Parsing Semantic Parts of Cars Using Graphical Models and Segment Appearance Consistency

Wenhao Lu[1]
yourslewis@gmail.com
Xiaochen Lian[2]
lianxiaochen@gmail.com
Alan Yuille[2]
yuille@stat.ucla.edu

[1] Department of Electrical Engineering
Tsinghua University

[2] Department of Statistics
University of California, Los Angeles

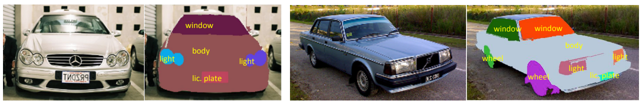
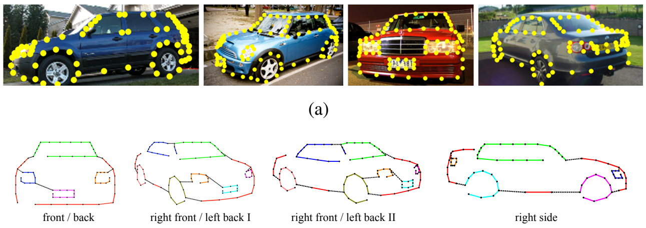


Figure 1: The inputs (left) are images of a car taken from different viewpoints. The outputs (right) are the segmentation of car parts.

We attempt to parse cars into wheels, lights, windows, license plates and body, as illustrated in Figure 1. We formulate the problem as landmark identification. We first select representative locations on the boundaries of the parts to serve as landmarks. They are selected so that locating them yields the silhouette of the parts, and hence enables us to do object part segmentation (see Figure 2(a)). We use a mixture of graphical models to deal with different viewpoints so that we can take into account how the visibility and appearance of parts alter with viewpoint (see Figure 2(b)). We then use a mixture of graphical models to deal with different viewpoints so that we can take into account how the visibility and appearance of parts alter with viewpoint.



(a)

(b)

Figure 2: (a) The landmark annotations for cars of some viewpoints. (b) The proposed mixture-of-trees model. The landmarks connected by the solid lines of same colors belong to the same semantic parts. The black dashed lines show the links between different parts.

A novel aspect of our graphical model is that we couple the landmarks with the segmentation of the image to exploit the image contents when modeling the pairwise relation between neighboring landmarks. In the ideal case where part boundaries of the cars are all preserved by the segmentation, we can assume that the landmarks lie near the boundaries between different segments. Each landmark is then associated to the appearance of its two closest segments. This enables us to associate appearance information to the landmarks and to introduce pairwise coupling terms which enforce that the appearance is similar within parts and different between parts. We call this segmentation appearance consistency (SAC) between segments of neighboring landmarks. However, in practice, it is always impossible to capture all part boundaries using single level segmentation. Instead we couple the landmarks to a hierarchical segmentation of the image. We treat the level f the hierarchy for each part as a hidden variable, which is chosen *dynamically* during inference/parsing. By doing this, our model is able to automatically select the most suitable segmentation level for each part while parsing the image.

The model for each viewpoint is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes $\mathcal{V}$ correspond to landmark points. They are divided into subsets $\mathcal{V} = \bigcup_{p=1}^N \mathcal{V}_p$, where N is the number of parts and $\mathcal{V}_p$ consists of landmarks lying at the boundaries of semantic part p. The edge structures $\mathcal{E}$ are manually designed (see Figure 2(b)). Each node has pixel position of landmark $l_i = (x_i, y_i)$. The set of all positions is denoted by $\mathbf{L} = \{l_i\}_{i=1}^{|\mathcal{V}|}$. We denote by p_i the indicator specifying which part landmark i belongs to, and by $h(p)$ the segmentation level of part p. Then the segment pair of node i, $\mathbf{s_i}$, can be seen as the function of $h(p_i)$, which we denote by $\mathbf{s}_{i,h}$ for simplicity. Similar to the definitions of $\mathbf{L}$, we have $\mathbf{H} = \{h(p_i)\}_{i=1}^N$ and $\mathbf{S}(\mathbf{H}) = \{\mathbf{s}_{i,h}\}_{i=1}^{|\mathcal{V}|}$. The score function of the model for viewpoint v is

$$S(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) = \phi(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) + \psi(\mathbf{L}, \mathbf{H}, v \mid \mathbf{I}) + \beta_v \tag{1}$$

In the following we omit v for simplicity. The unary terms $\phi(\mathbf{L}, \mathbf{H} \mid \mathbf{I})$ is expressed as:

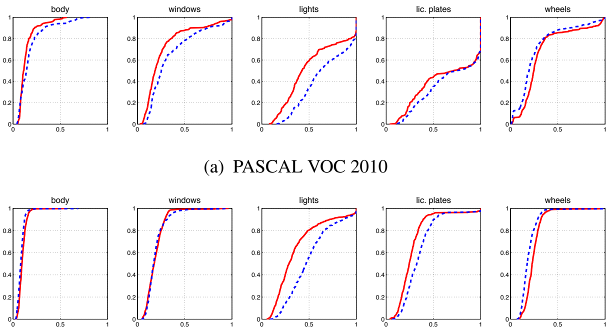
$$\phi(\mathbf{L}, \mathbf{H} \mid \mathbf{I}) = \sum_{i \in \mathcal{V}} \left[\mathbf{w}_i^f \cdot f(l_i \mid \mathbf{I}) + w_i^e e(h(p_i), l_i \mid \mathbf{I}) \right] \tag{2}$$

$\mathbf{w}_i^f \cdot f(l_i \mid \mathbf{I})$ measures the appearance evidence for landmark i at location l_i, where $f(l_i \mid \mathbf{I})$ is the HOG feature vector. The term $e(h(p_i), l_i \mid \mathbf{I})$ penalizes landmarks being far from edges. The binary term $\psi(\mathbf{L}, \mathbf{H} \mid \mathbf{I})$ is:

$$\psi(\mathbf{L}, \mathbf{H} \mid \mathbf{I}) = \sum_{(i,j) \in \mathcal{E}} \mathbf{w}_{i,j}^d \cdot d(l_i, l_j) + \sum_{\substack{(i,j) \in \mathcal{E} \\ p_i = p_j}} \mathbf{w}_{i,j}^A \cdot A(\mathbf{s}_{i,h}, \mathbf{s}_{j,h} \mid \mathbf{I}) \tag{3}$$

$d(l_i, l_j) = (-|x_i - x_j - \bar{x}_{ij}|, -|y_i - y_j - \bar{y}_{ij}|)$ measures the deformation cost for connected pairs of landmarks, where $\bar{x}_{ij}$ and $\bar{y}_{ij}$ are the anchor (mean) displacement of landmark i and j. We adopt L1 norm to enhance our model's robustness to deformation. In the second term of Equation 3, $A(\mathbf{s}_{i,h}, \mathbf{s}_{j,h} \mid \mathbf{I}) = (\alpha(s_{i,h}^1, s_{j,h}^1 \mid \mathbf{I}), \alpha(s_{i,h}^1, s_{j,h}^2 \mid \mathbf{I}), \alpha(s_{i,h}^2, s_{j,h}^1 \mid \mathbf{I}), \alpha(s_{i,h}^2, s_{j,h}^2 \mid \mathbf{I}))$ is a vector storing the pairwise similarity between segments of nodes i and j. This, together with the strength term $\mathbf{w}_{ij}^A$, models the SAC. Finally, β is a mixture-specific scalar bias. The parameters of the score function are $\mathcal{W} = \{\mathbf{w}_i^f\} \cup \{w_i^e\} \cup \{\mathbf{w}_{ij}^d\} \cup \{\mathbf{w}_{ij}^A\} \cup \{\beta\}$.

We validate our approach on a subset of PASCAL VOC2010 car images (VOC10) [1] and 3D car (CAR3D) [2]. The comparison with [3] are shown in Figure 3.



(a) PASCAL VOC 2010

(b) CAR3D

Figure 3: Cumulative segmentation error distribution for parts. X-axis is the average segmentation error normalized by image width, and Y-axis is the fraction of the number of testing images. The red solid lines are the performance using SAC and the blue dashed lines are from [3].

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[2] Silvio Savarese and Fei-Fei Li. 3d generic object categorization, localization and pose estimation. In *ICCV*, pages 1–8, 2007.

[3] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.