ACTIVE SEMANTIC LABELING OF STREET VIEW POINT CLOUDS

Yang Zhou\textsuperscript{1,2} Shuhan Shen\textsuperscript{1,2} Zhanyi Hu\textsuperscript{1,2}

1. NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China
2. University of Chinese Academy of Sciences, Beijing 100049, P. R. China

ABSTRACT

Semantic 3D models have shown their importance in many fields such as autonomous driving. However, it remains a tough task to assign semantic labels to various scenes. In this paper, we propose an Active Learning based method for semantic labeling of street view point clouds with a small amount of annotated data samples. The proposed method takes a point cloud and registered images as the input, and yields a point cloud with semantic labels. We iteratively fine-tunes a network with the ever-enlarging training set to exploit the semantic information of the scene, and fuse the semantic labels in 3D space. To deal with the imbalanced data in street view scenes, a label biased criterion for query selection is proposed to help select images to efficiently improve the performance of the network and the quality of the semantic model. Experimental result shows that the proposed method demands limited human labor and works well in assigning semantic labels to the imbalanced scenes like street view scenes.

Index Terms— Semantic, Street View, Active Learning

1. INTRODUCTION

Autonomous driving has gained its popularity in recent years. Two of the major tasks of autonomous driving are scene parsing and self localization. These two address the problems of what are in the scene and where the vehicle is. To help researchers validate their methods, a number of datasets containing urban street views have been publicly available, such as KITTI [1], Cityscapes [2], TorontoCity [3], RobotCar [4] and ApolloScape [5]. Generally for the task of scene parsing and self localization, both 2D annotation for images and 3D annotation for point clouds are necessary. However, till now only ApolloScape [5] provide 3D annotation for point clouds, while 2D annotation for images prevails in most datasets. The main reason for the lack of 3D annotation is that there are few effective tools for manual annotation in 3D space, and current deep learning based labeling pipeline like [6, 7] cannot deal with such large-scale 3D models. Though directly labeling 3D models remains a challenging problem, 2D image understanding has been through a huge boost due to the improvement of deep learning techniques. Therefore, possible ways to acquire semantic 3D models could be either jointly optimizing the 3D structure and semantic label assignment [8, 9, 10] or assigning semantic labels to existing 3D models [11, 12, 13]. In this work, we aim to assign semantic labels to existing 3D models, especially point clouds of street views.

In order to acquire trusted image segmentation, a well-performed segmentation Convolutional Neural Network (CNN) is essential. Usually a segmentation CNN pre-trained on public datasets could not be seamlessly transferred to arbitrary datasets, unless it is fine-tuned with a substantial number of annotated data samples. However, for many sophisticated supervised learning tasks, labeled data samples could be very difficult or expensive to obtain. Thus, reducing the annotation cost is meaningful. In doing so, we propose a novel method to label 3D point clouds with the Active Learning (AL) technique. AL is an established way to reduce the labeling workload while dealing with a large number of unannotated data samples. By asking queries in the form of unlabeled data samples to be labeled by an oracle (e.g., a human annotator), the active learner achieve high accuracy using as few labeled data samples as possible [14]. Our method takes the 3D point cloud and registered images as the input. We fine-tune a pre-trained CNN for image semantic segmentation with a small training set which is enlarged with the images chosen from the unannotated image set and annotated manually. The semantic labels in each image are back-projected to and fused on the 3D model, and then the query images are selected with the supervision of the 3D semantic model for annotation and incorporated into the training set for next iteration. The main contributions of our work are 1) an active learning based framework for semantic labeling of street view point clouds and 2) a query selection criterion for improving the performance of the active learner in dealing with imbalanced data.

2. RELATED WORK

Semantic labeling of 3D models has been a hot topic in recent years. Lafarge et al. [15] and Liu et al. [16] focus on segmenting 3D meshes only by exploiting the geometry information of the 3D model. The mesh is finally represented as a set of patches clustered according to their geometrical similarities. Armeni et al. [17] use a template based method to parse indoor laser point clouds. Parameterized templates like chairs are used to match each part of the voxelized scene. Semantic
labels are then assigned to the 3D model. Häne et al. [8] and
the extensive works [10, 9] deal with the problem of simulta-
neous 3D reconstruction and segmentation. The semantic
labels are acquired by segmenting images with a pre-trained
decision tree, and then mapped to the 3D space. Finally a
semantic 3D model is generated jointly with depth maps and
label images. Bláha et al. [18] also deal with both semantic
labeling and geometry of the 3D model. But it starts from an
existing model and iteratively refine the label configuration
and the geometry of the 3D model. Valentin et al. [11] and
Rouhani et al. [12] regard the semantic labeling of a mesh as
an energy minimization problem defined on the Conditional
Random Field (CRF). The unary potential is acquired with a
classifier taking both image (or texture) information and geo-
metric information as the input. By minimizing the energy,
the final label configuration of the 3D model is yielded. M-
cCormac et al. [13] propose a method to assign semantic la-
bels to the 3D model generated by a Simultaneous Localiza-
tion and Mapping (SLAM) system. The segmentation of each
video frame is acquired with a pre-trained CNN, and used to
update the distributions of the elements in the 3D model.

While deep learning techniques become more and more
popular, some work [6, 7, 19] start to exploit the potential
of deep learning in directly processing 3D models. Qi et al.
[6] and their extensive work [7] use a Recurrent Neural Net-
work (RNN) based architecture to classify the unstructured
3D point clouds. The network is not complex but the perfor-
ance is quite impressive. Dai et al. [19] propose a method
which takes the voxel grid as the input, and output completed
scenes with semantic labels. The network in [19] is simple
and it is used hierarchically to complete and segment a vox-
elized scene. It performs well on the simulated datasets.

For most of the methods mentioned above, a pre-trained
classifier is necessary, which means a remarkable amount of
annotated data samples is essential. However, this condition is
not always met in real world cases. For those with limited an-
notated data samples, AL is a good solution. Plenty of work
[20, 21, 22, 23, 24, 25, 26] have been published to validate
the effectiveness of AL method in reducing annotation cost.
Zhou et al. [20] propose an AL based method for the classi-
fication of medical images. The criterion for query selection
consists of the entropy and diversity of the candidate images.
Konyushkova et al. [22] propose a method to segment med-
ical images of a volume. The images are oversegmented and
piled up for supervoxels, and then the feature uncertainty and
the geometric uncertainty are used to guide the query selec-
tion process. Zhou et al. [26] propose an AL based method
for semantic labeling of 3D meshes aligned with calibrated
images. The observation uncertainty and the observation di-
vergence are introduced to help select queries to improve the
performance of the CNN and the quality of the semantic mod-
el. In this paper, we follow the pipeline in [26] and focus on
the semantic labeling of 3D models, especially point clouds
of street views.

3. PROPOSED METHOD

The pipeline of the proposed method is shown in Fig. 1. We
follow the pipeline in [26]. The proposed method takes a 3D
point cloud and registered images as the input, and outputs
a semantic point cloud in which each point is attached with
a semantic label. Since a street view point cloud consists of
surface points of various complex objects like trees, it is not
suitable for it to be transformed to a watertight mesh as in
[26]. Therefore, to begin with, we create a 3D voxel grid
over the input cloud and then the 3D model is represented
as a voxel set consisting of voxels in which at least one 3D
point locates. The voxelized representation of the 3D model
reduces the memory cost and makes it convenient to exploit
the pair-wise constraints of the neighboring elements in 3D
space. Within the AL framework, the proposed method iter-
atively performs the following three steps: 1) Fine-tuning a se-
matic segmentation CNN with the ever-enlarging annotated
image set. Then pixel-wise semantic labels for unannotated
images are acquired using the fine-tuned CNN. 2) The label-
s of pixels in all images are back-projected to the 3D voxels
using the calibrated camera parameters. Then a Markov Ran-
dom Field (MRF) optimization is applied to fuse the labels
and give each voxel a single label by taking both 2D semantic
labels and 3D geometry into consideration. 3) After that, we
use the fused semantic 3D model as a supervisor to select sev-
eral query images for annotation. After being annotated, these
images are then incorporated into the training set for next it-
eration. The process of above three steps continues until the
label configuration of the model becomes steady.

3.1. Fine-tuning

In order to obtain a CNN for the segmentation task on a cus-
tomized set of images with limited annotated samples, we
choose to fine-tune other than to train a CNN from scratch.
The initial training set is given manually. And by iteratively
choosing and annotating images from the unannotated image
set, the training set enlarges after each iteration. To improve
the robustness of the CNN, the training set is augmented with traditional data augmentation techniques [27, 28]. In our experiments, we choose DeepLab V2 [29] for image segmentation. To start with, DeepLab V2 is pre-trained on COCO [30]. Then DeepLab V2 is continuously fine-tuned with the ever-enlarging augmented training set. Note that DeepLab V2 could be replaced with other segmentation CNNs if needed.

In street view datasets, the data samples are often imbalanced. Take the Road 1 in ApolloScape [5] as the example, the total amount of class “vegetation” is over 1,000 times of that of the class “traffic cone” in both pixel and voxel level. This could be harmful for the performance of the CNN to classify the image pixels belonging to any classes. To deal with the imbalance, we introduce a label bias factor \( b(b \in [1, +\infty)^{1 \times |C|}) \) in which each element represents the weight of that class due to its rareness. \( b \) is acquired directly from the training set. After the regular data augmentation, we count the number \( n_l \) of samples of each class \( l \). Then we set \( \text{th entry in } b \) as \( \max(1, \frac{n_{\text{mean}}}{\min(n_m, \forall m \in C)}) \), and apply extra data augmentation according to \( b \). The larger the label bias for one class is, the more extra augmented image samples of that class will be.

### 3.2. Semantic Fusion

After predicted by the CNN, all unannotated images get their pixel-wise semantic labels. The labels of the pixels in both annotated images and unannotated images are back-projected pixel-wise semantic labels. The labels of the pixels in both classes respectively form the unary term and the pair-wise term of the Gibbs energy on the MRF. By minimizing the energy, the label configuration of the 3D model will be yielded.

#### 3.2.1. Voxel Likelihood

Similar to that in [26], we estimate the likelihood distribution of each voxel with each entry representing the probability of assigning the label to the voxel. To alleviate the influence of the voxel size, we project the original 3D point cloud to obtain the depth maps for all images. Then the corresponding voxel of each pixel is found via locating the corresponding 3D surface, the normal of each point meaningfully represents the normal of local surface, which could be estimated using PCL\(^1\). Generally, if the normals of two neighbor points have similar directions, the two points should be given the same semantic label; otherwise not. We transfer the constraint to the voxel field by giving each voxel the average normal of the normals of the points located in the voxel. Let \( n_v \) be the normal of voxel \( v \), the geometric constraint of adjacent voxels \((v, u)\) is given in the form of a generalized Potts formulation

\[
V_{v, u}(l_v, l_u) = \begin{cases} 
1 & \text{if } l_v \neq l_u \\
\sin^2(\langle n_v, n_u \rangle) & \text{if } l_v = l_u 
\end{cases} \quad (2)
\]

where \( \langle \ast, \ast \rangle \) means the angle between two vectors. The voxels \( v \) and \( u \) are adjacent voxels, i.e. they should share a face.

#### 3.2.2. Geometric Constraint

As it is put in the beginning of this section, we voxelize the 3D space to utilize the geometric information of the 3D model. Since the point cloud only consists of points on the scene surface, the normal of each point meaningfully represents the normal of local surface, which could be estimated using PCL\(^1\). Generally, if the normals of two neighbor points have similar directions, the two points should be given the same semantic label; otherwise not. We transfer the constraint to the voxel field by giving each voxel the average normal of the normals of the points located in the voxel. Let \( n_v \) be the normal of voxel \( v \), the geometric constraint of adjacent voxels \((v, u)\) is given in the form of a generalized Potts formulation

\[
V_{v, u}(l_v, l_u) = \begin{cases} 
1 & \text{if } l_v \neq l_u \\
\sin^2(\langle n_v, n_u \rangle) & \text{if } l_v = l_u 
\end{cases} \quad (2)
\]

where \( \langle \ast, \ast \rangle \) means the angle between two vectors. The voxels \( v \) and \( u \) are adjacent voxels, i.e. they should share a face.

#### 3.2.3. Energy Minimization

The voxel labeling problem is regarded as an energy minimization problem on a MRF. Similar to [15, 26], the Gibbs energy of the posterior probability distribution of the MRF is

\[
E(l) = \sum_{v \in V} D_v(l_v) + \lambda \sum_{(v, u) \in A} V_{v, u}(l_v, l_u) \quad (3)
\]

where \( V \) is the voxel set; \( A \) is the adjacent voxel set; \( \lambda \) is a constant (\( \lambda = 0.2 \) in our experiments). Here, \( D_v(l_v) = 1 - d_v^{l_v} \) where \( d_v \) is defined in Eq. 1, and \( V_{v, u}(l_v, l_u) \) is defined in Eq. 2. Finally, the energy \( E \) is minimized with the \( \alpha \)-expansion algorithm [31], and a semantic 3D model is generated in which each voxel has a semantic label.

### 3.3. Query Selection

The key issue in an AL method is to select proper queries to boost the performance of the active learner. Since a temporal semantic 3D model is acquired, we could use it to help select proper candidate images for annotation. The reason of this practice is that the semantic labels of the voxels in 3D model are more reliable than those of the pixels in images, as the semantic 3D model is acquired with the semantic fusion process which combines both 2D and 3D information together. In [26], a query selection criterion consisting of observation uncertainty and observation divergence is introduced. The observation uncertainty measures the differences of the semantic labels of objects in images and in the 3D model, and is formulated as

\[
T_{u, I_v} = \sum_{i \in I_v} \sum_{v \in V_i} \Gamma(a_v, i, v, I_v) U_{v, i} \quad (4)
\]

\(^1\)http://pointclouds.org/
where

\[ U_{v,i} = \Psi(o_{v,i}, \hat{d}_v) \]  \hspace{1cm} (5)

\( I_s \) is the query image set; \( a_{v,i} \) is a reliability factor that represents how reliable voxel \( v \) is viewed by image \( i \); \( \Gamma \) guarantees that each voxel contributes to \( T_{U,I_s} \) only once; \( o_{v,i} \) and \( \hat{d}_v \) are the distributions of voxel \( v \) in image \( i \) and in the 3D space; \( \Psi \) is the Total Variation Distance of two distributions. The observation divergence measures the diversity of the semantic labels of objects in images, and is formulated as

\[ T_{D,I_s} = m^T R m \]  \hspace{1cm} (6)

where

\[ R(v,u) = \sum_{i \in I_s} \sum_{j \in I_s} \Gamma(a_{v,i},v,I_s)\Gamma(a_{u,j},u,I_s)\Psi(o_{v,i},o_{u,j}) \]  \hspace{1cm} (7)

\( m \in \{0,1\}^{1|I_s| \times 1} \), in which each entry represents whether a voxel is covered by \( I_s \). The query selection criterion of [26] is given as

\[ \max_{I_s} T_{U,I_s} + \beta \sqrt{T_{D,I_s}} \]  \hspace{1cm} (8)

\[ s.t. I_s \subset I_{a,t} \text{ and } |I_s| = k \]

where \( \beta \) is the balance factor; \( k \) is the batch size.

However, when dealing with general street view datasets containing imbalanced classes, the candidate images selected by [26] tend to ignore the classes with minor data samples (denoted as minor classes), resulting in low performance of label assignments of voxels that truly belong to minor classes. The reason is that the observation uncertainties of the voxels that belong to those classes are often low due to their small reliability factors \( a_{v,i} \). Since \( a_{v,i} \) represents how reliable the voxel \( v \) is viewed by image \( i \), the better the voxel can be viewed by the candidate images, the larger the reliability factor \( a_{v,i} \) of that voxel are. Therefore, to solve this problem, we should increase the reliability factors \( a_{v,i} \) of the voxels belonging to minor classes. In doing so, we apply the label bias factor \( b \) in Section 3.1 to the labels of the pixels that lies in the projection of a voxel. Thus, we acquire the bias weight factor \( b_{v,i} \) of the voxel \( v \) in image \( i \) as

\[ b_{v,i} = \frac{\sum_{p \in \Omega_{v,i}} b_{l_p}}{\sum_{p \in \Omega_{v,i}} 1} \]  \hspace{1cm} (9)

Then we use \( a_{v,i} \cdot b_{v,i} \) as the weight of voxel \( v \) in image \( i \) instead of \( a_{v,i} \) in Eq. 4 & 7. In this way, we adaptively change the weights of each voxel in different views. If the voxel viewed by an image is classified partially or fully as the object of minor classes, the weight of that voxel is increased, thus images containing objects of minor classes could be more likely to be selected.

4. EXPERIMENTAL RESULT

In this section, we show our experimental result on ApolloScape [5] dataset, as to the best of our knowledge, ApolloScape is the only dataset containing both 2D and 3D annotations for street view scenes. We choose Road 1 and Road 2 in the task of Scene Parsing in ApolloScape. Road 1 is a straight one and Road 2 is a loop one. Both of them contain over 10,000 street view images and are about 3km in length.

4.1. Data Preparation

Road 1 and Road 2 contain color images and their corresponding calibration parameters, annotation images and depth maps. We acquire the point cloud and its ground truth semantic labels of each scene by joining the recovered point clouds of all of the images with camera parameters, depth maps and annotation images. The voxel size is set 0.2m × 0.2m × 0.2m, and the total number of voxels is 13.1M and 4.5M in Road 1 and Road 2 respectively. Each voxel is given a label that occurs mostly among those of the points located in the voxel. Since the point cloud contains only static objects, we ignore the pixels classified as moving objects (e.g., car) in semantic fusion process and query selection process.

4.2. Quantitative Evaluation

We set \( \beta = 1 \) and \( k = 8 \) in Eq. 8 and manually give two annotated images to start. The final result of Zhou et al. [26] and our method is shown in Fig. 2 and Table 1. From Fig. 2 we can find that the final result of our method is more accurate than that of [26], especially when dealing with traffic signs. The quantitative result is given in Table 1. Noted that class “sky” and those of moving objects are not presented in the table, as they are not in the point cloud. In Table 1, although the two methods start with the same initial training set, they perform differently in the first iteration because of the label biased data augmentation in the fine-tuning process. From Table 1 we can find that our method rapidly improves the IoUs of all of the classes especially the minor classes. Exceptions are classes like “traffic cone”, as they are too few to be selected. Compare to [26], our method is more capable of assigning semantic labels to the imbalanced street view scenes.

5. CONCLUSION

In this work, we propose an Active Learning based method for semantic labeling of 3D models, especially point clouds of street views. The proposed method takes a point cloud and registrated images as the input and yields a semantic 3D model with limited queries. To deal with the imbalance of object classes in typical street view scenes, we oversample the images with minor classes during the fine-tuning process. And in query selection process, we give adaptive weights to the voxels viewed by selected candidate images according to the image segmentation result. In this way, we iteratively improve the label assignment of the 3D model in terms of both major and minor classes.
Fig. 2. The final result of Zhou et al. [26] and our method on Road 1 and Road 2. (a) color specifications; (b) the sample image and corresponding label image of two scenes; the overview (c) and the models of the two methods (d)(e) and the ground truth (f) on Road 1 (upper) and Road 2 (lower).

Table 1. The result of Zhou et al. [26] and our method on Road 1 and Road 2. Column P means the proportion of each class in 3D model; column 1 - 5 means different iterations; row overall means the overall accuracy which is defined as the ratio of the amount of voxels with correct labels to the total amount of voxels; other numbers means the IoU of each class in voxel level.

<table>
<thead>
<tr>
<th>Class</th>
<th>Zhou et al. [26]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>road</td>
<td>0.10</td>
<td>0.21</td>
</tr>
<tr>
<td>sidewalk</td>
<td>0.21</td>
<td>0.16</td>
</tr>
<tr>
<td>traffic cone</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>fence</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>pole</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>traffic light</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>wall</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>dustbin</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>billboard</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>vegetation</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>overall</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Acknowledgement

This research was funded by National Natural Science Foundation of China grant numbers 61632003, 61873265 and 61573351.

6. REFERENCES


[8] Christian Häne, Christopher Zach, Andrea Cohen, and


