Multi-Modal Compositional Object Representation

February 13, 2015

1 Introduction

The objective of this task is to represent an object with visual and non-visual features such as haptics features. The multi-modal representation will allow us to infer visual information based on haptics information and vice-verse. In this context, we considered multi-modal object recognition problem. The non-visual features then help in object recognition in the lack of visual observations. We have developed a 3D compositional part-based object representation and as the non-visual features we make use of tactile information after grasping an object.

2 3D Compositional Object Model

In this section, we briefly explain our part-based compositional object model. We learn our model based on view-based RGB-D pointclouds. We have a bottom-up compositional model which starts from the object points and reaches to the object category level. Our model consists of three levels: patches, parts and objects as shown in Figure 1. There is two main concepts in each level, composition and representation. More precisely, each level is composed of the elements of the lower levels, but its representation might be different than its individual composed low level elements. For representation, we extract features in each level based on its composed elements.

From Points to Patches Initially, the points form low-level patches. As can be seen in Figure 1, they are denoted by $\mathbf{q}$. These patches construct the lowest level of our compositional model. We extract scale-invariant features based on surface curvatures from the patches, and represent each patch with the extracted feature. So, a patch is represented by the extracted feature but is composed of the input points. We then perform clustering based on the extracted feature and we obtain a set of patch clusters $Q = \{q_1, q_2, \ldots, q_n\}$ which gives us the patch types in our dataset. Then $p(\mathbf{q}|q)$ denotes the observation probability for patch $\mathbf{q}$ based on patch type q.

From Patches to Parts Object parts, as shown in Figure 1 with $\mathbf{r}$, are composed of the low-level patches. We represent each object part as a histogram of the patch types which compose it. That is we compute the observation likelihood for all the observed patches which belong to a part by computing $p(\mathbf{q}|q)$. A patch can be assigned to multiple patch types. We then make a normalized histogram based on them. We then perform clustering based on our object parts histograms to obtain different part types $R = \{r_1, r_2, \ldots, r_n\}$.

From Parts to Object An object is composed of a set of parts in a specific configuration. Therefore, our objective is to collect two information: 1) occurrence of a certain part for an object class, 2) connection between the parts.

For the first information, we segment an object to its constitutes parts $\mathbf{R} = \{\mathbf{r_1}, \ldots, \mathbf{r_n}\}$ and we recognize the parts based on the part types. That is we compute $\forall \mathbf{r} \in \mathbf{R},\ r \in R\ p(\mathbf{r}|r)$ and a part can be matched to more than one type. From the obtained part types, we collect statistics of occurrence of the part types for a certain object class $p(r_1, \ldots, r_{N_R}|o)$.

The second information cannot be obtained at the part level. The reason is that we represented a part as a histogram of patch types. This representation is not local enough to give us the connectivity information. In order to overcome this problem, we make use of the compositional nature of our model, and we use the patches which consists the parts. However, we do not use all the constitutes patches, but only the adjacent bounding patches between the pair of adjacent parts. We then recognize these adjacent bounding patches based on our learned patch types and we collect statistics of their co-occurrence for a certain object class. That is we compute $p(q_1, q_2|o)$ for all the patch pairs which are adjacent but belong to the different parts.

3 Multi-Modal Representation

The objective in multi-modal representation is to represent an object with tactile and visual information. We considered multi-modal object recognition problem. In Section 2, we noted that an object is represented by its parts and parts connectivites. For recognition purpose, we gathered information based on the probability of occurrence of certain *part types* for an object class and the adjacent bounding *patch types*. In multi-modal representation for object recognition, the tactile and visual information are joint together. Since, we perform recognition in the part level, we add tactile information in the part level as well. Hence, the tactile information is added at the part level of our compositional model.

The visual information comes from our compositional model and tactile information from grasping. We perform multiple grasps on object parts (from our *IKEA* kitchen object dataset) and we obtain tactile features from that. The first step is to quantize the tactile features. To this aim, we perform an agglomerative clustering based on all the extracted tactile features from all the objects parts. This form our tactile feature types $T = \{t_1, \ldots, t_m\}$

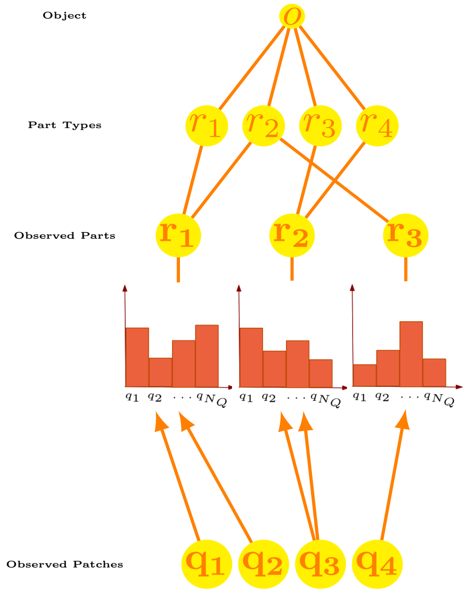


Figure 1: A schematic diagram for the compositional model for object class o. For each object, the observed patches $\mathbf{Q} = \{\mathbf{q_1}, \ldots, \mathbf{q_4}\}$ are recognized to patch types $Q = \{q_1, \ldots, q_{N_Q}\}$. Each patch belong to only one object part. An object part is represented as a histogram of patch types based on its constitutes patches. An Object from class o consists of a set of parts $\mathbf{R} = \{\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}\}$ which are recognized to the part types $R = \{r_1, \ldots, r_4\}$. For object recognition we learn the $p(r_1, \ldots, r_{N_R}|o)$ as well as parts connectivity based on adjacent bounding patches $p(q_1, q_2|o)$.

The second step is to associate tactile features to object parts. From our robotic part-based grasping experience, we have three information. 1) We know the object label t. 2) Thanks to our compositional representation, the grasped part is matched to the respective part type(s) $\{r_1, \ldots, r_m\}$. 3) The obtained tactile feature is matched to its corresponding tactile feature type(s) $\{t_1, \ldots, t_n\}$. We then collect the joint tactile-part co-occurrence for object class o, that is $p(r, t|o)$. We use this information instead of a part only information $p(r|o)$ in object recognition task.

4 Object Recognition Using Visual-Tactile Features

In this section, we explain the probabilistic object recognition formulation based on tactile and visual features. We extract tactile features from the tactile readings and the visual features from the object parts. As described in Section 2, the feature for each part is a histogram of its constitutes patch types. It should be noted that the same formulation can be applied for instance-level object recognition.

As mentioned in Section 2, for object recognition we are interested in the parts which compose an object and the connectivity between them. Therefore the input data to our object recognition system is the part segmented object. Since, the parts are composed of patches and the patches from points, we also have adjacency information between the object parts. We then use the parts and their adjacency information for object classification task.

We are interested in recognizing class o for the observed parts $\mathbf{R}$ in the configuration $\mathbf{S}$. $\mathbf{S}$ is a binary adjacency matrix in terms of the neighborhood between each two part. More precisely, each row in $\mathbf{S}$ indicates the adjacent parts for a certain object part. In addition, we perform multiple grasps on each object part and extract tactile feature from them. We denote these tactile features with $\mathbf{T}$. In the following formulation, we indicate each individual part cluster with r which belongs to a set R of all the parts clusters. Likewise, each individual patch cluster is shown with q which belongs to a set of patch clusters Q. Finally, each tactile cluster is written as t which belongs to the set T of all the tactile clusters.

$$o_{\mathbf{R},\mathbf{S},\mathbf{T}} = \underset{o}{\operatorname{argmax}}\, p(\mathbf{R},\mathbf{S},\mathbf{T}|o) \tag{1}$$

having an independent assumption between the object parts, the above conditional probability can be re-written as follows,

$$p(\mathbf{R},\mathbf{S},\mathbf{T}|o) = \prod_{\mathbf{i}} p(\mathbf{r_i},\mathbf{t_i}|\mathbf{s_i},o)p(\mathbf{s_i}|o) \tag{2}$$

We consider a uniform probability distribution for $p(\mathbf{s_i}|o) = \frac{1}{N_\mathbf{S}}$, where $N_\mathbf{S}$ is the number of all the configurations.

$$p(\mathbf{R},\mathbf{S},\mathbf{T}|o) = \prod_{\mathbf{i}} \frac{p(\mathbf{r_i},\mathbf{t_i}|\mathbf{s_i},o)}{N_\mathbf{S}} \tag{3}$$

$\mathbf{s_i}$ is the configuration or adjacency vector for the part $\mathbf{r_i}$. Hence, it contains all the adjacent parts to $\mathbf{r_i}$. Therefore, we marginalize over all the parts that vector $\mathbf{s_i}$ gives us,

$$p(\mathbf{r_i},\mathbf{t_i}|\mathbf{s_i},o) = \sum_{\mathbf{r}\in\mathbf{s_i}} p(\mathbf{r_i},\mathbf{t_i},\mathbf{t}|\mathbf{s_i},o) \tag{4}$$

We compute the conditional probability of the pair parts $p(\mathbf{r_i},\mathbf{r},\mathbf{t_i}|\mathbf{s_i},o)$ independent of the given configuration $\mathbf{s_i}$. Hence, we drop it from Eqn 4,

$$p(\mathbf{r_i},\mathbf{t_i}|\mathbf{s_i},o) = \sum_{\mathbf{r}\in\mathbf{s_i}} p(\mathbf{r_i},\mathbf{r},\mathbf{t_i}|o) \tag{5}$$

In order to compute the conditional probability of the adjacent patches $\mathbf{r_i}$ and $\mathbf{r}$, we make use of their adjacent boundary patches. Patch $\mathbf{q_j}$ belong to the part $\mathbf{r_i}$ and is the neighbor of the part $\mathbf{r_j}$ as denoted by $N_\mathbf{r}$ in the following equation. Therefore, we marginalize over these patches,

$$p(\mathbf{r_i}, \mathbf{r}, \mathbf{t_i}|o) = \sum_{\mathbf{q_j} \in \mathbf{r_i}, \mathbf{q_j} \in N_\mathbf{r}} p(\mathbf{r_i}, \mathbf{r}, \mathbf{t_i}|o, \mathbf{q_j}) p(\mathbf{q_j}|o) \tag{6}$$

We can then factorize the first term in the summation of Eqn 6,

$$p(\mathbf{r_i}, \mathbf{r}, \mathbf{t_i}|o) = \sum_{\mathbf{q_j} \in \mathbf{r_i}, \mathbf{q_j} \in N_\mathbf{r}} p(\mathbf{r}|o, \mathbf{q_j}, \mathbf{r_i}, \mathbf{t_i}) p(\mathbf{r_i}, \mathbf{t_i}|o, \mathbf{q_j}) p(\mathbf{q_j}|o) \tag{7}$$

We consider that the conditional probability of $\mathbf{r_i}, \mathbf{t_i}$ is independent of $\mathbf{q_j}$ when the object class is given,

$$\begin{aligned}
p(\mathbf{r_i}, \mathbf{t_i}|o, \mathbf{q_j}) &= \frac{p(\mathbf{r_i}, \mathbf{t_i}, o, \mathbf{q_j})}{p(o, \mathbf{q_j})} &(8)\\
&= \frac{p(\mathbf{r_i}, \mathbf{t_i}, \mathbf{q_j}|o) p(o)}{p(o, \mathbf{q_j})} &(9)\\
&= \frac{p(\mathbf{r_i}, \mathbf{t_i}|o) p(\mathbf{q_j}|o) p(o)}{p(o, \mathbf{q_j})} &(10)\\
&= p(\mathbf{r_i}, \mathbf{t_i}|o) &(11)
\end{aligned}$$

Furthermore, we consider that the conditional probability of a part $\mathbf{r}$ is independent of the tactile features of $\mathbf{r_i}$. Therefore, we obtain the following formulation for computing the conditional probability of the two adjacent part,

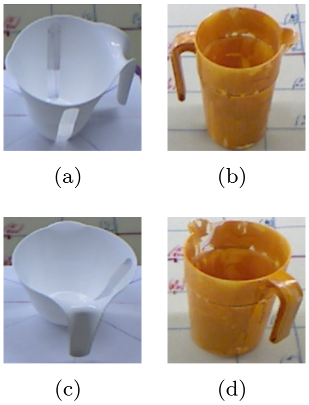
$$p(\mathbf{r_i}, \mathbf{r}, \mathbf{t_i}|o) = \sum_{\mathbf{q_j} \in \mathbf{r_i}, \mathbf{q_j} \in N_\mathbf{r}} p(\mathbf{r}|o, \mathbf{q_j}, \mathbf{r_i}) p(\mathbf{r_i}, \mathbf{t_i}|o) p(\mathbf{q_j}|o) \tag{12}$$

The only difference with our probabilistic vision based formulation is the computation of the joint probability of a part $\mathbf{r_i}$ and its tactile feature $\mathbf{t_i}$. So, we only derive that here. In order to compute this probability we marginalize over our parts clusters r,

$$p(\mathbf{r_i}, \mathbf{t_i}|o) = \sum_r p(\mathbf{r_i}, \mathbf{t_i}|o, r) p(r|o) \tag{13}$$

In order to compute $p(\mathbf{r_i}, \mathbf{t_i}|o, r)$, we need to make use of the object-part-tactile co-occurrence table that we collected in the training phase. That is $p(r, t|o)$ where t is the tactile feature type and r is the part type which $\mathbf{t_i}$ and $\mathbf{r_i}$ can be matched to them. To this aim, we marginalize over all the tactile clusters t,

$$\begin{aligned}
p(\mathbf{r_i}, \mathbf{t_i}|r, o) &= \sum_t p(\mathbf{r_i}, \mathbf{t_i}, t|r, o)\\
&= \sum_t \frac{p(\mathbf{r_i}|\mathbf{t_i}, r, t, o) p(\mathbf{t_i}|r, t, o) p(r, t|o)}{p(r|o)}
\end{aligned} \tag{14}$$



(a) (b)

(c) (d)

Figure 2: Figures 2(a), 2(b) show an example of object recognition problem with the fixed view-point and scale-invariant features. The difference between two objects can only be perceived when the tactile features of their parts are considered.

We compute the conditional probability of a part $\mathbf{r_i}$ given a part type r independent of its tactile information (tactile feature and tactile cluster). In the same way, we compute the conditional probability of a tactile feature $\mathbf{t_i}$ given a tactile type t independent of the visual information. Therefore, we get, We integrate Eqn 14 into Eqn 13 which yields,

$$p(\mathbf{r_i}, \mathbf{t_i}|o) = \sum_r \sum_t p(\mathbf{r_i}|r)p(\mathbf{t_i}|t)p(r,t|o) \tag{15}$$

The conditional probability of $\mathbf{r_i}$ based on a part cluster r denotes simply the observation likelihood for clusters r. This probability is zero if $\mathbf{r_i}$ cannot be matched to r otherwise this is the inverse of the number of clusters that $\mathbf{r_i}$ can be matched to them $NC_{\mathbf{r_i}}$. For each cluster we compute its centerμ_r and we keep a distance threshold. Then, matching to a cluster means if the distance between the cluster center μ_r and the part $\mathbf{r_i}$ is less than the determined threshold. This is our simplified assumption at the moment, in the future we will use Gaussian Mixture Models for assigning the observation likelihood,

$$p(\mathbf{r_i}|r) = \begin{cases} 0, & \text{if } dist(\mathbf{r_i}, \mu_r) > t \\ \frac{1}{NC_{\mathbf{r_i}}}, & \text{otherwise} \end{cases} \tag{16}$$

The conditional probability of $\mathbf{t_i}$ given tactile cluster t indicates if the tactile feature $\mathbf{t_i}$ can be matched to cluster t. We compute $p(\mathbf{t_i}|t)$ as we computed in Eqn 16. $p(r,t|o)$ is computed from the object-part-tactile co-occurrence table. We now, integrate this joint multi-modal probability into our object class conditional probability. Skipping the detail of the formulation, we finally get Eqn 4 and we integrate this joint probability into this final formulation,

$$p(\mathbf{R},\mathbf{S},\mathbf{T}|o) = \prod_{\mathbf{i}} \frac{1}{N_{\mathbf{S}}} p(\mathbf{r_i},\mathbf{t_i}|o) \sum_{\mathbf{r}} p(\mathbf{r}|o) \sum_{\mathbf{q_j}\in\mathbf{r_i},\mathbf{q_j}\in N(\mathbf{r})} \sum_{\mathbf{q_k}\in\mathbf{r},\mathbf{q_k}\in N(\mathbf{q_j})}$$

$$\sum_{q_1\in Q} \sum_{q_2\in Q_2} p(\mathbf{q_k}|q_2) p(q_2,q_1|o) p(\mathbf{q_j}|q_1)$$

This multi-modal representation help us when the visual features are not enough to distinguish object classes or object instances. That is due to the fact that we have a view-dependent acquisition. And some object parts or patches are not clearly visible from one view-point. An example is given in Figure 2. The two different objects in Figure 2(a), 2(b) seem visually similar, especially the handles look similar from one view-point. As shown in Figure 2(c), 2(d), their difference in the shape of the handles can be seen from another view-point. There is the use case of the tactile features which can discriminate between these two objects without changing the view-point. In addition to that, our visual features are only dependent to the curvature of object surface and are invariant to scale. These shortcomings can be handled by integrating tactile features.