Energy Formulation of PanoContext

Yinda Zhang

July 17, 2014

1 Parameter Space

Given an image I, the output of system is a big 3D box for the room and a set of 3D small boxes for major objects inside. We take the Manhattan World Assumption, so that all boxes are perfectly aligned with three vanishing directions. A box can be represented with a 6-dim vector $B = (x, y, z, l, m, n)$, in which x, y, z is the smaller value of coordinates in 3 dimensions, while l, m, n is the size of the box in 3 dimensions.
Consider the we M categories of objects, and each category i is allowed to have no more than N_i instances, so that a whole room can be represented as

$$\mathcal{B} = \{B_{room}, B_i^{j_i} | i = 1, 2, \ldots, M; j_i = 1, 2, \ldots, N_i\} \tag{1}$$

In practice, the number of instances of each category can be different from data to data. We model this by adding a existent indicator for each box. So the whole representation for a room becomes

$$\begin{aligned} \mathcal{P} = \{\mathcal{B}, \mathcal{C}\} &= \{B_{room}, B_i^{j_i}, C_i^{j_i} | i = 1, 2, \ldots, M; j_i = 1, 2, \ldots, N_i\} \\ C_i^{j_i} &= \begin{cases} 1, \text{the } j_i\text{-th instance of the i-th category exists} \\ 0, \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

2 Energy Definition

The energy consists of two terms $E^{\text{scene}} = E^{\text{image}} + E^{\text{context}}$, in which E^{image} measures how well the image I can be explained by $\mathcal{P}$, and E^{context} measures how semantically reasonable of $\mathcal{P}$. In other words, E^{image} is the bottom up term, which test whether the boxes are well supported by some image evidence, e.g. Are there line segments on the edge of boxes? Is the image based normal estimation consistent with boxes? On the other hand, E^{context} is the top down term, which captures a variety of context information among boxes, e.g. whether the nightstand is placed beside the bed. We will introduce each term specifically in the rest of this section.

2.1 Bottom up term

The bottom up energy term is defined as

$$E^{\text{image}}(\mathcal{P}, I) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} C_i^j \cdot \left[E^{\text{det}}(B_i^j) + E^{\text{seg}}(B_i^j) + E^{\text{norm}}(B_i^j) \right] + E^{\text{norm}}(BG) \tag{3}$$

E^{det} captures information from rectangle detector. Two or three surfaces of a box will be visible in image depending on the viewing direction. We use d_i^j to denote the maximal value of rectangle detector scores from all visible surfaces of the box B_i^j. So the detection term is defined as

$$E^{\text{det}}(B_i^j) = \lambda_{\text{det}} \cdot e^{-d_i^j} \tag{4}$$

E^{seg} measures the consistency of the projection of box with the color image segmentation. We run graph based image segmentation with multiple parameters, and compute the intersection over union (IOU) between projection of a box and a segment. s_i^j is the maximal IOU for box B_i^j. Like E^{det}, the segmentation term is defined as

$$E^{\text{seg}}(B_i^j) = \lambda_{\text{seg}} \cdot e^{-s_i^j} \tag{5}$$

E^{norm} measures the consistency between surface normal direction estimation from box and image. With orientation map(OM) and geometric context(GC), we can have a rough estimation of pixelwise normal estimation on panorama image. On the other hand, the normal can be also estimated from a box. We compute the consistency score between two normal estimation as n_i^j for B_i^j, so the normal term is defined as

$$E^{\text{norm}}(B_i^j) = \lambda_{\text{norm}} \cdot e^{-n_i^j} \tag{6}$$

With all the boxes in the scene, we can get all region not covered by any box as the background region. The normal estimation of the background region can be obtained according to B_{room}. We add a term $E^{\text{norm}}(BG)$ to capture normal consistency for background region.

2.2 Top down term

The top down term is defined as

$$E^{\text{context}}(\mathcal{P}) = E^{\text{num}}(\mathcal{C}) + E^{\text{unary}}(\mathcal{B}, \mathcal{C}) + E^{\text{binary}}(\mathcal{B}, \mathcal{C}) + E^{\text{align}}(\mathcal{B}, \mathcal{C}) \tag{7}$$

E^{num} measures the likelihood of number of instances per category base on the probability distribution P^{num} from dataset. For example, most of the bedrooms

in our dataset contain only 1 or 2 beds. Then $P_{\text{bed}}^{\text{num}}(\sum_{j=1}^{N_{\text{bed}}} C_{\text{bed}}^{j} = 1 \text{ or } 2)$ will be large, and be small for other value. So the instance number term will be

$$P_{\text{all}}^{\text{num}} = \prod_{i=1}^{M} P_i^{\text{num}}(\sum_{j=1}^{N_i} C_i^j)$$
$$E^{\text{num}}(\mathcal{C}) = \lambda_{\text{num}} \cdot e^{-P_{\text{all}}^{\text{num}}} \tag{8}$$

E^{unary} takes the score of semantic classifier into account. We train the semantic classifier using random forest with basic 3D information of boxes as feature. Let $S_i(B_i^j)$ be the confidence of a box B_i^j belonging to class i.

$$S_{\text{all}} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} C_i^j \cdot S_i(B_i^j)$$
$$E^{\text{unary}}(\mathcal{B}, \mathcal{C}) = \lambda_{\text{unary}} \cdot e^{-S_{\text{all}}} \tag{9}$$

E^{binary} encodes distance constraints between a pair of boxes. To write the formulate in a easier way, we assign each box a unique ID, i.e. B_i, and use C_i as its existence indicator and T_i as its semantic category. From dataset, we can build non-parametric model for any pair of categories (Notice that two categories can even be same). When given a pair of boxes B_i and B_j, we can measure the likelihood of their co-occurrence $P_{(t_i,t_j)}^{pair}(B_i, B_j)$ by computing the distance to the nearest point in the non-parametric model (refer to paper for more details). When one of the box is the big room, $P_{(room,t_j)}^{pair}(B_{room}, B_j)$ does not allow B_j to be partially outside the B_{room}. The binary term captures pairwise constraints between any pair of objects, as

$$B_{\text{all}} = \sum_{i=1}^{N} \sum_{j=1}^{N} C_i C_j \cdot P_{(t_i,t_j)}^{pair}(B_i, B_j)$$
$$E^{\text{binary}}(\mathcal{B}, \mathcal{C}) = \lambda_{\text{binary}} \cdot e^{-B_{\text{all}}} \tag{10}$$

$E^{\text{align}}(\mathcal{B}, \mathcal{C})$ is defined by room alignment score, which grasps scene structure that cannot be represented by pairwise constraints. We match a room with all rooms in training data, and take the smallest matching cost. So the room alignment term is defined as

$$A_{\min} = \min_k A(\mathcal{P}, \mathcal{P}_k)$$
$$E^{\text{align}}(\mathcal{B}, \mathcal{C}) = \lambda_{\text{align}} \cdot e^{A_{\min}} \tag{11}$$

in which $\mathcal{P}_k$ is the k-th room in training data, $A(\mathcal{P}_i, \mathcal{P}_j)$ is the matching cost of two rooms, and the detailed definition can be found in our paper.

3 Analysis

For an input image I, the most optimal solution is the one minimizing the whole energy cost:

$$\mathcal{P}_{\text{best}} = \arg\min_{\mathcal{P}} E^{\text{scene}}(I, \mathcal{P}) \tag{12}$$

However, the optimization can be difficult due to the property of the energy function and the huge solution space. The energy function is highly non-linear and could have many bad local minimums. Also a bunch of parameters (the λs) are unknown. The good thing is that our current sampling based system can provide some good solutions as initialization for the optimization. Our ranking function can be taken as a rough suggestion for gradient descending direction during the search.