $K_{c,s}$ is the curvature of the level lines of the distance function to the connected component c at scale s, nc is the number of connected components at scale s, γ_s and β_s are scale-related parameters which are fixed as proposed in [16]. In order to keep these features local and avoid overlapping with other boundaries, the distance function is clipped at a distance 5. In order to generate a multi-scale local feature we combine the local features according to (1) by computing an average of the normalized features at several scales, as in [16]. In this work, we integrate the features from scales 1 to 5. Figure 5 illustrates with an example the behaviour of this multi-scale features. As it can be seen in Fig. 5 right, the part of the cue that is perceptually closer to the observer has a higher multi-scale feature value.

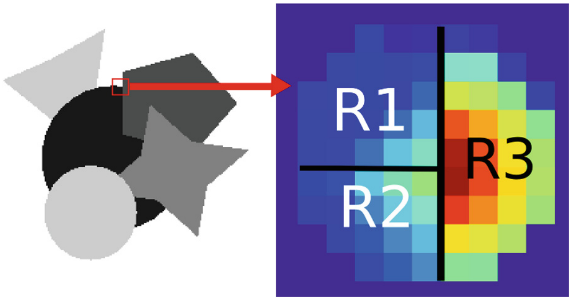

Fig. 5. Multi-scale features obtained after averaging the features E_s (1) of the first five scales.

Finally, to estimate the local depth order induced by a certain T-junction, first a representative value of the multi-scale depth features is computed for each region (e.g. R1, R2, and R3 in Fig. 5 right) in the neighborhood of the T-junction given by a disk of radius 5. The representative value is computed by applying either the median or max operators on the features of the respective region (i.e. R1, R2, R3). In Sect. 4 we compare the performance of both operators. The region with a higher representative local feature value is assigned to be in front of the other two neighboring regions (R3 in front of R1 and R2 in the example of Fig. 5).

Ground Contact Cue. The physical restriction of the real world imposes that every non-flying, non-wall-mounted, object should be connected to the ground. A reasonable extension of this phenomena suggests that every object in the scene is either directly connected to the ground or is occluded by other objects, and at least one of which is connected to the ground. Making reasonable assumption that the ground plane can be estimated, or manually segmented, the lowest point of contact of each object to the ground can be easily used as an extra cue for depth order. As ground contact cues can establish an order between non-adjacent objects they can complement the occlusion relations computed from T-junction and convexity. This kind of cue was previously used in [20].

3.2 From Local to Global Depth

In order to establish a global depth order given by the local cues we use an approximation of rank aggregation [2] similar to the one used in [3] for photo-sequencing. To do so, we construct a weighted graph $G(U, E)$ to represent the partial order between pairs of shapes (objects), which are represented in the graph by the nodes in U. The graph is constructed by placing a directional edge $e(i, j) \in E$ connecting the node i to node j if the local cues relating the objects suggest that object i is in front of object j (represented here by $i >> j$). The weight of the edge gathers up the local depth order cues. Each convexity cue indicates a depth order relation between two nodes (e.g. $i << j$) and each T-junctions indicates a relation between three nodes using two edges (e.g. $i << j$, $i << k$). The weight of the edge $e(i, j)$ between nodes i and j is proportional to the number of local cues indicating the local order $i >> j$, which can be interpreted as proportional to the number of votes for the local order $i >> j$. This weight corresponds to the probability that $i >> j$. In such a graph, a random walk after a sufficient time (in the steady state) will reach the sink of the graph (or of a sub-graph) which represents the object (or objects) perceptually furthest from the viewer. Repeating this process iteratively while in each iteration removing the sink node (or nodes) from the previous iteration will provide us with the global depth order. In particular the iteration number in which a set of nodes is removed reveals the global order of this set of nodes. For illustration of this process see Fig. 2 – step 3.

The steady state can be computed using an eigenvector analysis of M, the transition state matrix associated to the graph. The elements of M are the probabilities of moving from one state (node) to another. To construct the matrix M with non-negative entries, we initially form a matrix V collecting the votes, where the rows and columns indices correspond to the index of each associated connected component. Thus, an image with N shapes will produce an $N \times N$ matrix V. The i, j-th element of matrix V, $V(i, j)$, collects the number of votes (local cues) that agree with the partial ordering $i >> j$.

Once the matrix V is filled, we compute the matrix M which specifies the probability that $i >> j$. Firstly, the cycles of length two which may have been introduced by conflicting cues are removed. We follow the method proposed in [3] to resolve these conflicts. In particular, $M(i, j) = 1 - \frac{V(j,i)}{V(i,j)}$, and $M(j, i) = 0$ if $V(i, j) > V(j, i)$. The rest of the cycles do not need to be removed since the rank aggregation method automatically solves them. Finally, the rows of M are normalized to 1 in order to get transition probabilities.

After an initial depth order is computed from occlusion cues, it is possible to refine the depth for those segments which are assigned the same depth level due to lack of occlusion relations. To this end, a ground contact cue is used. For each set of objects placed at the same depth, the lowest ground contact point is computed as the point with lowest y-coordinate of common boundary between objects and the ground plane (which has been pre-segmented). Subsequently, the order is refined so that objects whose ground contact is closer to the lower border of the image are assigned a higher (i.e. closer) depth level.

4 Experimental Results

This section presents three different experiments with different kind of data designed to evaluate and illustrate various aspects of the proposed method. An initial experiment is first presented as a proof of concept using synthetic images with the following parameters: $thr = 0.15$ for convexity cue detection, and median as T-junction feature operator. The goal of the second experiment is twofold: first, to present an experimental study of different parameter settings to find the best performance and fix the parameter values for the rest of the experiments and, second, to provide a quantitative comparison of the proposed method and the most recent state-of-the-art methods [17,20,21]. This experiment is done using a dataset of 52 images proposed by [20]. For both the first and second experiments the ground truth segmentation is available, whereas in the third experiment the segmentation is done using an interactive tool [26].

Figure 6 illustrates the results of applying the proposed method to a small set of synthetic images. The first row shows the input images and the second row shows the global depth order images with convexity and T-junctions cues superimposed on them, respectively. The local depth order is illustrated in each cue, where green indicates the section perceived to be closer to the observer. As for global depth order, the grey values indicate global depth order, particularly the brighter areas are closer to the observer. As it can be seen all T-junction cues indicate a correct local depth order, whereas some of the convexity cues are incorrect or inconclusive (marked as yellow). However, the T-junctions cues are able to compensate these errors and create a globally consistent depth order that complies with human depth perception.

In the first part of the second experiment the proposed method is evaluated under different parameter settings with the dataset proposed by [20]. Figures 7 and 8 illustrate these results. The horizontal axis denotes the parameter thr

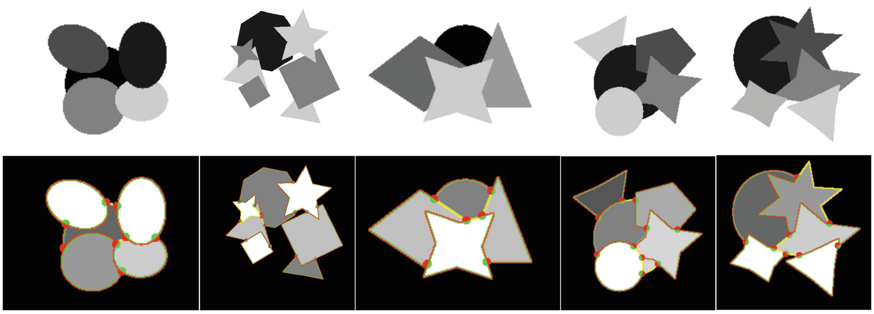

Fig. 6. Experiments with synthetic images: estimated global depth ordering (brighter gray levels indicate closer objects). The automatically detected local depth cues, convexities and T-junctions establish a local depth order (green areas are estimated to be in front of red areas). Inconclusive convexity cues are marked in yellow. (Color figure online)

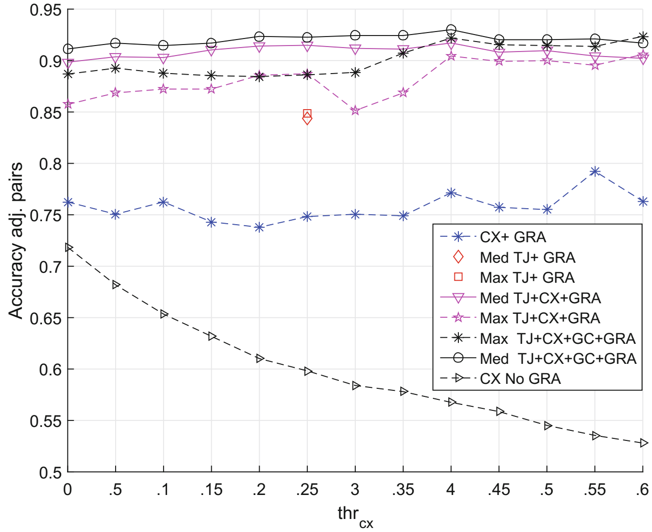

Fig. 7. Accuracy of local depth order between adjacent pairs of shapes.

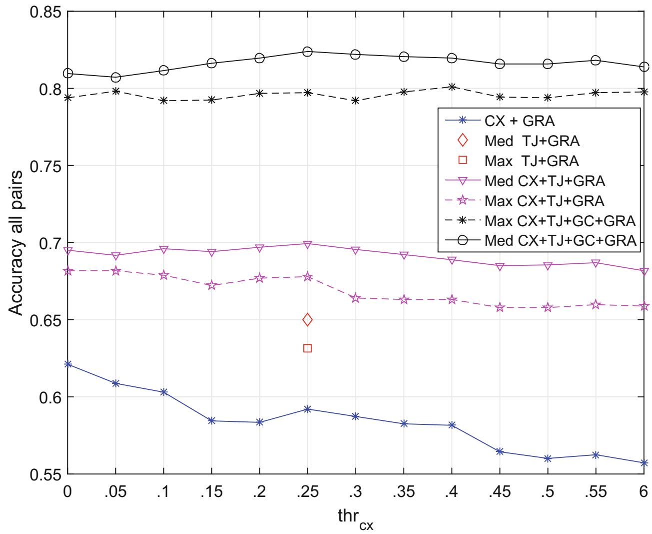

Fig. 8. Accuracy of global depth order between all pairs of shapes.

that defines the threshold $thr_{CX} = L \cdot \pi \cdot thr$ applied to the difference of defect areas. Then, as the values in the horizontal axis increase the threshold thr_{CX} increases and more boundaries become inconclusive, meaning that the sensitivity for detecting global convex boundaries decreases (see also Fig. 4). In Fig. 7 the vertical axis indicates the accuracy as the percentage of pairs of adjacent shapes which have been assigned a correct local depth order. Whereas in Fig. 8 the vertical axis indicates the accuracy as the percentage of pairs of all shapes which have been assigned a correct global depth order. These accuracy measures are identical to the measures of performance evaluation in [20]. The legend of Figs. 7 and 8 indicate the operator for the T-junction (*median* or *max*), the type of local cue (T-junction (TJ), convexity (CX), ground contact cue (GC) or combination of these), and whether or not a global rank aggregation was used ("GRA" or "$No\,GRA$"). Both Figs. 7 and 8 indicate that the best performance is achieved when the depth order induced by the T-junction cues is computed using the *median* operator and is combined with the depth order induced by the convexity cue using rank aggregation, denoted as "$Med\,TJ{+}CX{+}GRA$". Thus, achieving a performance of 91.49% accuracy in local depth order estimation and 69.94% in global depth order estimation. On the other hand, using the *max* operator slightly decreases the performance to 89% and 67.8% for local and global depth estimation, respectively. The decrease in accuracy of the global order with respect to the local order can be explained by the fact that the proposed method can only infer depth relations between objects connected by a path in the graph. It should also be noted that the performance of the *max* operator is slightly less stable. Further, it can be seen that the contribution of T-junctions is significant for both global and local depth estimation as they improve the performance compared to when only convexities are used (16% increase for local depth estimation and 19% increase in accuracy of global depth estimation). The red square and diamond in the Figs. 7 and 8 highlight the performance of using only T-junctions (the parameter thr does not affect this computation). As expected, T-junctions seem to be a more reliable cue than convexities as they consistently achieve a higher accuracy. Figure 7 illustrates how the global integration of convexity cues using rank aggregation improves the performance of local depth estimation between adjacent pairs of shapes, namely, the performance increases from 59% to 75%. Further, both Figs. 7 and 8 indicate that using the ground contact point to refine the depth order improves the accuracy of depth estimation, particularly among all pairs (see Fig. 8).

Finally, observing the two lower curves the in Fig. 7 we can see that, while the accuracy of "$CX + No\ GRA$" decreases as the threshold increases, the accuracy of "$CX + GRA$" remains relatively stable. This indicates that most of the convexity cues in the dataset are conclusive (i.e. comply with human depth perception) and increasing the threshold will lead to less cues and thus less accuracy. However, it is interesting to note that the global integration is able to compensate for the removal of cues that did not satisfy the threshold and stabilize the performance. It can be seen that the best operation point for the threshold of the global convexity is the mid-range value $thr = 0.25$, where the average of the two accuracy measures is the highest. While the effect of *thr* is not significant it leads to a slight increase in the performance of the global depth estimation (see Fig. 8).

In the second part of the second experiment, the proposed method is compared with the state-of-the-art [17,20,21] with the accuracy measures presented in [20]. According to the results obtained in the previous analysis, we fix the parameters to the following values: $thr = 0.25$ and *median* as the operator in the depth order estimated from the T-junctions. To this end, we follow the experimental setup suggested by [20] on their proposed depth ordering dataset. The results in Table 1 show that using a combination of T-junction, convexity and ground contact (GC) cues achieves the highest performance. As it can be seen, the proposed method, both with and without the ground contact cue, outperforms all of the state-of-the-art methods in the adjacent pairs case and, in the

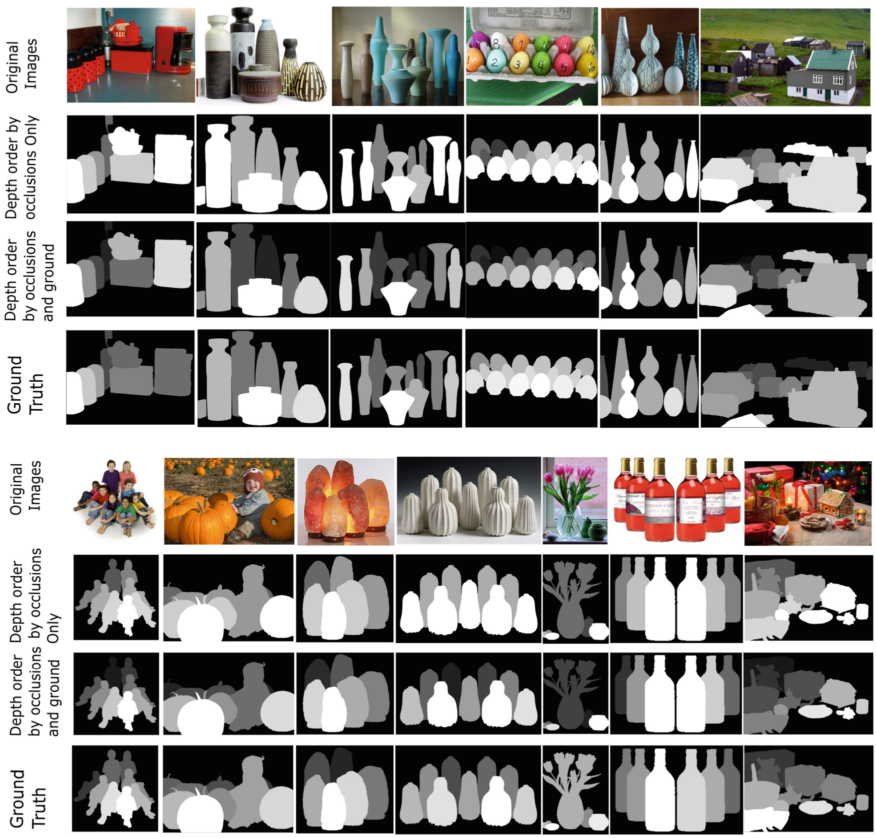

Fig. 9. Depth ordering results using the proposed method (using occlusion cues only and both occlusion and ground contact cues) on near-view scenes from the dataset by [20].

all pairs case, the proposed method performs superior to [17,21] and very close to [20]. If we do not use the ground contact cue, our methods falls short of [20] in the all pairs case. This is mainly due, as previously noticed, to the fact that our proposed method cannot infer depth relations between objects that are not connected with a path in our graph i.e. there are no transitive relations to be used to infer a global depth order. In contrast, when we add the ground contact cue, as in [20], as a post-processing of the depth order it helps to infer new depth relations when the other cues (T-junction, convexity) are not present. The improvement is illustrated in Table 1 where the accuracy increases from 69.94 to 82.59 in the all pairs case. A qualitative comparison of the best performance of the proposed method can bee seen in Fig. 9. The improvements of the depth refinement step using the ground contact point are evident in here.

Table 1. Depth order accuracy.

	Adj. pairs	All pairs
Jia et al. [21]	79.84	29.88
Palou and Salembier [17]	43.85	43.56
Zeng et al. [20]	82.66	**84.60**
Our method: CX + GRA	74.83	59.21
Our method: TJ + GRA	84.39	65.3
Our method: TJ + CX + GRA	**91.49**	69.94
Our method: TJ + CX + GRA + GC	**92.11**	**82.59**

Finally, to show how the proposed method may be used as a real world application, the interactive segmentation tool [26] has been used to segment some images from the Berkeley dataset [27] and the global depth order of the segmented objects is estimated with the proposed method. As it can be seen in Fig. 10 the order of the segmented objects is correct in most of the cases.

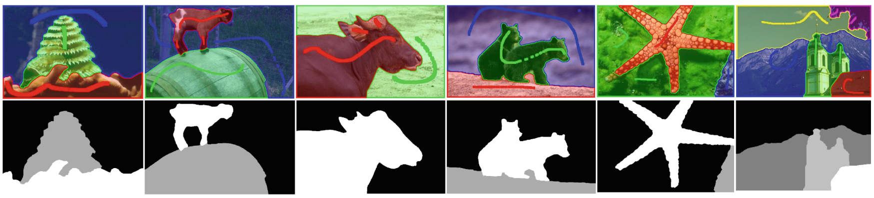

Fig. 10. Using interactive segmentation [26] and the proposed method to create a depth ordering of objects in the scene.

5 Limitations and Assumptions

Estimating depth from a single image is a very challenging and under-determined problem. It is necessary to make suitable assumptions to make the problem tractable. Our first assumption is that a good segmentation is available where the boundaries of the segmentation regions coincide with the actual object boundaries. As the method is based on a convexity cue defined on boundaries and T-junctions (which are points at the intersection of boundaries), a deficient segmentation leads to significant depth artifacts in the estimated depth order. A second limitation may be noticed in one of the examples in Fig. 11: the one in box

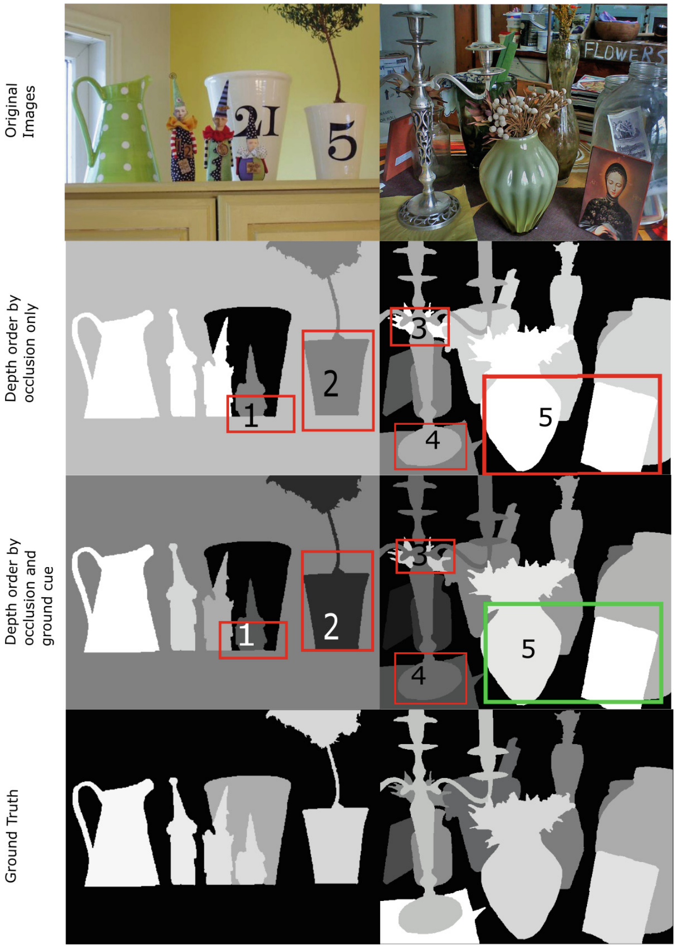

Fig. 11. Due to some limitations of the proposed approach, the violation of certain assumptions leads to errors in the estimated depth order which have been delimited with red boxes (see Sect. 5 for a detailed explanation of these problems). (Color figure online)

1 of the left image. The T-junction and convexity cues that are detected on the ground contact of the object indicate incorrect depth order. In some cases, there exist other cues that compensate for these mistakes, either directly or indirectly using the transitivity property of the graph. However, this is not the case in the aforementioned example. Another limitation is inaccuracies in our convexity detector which can be seen in Fig. 11 box 2, a misinterpretation of convexities in cases where a long narrow shape is next to two concavities. Figure 11 box 3 shows the bias of the proposed method to interpret small convex objects to be in front of their neighbouring shapes (this may happen also in visual holes, such as windows or arch bridges). A more general limitation is that objects in the scene should be approximated with fronto-parallel planes to the camera. When this assumption does not hold it may lead to misinterpretation of local cues and thus misestimation in the order of objects. An example of this can be found in Fig. 11 right, box 4. In this case, since the two objects sharing the same border cannot be approximated with fronto-parallel planes, the algorithm misestimates the depth order. Another limitation marked in the box 5, illustrating that, if the ground cue is not used or is not available, the non-adjacent objects can be placed in the same depth level.

Fortunately, in some cases there are solutions to deal with the aforementioned limitations. The problem illustrated in box 5 can be resolved using the ground contact point as shown in the third row right column of Fig. 11. The non-fronto parallel problem can be resolved by ground separation in simple cases. In cases where there are more than one non-fronto parallel planes in the image, a geometric context method, based for instance on surface normal extraction, may be used to guide the depth estimation. The problem related to visual holes can be addressed using a semantic labelling method that is able to identify the visual hole; for example by classifying areas like the sky which are always in the back.

6 Discussion and Conclusion

Inspired by the human vision capability to perceive depth using monocular cues, we proposed a method for the detection and integration of T-junction and convexity cues that is able to obtain a globally consistent depth order. The proposed method computes partial depth orders using multi-scale features, then, integrates them using a rank aggregation method that resolves conflict. This allows to simultaneously compensate for incorrect partial depth orders introduced by invalid cues and also harnesses the transitivity between the cues to obtain a global order from partial orders. The proposed method is applicable to any scene that complies with the dead leaves model and does not require training. In the presense of a segmented ground plane, the contact point with the ground can be used as an additional cue to refine the depth order estimated using occlusion cues. For future work we propose to extend the method to images containing non fronto-parallel objects using other monocular and binocular cues that may be integrated in the rank aggregation step as additional votes for partial depth orders.

Acknowledgements. The authors acknowledge partial support by the MINECO/FEDER project with reference TIN2015-70410-C2-1-R, the MICINN project with reference MTM2012-30772, and by GRC reference 2014 SGR 1301, Generalitat de Catalunya.

References

1. Matheron, G.: Modèle séquentiel de partition aléatoire. Technical report, CMM (1968)
2. Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: Proceedings of the 10th International Conference on World Wide Web, pp. 613–622. ACM (2001)
3. Basha, T., Moses, Y., Avidan, S.: Photo sequencing. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 654–667. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33783-3_47
4. Guzmán, A.: Decomposition of a visual scene into three-dimensional bodies. In: Proceeding AFIPS 1968 (Fall, part I) (1968)
5. Malik, J.: Interpreting line drawings of curved objects. Int. J. Comput. Vis. **1**(1), 73–103 (1987)
6. Rubin, N.: Figure and ground in the brain. Nat. Neurosci. **4**, 857–858 (2001)
7. Kanizsa, G.: Organization in Vision: Essays on Gestalt Perception. Praeger, New York (1979)
8. Marr, D.: Vision: A Computational Approach. Freeman & Co., San Francisco (1982)
9. Nitzberg, M., Mumford, D.: The 2.1-D sketch. In: Proceedings of Third International Conference on Computer Vision, pp. 138–144. IEEE (1990)
10. Nitzberg, M., Mumford, D., Shiota, T.: Filtering, Segmentation and Depth. LNCS, vol. 662. Springer, Heidelberg (1993)
11. Gao, R.-X., Wu, T.-F., Zhu, S.-C., Sang, N.: Bayesian inference for layer representation with mixed markov random field. In: Yuille, A.L., Zhu, S.-C., Cremers, D., Wang, Y. (eds.) EMMCVPR 2007. LNCS, vol. 4679, pp. 213–224. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74198-5_17
12. Palou, G., Salembier, P.: Occlusion-based depth ordering on monocular images with binary partition tree. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1093–1096. IEEE (2011)
13. Esedoglu, S., March, R.: Segmentation with depth but without detecting junctions. J. Math. Imaging Vis. **18**, 7–15 (2003)
14. Pao, H., Geiger, D., Rubin, N.: Measuring convexity for figure/ground separation. In: 1999 The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 948–955. IEEE (1999)
15. Dimiccoli, M., Morel, J.M., Salembier, P.: Monocular depth by nonlinear diffusion. In: Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, pp. 95–102. IEEE (2008)
16. Calderero, F., Caselles, V.: Recovering relative depth from low-level features without explicit t-junction detection and interpretation. Int. J. Comput. Vis. **104**, 38–68 (2013)
17. Palou, G., Salembier, P.: Monocular depth ordering using t-junctions and convexity occlusion cues. IEEE Trans. Image Process. **22**, 1926–1939 (2013)

18. Burge, J., Fowlkes, C., Banks, M.: Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. J. Neurosci. **30**, 7269–7280 (2010)
19. Dimiccoli, M., Salembier, P.: Hierarchical region-based representation for segmentation and filtering with depth in single images. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 3533–3536 (2009)
20. Zeng, Q., Chen, W., Wang, H., Tu, C., Cohen-Or, D., Lischinski, D., Chen, B.: Hallucinating stereoscopy from a single image. In: Computer Graphics Forum, vol. 34, pp. 1–12. Wiley Online Library (2015)
21. Jia, Z., Gallagher, A., Chang, Y., Chen, T.: A learning-based framework for depth ordering. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 294–301. IEEE (2012)
22. McDermott, J.: Psychophysics with junctions in real images. Perception **33**, 1101–1127 (2004)
23. Caselles, V., Coll, B., Morel, J.: Topographic maps and local contrast changes in natural images. Int. J. Comput. Vis. **33**, 5–27 (1999)
24. Serra, J.: Introduction to mathematical morphology. Comput. Vis. Graph. Image Process. **35**(3), 283–305 (1986)
25. Dimiccoli, M., Salembier, P.: Exploiting t-junctions for depth segregation in single images. In: Acoustics, Speech and Signal Processing, pp. 1229–1232 (2009)
26. Santner, J., Pock, T., Bischof, H.: Interactive multi-label segmentation. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010. LNCS, vol. 6492, pp. 397–410. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19315-6_31
27. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of 8th International Conference on Computer Vision, vol. 2, pp. 416–423 (2001)