Multiple View Semantic Segmentation for Street View Images

Jianxiong Xiao and Long Quan {csxjx, quan}@cse.ust.hk The Hong Kong University of Science and Technology

Abstract

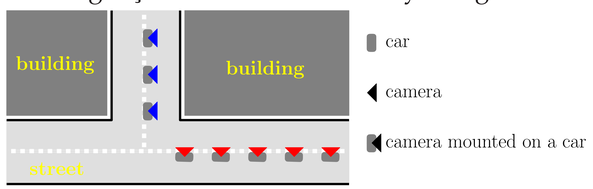
We propose a simple but powerful multi-view semantic segmentation framework for images captured by a camera mounted on a car driving along streets.

In our approach, a pair-wise Markov Random Field (MRF) is laid out across multiple views. Both 2D and 3D features are extracted at a super-pixel level to train classifiers for the unary data terms of MRF. For smoothness terms, our approach makes use of color differences in the same image to identify accurate segmentation boundaries, and dense pixel-to-pixel correspondences to enforce consistency across different views. To speed up training and to improve the recognition quality, our approach adaptively selects the most similar training data for each scene from the label pool.

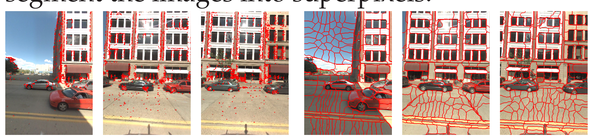
Furthermore, we also propose a powerful approach within the same framework to enable large-scale labeling in both the 3D space and 2D images. We demonstrate our approach on more than 10,000 images from Google Maps Street View.

Preprocessing

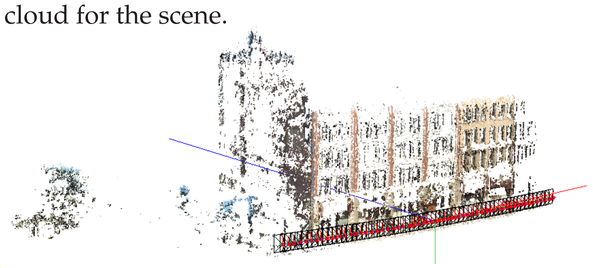
Data is captured by a camera that usually faces the building façade and moves laterally along streets.



We compute pixel-to-pixel correspondences to obtain feature tracks across multiple view, and oversegment the images into superpixels.

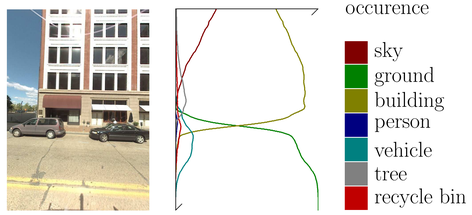


We use Structure from Motion to obtain a 3D point cloud for the scene.

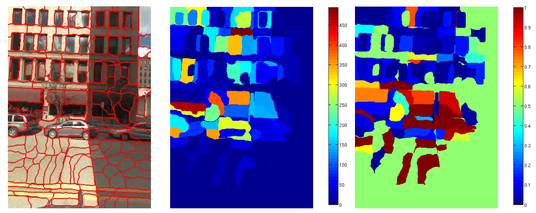


Unary Potential

For each superpixel, we extract 2D features from texture, color, size, shape and pixel position in the image.



We extract 3D features as the feature track density in SFM and the dot product between the estimated normal direction and $-y$ direction.



We apply the AdaBoost classifier that we have learned for each class, l, to the descriptors. The estimated confidence value can be reinterpreted as a probability distribution using softmax transformation:

$$P_i\left(l_i|\mathbf{f}_i^A,\mathbf{f}_i^P,\mathbf{f}_i^G\right)=\frac{\exp\left(H\left(l_i,\mathbf{f}_i^A,\mathbf{f}_i^P,\mathbf{f}_i^G\right)\right)}{\sum_l\exp\left(H\left(l,\mathbf{f}_i^A,\mathbf{f}_i^P,\mathbf{f}_i^G\right)\right)},$$

where $H\left(l,\mathbf{f}_i^A,\mathbf{f}_i^P,\mathbf{f}_i^G\right)$ is the output of the AdaBoost classifier for class l. We then define the unary potential as $\psi_i\left(l_i\right)=-\log P_i\left(l_i|\mathbf{f}_i^A,\mathbf{f}_i^P,\mathbf{f}_i^G\right)$.

Smoothness

For edges in the same image, the smoothness cost is defined as

$$\psi_{ij}\left(l_i,l_j\right)=\left[l_i\neq l_j\right]\cdot g\left(i,j\right),$$

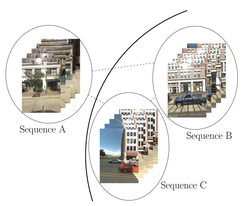
where $g\left(i,j\right)=\frac{1}{\zeta\|\mathbf{c}_i-\mathbf{c}_j\|^2+1}$ and $\|\mathbf{c}_i-\mathbf{c}_j\|^2$ is the $L2$-Norm of the RGB color difference of two superpixels, p_i and p_j.

For edge across two images, the smoothness cost is defined as

$$\psi_{ij}\left(l_i,l_j\right)=\left[l_i\neq l_j\right]\cdot\lambda\left|\mathcal{T}_{ij}\right|g\left(i,j\right)$$

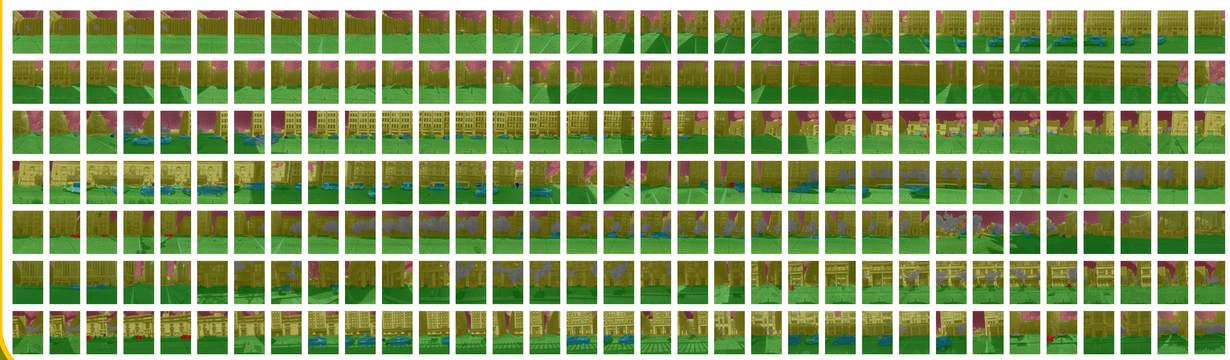
where $\mathcal{T}_{ij}=\left\{\mathbf{t}=\langle\mathbf{x},\left(x_i,y_i,i\right),\left(x_j,y_j,j\right),\ldots\rangle\right\}$ is the set of all feature tracks with projection (x_i,y_i) lying inside the superpixel, p_i, in image I_i, and projection (x_j,y_j) lying inside the superpixel, p_j, in image I_j. This definition encourages two superpixels with more feature track connections to have the same semantic segmentation label, since the cost to have different labels is high due to large $|\mathcal{T}_{ij}|$.

Adaptive Training



We cluster the 40 labeled sequences in the pool based on affinity, defined as the minimal Gist distance between any image in a sequence and any image in another sequence. We then learn 7 models respectively for each cluster. Given a testing sequence, we choose the model trained from most similar cluster for prediction.

Results



Large-scale Labeling

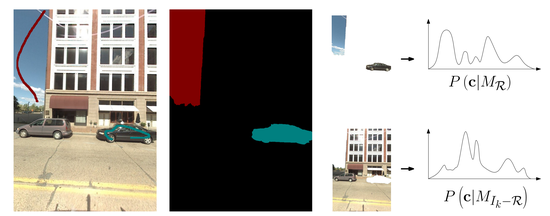
We let the user label the 3D points in the 3D space. Using labels of 3D points, we are able to segment the 2D images at the same time.

$$P_i^{3D}\left(l_i=l\right)\propto\left|\mathcal{T}^l\cap\mathcal{T}_i\right|+\frac{\left|\mathcal{T}^{\text{unknown}}\cap\mathcal{T}_i\right|}{n},$$

where $\mathcal{T}^l=\{\mathbf{t}|\mathbf{t}$ is a track labeled as class l by the user in 3D$\}$, $\mathcal{T}^{\text{unknown}}$ is the set of feature tracks that have no label information from the user. The unary potential is defined to be

$$\psi_i\left(l\right)=-\frac{\left|\mathcal{T}_i-\mathcal{T}^{\text{unknown}}\right|+\epsilon}{H\left(P_i\left(\cdot\right)\right)+\varepsilon}\log P_i\left(l\right),$$

where $\left|\mathcal{T}_i-\mathcal{T}^{\text{unknown}}\right|$ is the number of labeled feature tracks with projections in superpixel p_i, $H\left(P_i\left(\cdot\right)\right)$ is the entropy of the distribution $P_i\left(\cdot\right)$. For region with few 3D points, such as the sky, the user can do labeling in 2D.



When a superpixel in one image is covered by the strokes drawn by the user to be class l, the corresponding unary potential is set to $\psi_i\left(l_i=l\right)=-\infty$ and $\psi_i\left(l_i\neq l\right)=+\infty$. For a superpixel without strokes, we determine the likelihood that the label information comes from the 2D color from

$$P_i^{col}\left(\mathbf{c}_i\right)=\frac{P\left(\mathbf{c}_i|M_{\mathcal{R}}\right)}{P\left(\mathbf{c}_i|M_{\mathcal{R}}\right)+P\left(\mathbf{c}_i|M_{I_k-\mathcal{R}}\right)}.$$

The probability is defined accordingly by

$$P_i\left(l\right)=P_i^{col}\left(\mathbf{c}_i\right)P_i^{2D}\left(l\right)+\left(1-P_i^{col}\left(\mathbf{c}_i\right)\right)P_i^{3D}\left(l\right),$$

where $P_i^{2D}\left(l\right)$ is the color likelihood computed from pixel colors in the stroke-covered regions of image belonging to class l.