Unsupervised Multimodal Feature Learning for Semantic Image Segmentation

Deli Pei, Huaping Liu, Yulong Liu and Fuchun Sun

Abstract—In this paper, we address the semantic segmentation problem using single-layer networks. This network is used for unsupervised feature learning for the available RGB image and the depth image. A significant contribution of the proposed approach is that the dictionary is selected from the existing samples using the $L_{2,1}$ optimization. Such a dictionary can capture more meaningful representative samples and exploit intrinsic correlation between features from different modalities. The experimental results on the public NYU dataset show that this strategy dramatically improves the classification performance, compared with existing dictionary learning approach. In addition, we perform experimental verification using the practical robot platforms and show promising results.

I. INTRODUCTION

SEMANTIC image segmentation is a fundamental and challenging problem in computer vision, which aims to assign each pixel in an image a pre-defined semantic label. It can be seen as an extension of the traditional object detection which aims at detecting prominent objects in the foreground of an image, with closed relation to some other fundamental computer vision tasks such as image segmentation and image classification. Semantic segmentation has many applications in practice, including scene understanding, robot navigation, and content based image retrieval et al. However, currently most of the work focuses on outdoor environment and only the optic image is considered [1], [2], [3].

Recently, the depth camera received great attentions from industrial to academic fields. Microsoft provided a popular mass-production consumer electronics device Kinect, which captures RGB images along with corresponding depth information. Such RGB-D cameras have been extensively used for human body pose estimation [4], object recognition [5], and 3D modeling and interaction [6]. In addition, Henry et al [19] investigated how such cameras can be used in the context of robotics, specifically for building dense 3D maps of indoor environments. In [17], A related work also showed promising results labeling 3D point clouds from merged Kinect frames.

RGB-D cameras also received great attention from the community of semantic image segmentation. Ref. [9] adopted Kinect for RGB-D scene labeling and provided a new NYU dataset containing 13 semantic categories over 7 scene types with encouraging results by combining SIFT features and 3D prior in Markov Random Fields (MRF). Ren et al [14] developed a scene labeling approach that combines rich RGB-D features and contextual models using MRFs and hierarchical segmentation. Both work focus on utilizing context information to improve the performance but neglects the intrinsic relation between the low-level features of different modal information.

On the other hand, much recent work in machine learning has focused on learning good feature representations from unlabeled input data for higher-level tasks such as classification and recognition. Current solutions typically learn multi-level representations by greedily "pre-training" several layers of features, one layer at a time, using an unsupervised learning algorithm [7], [8]. Such learning methods have been extended to multimodal learning to explore complementary information from different modalities [18], [15], [16]. In a multimodal setting, data consists of multiple input modalities, each modality having a different kind of representation and correlative structure. Joint representations can be learned by fusing data from different modalities, which capture complementary information of objects and scenes.

In this work, we propose an unsupervised multimodal feature learning algorithm for semantic image segmentation. Different from traditional dictionary learning by K-means clustering, we select representative samples from actual data points by $L_{2,1}$ optimization of coefficient matrix. Moreover, by choosing same data points for multiple modalities as dictionary atoms, we are able to exploit intrinsic correlation between different modalities and result in more representative features. Fig. 1 describes the framework of the proposed method.
The contributions of this paper are summarized as follows:

1) An unsupervised multimodal feature learning method using single-layer networks is developed to effectively fuse the RGB and depth information. Different from traditional approaches which learn dictionaries separately for each modality, we propose a joint dictionary selection method to select same actual data points for different modalities to explore intrinsic correlation.

2) A hierarchical dictionary selection schema is proposed to improve the efficiency of traditional dictionary selection on large scale training set.

3) The proposed unsupervised multimodal feature, combined with linear classifier, is used for semantic image segmentation, where RGB image and depth information are available for each frame.

4) We utilized the developed semantic image segmentation on a robotic experiments with a mobile robot equipped with a Kinect camera.

The organization of this paper is as follows: We first briefly review the single-layer networks framework and detail our proposed dictionary selection model in section 2. Section 3 shows the experimental results on NYU Depth dataset and we conclude our work in Section 4.

II. UNSUPERVISED FEATURE LEARNING

We start by introducing a typical single-layer network, which is more efficient compared with multi-layer networks. As shown in [10], single-layer networks can achieve highly comparable performance even the state-of-the-art results when the parameters are deliberately tuned. Then we present our unsupervised multimodal feature learning approach based on single-layer networks.

A. Single-layer Networks

To learn the feature representation, we first collect densely sampled patches with certain overlap from training images, \( X = [x_1, ..., x_N] \), where \( x_i \in \mathbb{R}^d \) is the vector of raw pixel values of patch \( i \) and \( N \) is the number of training patches. Then some preprocessing steps such as normalization and whitening are applied to patch vectors before feature mapping.

Several different unsupervised learning algorithms are considered in recent literature, including sparse auto-encoder, sparse restricted Boltzmann machine, K-means clustering and Gaussian mixture. As shown in [10], K-means clustering with triangle activation function achieves best performance with deliberately chosen parameters in some image classification tasks, which is also adopted in our paper as a baseline method.

Given the dictionary learned by K-means clustering, \( C = [c_1, ..., c_K] \) where \( c_i \in \mathbb{R}^d \) and \( K \) is the size of dictionary, we compute the new feature vector from raw feature \( x \) by the following triangle activation function:

\[
    f_i(x) = \max\{0, \mu(x) - \|x - c_i\|_2\},
\]

where \( \mu(x) = \frac{1}{K} \sum_{i=1}^K \|x - c_i\|_2 \) is the mean of distances between \( x \) and \( K \) centroids.

The above procedure can be represented by a single-layer network. In this network, the hidden units correspond to the dictionary atoms, therefore the dictionary plays a very important role in unsupervised feature learning. In the following section, we will detail the proposed dictionary learning method.

Moreover, to fuse features from different modalities (e.g. RGB and depth), traditional feature fusion frameworks learn separate dictionaries for each modalities [16] and simply concatenate them together to form the final representation. This fusion strategy, although straightforward, ignores the intrinsic correlation between different modalities.

B. Dictionary Selection

In this section, we present the proposed multimodal dictionary selection algorithm. Instead of computing representative centroids by clustering, which strongly depends on initialization, we adopt Sparse Modeling Representative Selection (SMRS) algorithm [12] to find representative samples in training set as dictionary for sparse coding. Moreover, to explore the intrinsic correlation among different modalities, we assume the dictionary atoms from different modalities are correspond to each other in contrast to the independent dictionaries learned by K-means. That is to say, dictionary atoms of different modalities are extracted from patches of same location with different raw pixel representation, e.g. RGB values and depth values. This allows us to select more consistent representative samples (dictionary atoms) for feature fusion.

Considering a set of patches randomly sampled from training frames, their raw RGB features are \( R = [r_1, ..., r_N] \) where \( r_i \in \mathbb{R}^{M \times N} \) is the feature vector of \( M \times N \) patch \( i \) after pre-processing, and \( D = [d_1, ..., d_n] \) are the depth feature where \( d_i \in \mathbb{R}^{M \times N} \) is the depth vector of patch \( i \) after pre-processing. The objective of finding representative samples from actual data points can be formulated as follows:

\[
    \min_{S} \lambda \|S\|_{1,2} + \frac{1}{2} \|X - XS\|_F^2
\]

\[
\text{s.t. } 1^T S = 1^T.
\]

where \( X = \begin{bmatrix} R \\ D \end{bmatrix} = [x_1, ..., x_n] \) is the joint raw feature representation by concatenating RGB feature and depth feature, \( S \in \mathbb{R}^{N \times N} \) is the sparse coefficient matrix for matrix reconstruction and \( \|S\|_{1,2} = \sum_{i=1}^N \|s_i\|_2 \) is the sum of \( \ell_2 \) norms of the rows of \( S \), and \( \lambda \) is the regularization parameter and the value is typically in [2, 50].

By minimizing the reconstruction error of each patch as a linear combination of all the other data, representative samples can be chosen by sorting the \( \ell_2 \) norms of row \( s_i \) in coefficient matrix \( S \), which is the weight of patch \( x_i \), proportional to the contribution in reconstructing other patches. The \( L_{2,1} \) norm constraint is introduced to avoid trivial solution and the affine constraint \( 1^T S = 1^T \) is used to enforce global translation invariant.
Theoretically, this problem can be solved directly by Alternating Direction Method of Multipliers (ADMM) optimization framework [12], [13]. However, due to the heavy computation load of the algorithm, it is not feasible to solve this problem directly on large scale training set, e.g. over 100,000 training samples, which is necessary to learn a dictionary of thousands atoms. Here we propose a hierarchical dictionary selection scheme to solve this problem while obtaining a sub-optimal solution. By dividing the training set into groups, representative samples for each group are selected by solving this sub-question. Then the representative sample of groups are then put together to select representative samples for the whole training set.

Given the optimal solution \( \mathbf{S}^* \), we rank the training patches by \( \ell_2 \) norms of their corresponding rows where

\[
\| \mathbf{s}^{(1)} \|_2 \geq \| \mathbf{s}^{(2)} \|_2 \geq \ldots \geq \| \mathbf{s}^{(K)} \|_2. \tag{3}
\]

Then we can select the first \( K \) patch vectors as representative samples as dictionary atoms. Since each patch vector consists of two parts, we now decompose the representative samples into two dictionaries, \( \mathbf{C}^{RGB} = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_K] \) for RGB features and \( \mathbf{C}^{D_p} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K] \) for depth features.

C. Sparse Coding

In contrast to joint dictionary selection, we compute the sparse representation separately for each modality. Following the approach in [10], we use triangle activation function to map the raw pixel feature to the new feature vector:

\[
\begin{align*}
f_k (r) &= \max \{ 0, \mu (r) - \| r - r_i \|_2 \} \\
g_k (d) &= \max \{ 0, \mu (d) - \| d - d_i \|_2 \}
\end{align*} \tag{4}
\]

where \( \mu (r) = \frac{1}{K} \sum_{j=1}^{K} \| r - r_j \|_2 \) is the mean of distances between \( r \) and \( K \) dictionary atoms of RGB feature and \( \mu (d) = \frac{1}{K} \sum_{j=1}^{K} \| d - d_j \|_2 \) is the mean of distances between \( d \) and \( K \) dictionary atoms of depth feature.

Given the raw feature representation of patch \( x = [r^\top d^\top]^\top \), the unsupervised feature can be computed as following:

\[
\mathbf{y} = \mathbf{F}(\mathbf{x}) = [f_1 (r), ..., f_K (r), g_1 (d), ..., g_K (d)]^\top. \tag{5}
\]

Although we use the same strategy to concatenate feature vector into final feature, our feature is more consistent among different modalities because of our proposed dictionary selection, which explores their intrinsic correlations by additional location constraint.

III. SEMANTIC IMAGE SEGMENTATION

In this section, we apply the learned multimodal feature in semantic image segmentation task, and the framework is shown in Fig. 1.

A. Region Feature

Image segmentation is applied to get initial regions first and their labels are predicted instead of labeling pixel directly. Although using pixels as labeling unit is very straightforward and does not involve extra efforts, pixel itself contains limited and ambiguous information that cannot always be discriminative enough to determine its correct label. On the contrary, features of regions generated by unsupervised segmentation algorithms are not only more informative but also more robust on noise or other variance.

To obtain aligned feature representation of regions, we first decompose the image into patches with overlap and learn unsupervised feature for each patch in Equation (5), and then we use max pooling to summarize the patch features within a region and obtain the feature representation for the region. Fig. 2 describes the above procedure.

B. Region Prediction

To predict the labels of regions, the linear multi-class SVM in Liblinear package [20] is adopted due to its advantages in speed and good performance.

In the case of multiple classes, we use a one-versus-all strategy to train \( T \) binary classifiers. Given the training data \( \mathbf{X} = \{ \mathbf{x}_i, i = 1, ..., N \} \) and their corresponding labels \( Q = \{ q_i, i = 1, ..., N \} \), we solve the following unconstrained convex optimization problem for each class \( p \in \{ 1, ..., T \} \):

\[
\min_{p} \frac{1}{2} \omega_p^\top \omega_p + C \sum_i \log(1 + \exp(-l_{i,p}^p \omega_p^\top \mathbf{x}_i)), \tag{6}
\]

where \( \omega_p \) is the classifier parameter for class \( p \), \( l_{i,p}^p \) is the binary class label for \( \mathbf{x}_i \), and \( l_{i,p}^p = 1 \) if \( q_i = p \), otherwise \( l_{i,p}^p = -1 \). The cost parameter \( C \) is chosen manually where a larger \( C \) corresponds to the assignment of higher penalties to errors.

In the predicting stage, given a region feature \( \mathbf{x} \), its class label is predicted by

\[
l = \max \omega_p^\top \mathbf{x}. \tag{7}
\]

That is to say, we assign the class with highest probability to that region.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of our proposed method, we compare the performance of different multimodal feature representations computed by different dictionary learning strategies on NYU depth dataset [1], and discuss the effect of
\[ q_i = \frac{1}{T} \sum_{j=1}^{N} u_j \delta(q_i, q_j) \]
Quantitative Results of different features. The sizes of dictionaries learned by K-means and our method are 3000.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>40.1 ± 3.0</td>
<td>42.3 ± 2.4</td>
<td>43.8 ± 2.6</td>
</tr>
<tr>
<td>Depth</td>
<td>39.3 ± 2.2</td>
<td>40.1 ± 2.0</td>
<td>43.2 ± 1.2</td>
</tr>
<tr>
<td>Depth+RGB</td>
<td>45.8 ± 2.6</td>
<td>47.8 ± 2.1</td>
<td>50.5 ± 2.4</td>
</tr>
</tbody>
</table>

D. Application on Robot

We evaluate the proposed method on the practical PowerBot mobile robot platform which is equipped with a Kinect camera (see Fig.7 (a)). We teleoperate the mobile robot to walk around the laboratory and collect the RGB-D frame pairs. For each frame, we run the proposed method to predict region labels and some selected results are shown in Fig. 7 (b). It can be seen that we correctly labelled supporting objects such as wall and floor as well as foreground objects including sofa and picture, which shows a very promising results for robot navigation.

V. CONCLUSION

We present an unsupervised multimodal feature learning algorithm for semantic image segmentation. Different from traditional dictionary learning algorithms with K-means clustering, we select representative samples from actual data points as dictionary atoms, and these data points are identical for multiple modalities. This allows us to exploit intrinsic correlation between different modalities, and therefore obtain more powerful dictionaries for semantic image segmentation. Experimental results on public NYU Depth dataset show that our proposed unsupervised feature outperform the feature computed by traditional dictionary learning with K-means clustering.

For the future work, it would be very interesting to incorporate contextual information extracted multiple modalities in semantic image segmentation.

VI. ACKNOWLEDGEMENTS

This work is jointly supported by the National Key Project for Basic Research of China (2013CB329403), the National Natural Science Foundation of China (Grants No: 61075027, 91120011, 61210013) and the Tsinghua Self-innovation Project (Grant No:20111081111).
REFERENCES


