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 novel aspect of our graphical model is that we couple the landmarks connected by the edge structures $E$ to the segmentation of the image to exploit the image contents during inference/parsing. 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.

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.

The model for each viewpoint is represented by $G = (V, E)$. The nodes $V$ correspond to landmark points. They are divided into subsets $V = \bigcup_{p=1}^{N} V_p$, where $N$ is the number of parts and $V_p$ consists of landmarks lying at the boundaries of semantic part $p$. The edge structures $E$ 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 $L = \{l_i\}_{i=1}^{\mid V \mid}$. We denote by $p_i$ the indicator specifying which part landmark $l_i$ belongs to, and by $h(p_i)$ the segmentation level of part $p_i$. Then the segment pair of node $l_i$, $s_i$, can be seen as the function of $h(p_i)$, which we denote by $s_{i,h}$ for simplicity. Similar to the definitions of $L$, we have $H = \{h(p_i)\}_{i=1}^{\mid V \mid}$ and $S(H) = \{s_{i,h}\}_{i=1}^{\mid V \mid}$. The score function of the model for viewpoint $v$ is

$$S(L, H, v \mid I) = \phi(L, H, v \mid I) + \psi(L, H, v \mid I) + \beta,$$  

where $\phi(L, H, v \mid I)$ is the unary term, $\psi(L, H, v \mid I)$ is the pairwise term, and $\beta$ is a global term. In the following we omit $v$ for simplicity. The unary term $\phi(L, H, v \mid I)$ is expressed as

$$\phi(L, H, v \mid I) = \sum_{i \in V} \left[ w^f_i \cdot f(l_i \mid I) + w^e_i \cdot \phi(h(p_i), l_i \mid I) \right]$$

The score function of the model for viewpoint $v$ is

$$S(L, H, v \mid I) = \phi(L, H, v \mid I) + \psi(L, H, v \mid I) + \beta,$$  

and $\psi(L, H, v \mid I)$ is the pairwise term, which enforces that the appearance is similar within parts and different between parts. $\beta$ is a global term. The term $\phi(h(p_i), l_i \mid I)$ penalizes landmarks being far from edges. The binary term $\psi(L, H, v \mid I)$ is:

$$\psi(L, H, v \mid I) = \sum_{(i,j) \in E} w^{ij}_{d}(l_i, l_j) + \sum_{(i,j) \in E} w^{ij}_{v} \cdot A(s_{i,h}, s_{j,h} \mid I)$$

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

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].

