Computer Vision and Image Understanding 117 (2013) 42–55

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Occlusion cues for image scene layering ☆

Xiaowu Chen, Qing Li *, Dongyue Zhao, Qinping Zhao

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

ARTICLE INFO

Article history:
Received 4 January 2011
Accepted 2 October 2012
Available online 23 October 2012

Keywords:
Human perception
Occlusion cues
Occlusion prediction
Layering

ABSTRACT

To bring computer vision closer to human vision, we attempt to enable computer to understand the occlusion relationship in an image. In this paper, we propose five low dimensional region-based occlusion cues inspired by the human perception of occlusion. These cues are semantic cue, position cue, compactness cue, shared boundary cue and junction cue. We apply these cues to predict the region-wise occlusion relationship in an image and infer the layer sequence of the image scene. A preference function, trained with samples consisting of these cues, is defined to predict the occlusion relationship in an image. Then we put all the occlusion predictions into the layering algorithm to infer the layer sequence of the image scene.

The experiments on rural, artificial and outdoor scene datasets show the effectiveness of our method for occlusion relationship prediction and image scene layering.

1. Introduction

It is well known that an image is not a direct representation, but a projection of the 3D world. The understanding of an image scene includes not only the comprehension of the regions presented in the 2D image plane, but also the 3D spatial layout of the regions in the real world. We humans can immediately grasp the spatial relationship of the scene. Our perception includes the immediately visible portions as well as the estimation of the entire space. It can be seen from Fig. 1 (the input image is taken from the LHI dataset [1]) that our perception includes not only the visible textures of the rhinoceros, grass, ground, tree and sky, but also includes the estimation of the spatial relationship. For example, the rhinoceros standing on the ground is in front of the tree and grass. The ultimate goal of computer vision is to provide the computer with the same spatial understanding so that it can see the world as humans do.

Modern computer vision techniques, such as image segmentation, object detection and recognition and depth estimation, can be used to recover a lot of useful information about the image. Although a great progress has been made, it still remains extremely challenging for current computer vision systems to understand scenes as humans do. When the real world is projected into the image plane, occlusions will frequently occur between the objects that are spatially separated. In fact, almost every object in the image is occluded by, and (or) occludes other objects. Understanding occlusion helps us to comprehend the 3D spatial relationship. Thus, to bring computer vision closer to human vision, we need to make the computer understand the occlusion relationship. Recently there are some works paying attention to this topic such as the 2.1 D sketch [2] and the occlusion boundary detection [3,4]. They mainly focus on the occlusion boundary identification, but ignore the recognition of the global image. In addition, there is still a large gap between human perception and occlusion reasoning. How to bridge this gap is our main concern in this paper.

Previous works on this topic used a wide variety of perception features to recover the occlusion relationship. Inspired by these works, we use occlusion features to express the human perception rules. There are many low-level and high-level features used in the occlusion reasoning, figure/ground assignment and depth ordering, such as texture features, color features and gestalt cues. In psychophysics and cognition, gestalt cues, including size, convexity, symmetry, parallelism, surroundedness and lower-region, are considered to be useful and important. Inspired by human perception, we propose five low dimensional region-based occlusion cues (referred to as 'five cues' below), which are semantic cue, position cue, compactness cue, shared boundary cue and junction cue.

In this paper, we describe our five cues and show their capabilities of occlusion relationship prediction. The experimental results show that our cues are efficient, and thus demonstrate that our cues can recover occlusion relationship to some extent. We apply all the occlusion predictions in an image to infer the layer sequence of the image scene. Our layer sequence indicates the layer partition of regions, and each region is assumed to be assigned a unique layer. Fig. 1 shows the overview of our method. Since our cues are based on the regions in the image, we first preprocess an input image to get its semantic label map. Our cues can work well regardless of how to get the semantic label map. Thus semantic

☆ This paper has been recommended for acceptance by J.K. Tsotsos.

* Corresponding author.
E-mail address: liqing@vrlab.buaa.edu.cn (Q. Li).

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2012.10.001

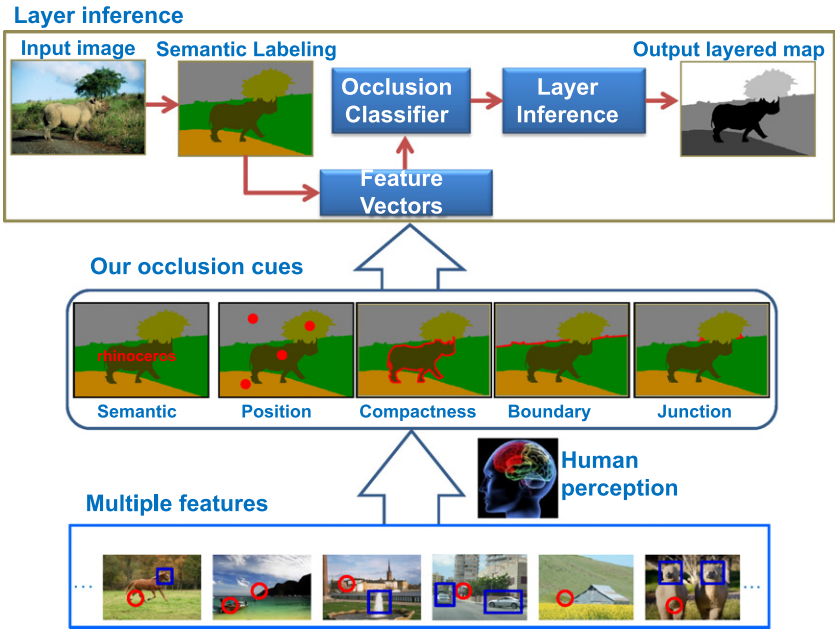

Fig. 1. The overview of our method. Inspired by the human occlusion perception, we propose five occlusion cues. We apply these cues to predict the occlusion relationship in an image and use the predictions to infer the layer sequence of the image scene. In the layer inference process, we first get the semantic label map of a given input image since our cues are based on the regions. Then we compute values of the occlusion cues to form feature vectors. Next, the trained occlusion classifier is used to predict the occlusion relationship. Finally, we put all the predictions of the input image into the layer inference algorithm and get the layer sequence which is visualized as a layer map with the blacker region in the front.

labeling work is not the main concern in this paper. In the section of experiments, we perform scene layering based on both the semantic label maps obtained manually and by semantic labeling methods, and both give promising results.

The main contributions of this paper include: (1) We propose five region-based cues to describe the occlusion relationship in an image, which are to some extent efficient to indicate the occlusion relationship in image scenes. (2) Then we analyze the interactions of these five cues as their importance varies in different datasets. (3) Finally, we give a selection scheme of these cues when they are applied to occlusion perception problem.

The remainder of this paper is organized as follows. We review the relevant literatures in Section 2. Our occlusion cues are described in Section 3. Section 4 introduces a previous semantic labeling work done by us and describes our layering inference procedure. To test the effectiveness of our cues and validate our results, three types of experiments are performed in Section 5. Experimental verification of our cues is demonstrated in Section 5.2, and followed by the analysis of the interactions between cues in Section 5.3. Scene layering results on multiple datasets are shown in Section 5.4. Section 6 is a brief conclusion.

2. Related works

The last decade has seen an increase of interest in occlusion reasoning. In computer vision, the study of occlusion reasoning has been largely confined to the context of stereo, motion and other multi-view problems [5–7]. For single-view tasks, the study of occlusion has been focused on how to recover the 3D information hidden in the 2D image plane, such as the 2.1D sketch [2], occlusion recovery and contour completion [8,9,3,10,11], image segmentation and depth recovery [4,12–16], and 3D layout recovery and modeling [17,18]. Nitzberg and Mumford [2] formulated the 2.1D sketch problem as an energy minimization problem. The goal of the 2.1D sketch is to recover the partial occluding order relation and to complete the occluded contour simultaneously. Moreover, it keeps multiple reasonable solutions accounting for the intrinsic ambiguity caused by occlusion. Wang et al. [5] proposed the concept of layered representation for image coding and motion analysis. Since then, there have been a certain number of research works on the 2.1D sketch problem.

Ren et al. [4] presented an approach for figure/ground assignment. Yang et al. [12] used the layer order to label the pixels in the image. Yu et al. [13] integrated occlusion cues with figure/ground segregation by using hierarchical Markov random field. Following the energy function defined by Nitzberg and Mumford [2], Esedoglu and March [15] proposed to segment an image with depth information but without detecting junctions. Hoiem et al. [3] proposed a method to recover occlusion boundaries by learning a CRF model. Liu et al. [16] performed a semantic segmentation of the scene and used semantic labels to guide the 3D reconstruction. Hedau et al. [17] used a parametric 3D 'box' to model the global room space, and introduced a structured learning algorithm to choose the set of parameters. Saxena et al. [18] created 3D models by using a MRF to infer a set of 'plane parameters' that can capture both the 3-D location and 3-D orientation of small homogeneous patches in an image.

In all these cases above, they either focus on the perceptual completion caused by occlusion, or on the application of occlusion reasoning such as segmentation and 3D modeling, but ignore the potential perceptual rules. We attempt to find some rules that can make computer reason occlusion relationship as humans do. Some of the above literatures have used the cues such as color and texture [3,16], surface layout [3], boundary [7,10], contour [8] and junction [19] to reason the occlusion or estimate the pixel depth in an image. Other literatures have focused on how to bridge the gap between computer vision and human perception [20–22]. They concentrate on multiple features or cues, such as shape, curvature and orientation, to obtain the region perception. According to these works, the occlusion cues may be a starting point to the addressing of our occlusion perception problem.

Fig. 2. The semantic label map. Each pixel is assigned its unique semantic label (visualized in color).

3. Region-based occlusion cues

Based on the rules of human perception, we propose five low dimensional cues for region-level occlusion reasoning and experimentally analyze the performance of our five cues comparing with other features. More experimental details are presented in Section 5.2. In this section, we describe the formulation of the five cues.

3.1. Semantic cue

In the semantic label map, each pixel of the image is assigned a semantic label from a predefined label set. As shown in Fig. 2, each semantic region is visualized in specific color.[1] Occlusions usually occur between regions of different categories, thus the semantic information can be useful for recovering occlusion relationship. For example, in rural scene, cow usually occludes the grass. Thus, the occlusion presence frequency can reveal the occlusion relationship between semantic regions in the scene. We evaluate the relative occlusion relationship between different semantic classes, and use the semantic information of regions as a cue for occlusion perception problem. Here, we add up the occlusion relationship to form the corresponding histogram in the training set. In Eq. (1), given region R_i and R_j with their semantic labels, the histogram value P is used to define the model of the semantic cue $S_{\langle R_i,R_j\rangle}$.

$$S_{\langle R_i,R_j\rangle} = P(label_{R_i}, label_{R_j}) \tag{1}$$

3.2. Position cue

The position of the region in the image scene indicates the depth of the region in the original 3D world scene. The shorter the distance between an object and the camera is, the lower position the object located at in the image. For example, the infinite sky is usually located at the top of the image, and the ground is generally at the bottom of the image. In [3], position is also considered. Inspired by this, we take position feature as an important cue for occlusion prediction. As shown in Fig. 3, the plant is located at the front, the building is at the second level, and the sky is at the final level on the top of the image. However, in the case of a overhanging tree branch, the position cue will fail. Given two regions R_i, R_j, we utilize the gravity height to define the position cue model $Pos_{\langle R_i,R_j\rangle}$ so as to characterize the figure-ground relation and the positions of regions as given by Eq. (2), where $\overline{y_i}, \overline{y_j}$ is the average height of R_i, R_j, and H is the height of the image.

$$Pos_{\langle R_i,R_j\rangle} = 1/(1+\exp(\overline{y_j} - \overline{y_i}/H)) \tag{2}$$

[1] For interpretation of color in Figs. 2 and 8–10, the reader is referred to the web version of this article.

3.3. Compactness cue

Compactness, along with other measures such as area, rectangularity and direction, is one of the region-based descriptors. The bigger the compactness of one region is, the more regular the region is. We can formulate the compactness occlusion cue by the contour compactness property. As shown in Fig. 4, regions in these images which are not occluded by others, such as sheep, car and cat, are located at the first layer, thus the contours of them are smooth and regular; the contours of other occluded regions appear to be irregular. If one region is located at the anterior layer, its contour could be regular and its area could be compact. Thus the compactness of region can be one of the cues for occlusion perception problem. According to the mathematical definition of compactness [23], our model of the compactness cue is given by:

$$Com_R = \exp\left\{-\alpha \cdot \frac{L^2}{A}\right\} \tag{3}$$

where L is the contour length of region R and A is the area of R. The weighting parameter α is set to be 0.05.

3.4. Shared boundary cue

Boundary is accepted as a useful cue for occlusion reasoning in previous works [3,4]. We take boundary as an important perceptual cue in our occlusion relationship prediction. When region R_i occludes region R_j, the shared boundary between them appears to be integrally convex towards R_i. As shown in Fig. 5, if the shared boundary turns to be more convex apparently, it is more likely that one region occludes the other. Notice that in some situations, such as the underneath of animals or bridges, the boundary cue may fail. We use the convexity of the shared boundary as one of important occlusion cues to measure the occlusion relationship between adjacent regions. The convexity of boundary can be described by the curvature mathematically. To describe the relation between the occlusion and the shared boundaries [24,25], we define the function utilizing curvature, as shown in the following equation:

$$g(\overrightarrow{L}) = 1\Big/\left(1+\exp\left(-\int_l \kappa ds/l\right)\right) \tag{4}$$

where κ is the curvature, and l is the length of the shared curve $\overrightarrow{L}$. According to the model above, we then define the model of the shared boundary cue in the following equation:

$$Bry_{\langle R_i,R_j\rangle} = \sum_{i=1}^{N} g(\overrightarrow{L}_i)\Big/N \tag{5}$$

where N is the sum of the shared curves between R_i and R_j.

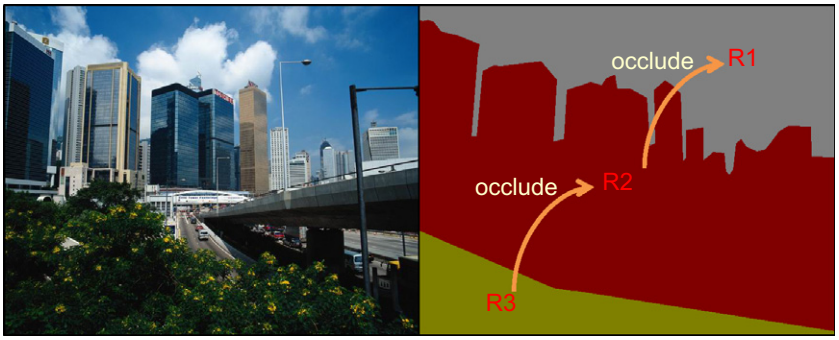

Fig. 3. Illustration of the position cue. The plant located in the lower part of the image is in front of buildings.

Fig. 4. Illustration of the compactness cue. Regions which have more regular contours may occlude other regions.

Fig. 5. Illustration of the shared boundary cue. When two regions are adjacent, the convexity of the boundary can show the occlusion to some extent.

3.5. Junction cue

The importance of junction has been emphasized by Gestalt psychologists [26] and investigated in human vision [19]. Junctions are the atoms of many complex processes or tasks, such as depth and motion estimation, segmentation and recognition. This feature can provide useful local information about geometric properties and occlusions. As shown in Fig. 6, when three regions are adjacent and occluded, one region generally occlude the other two, which results in junction. As junction is demonstrated in a variety of patterns, we need to extract the appropriate variables to describe it accurately.

To describe the junction cue [27,28], we simplify the curves of junction into line vectors, and use the angles between lines as the variables. As shown in Fig. 6, the short curves and center of the junction is detected, and then the short curves are transformed into vectors according to the average position of the curves. Finally we compute the angles between the vectors and use the three angle values to describe the junction $J_t = (\theta_1, \theta_2, \theta_3)$. For regions R_i, R_j, R_k with junctions, the model of junction cue is given by Eq. (6), where $arccos(\theta_{R_i})$ is the corresponding angle of region R_i and the denominator is the sum of the three angles.

$$Jun_{\langle R_i,R_j\rangle} = \frac{arccos(\theta_{R_i})}{\sum arccos\left(J_t\left(\theta_{R_i},\theta_{R_j},\theta_{R_k}\right)\right)} \tag{6}$$

4. Layer inference

Based on the simple intuition that the layer sequence of the image scene is the representation of occlusion phenomenon, we attempt to obtain the sequence to verify the effectiveness of our

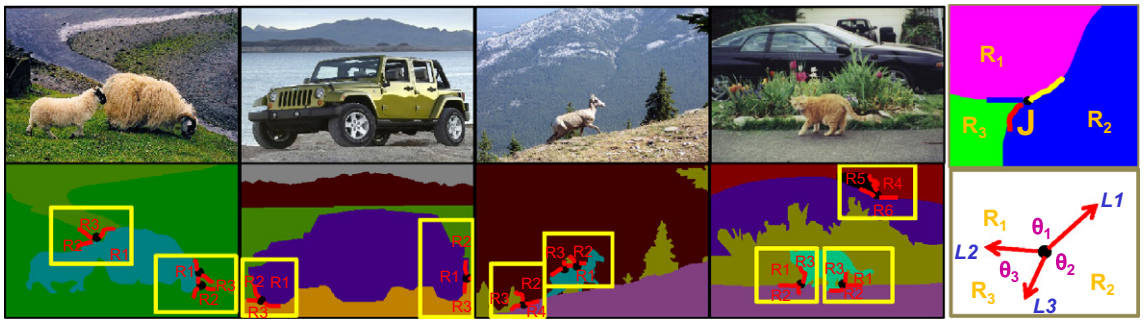

Fig. 6. Illustration of the junction cue. When three regions are adjacent, their junction can characterize the occlusion phenomenon.

occlusion cues. In this section we describe the occlusion prediction and layer inference.

4.1. Semantic labeling

Since our cues are region-based, we need to do semantic labeling work at first. We find that the occlusion cues can work well on the semantic label map given either by users or by semantic labeling methods. We briefly describe the semantic labeling work as it is not the main concern in this paper. Here we use our previous work [29] to get the semantic label map while some other works can also be taken into consideration [16,30]. Our previous work propose a fast geodesic propagation algorithm that integrates recognition proposal and image compatibility into a graphical representation. The geodesic distance is defined on a hybrid manifold, combining the color and boundary features with the recognition proposal map. Based on the geodesic distance, the semantic labeling is simultaneously propagated from the initial seeds of all classes to the rest of image pixels.

We take scene layering experiments with the semantic label map given in two ways, by manual labeling and by semantic labeling methods. The details of the experiments are introduced in Section 5.4.

4.2. 2D representation

The proposed work makes the assumption that the semantic segmentation result of an image scene has been obtained, though the semantic labeling work itself is a very hard problem in computer vision today. We define the 2D representation of the image to organize features in the preparation phase. As shown in Fig. 7, the 2D representation W_{2D} of the image consists of a set of 2D regions V_R, a set of semantic labels S_R and a set of junctions J_T.

$$W_{2D} = (V_R, S_R, J_T) \tag{7}$$

The region set $V_R = (R_1, R_2, \ldots, R_N)$ contains all the information of regions such as area, contour, position and shared boundary curves. The semantic label set $S_R = (S_{R_1}, S_{R_2}, \ldots, S_{R_N})$ consists of the region semantic labels. The junction set $J_T = (J_1, J_2, \ldots, J_t)$ consists of junctions detected from the semantic label map. The region set, semantic label set and junction set are firstly extracted from the image to make preparation for the inference process.

4.3. Layered representation

Since our desired representation is the reasonable layer sequence, we define the solution structure of the layered representation W_L as a layer sequence, which indicates the layer partition of regions.

$$W_L = (R_{L_1}, R_{L_2}, \ldots, R_{L_N}) \tag{8}$$

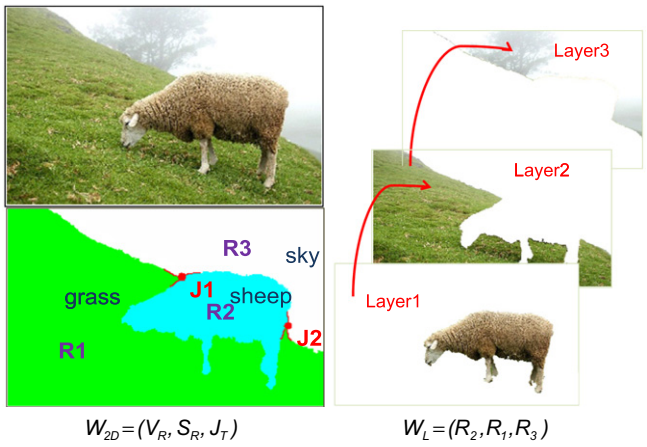

Fig. 7. Illustration of the 2D representation and the layered representation. The input image scene is on the upper left; the semantic label map is on the lower left, which includes the 2D regions, the semantic labels and junctions; images on the right show the layered representation of the given image scene.

where R_{L_i} indicates the region which is assigned layer L_i. Our objective is to infer the optimal layer sequence according to the 2D information of the image. Here, we suppose each region is assigned a unique layer. In Fig. 7, the layered representation of the input image is on the right column, indicating in which layer an region is located.

4.4. Occlusion prediction using Adaboost

After the preparation of cues, the next stage of our method is to get the occlusion relationship predictions and the confidences of these predictions. We define a preference function PREF which is a combination of a set of binary indicator functions. PREF interprets the primitive features as a score that shows the possibility of corresponding occlusion with its features. The preference function is formulated in Eq. (9), where $R_i \succ R_j$ means that region R_i occludes region R_j and $R_i \prec R_j$ means that region R_j occludes region R_i.

Table 1
Cues for comparison.

Cues	Dim	Cues	Dim
F1: Semantic	1	F5: Junction	1
F2: Position	1	F6: Appearance	51
F3: Compactness	2	F7: Boundary Contrast	4
F4: Boundary	1	F8: Shape	17

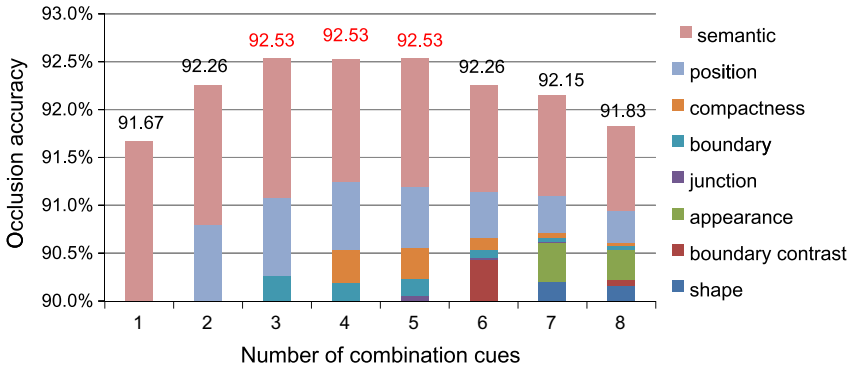

Fig. 8. The accuracy of occlusion prediction with different number of combination cues. The *x*-axis indicates the number of combination cues and the *y*-axis indicates the accuracy. There are many combinations with the same number of cues. Each bar demonstrates the accuracy of the best combination among these combinations and the corresponding components. The color legend of the eight type of cues is listed on the right. For example, there are 28 combinations which are formed by two types of cues out of the eight types of cues. The best accuracy percentage of these 28 combinations is listed above the 2nd bar. Meanwhile, the 2nd bar shows that the best among the 28 combinations consists of our semantic and position cues. The importance of these two components are visualized in the form of length. Our semantic cue takes a more important role than our position cue in this combination. Comparing these eight bars, we can see that our five cues perform well.

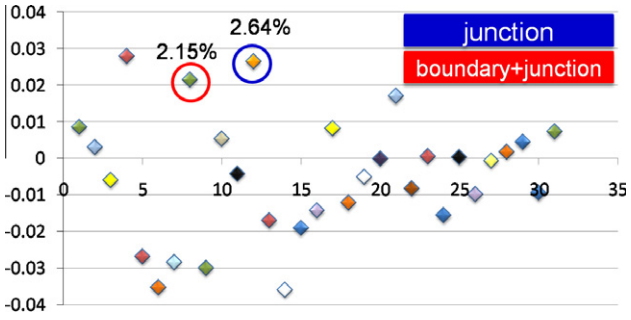

Fig. 9. Accuracy differences of 31 combinations. This figure shows the accuracy differences between adjacent and non-adjacent regions in all the combinations of our five cues (31 combinations). Each point illustrates the accuracy difference between adjacent and non-adjacent regions in the case of the same combination. We get the difference by subtracting the accuracy of adjacent regions with the accuracy of non-adjacent regions. Two combinations which improve the accuracy significantly are denoted in color circles. The blue one has only the junction cue and the red one has the combination of the boundary and junction cues. Difference of accuracy is listed above the point.

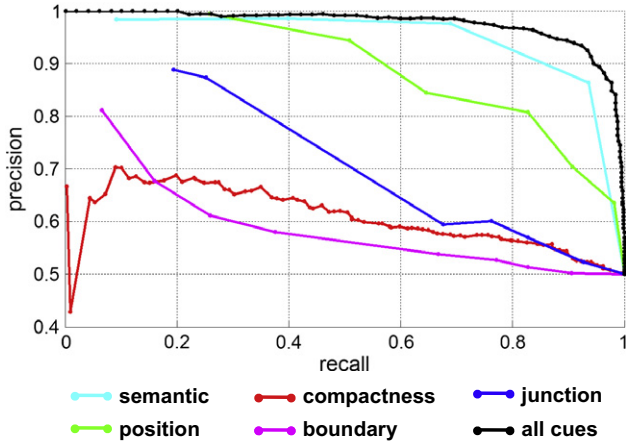

Fig. 10. Occlusion prediction comparison of five PR curves of only one cue with the PR curves of all five cues on the dataset of 200 rural images.

$$\mathrm{PREF}(R_i,R_j) \stackrel{def}{=} \begin{cases} score > 0, & R_i \succ R_j \\ score < 0, & R_i \prec R_j \end{cases} \tag{9}$$

Larger magnitude of the score indicates higher possibility of this occlusion relationship. There are many algorithms to learn the preference function [31]. We choose Adaboost [32], an algorithm to learn strong classifier by using a set of weak binary classifier linearly, to learn our PREF function. The form of Adaboost classifier is as:

$$\mathrm{PREF}(R_i,R_j) = h(x) = \sum_{t=1}^{T} \alpha_t h_t(x), h_t(x) = 1[f_t(x) > \theta_t] \tag{10}$$

where $h(x)$ is the confidence score of the prediction, $h_t(x)$ is an indicator weak classifier and is chosen to be decision stump, and θ_t is the threshold for weak classifier t. Algorithm 1 shows our revised version of Adaboost algorithm, which obtains the output of the final classification as well as the score of prediction confidence. For example, if PREF(R_i, R_j) and PREF(R_j, R_i) are both positive and the magnitude of PREF(R_i, R_j) is larger than that of PREF(R_j, R_i), we can conclude that it is more possible that region R_i occludes region R_j.

Algorithm 1. Revised version of Adaboost algorithm

Require: $(fv_1, y_1)\cdots(fv_m, y_m)$ where $fv_i \in FV$, $y_i \in Y = \{-1,+1\}$
Ensure: the final hypothesis $H(fv_i)$ and PREF scores PREF(fv_i).
1: Initialize $D_1(i) = 1/m$.
2: Update the Weighting.
for each t, $t \in \{T\}$ **do**
(1) Train weak learner using distribution D_t.
(2) Get weak hypothesis h_t: $FV \to \{+1,-1\}$ with error

$$\varepsilon_t = \Pr_{i,D_t}(h_t(fv_i) \neq y_i).$$

(3) Choose $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.
(4) $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(fv_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(fv_i) \neq y_i \end{cases} = \frac{D_t(i)\exp(-\alpha_t y_i h_t(fv_i))}{Z_t}$

where Z_t is a normalization factor.
end for
3: Then output the final hypothesis and PREF scores:
(1) $H(fv_i) = sign\left(\sum_{t=1}^{T}\alpha_t h_t(fv_i)\right), \quad i = 1\cdots m.$
(2) $\mathrm{PREF}(fv_i) = \sum_{t=1}^{T}\alpha_t h_t(fv_i), \quad i = 1\cdots m.$

According to the equations in Section 3, we can compute the response values which indicate the occlusion relationship of two regions. We then form a 6-dimension feature vector consisting of these response values of five cues. For region R_i and R_j, the feature vector FV is symbolized as below:

$$FV\langle R_i,R_j\rangle = \left(S_{\langle R_i,R_j\rangle}, Pos_{\langle R_i,R_j\rangle}, Com_{R_i}, Com_{R_j}, Bry_{\langle R_i,R_j\rangle}, Jun_{\langle R_i,R_j\rangle}\right) \tag{11}$$

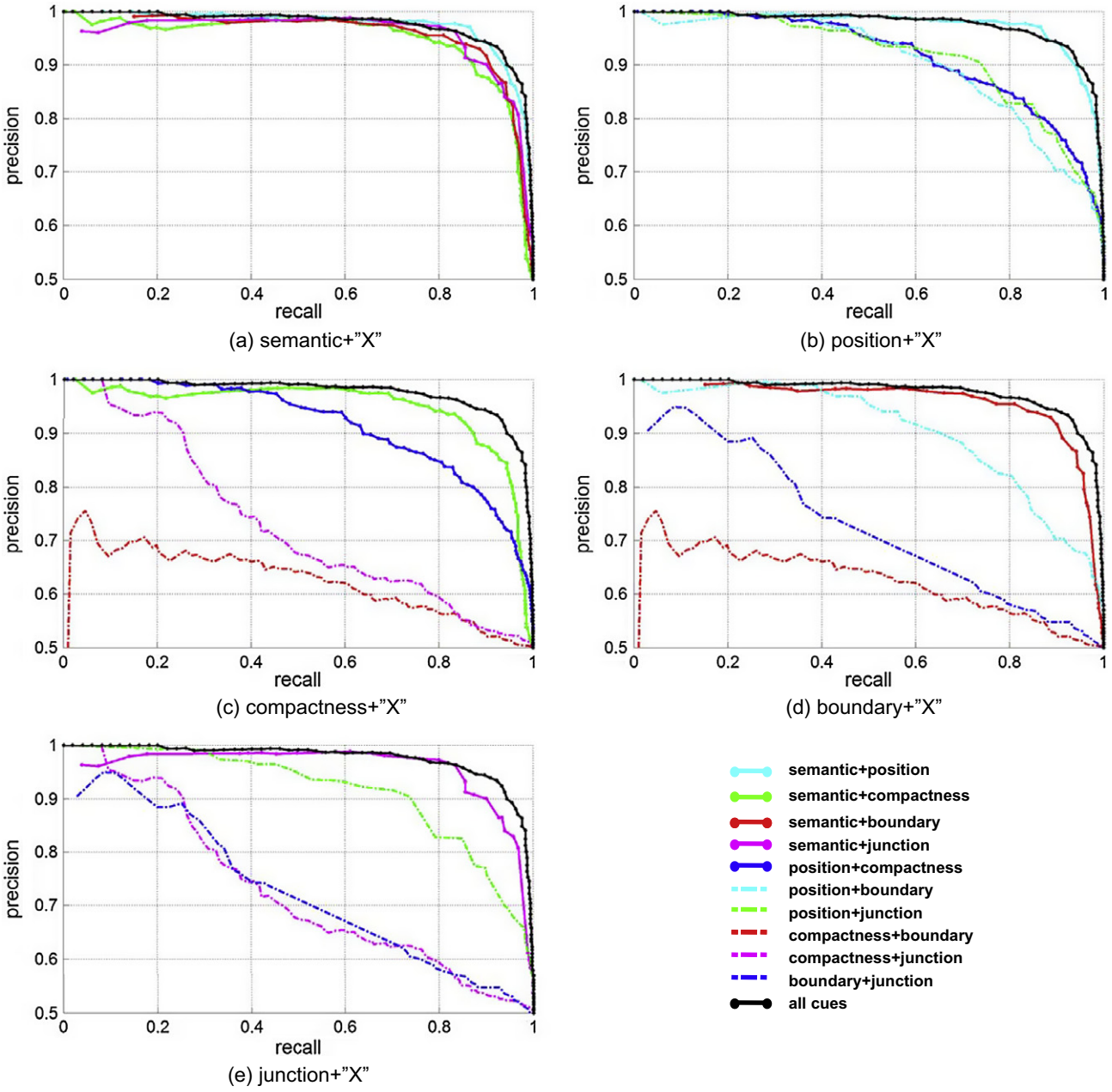

Fig. 11. Occlusion prediction comparison of the PR curves of two cues with the PR curves of all five cues on the dataset of 200 rural images.

where $S_{\langle R_i,R_j\rangle}$, $Pos_{\langle R_i,R_j\rangle}$, $Bry_{\langle R_i,R_j\rangle}$ and $Jun_{\langle R_i,R_j\rangle}$ are the response values of the semantic cue, position cue, shared boundary cue and junction cue; Com_{R_i} and Com_{R_j} are the compactness response values of region R_i and R_j respectively. $FV\langle R_i, R_j\rangle$ is different from $FV\langle R_j, R_i\rangle$, and their response values are also different. $FV\langle R_i, R_j\rangle$ is considered to be a positive occlusion relationship sample if R_i occludes R_j in an image, otherwise it is a negative sample.

4.5. Inference of layer sequence from occlusions

Many methods generally make use of some mechanism to search the solution in the whole solution space, giving rise to low searching speed. Therefore, we use a permutation algorithm to search for the layer sequence to improve the efficiency. Our desired optimal solution W_L is the order that can achieve the best agreement with the agreement function [31], and the function is defined as:

$$W_L = \rho^* = \mathrm{MAX}_{\rho\in P}\{\mathrm{AGREE}(\rho, \mathrm{PREF})\} \tag{12}$$

$$\mathrm{AGREE}(\rho, \mathrm{PREF}) = \sum_{(R_i,R_j:\rho(R_i)\succ\rho(R_j))} \mathrm{PREF}(R_i, R_j) \tag{13}$$

where ρ is an order of region instances in an image. If and only if region R_i is in front of region R_j in this order, then $\rho(R_i) \succ \rho(R_j)$. The agreement of a fixed order is the total sum of all PREF values of this order under the condition that $\rho(R_i) \succ \rho(R_j)$. Then how to find an optimal scene layer sequence is equal to how to find an order that can achieve the best agreement with this function.

Algorithm 2. Permutation algorithm

Require: An region instance set V of an image; preference function PREF
Ensure: An approximate optimal layer sequence W_L
1: Let P be the set of all permutations of instances in V.
2: For each $p \in P$ do $\pi(p) = \sum_{(R_i,R_j \in V) \cap (R_i \neq R_j)} \mathrm{PREF}(R_i, R_j)$.
3: Let $\pi(p^*) = \arg\max_{p \in P} \pi(p)$.
4: Let $W_L = p^*$.

We adopt a brute-force algorithm that enumerates all permutations of the region instances in an image to find the approximate optimal layer sequence. This algorithm, as shown in Algorithm 2, is based on the idea that the optimal solution must be the one that can maximize the sum of the confidence of all occlusions in an image. We can deterministically find the solution for image scene layering with less searching time.

5. Experiments

The experiments consist of three parts. The first part is the verification of our five cues. The second part demonstrates the interactions of these five cues. The third part shows our scene layering results on multiple datasets.

5.1. Datasets

We use three datasets: the datasets of rural image scene, artificial scene and outdoor scene. The rural scene dataset, taken from Ref. [1], is composed of 200 images with 17 categories, such as sky, water, horse, grass, and dog. The dataset of artificial scene [1] consists of 250 images and the dataset of outdoor scene consists of 645 images (from the Internet). Both of these two datasets include 59 categories.

All the images are assigned the ground truth of semantic label and occlusion layer manually. For an image, we manually assign each semantic region a unique occlusion layer according to human occlusion perception. See the layered representation in Fig. 7 for example. We assign the sheep the first layer, though part of the grass is more closer than the sheep in depth. Then the second layer is the grass and the third layer is the sky. The occlusion ground truth can be obtained according to the layer number. For example, region in layer 1 occludes regions in layer 2 and layer 3. We assume that any two regions have an occlusion relationship and our performance metric of occlusion prediction is segment-wise over regions.

5.2. Capabilities of cues

We randomly select a subset from each dataset used in this paper and form these subsets into a large dataset, on which the verification of our five cues is performed. We find that our cues can do better than others to some extent. There are totally eight types of cues for comparison as shown in Table 1. The appearance cue, shape cue and boundary contrast cue are from previous works. Note that some of our five cues are inspired by previous works such as the position cue and junction cue while their formulations are different.

The appearance features include the mean and standard deviation of 17 raw filter responses used in [33], the mean and standard deviation of boosted classifier scores for each pixel, and the log-determinant of all the deviations. The boundary contrast features measure the contrast along the boundary of the region relative to its interior as did in [34]. The shape features consist of region area, region size, perimeter length, residual to boundary lines and first- and second-order shape moments. Features F6–F8 are high dimensional and sensitive to the low level features of images.

We apply random forests [35] to predict the importance of each type of cue on identifying occlusions and the influence of the number of cues. Fig. 8 demonstrates the best occlusion prediction accuracy of different number of combinations cues. We can see that the combinations of three, four and five cues perform better than others. There are many combinations with the same number of cues. In the combinations of three type of cues, our semantic, position and boundary cues perform the best. In the combinations of four type of cues, our semantic, position, compactness and boundary cues perform the best. In the combinations of five type of cues, our five cues perform the best. Since our cues perform better, we take them as occlusion cues. In addition, our low-dimensional cues can be extracted conveniently. Note that our compactness cue is more important than boundary cue in the 4th and 5th bars, though the 3rd, 4th and 5th bars have the same accuracy. Considering the dataset bias, we take compactness and junction cues into consideration as their contributions may be different in other datasets. Besides, we find that the junction cue works efficiently in the case of predicting occlusion relationship between adjacent regions as shown in Fig. 9. That is another reason that why we still take the junction cue as one of our five occlusion cues. Since our five cues can make totally 31 combinations, there are totally 31 points in Fig. 9. Each point illustrates the accuracy difference between adjacent and non-adjacent regions in the case of same combination. Two combinations which improve the accuracy significantly are denoted in color circles, both containing the junction cue. The blue one has only the junction cue and the red one has the combination of boundary cue and junction cue.

5.3. Interactions of the five cues

To outline the interactions of the five cues and the influence of the nature of cues in the process of occlusion prediction, we perform experiments to find the interactions of the cues by using different combinations. As the interactions of the cues may vary in different datasets, we perform experiments on two datasets: the rural scene dataset (Figs. 10 and 11) and the artificial scene dataset (Figs. 12 and 13).

We firstly test the importance of each cue on the rural scene dataset. Fig. 10 shows the occlusion prediction comparison of five PR curves of only one cue with PR curve of all five cues. We can see that the semantic cue is more important than the other four. The

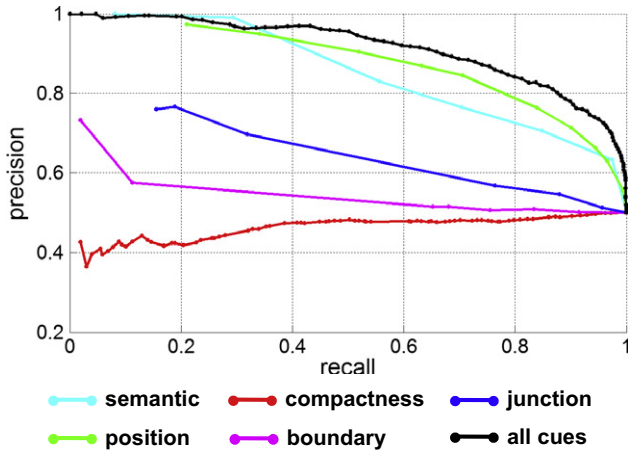

Fig. 12. Occlusion prediction comparison of the five PR curves of only one cue with the PR curve of all five cues on the dataset of 250 artificial images. The most important one is position cue.

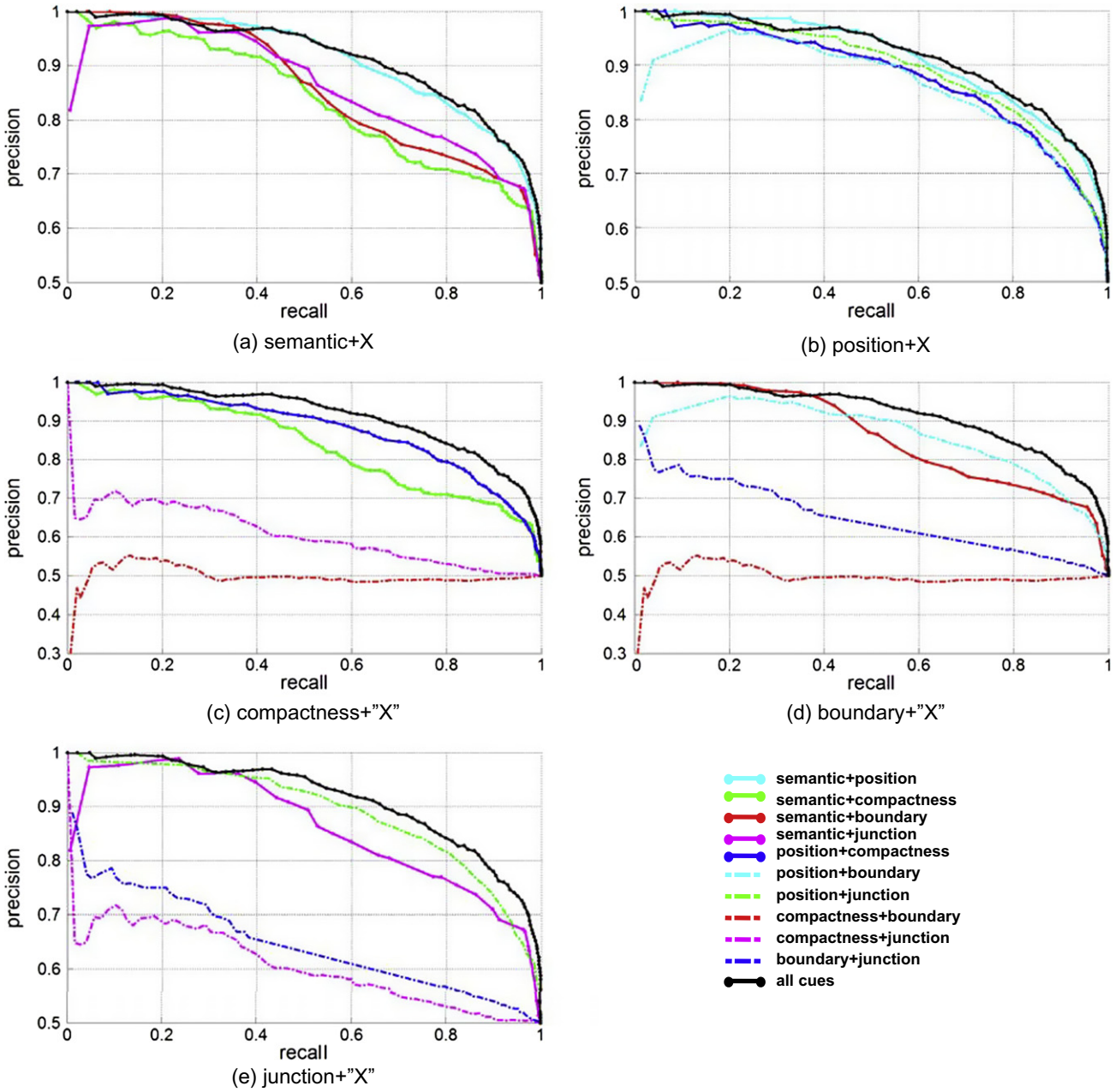

Fig. 13. Occlusion prediction comparison of the PR curves of two cues with the PR curve of all five cues on the dataset of 250 artificial images.

Table 2
Interaction of five cues tested on rural dataset.

	Semantic	Position	Compactness	Boundary	Junction
Semantic	–	↑	↑	↑	↑
Position	↑	–	↑	↑	↑
Compactness	→	↑	–	↑	→
Boundary	↑	→	↑	–	→
Junction	↑	↑	↑	↑	–

Table 3
Interaction of five cues tested on artificial dataset.

	Semantic	Position	Compactness	Boundary	Junction
Semantic	–	↑	↑	↑	↑
Position	↑	–	↑	↑	↑
Compactness	↓	→	–	↓	↓
Boundary	→	→	↑	–	→
Junction	↑	↑	↑	↑	–

second important cue is the position and the third important cue is the junction. It is hard to tell the difference between compactness and boundary. We perform the following experiments to make clear the interrelations of these five cues.

In Fig. 11, we compare the performance of different combinations of two cues. (a) demonstrates the comparison of 'semantic + *X*' curves, where *X* indicates the other four cues. The 'semantic + position' curve is better than the others and 'semantic + compactness' curve is worse. (b) Comparison of the 'position + *X*' curves, where *X* indicates the other four cues. The 'semantic + position' curve performs better than the other three curves except 'all cues' curve. Thus we conclude that the interac-

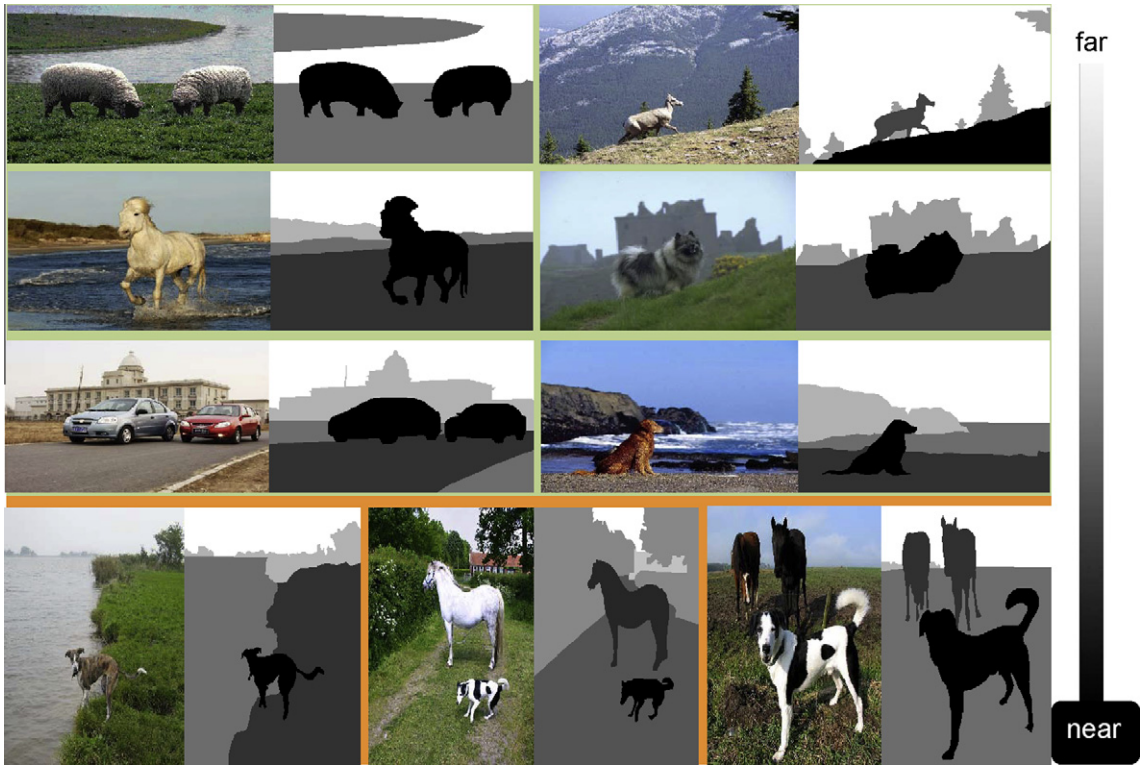

Fig. 14. Some scene layering results of rural scene dataset. The grayscale of regions stands for the layer of them. The blacker the region is, the more forward its corresponding layer is. The left column is the initial image, and the corresponding layered map is on the right column.

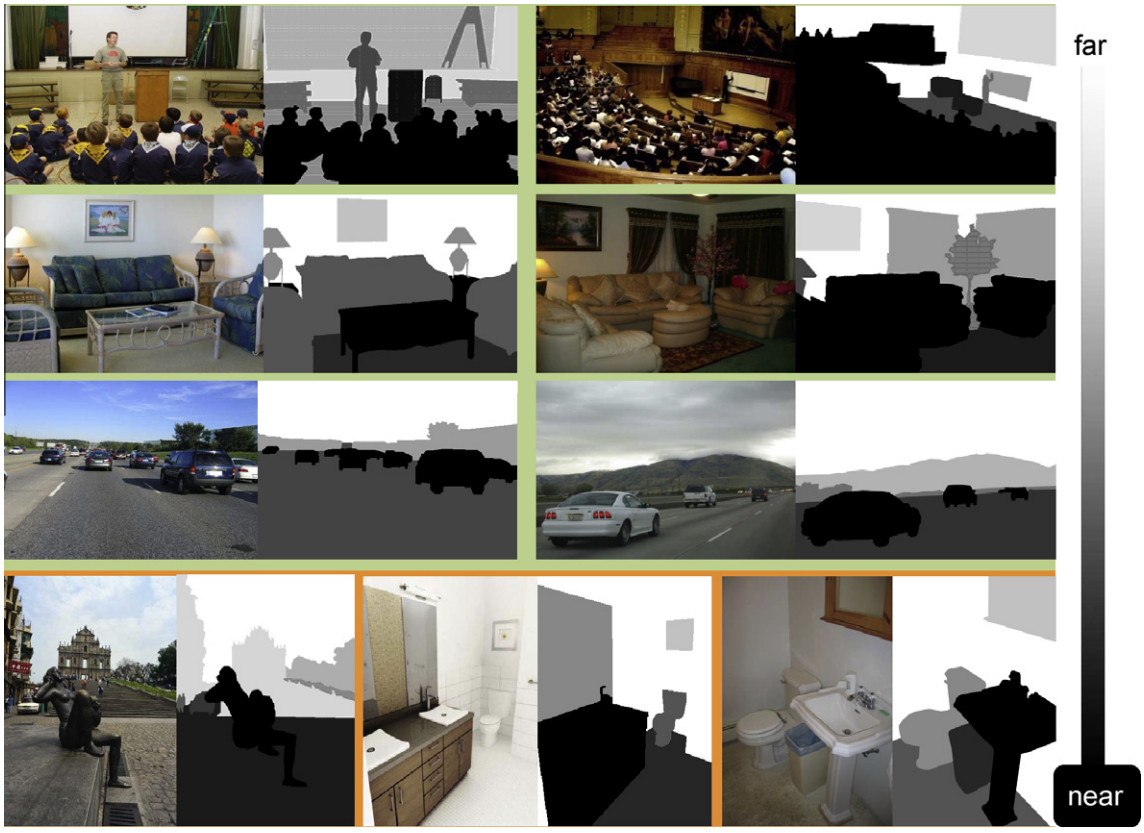

Fig. 15. Some scene layering results of the artificial scene dataset. The corresponding layer of the blacker region is more forward.

tions of the compactness, boundary and junction cues on the position cue are almost the same. (c) Comparison of the 'compactness + X' curves. We can obtain the order of the interactions of four cues on the compactness cue. From the most influential to

Fig. 16. Some scene layering results of the outdoor scene dataset. The corresponding layer of the blacker region is more forward.

the least, the order of the interactions is the semantic, position, junction and boundary cues. (d) Comparison of the 'boundary + *X*' curves. The influence order of the interactions is the semantic, position, junction and compactness cues. (e) Comparison of the 'junction + *X*' curves. The 'semantic + junction' curve performs better than the 'position + junction' curve, while the interactions of the compactness and boundary cue on the junction cue are hard to differentiate. Then in the case of the combination of two cues in the rural scene dataset, we can select semantic and position cues for occlusion prediction. Thus we can get the similar performance with the case of all cues and reduce the cost of computation.

Comparing Fig. 10 with Fig. 11, we get the interactions of these five cues on the rural scene dataset concluded in Table 2. Here, we use '↑' to indicate the enhancing effect, '↓' to indicate the weakening effect, '→' to indicate no apparent variation, and '–' to indicate no such experiments. For instance, 'compactness → semantic' means that the combination of the semantic cue with the compactness cue will result in no apparent variation from the case when only the semantic cue is used. The position cue companied with the compactness cue can lead to better performance than the case of only the position cue, as displayed by 'compactness ↑ position'. We give this scheme of the interactions of cues, not only to view their importance intuitively, but also to help researchers select the cues they need. Comparisons of three cues and four cues are shown in Supplementary materials.

Fig. 12 shows the comparison of five PR curves of one cue with the PR curve of all cues on artificial dataset. The most important one is the position cue. That is different from the rural scene. The second important cue is the semantic cue. The importance sequence of remaining cues is the junction, boundary, and compactness cues.

In Fig. 13, we compare the performance of different combinations of two cues on the artificial dataset. (a) Comparison of 'semantic + *X*' curves. The 'semantic + position' curve is most similar to the 'all cues' curve, and the 'semantic + compactness' curve is the worst one. (b) Comparison of the 'position + *X*' curves. Although they are almost similar to the 'all cues' curve, we can still see that the 'semantic + position' curve is slightly better than the others, while the 'position + boundary' curve is slightly worse. Then the most influential cue to the position cue is the semantic cue, while the least is the boundary cue. (c) Displays the 'compactness + *X*' curves. We can conclude the order of interactions of four cues on the compactness cue. The order is the position, semantic, junction and boundary cues. (d) Displays the 'boundary + *X*' curves. Although the influence of the semantic and position cues on the boundary cue is hard to differentiate, they all perform better than the junction and compactness cues. (e) Displays the 'junction + *X*' curves. It is apparent that the 'position + junction' curve gets similar performance to 'all cues'. The influential order is the position, semantic, boundary and compactness cues. Then in the

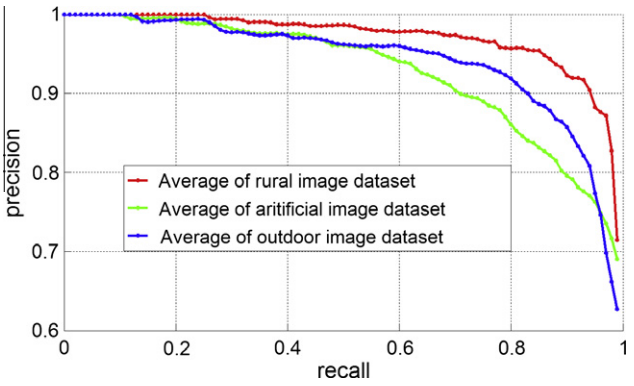

Fig. 17. The PR curve of the occlusion classification on three datasets. The average classification accuracies are 92.9%, 82.7% and 87.4% in the datasets of rural, artificial and outdoor image scene.

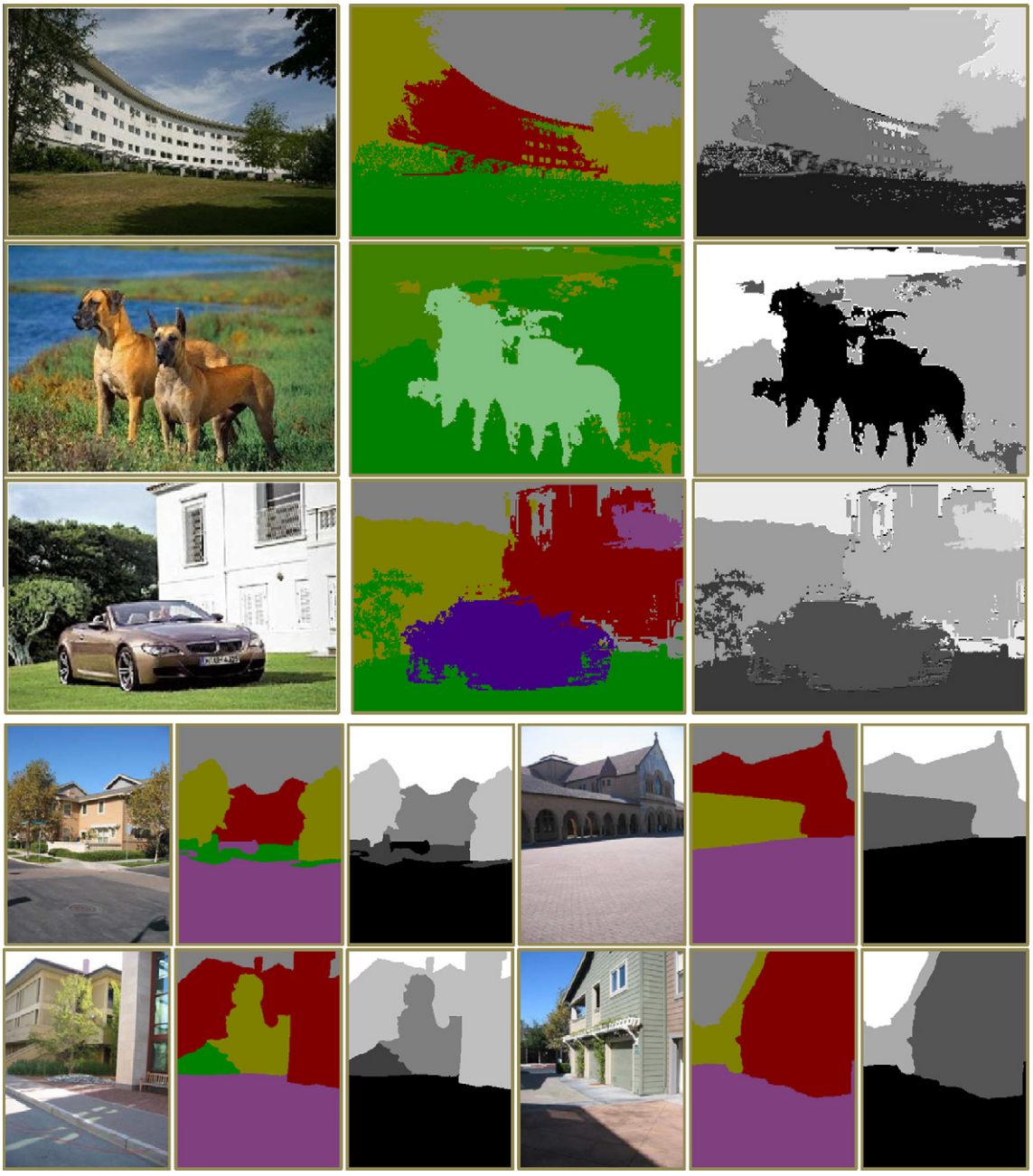

Fig. 18. Our scene layering results given predicted semantic labels. Semantic label maps from row one to row three are given by our previous work [29] and the last two rows are given by Liu et al. [16]. From left to right: the input image, semantic label map and our layering result.

case of the combination of two cues in the artificial dataset, we can select the semantic and position cues. Thus we can get the similar performance with the case of all cues and reduce the cost of computation.

Comparing Fig. 12 with Fig. 13, we conclude the interactions of these five cues on the artificial dataset as shown in Table 3. The symbols are the same with those in Table 2, while the interactions are slightly different. Obviously, the influence of the compactness cue on other four cues varies between these two datasets.

5.4. Scene layering of multiple datasets

To evaluate the performance of the scene layering method proposed, we test it on datasets of rural image scene, artificial scene and outdoor scene. The layering demo is shown in Supplementary video.

At first, we test our method on the dataset of rural scene. The region number in each image of this dataset is in the interval [3,8], thus there are totally 2926 occlusion pairs in this dataset. In this part of experiments, we randomly select 100 images for the learning process and the rest 100 images for the testing process.

Some results of the scene layering are shown in Fig. 14. The initial image is on the left with its corresponding layered map on the right. The grayscale of the region stands for the layer of it in the layered maps, where the blacker region is assigned a more forward layer. The results demonstrate that our method can handle the images of rural scene and our cues have the capacity to characterize the occlusion relationship and layer information.

Meanwhile, we test our method on the datasets of artificial scene and outdoor scene. We randomly select 125 images for training and the rest 125 images for testing in the dataset of artificial

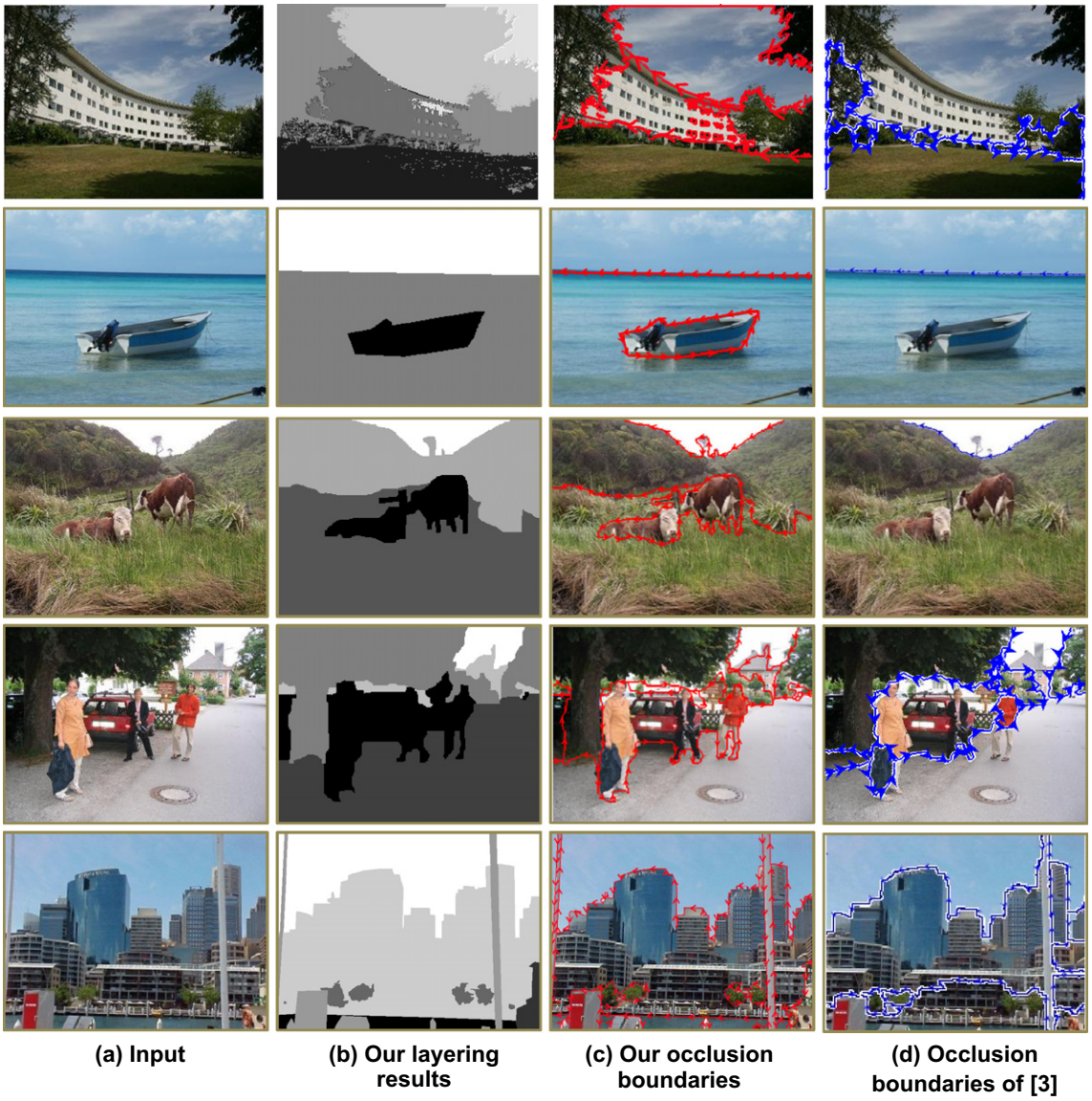

Fig. 19. Comparison of occlusion identification with Hoiem et al. [3]. (a) The input image. (b) Our scene layering result. (c) Our occlusion result. We visualize our occlusion prediction results in the same form of Hoiem et al. [3]. Red lines denote occlusion boundaries, arrows indicate which region (left) is in front. The region on the left side of an arrow occludes the region on the right side. (d) Occlusion results of Hoiem et al. [3].

scene, 345 images for training and 300 images for testing in the dataset of outdoor scene. There are totally 6420 occlusion pairs in the artificial dataset and 8490 occlusion pairs in the outdoor dataset. Results of these two datasets are displayed in Figs. 15 and 16. The input images are on the left column and their corresponding layered maps are on the right. A blacker region is assigned a more forward layer. Running 5-fold cross-validation, the PR curves of the three datasets displayed in Fig. 17 show the average performance with 50 rounds of Adaboost. The average classification accuracies are 92.9%, 82.7% and 87.4% in the datasets of rural, artificial and outdoor scene.

Besides, we present the layering results of our method based on our previous semantic labeling work and other semantic labeling work [16] (Fig. 18). Our scene layering method can work well given the predicted semantic labels.

Hoiem et al. [3] proposed a method to recover and label the occlusion boundary using many cues. Our work is similar with theirs. Note that Hoiem et al. [3] identifies the boundaries and reasons occlusion relationship simultaneously while our occlusion reasoning is based on the regions in the image. Thus, we compare our work with theirs and visualize our occlusion prediction in the same form of Hoiem et al. [3], as shown in Fig. 19. The location of occlusion boundary using our method is more precise than [3] as displayed in the top three rows. Moreover, our occlusion classification is more accurate than [3]. Take image in the fourth row for example, our occlusion prediction of the boundary around the tree and sky is correct while [3] gets wrong prediction.

6. Conclusions

In this paper, we attempt to enable the computer to understand the 3D world behind the 2D image plane as humans do. Inspired by human occlusion perception, we propose five cues to indicate the occlusion relationship, and show their capabilities of occlusion prediction through experiments. Based on the semantic label map of the image, we predict the occlusion relationship with our five cues and infer layer sequence of the image scene. As the experimental results shown in Section 5, the importance and interactions of these five cues vary in different datasets, and we thus give the selection scheme of cues not only to view their importance

intuitively, but also for researchers to select the cues they need. Promising experimental results demonstrate that the proposed layering method can be used in the datasets of rural, artificial and outdoor scenes. The main limitation of our approach is that our layering method work on the basis of semantic labeling image, while actually the semantic labeling itself is a challenging problem in computer vision that rich literatures have concentrated on today. This limitation should be overcome by giving the semantic label map simultaneously with the scene layered map. It should also be noted that we propose only five cues for the computer to understand the occlusion relationship, but there may be other cues to be discovered. As [3] has made use of 3D surface cues, which are also under our consideration for future work, we should also take into account geometry features. Then we can combine the work of semantic labeling together with occlusion prediction by using more promising cues. Furthermore, the motion-based features can be extended when the perception of occlusion is applied to video sequences.

Acknowledgments

We would like to thank the anonymous reviewers for helpful suggestions, Yi Liu for data processing. This work was partially supported by NSFC (60933006), 863 Program (2012AA011504), R&D Program (2012BAH07B01), ITER (2012GB102008) and BUAA (YWF-12-RBYJ-035).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cviu.2012.10.001.

References

[1] B. Yao, X. Yang, S.-C. Zhu, Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks, in: Proceedings of Energy Minimization Methods in CVPR, 2007, pp. 169–183.
[2] M. Nitzberg, D. Mumford, The 2.1-d sketch, in: IEEE International Conference on Computer Vision, 1990, pp. 138–144.
[3] D. Hoiem, A. Stein, A.A. Efros, M. Hebert, Recovering occlusion boundaries from a single image, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[4] X. Ren, C.C. Fowlkes, J. Malik, Figure/ground assignment in natural images, in: Proceedings of 9th European Conference on Computer Vision, 2006, pp. 614–627.
[5] J. Wang, J.Y.A. Wang, Edward, H. Adelson, Representing moving images with layers, IEEE Trans. Image Process. 3 (1994) 625–638.
[6] M.J. Black, D.J. Fleet, Probabilistic detection and tracking of motion discontinuities, in: IEEE International Conference on Computer Vision, 1999, pp. 551–558.
[7] A. Stein, D. Hoiem, M. Hebert, Learning to find object boundaries using motion cues, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[8] E. Saund, Perceptual organization of occluding contours of opaque surfaces, Comput. Vis. Image Understand. 76 (1999) 70–82.
[9] L.R. Williams, A.R. Hanson, Perceptual completion of occluded surfaces, Comput. Vis. Image Understand. 64 (1996) 1–20.
[10] A.N. Stein, M. Hebert, Occlusion boundaries from motion: low-level detection and mid-level reasoning, Int. J. Comput. Vis. 82 (2009) 325–357.
[11] J. Wang, E. Gu, M. Betke, Mosaicshape: stochastic region grouping with shape prior, in: Proceedings of Computer Vision and Pattern Recognition, 2005, pp. 902–908.
[12] Y. Yang, S. Hallman, D. Ramanan, C. Fowlkes, Layered object detection for multi-class segmentation, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 3113–3120.
[13] S. Yu, T.S. Lee, T. Kanade, A hierarchical Markov random field model for figure-ground segregation, in: Proceedings of Energy Minimization Methods in CVPR, 2001, pp. 118–133.
[14] A. Torralba, A. Oliva, Depth estimation from image structure, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1226–1238.
[15] S. Esedoglu, R. March, Segmentation with depth but without detecting junctions, J. Math. Imag. Vis. 18 (2002) 7–15.
[16] B. Liu, S. Gould, D. Koller, Single image depth estimation from predicted semantic labels, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 1253–1260.
[17] V. Hedau, D. Hoiem, D. Forsyth, Recovering the spatial layout of cluttered rooms, in: IEEE International Conference on Computer Vision, 2009, pp. 1849–1856.
[18] A. Saxena, M. Sun, A.Y. Ng, Learning 3-d scene structure from a single still image, in: IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[19] M. Dimiccoli, P. Salembier, Exploiting t-junctions for depth segregation in single images, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2009, pp. 1229–1232.
[20] S. Zheng, Z. Tu, A.L. Yuille, Detecting object boundaries using low-, mid-, and high-level information, in: Proceedings of Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[21] A. Desolneux, L. Moisan, J. michel Morel, Computational gestalts and perception thresholds, J. Physiol. 97 (2002) 311–324.
[22] D. Kersten, P. Mamassian, A. Yuille, Object perception as bayesian inference, Ann. Rev. Psychol. 55 (2004) 271–304.
[23] Bribiesca, Ernesto, An easy measure of compactness for 2d and 3d shapes, Pattern Recognit. 41 (2008) 543–554.
[24] T. Roussillon, L. Tougne, I. Sivignon, Robust decomposition of a digital curve into convex and concave parts, in: International Conference on Pattern Recognition, 2008, pp. 1–4.
[25] H. Liu, L.J. Latecki, W. Liu, A unified curvature definition for regular, polygonal, and digital planar curves, Int. J. Comput. Vis. 80 (2008) 104–124.
[26] W. Metzger, Gesetze des Sehens, Verlag Waldemar Kramer, 1975.
[27] M. Maire1, P. Arbelaez, C. Fowlkes, J. Malik, Using contours to detect and localize junctions in natural images, in: Proceedings of Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[28] M.A. Cazorla, F. Escolano, Two bayesian methods for junction classification, IEEE Trans. Image Process. 12 (2003) 317–327.
[29] X. Chen, D. Zhao, Y. Zhao, L. Lin, Accurate semantic image labeling by fast geodesic propagation, in: Proceedings of the International Conference on Image Processing, 2009, pp. 4021–4024.
[30] J. Xiao, L. Quan, Multiple view semantic segmentation for street view images, in: Proceeding of International Conference on Computer Vision, 2009, pp. 686–693.
[31] W.W. Cohen, R.E. Schapire, Y. Singer, Learning to order things, J. Artif. Intell. Res. 14 (1999) 243–270.
[32] Y. Freund, R.E. Schapire, A short introduction to boosting, J. Jpn. Soc. Artif. Intell. 14 (1999) 771–780.
[33] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, Int. J. Comput. Vis. 81 (2009) 2–23.
[34] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Proceeding of International Conference on Computer Vision, 2009.
[35] L.B. Statistics, L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.