Unsupervised Learning of Dictionaries of Hierarchical Compositional Models

Jifeng Dai[1,4], Yi Hong[2], Wenze Hu[3], Song-Chun Zhu[4], and Ying Nian Wu[4]
[1]Tsinghua University, China
daijifeng001@gmail.com
[2] WalmartLab, USA
yihongucla@gmail.com
[3]Google Inc, USA
wenzehu@google.com
[4]University of California, Los Angeles (UCLA), USA
{sczhu, ywu}@stat.ucla.edu

Abstract

This paper proposes an unsupervised method for learning dictionaries of hierarchical compositional models for representing natural images. Each model is in the form of a template that consists of a small group of part templates that are allowed to shift their locations and orientations relative to each other, and each part template is in turn a composition of Gabor wavelets that are also allowed to shift their locations and orientations relative to each other. Given a set of unannotated training images, a dictionary of such hierarchical templates are learned so that each training image can be represented by a small number of templates that are spatially translated, rotated and scaled versions of the templates in the learned dictionary. The learning algorithm iterates between the following two steps: (1) Image encoding by a template matching pursuit process that involves a bottom-up template matching sub-process and a top-down template localization sub-process. (2) Dictionary re-learning by a shared matching pursuit process. Experimental results show that the proposed approach is capable of learning meaningful templates, and the learned templates are useful for tasks such as domain adaption and image cosegmentation.

1. Introduction

Motivation. Learning dictionaries of representational units for describing natural images is one of the most fundamental problems in vision. One of the most successful framework for solving this problem is sparse coding [19, 20]. By enforcing sparsity of the coefficients in the linear representations of natural image patches, a dictionary of Gabor-like basis functions can be learned, so that each training image patch can be represented or approximated by a linear combination of a small number of basis functions selected from the dictionary. An important question then is how to continue to learn higher-level representational units based on the learned basis functions. Such representation units may capture frequently occurring patterns that are more specific about the objects and scenes. They may lead to sparser representations, and they may be useful for object recognition and scene understanding.

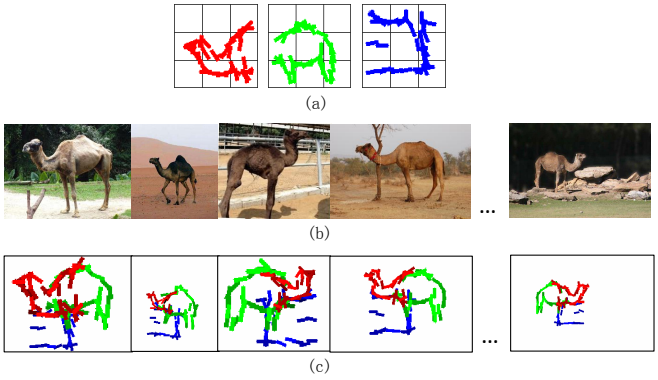

Figure 1: (a) Three hierarchical compositional templates learned from 24 camel images without annotation. Each of the templates is composed of 9 part templates, and each part template is composed of a set of Gabor wavelets. Each Gabor wavelet is illustrated by a bar at the same location and orientation and with the same length as that wavelet. (b) 5 samples of the input images. (c) Encoding of the sample images by the activated deformed templates.

Overview of our method. Our goal is to develop an unsupervised method for learning such representational units, based on the simple observation that the basis functions

(Gabor wavelets) selected for representing the training images form various compositional patterns, and the composition can be hierarchical. Thus we propose to model the representational units as hierarchical compositions of Gabor wavelets. Fig. 1 illustrates the basic idea. A small dictionary of three hierarchical compositional templates (head, body and leg) are learned from a set of input images of camels that are not a priori aligned or annotated. Each hierarchical compositional template is composed of a group of part templates that are allowed to shift their locations and orientations relative to each other. Each part template is in turn composed of a group of Gabor wavelets (illustrated by bars in Fig. 1) that are allowed to shift their locations and orientations relative to each other. These templates capture distinct and specific image patterns and the same template is matched to similar image patches in different images by spatial translations, rotations, scalings and deformations. We focus on unsupervised learning of dictionaries of such templates.

Experiments and performances. We have tested the proposed approach for image representation, domain transfer and cosegmentation. Experimental results show that the proposed approach is capable of learning meaningful representational units. Higher accuracies than previous approaches are achieved on the four domain benchmark [22] for domain transfer, and on the ImageNet [3] dataset for cosegmentation.

Prior work. Hierarchical compositional models are very popular for modeling patterns of objects, see [5, 23, 31, 27, 6, 30, 25] for some examples. Many existing approaches to learning hierarchical compositional models are usually supervised where the object bounding boxes are given [4, 30] or weakly supervised where images are labeled and roughly aligned [5, 25]. In this paper, we learn dictionaries of hierarchical compositional templates from unaligned natural images without annotations, which is more challenging. In comparison to our past work, [11] is concerned with learning templates with only one layer of deformations, while [25] is concerned with learning a single template instead of learning a dictionary of multiple templates.

Our work bears some similarities to [5, 31], which seek to organize the compositions of Gabor wavelets or edgelets into hierarchical structures. The hierarchical structures in [5, 31] are learned layer-by-layer in a bottom-up manner. Once the lower layers are learned, they are fixed in the subsequent learning of higher layers. In our iterative learning algorithm, the part templets are re-learned and the Gabor wavelets are re-selected in each iteration, so the learning is more top-down than bottom-up. Please refer to Fig. 3 for the iterative learning process.

Our work is also related to [1, 17, 15, 11, 26, 16, 18], where repeated patterns are learned from the input images. In [26, 16], a set of HOG templates are learned from multiple input images of the same object category. In [18], recurring tuples of visual words are extracted from single image with repetitive patterns. Unlike our method, they learn simple rigid templates from the images. Our method learns a dictionary of hierarchical deformable templates, which are more flexible.

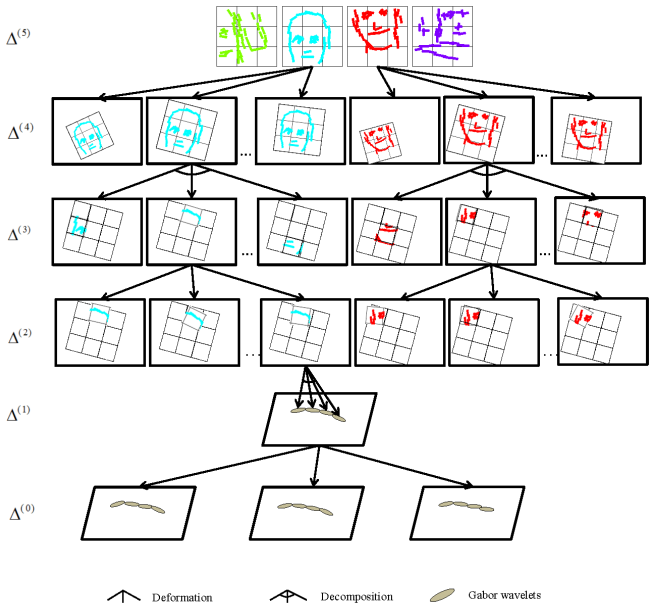

Figure 2: Visual elements in the layered dictionaries, which form a hierarchical compositional structure.

2. Hierarchical Compositional Model

We represent an image $\mathbf{I}$ using a dictionary of hierarchical compositional templates. As shown in Fig. 2, each hierarchical compositional template is decomposed into a set of shiftable part templates, and each part template is further decomposed into a set of shiftable Gabor wavelets. The representational units at different layers of the model can be organized in layered dictionaries $\Delta^{(j)}$. Table 1 defines the dictionaries and elements, their parameters and the allowed ranges of values, which we shall elaborate in the following.

2.1. Layered dictionaries

$\Delta^{(5)}$ is the dictionary of hierarchical compositional templates

$$\Delta^{(5)} = \{H^{(t)}, t = 1, ..., T\}. \quad (1)$$

The hierarchical templates capture the frequently occurring patterns in the input images.

$\Delta^{(4)}$ contains the spatially translated, rotated and scaled versions of the hierarchical compositional templates in $\Delta^{(5)}$ for image representation. For an image $\mathbf{I}$, we encode it by

Layer ID	Template type	Parameters	Deformation Range	Template Size
$\Delta^{(5)}$, $\Delta^{(4)}$	Hierarchical compositional template	Activated template t_k $\widetilde{L_k}$ = (position $\widetilde{X_k}$, scale $\widetilde{S_k}$, orientation $\widetilde{O_k}$)	$\widetilde{X_k} \in$ image domain Λ $\widetilde{S_k} = \{0.8, 1, 1.2\}$ $\widetilde{O_k} = \{-2, -1, 0, 1, 2\} \times \pi/16$	120×120 pixels
$\Delta^{(3)}$, $\Delta^{(2)}$	Part template	L = (position X, scale S, orientation O)	$\delta X = \{-2, 0, 2\} \times \{-2, 0, 2\}$ pixels $\delta O = \{-1, 0, 1\} \times \pi/16$	40×40 pixels
$\Delta^{(1)}$, $\Delta^{(0)}$	Gabor wavelet	l = (position x, scale s, orientation o)	$\delta x = \{-1, 0, 1\} \times \{-1, 0, 1\}$ pixels $\delta o = \{-1, 0, 1\} \times \pi/16$	13×13 pixels

Table 1: List of visual concepts used in our hierarchical compositional templates, their parameters, deformation ranges.

K activated templates, which are spatially translated, rotated and scaled copies of the hierarchical compositional templates picked from $\Delta^{(5)}$. Let $H^{(t_k)}_{\widetilde{L_k}}$ be the k-th activated template of type t_k, let $\widetilde{L_k} = (\widetilde{X_k}, \widetilde{S_k}, \widetilde{O_k})$ be its geometric attribute, where $\widetilde{X_k}$ is the location, $\widetilde{S_k}$ is the scale, and $\widetilde{O_k}$ is the orientation of the activated template. Then the set $\partial H^{(t)} = \{H^{(t_k)}_{\widetilde{L_k}}, t_k = t\}$ forms an equivalent class of $H^{(t)}$. $\Delta^{(4)}$ is the union of all possible activated templates:

$$\Delta^{(4)} = \cup \partial H^{(t)}, \quad \text{for} \quad H^{(t)} \in \Delta^{(5)}. \tag{2}$$

$\Delta^{(3)}$ denotes the dictionary of part templates of the activated hierarchical compositional templates in $\Delta^{(4)}$. Let $P^{(t,v)}$ denote the v-th part template within $dH^{(t)} \in \Delta^{(4)}$, then $dH^{(t)}$ can be decomposed into

$$dH^{(t)} = (P^{(t,v)}_{L_v}, v = 1, ..., V),$$

where V is the number of part templates, $L_v = (X_v, S_v, O_v)$ is the geometric attribute of the v-th part template, where X_v, S_v and O_v are the relative position, scale and orientation respectively. Here we fix the model structure by assigning 9 non-overlapping part templates arranged into a 3×3 grid to each hierarchical compositional template. Then $\Delta^{(3)}$ is the collection of all the part templates

$$\Delta^{(3)} = \{P^{(t,v)}, t = 1, ..., T, v = 1, ..., V\}. \tag{3}$$

$\Delta^{(2)}$ includes all the shifted part templates. We allow each $P^{(t,v)}$ in $\Delta^{(3)}$ to translate and rotate within a small bounded range to account for object deformations in different images. Let $\delta L = (\delta X, 0, \delta O)$ be the shift within the bounded range, then we derive a set of shifted part templates $\{P^{(t,v)}_{\delta L}\}$ for each $P^{(t,v)}$ in $\Delta^{(3)}$. Let $\partial P^{(t,v)}$ denote the equivalent class of $P^{(t,v)}$ subject to bounded shifts. Then $\Delta^{(2)}$ is the union of all the shifted part templates

$$\Delta^{(2)} = \cup \partial P^{(t,v)}, \quad \text{for} \quad P^{(t,v)} \in \Delta^{(3)}. \tag{4}$$

$\Delta^{(1)}$ contains the Gabor wavelets in the deformed part templates in $\Delta^{(2)}$. Following the active basis model in [28], the basis elements B are chosen to be Gabor wavelets at different positions and orientations. A deformed part template $dP^{(t,v)} \in \Delta^{(2)}$ is decomposed into a group of Gabor wavelets with zero mean and unit ℓ_2 norm

$$dP^{(t,v)} = (B_{l_{t,v,i}}, i = 1, ..., n),$$

where $B_{l_{t,v,i}}$ is the i-th Gabor wavelet with geometric attribute $l_{t,v,i} = (x_{t,v,i}, s_{t,v,i}, o_{t,v,i})$. Here the geometric attribute $l_{t,v,i}$ for each $B_{l_{t,v,i}}$ is not prefixed, but to be learned from the input images. Therefore, $\Delta^{(1)}$ is a set of Gabor wavelet elements decomposed from $\Delta^{(2)}$

$$\Delta^{(1)} = \{B_{l_{t,v,i}} \in dP^{(t,v)}, \quad dP^{(t,v)} \in \Delta^{(2)}\}. \tag{5}$$

$\Delta^{(0)}$ contains the shifted Gabor wavelets in $\Delta^{(1)}$, which ground the templates onto image pixels. For each wavelet $B_l \in \Delta^{(1)}$, we allow translations and rotations within bounded ranges and derive a shifted set $\partial B_l = \{B_{l+\delta l}\}$, where $\delta l = (\delta x, 0, \delta o)$. Then $\Delta^{(0)}$ is the union of all these shifted Gabor wavelets

$$\Delta^{(0)} = \cup \partial B_l, \quad \text{for} \quad B_l \in \Delta^{(1)}, \tag{6}$$

where each Gabor wavelet is the translated and rotated version of the original one.

As a summary, dictionaries $\Delta^{(j)}, j = 5, 4, 3, 2, 1, 0$ form a hierarchical compositional representation of the visual patterns. $\Delta^{(5)}$ is decomposed to $\Delta^{(3)}$, and $\Delta^{(3)}$ is decomposed to $\Delta^{(1)}$. $\Delta^{(4)}$ is the activated and shifted version of $\Delta^{(5)}$; while $\Delta^{(2)}$ and $\Delta^{(0)}$ are the shifted versions of $\Delta^{(3)}$ and $\Delta^{(1)}$ respectively.

2.2. Probabilistic modeling

Given an input image $\mathbf{I}$, we encode the visual patterns in $\mathbf{I}$ by K activated hierarchical compositional templates that

are shifted instances of the templates in the dictionary. For now, let us assume that these K templates as well as their part templates and Gabor wavelets do not overlap with each other. The issue of overlapping will be considered later, which will not add anything conceptually.

Let $\Lambda^{(k)}$ be the image domain covered by the k-th activated and deformed hierarchical compositional template $H^{(t_k)}_{\widetilde{L_k}} \in \Delta^{(4)}$ to encode image $\mathbf{I}$. Then the image domain Λ of $\mathbf{I}$ can be divided into

$$\Lambda = \Lambda^{(0)} \cup [\cup_{k=1}^{K} \Lambda^{(k)}], \tag{7}$$

where $\Lambda^{(0)}$ refers to the image domain not covered by any templates.

Each activated template is further divided into part templates, which are also allowed to shift to encode the input image. Let $\Lambda^{(k,v)}$ be the image domain covered by the shifted part template $P^{(t_k,v)}_{dL_{k,v}} \in \Delta^{(2)}$, where $dL_{k,v} = \widetilde{L_k} + L_v + \delta L_{k,v}$. Then the image domain $\Lambda^{(k)}$ covered by $H^{(t_k)}_{\widetilde{L_k}}$ is divided into

$$\Lambda^{(k)} = \cup_{v=1}^{V} \Lambda^{(k,v)}. \tag{8}$$

Each shifted part template $P^{(t_k,v)}_{dL_{k,v}}$ is further divided into shiftable Gabor wavelets to ground onto image pixels. Let $\Lambda^{(t,v,i)}$ be the image domain covered by the shifted Gabor wavelet $B_{dl_{k,v,i}} \in \Delta^{(0)}$, where $dl_{k,v,i} = dL_{k,v} + l_{t_k,v,i} + \delta l_{k,v,i}$. Then the image domain $\Lambda^{(k,v)}$ covered by $P^{(t_k,v)}_{dL_{k,v}}$ is divided into

$$\Lambda^{(k,v)} = \Lambda^{(k,v,0)} \cup [\cup_{i=1}^{n} \Lambda^{(k,v,i)}], \tag{9}$$

where $\Lambda^{(k,v,0)}$ refers to the empty pixels inside $\Lambda^{(k,v)}$ not occupied by the Gabor wavelets.

Let $\Lambda_S = \cup_{k,v,i} \Lambda^{(k,v,i)}$ denote the pixels covered by the Gabor wavelets in image $\mathbf{I}$, which correspond to the sketchable image areas, and let $\overline{\Lambda_S} = \{\Lambda^{(0)} \cup [\cup_{k,v} \Lambda^{(k,v,0)}]\}$ denote the pixels not covered by the Gabor wavelets, which correspond to the non-sketchable image areas. The image is divided into two components

$$\mathbf{I} = (\mathbf{I}(\Lambda_S),\ \mathbf{I}(\overline{\Lambda_S})).$$

The activation and deformation states of the dictionaries at different layers form the encoding $W = (t_k, \widetilde{L}_k, \delta L_{k,v}, \delta l_{k,v,i}, \forall k, v, i)$ of image $\mathbf{I}$. Here we define a probability model $p(\mathbf{I}|W)$ over W. Due to the tree structure of the hierarchical compositional model and the non-overlapping assumption, $p(\mathbf{I}|W)$ can be factorized as follows by assuming independence between the parts,

$$\begin{aligned} p(\mathbf{I}|W) &= p(\mathbf{I}(\overline{\Lambda_S}), \mathbf{I}(\Lambda_S)|W) \\ &= p(\mathbf{I}(\overline{\Lambda_S}))p(\mathbf{I}(\Lambda_S)|W) \\ &= p(\mathbf{I}(\overline{\Lambda_S})) \prod_{k,v,i} p\left(\mathbf{I}(\Lambda^{(k,v,i)})|B_{dl_{k,v,i}}\right). \end{aligned} \tag{10}$$

Following the active basis model [28], we take a reference model $q(\mathbf{I})$ for generic natural images, which can be factorized into the product of the patch probabilities $q\left(\mathbf{I}(\Lambda^{(k,v,i)})\right)$ as well as $q(\mathbf{I}(\overline{\Lambda_S}))$ under independence assumption.

We compute the probability ratio

$$\frac{p(\mathbf{I}|W)}{q(\mathbf{I})} = \frac{\prod_{k,v,i} p\left(\mathbf{I}(\Lambda^{(k,v,i)})|B_{dl_{k,v,i}}\right)}{\prod_{k,v,i} q\left(\mathbf{I}(\Lambda^{(k,v,i)})\right)}. \tag{11}$$

Since $p(\mathbf{I}(\overline{\Lambda_S}))$ uses the same model as $q(\mathbf{I}(\overline{\Lambda_S}))$, it is canceled in the ratio.

As each image patch $\mathbf{I}(\Lambda^{(k,v,i)})$ is still high dimensional, we project it to a one dimensional probability ratio along the response of basis function $B_{dl_{k,v,i}}$

$$r_{k,v,i} = \left\| \left\langle \mathbf{I}(\Lambda^{(k,v,i)}), B_{dl_{k,v,i}} \right\rangle \right\|^2,$$

and the latter follows a one-dimensional exponential distribution [28],

$$\begin{aligned} \frac{p\left(\mathbf{I}(\Lambda^{(k,v,i)})|B_{dl_{k,v,i}}\right)}{q\left(\mathbf{I}(\Lambda^{(k,v,i)})\right)} &= \frac{p(r_{k,v,i})}{q(r_{k,v,i})} \\ &= \frac{1}{Z_{t_k,v,i}} \exp\left\{\lambda_{t_k,v,i} h(r_{k,v,i})\right\}. \end{aligned} \tag{12}$$

The above model has four aspects.

- $q(r)$ is a histogram of filter responses pooled over a set of natural images. It has high probabilities near zero and has heavy tail.
- h is a sigmoid transform that saturates at τ:

  $$h(x) = \tau \left[2/(1 + e^{-2x/\tau}) - 1\right].$$

  It has high response when the patch coincides with an edge/bar feature in the image.
- $\lambda_{t,v,i}$ reflects the importance of the corresponding Gabor wavelet element in the learned shape template, and should be estimated by maximum likelihood so that the expectation $E_{\lambda_{t,v,i}}[h(r)]$ matches the corresponding observed mean response from covered image patches.
- $Z_{t,v,i}$ can be computed using numerical integration to normalize the 1D probability $p(r) = q(r)\frac{1}{Z_{t,v,i}} \exp\{\lambda_{t,v,i} h(r)\}$.

Let $\Theta = (\lambda_{t,v,i}, \Delta^{(j)}, \forall t, v, i, j)$ be the parameters for the hierarchical compositional model, the log-likelihood ratio of image $\mathbf{I}$ encoded by W is

$$\begin{aligned} l(\mathbf{I}|W, \Theta) &= \log \frac{p(\mathbf{I}|W)}{q(\mathbf{I})} \\ &= \sum_{k=1}^{K} \sum_{v=1}^{V} \sum_{i=1}^{n} \left[\lambda_{t_k,v,i} h(r_{k,v,i}) - \log Z_{t_k,v,i}\right]. \end{aligned} \tag{13}$$

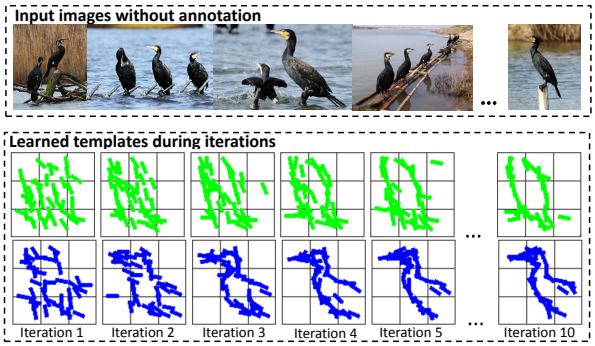

Figure 3: Learned hierarchical compositional templates during 10 iterations on 14 input images.

3. Unsupervised Learning

In this section, we present the unsupervised learning algorithm for learning a dictionary of the hierarchical templates from a set of unannotated images $\{\mathbf{I}^{(m)}, m = 1, \cdots, M\}$. The unsupervised learning algorithm seeks to minimize the total energy function $-\sum_{m=1}^{M} l(\mathbf{I}^{(m)}|W^{(m)}, \Theta)$ over $\{W^{(m)}, \forall m\}$ and Θ. The algorithm iterates the following two steps. (I) Image encoding: Given Θ, infer $W^{(m)}$ for each $\mathbf{I}^{(m)}$. (II) Re-learning: Given $\{W^{(m)}, \forall m\}$, estimate Θ. Examples of the learned templates during iterations are shown in Fig. 3.

3.1. Image encoding

During the image encoding step, we assume the dictionaries of templates and the corresponding parameters are known and fixed. Firstly, we calculate a series of template matching score maps SUM1, MAX1, SUM2, MAX2, SUM3 by a bottom-up template matching sub-process:

Procedure Template matching

Up-1 compute the Gabor wavelet matching score SUM1 of B on the image $\mathbf{I}$:

$$\text{SUM1}(l) = |\langle \mathbf{I}, B_l \rangle|^2.$$

Up-2 compute MAX1 by local maximization to account for the shifts of Gabor wavelets:

$$\text{MAX1}(l) = \max_{\delta l} \text{SUM1}(l + \delta l).$$

Up-3 for $t = 1, ..., T, v = 1, ..., V$, compute matching scores $\text{SUM2}_{t,v}$ of the part template $P^{(t,v)}$ on $\mathbf{I}$:

$$\text{SUM2}_{t,v}(L) = \sum_{i=1}^{n} [\lambda_{t,v,i} h(\text{MAX1}(L + l_{t,v,i})) - \log Z(\lambda_{t,v,i})].$$

Up-4 for $t = 1, ..., T, v = 1, ..., V$, compute $\text{MAX2}_{t,v}$ of $P^{(t,v)}$ by local maximization to account for shifts of part templates:

$$\text{MAX2}_{t,v}(L) = \max_{\delta L} \text{SUM2}_{t,v}(L + \delta L).$$

Up-5 for $t = 1, ..., T$, compute the matching score SUM3_t of the hierarchical compositional template $H^{(t)}$ on $\mathbf{I}$:

$$\text{SUM3}_t(\widetilde{L}) = \sum_{v=1}^{V} \text{MAX2}_{t,v}(\widetilde{L} + L_v).$$

Suppose the k-th activation of shape templates in the image $\mathbf{I}$ is known to be $H^{(t_k)}$, then the geometric attributes of $H^{(t_k)}$ and its part templates $P^{(t_k,v)}$ can be determined by a top-down template localization sub-process:

Procedure Template localization

Down-1 localize the hierarchical compositional template $H^{(t_k)}$ in the image $\mathbf{I}$:

$$\widetilde{L}_k = \arg\max_{\widetilde{L}} \text{SUM3}_{t_k}(\widetilde{L}).$$

Down-2 localize the part templates in the image $\mathbf{I}$:

$$\delta L_{k,v} = \arg\max_{\delta L} \text{SUM2}_{t_k,v}(\widetilde{L}_k + L_v + \delta L), \forall v$$

.

Finally, a template matching pursuit process is performed to sequentially select hierarchical compositional templates to encode $\mathbf{I}$.

Algorithm 1 Template matching pursuit

1: Initialize the maps of template matching scores $\text{SUM3}_t(\widetilde{L})$ for all $(\widetilde{L}, t)$ by the template matching sub-process. Let $k \leftarrow 1$.
2: Select the next best hierarchical compositional template by finding the global maximum of the maps: $t_k = \arg\max_t [\max_{\widetilde{L}} \text{SUM3}_t(\widetilde{L})]$.
3: Localize the selected template $H^{(t_k)}$ in the image $\mathbf{I}$ by the template localization sub-process and get $\{\widetilde{L}_k, \delta L_{k,v} \forall v\}$.
4: Let the selected arg-max template inhibit overlapping candidate templates. If the candidate template $H^{(t)}_{\widetilde{L}}$ overlaps with $H^{(t_k)}_{\widetilde{L}_k}$, then set $\text{SUM3}_t(\widetilde{L}) \leftarrow -\infty$.
5: Stop if all $\text{SUM3}_t(\widetilde{L}) \leq 0$. Otherwise let $k \leftarrow k + 1$, and go to Step 2.

In practice, it is desirable to allow some limited overlapping between the K hierarchical compositional templates that encode $\mathbf{I}$. If not, some salient patterns of $\mathbf{I}$ may fall through the cracks between the templates. On the other hand, we do not want to allow excessive overlap. Otherwise the learned part templates will be too redundant, and we will need a lot of them to encode different visual patterns. In practice, given a selected template $H^{(t_k)}_{\widetilde{L}_k}$, we set $\text{SUM3}_t(\widetilde{L}) \leftarrow -\infty$ if $\|\widetilde{X} - \widetilde{X}_{t_k}\|_2 \leq \rho D$, where D is the side length of hierarchial compositional templates, and ρD is the preset overlapping distance (default setting: $\rho = .4$).

3.2. Re-learning

Given current encoding $\{W^{(m)}, \forall m\}$ on images $\{\mathbf{I}^{(m)}, \forall m\}$, we re-learn the hierarchical compositional templates. For each template $H^{(t)}$, we extract the image patches covered by part templates within it. Let $\{\widehat{\mathbf{I}}_{u,t,v}, u = 1, ..., U\}$ be the cropped aligned image patches covered by $P^{(t,v)}$, the shared matching pursuit algorithm sequentially selects the Gabor wavelets and estimates the associated parameters. Each iteration seeks the maximal increase of the total log-likelihood. The algorithm is as follows.

Algorithm 2 Shared matching pursuit

1: Initialize $i \leftarrow 0$. For $u = 1, ..., U$, initialize the response maps $\widehat{R}_{u,t,v}(l) \leftarrow |\langle \widehat{\mathbf{I}}_{u,t,v}, B_l \rangle|^2$ for all l.

2: $i \leftarrow i + 1$. Select the next basis function by finding

$$l_{t,v,i} = \arg\max_{l} \sum_{u=1}^{U} \max_{\delta l} h(\widehat{R}_{u,t,v}(l + \delta l)),$$

where $\max_{\delta l}$ is local maximum pooling within the bounded range of perturbations.

3: For $u = 1, ..., U$, given $l_{t,v,i}$, infer the perturbations by retrieving the arg-max in the local maximum pooling of Step 2:

$$\delta l_{u,t,v,i} = \arg\max_{\delta l} \widehat{R}_{u,t,v}(l_{t,v,i} + \delta l).$$

Let $dl_{u,t,v,i} = l_{t,v,i} + \delta l_{u,t,v,i}$ and the response $r_{u,t,v,i} \leftarrow \widehat{R}_{u,t,v}(dl_{u,t,v,i})$. Then let the arg-max wavelet inhibit nearby wavelet by setting $\widehat{R}_{u,t,v}(l) \leftarrow 0$ if the correlation $|\langle B_l, B_{dl_{u,t,v,i}} \rangle|^2 > \epsilon$ (default: $\epsilon = .1$ to enforce the approximate orthogonality of Gabor wavelets).

4: Compute $\lambda_{t,v,i}$ by solving the maximum likelihood equation $E_{\lambda_{t,v,i}}[h(r)] = \sum_{u=1}^{U} h(r_{u,t,v,i})/U$. And derive the corresponding $Z_{t,v,i}$ by solving $p(r) = q(r) \frac{1}{Z_{t,v,i}} \exp\{\lambda_{t,v,i} h(r)\}$.

5: Stop if $\lambda_{t,v,i}[\sum_{u=1}^{U} h(r_{u,t,v,i})/U] - \log Z_{t,v,i} \leq 0$, else go back to Step 2.

4. Experiments

The code and results can be downloaded from http://www.stat.ucla.edu/~jifeng.dai/research/HCM.html.

4.1. Image representation

Fig. 4 shows experimental results of the proposed approach on image representation. Hierarchical compositional templates are learned from single image with repetitive patterns (Fig. 4 (a)), input images of the same object category (Fig. 4 (b)), and input images of different categories (Fig. 4 (c)) respectively. It can be seen that the learned hierarchical compositional templates are quite meaningful.

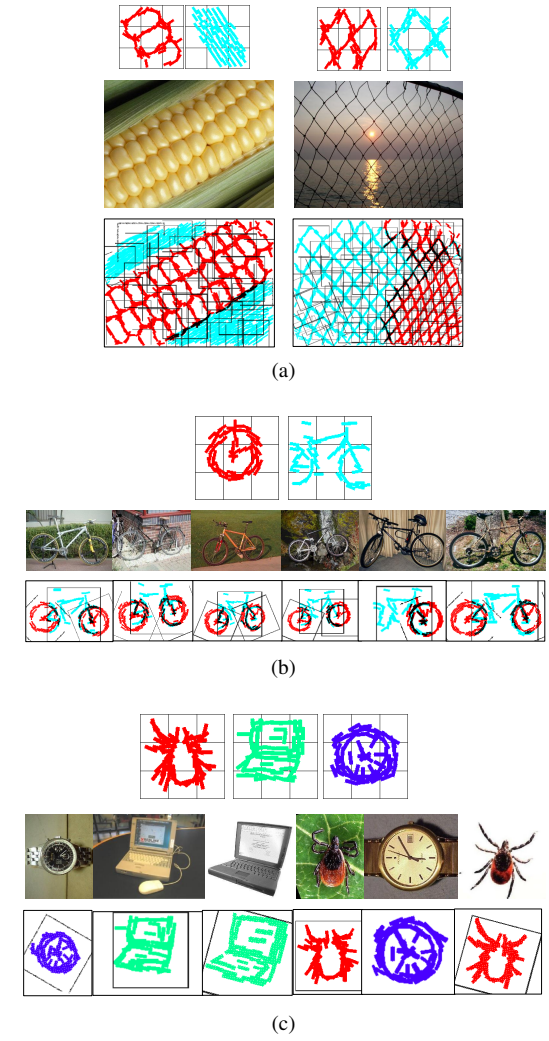

Figure 4: Learning hierarchical compositional templates and performing image encoding of: (a) single image with repetitive patterns; (b) images of the same object category (23 input images); (c) images of different object categories (30 input images).

4.2. Domain transfer

We also test the proposed approach for the task of domain transfer on the four domain dataset. The original dataset [22] contains 31 object categories where for each category, images are captured using high quality DSLR cameras, low quality webcams, and also collected from amazon.com. Using commonly used image descriptors, these images are considered as from different domains as classifiers trained from one domain will perform poorly on other domains. More recently, researchers further combine this dataset with the Caltech 256 dataset, making it a four domain dataset. We use the evaluation protocol in [8] to randomly select the training data from specific training and

Table 2: Classification accuracies on the four domain benchmark.

(a) Classification accuracies on single source four domains benchmark (C: caltech, A: amazon, D: DSLR, W: webcam)

Methods	C → A	C → D	A → C	A → W	W → C	W → A	D → A	D → W
Metric [22]	33.7 ± 0.8	35.0 ± 1.1	27.3 ± 0.7	36.0 ± 1.0	21.7 ± 0.5	32.3 ± 0.8	30.3 ± 0.8	55.6 ± 0.7
SGF [9]	40.2 ± 0.7	36.6 ± 0.8	37.7 ± 0.5	37.9 ± 0.7	29.2 ± 0.7	38.2 ± 0.6	39.2 ± 0.7	69.5 ± 0.9
GFK [8]	46.1 ± 0.6	55.0 ± 0.9	39.6 ± 0.4	56.9 ± 1.0	32.8 ± 0.1	46.2 ± 0.6	46.2 ± 0.6	**80.2 ± 0.4**
FDDL [29]	39.3 ± 2.9	55.0 ± 2.8	24.3 ± 2.2	50.4 ± 3.5	22.9 ± 2.6	41.1 ± 2.6	36.7 ± 2.5	65.9 ± 4.9
Landmark [7]	56.7	57.3	45.5	46.1	35.4	40.2	-	-
MMDT [10]	49.4 ± 0.8	56.5 ± 0.9	36.4 ± 0.8	64.6 ± 1.2	32.2 ± 0.8	47.7 ± 0.9	46.9 ± 1.0	74.1 ± 0.8
SDDL [24]	49.5 ± 2.6	**76.7 ± 3.9**	27.4 ± 2.4	**72.0 ± 4.8**	29.7 ± 1.9	49.4 ± 2.1	48.9 ± 3.8	72.6 ± 2.1
Our method	**68.3 ± 2.3**	57.4 ± 6.0	**52.7 ± 3.0**	54.8 ± 2.8	**42.2 ± 3.1**	**57.1 ± 3.5**	**60.1 ± 3.2**	79.7 ± 2.5

(b) Classification accuracies on multiple sources three domains benchmark

Source	Target	SGF [9]	RDALR [12]	FDDL [29]	SDDL [24]	Our method
DSLR, amazon	webcam	52 ± 2.5	36.9 ± 1.1	41.0 ± 2.4	**57.8 ± 2.4**	**57.8 ± 2.4**
amazon, webcam	DSLR	39 ± 1.1	31.2 ± 1.3	38.4 ± 3.4	56.7 ± 2.3	**60.6 ± 2.3**
webcam, DSLR	amazon	28 ± 0.8	20.9 ± 0.9	19.0 ± 1.2	24.1 ± 1.6	**34.6 ± 1.4**

Table 3: Correctly labeled pixel ratios on the ImageNet dataset.

	Ours	[13]	[2]	[21]
Average	**79.0**	77.1	76.9	71.0

testing domains, and learn a dictionary of 3 templates for each of the object category to construct a codebook. The templates as well as their parts in the codebook are then fed to the spatial pyramid matching method [14], which equally partition an image into 1,4, and 16 regions, and concatenates the maximum template and part matching scores at different image regions into a feature vector. We use this feature vector and the multi-class SVM to build image classifiers, and apply these classifiers in testing domains to evaluate the classification accuracy. Table 2 shows the classification results of the proposed approach and several recent approaches [22, 9, 8, 29, 7, 10, 24]. It can be seen that our method performs better than the other methods on 8 out of 11 sub tasks. Note that this accuracy is achieved using the learned hierarchical compositional templates by itself without integrating additional features. This experiment suggests that object representations learned using our method can be transferred to different domains efficiently.

4.3. Cosegmentation

In addition, we follow [2] and test the proposed approach for cosegmentation on the ImageNet dataset [3]. ImageNet is a challenging large-scale dataset of object classes. The original ImageNet dataset does not have ground-truth annotations of segmentation. In [13], a subset of ImageNet is labeled with ground-truth segmentations. The test set contains 10 random images from each of 446 classes, for a total of 4460 images. It is very challenging due to limited training examples per class, huge visual variability and cluttered backgrounds. In [13], 60k images annotated with bounding boxes, 440k images with class labels and the semantic structure of class labels in ImageNet are utilized to provide strong supervision for segmentation propagation.

We perform cosegmentation on the full test set without any additional supervision. 3 templates are learned on the images of each class, which help to align objects in different images as in [2]. In this way, the learned hierarchical compositional templates provide vital top-down information for cosegmentation. Segmentation accuracy is measured by the correctly labeled pixel ratio, following the criterion in [13]. As shown in Table 3, our approach delivers the state of the art average accuracy of 79.0%, which is 1.9%, 2.1%, and 8.0% higher than the supervised segmentation propagation algorithm in [13], the cosegmentation and cosketch approach in [2] and the Grabcut [21] baseline respectively. Note that the previous state of the art result in [13] is achieved with the help of abundant supervision, whereas our approach outperforms it without any additional supervision. Some image encoding and cosegmentation examples of the proposed approach are shown in Fig. 5.

5. Conclusion

We propose an unsupervised approach for learning hierarchical compositional templates as representational units for natural images. Experimental results show that the proposed approach is capable of learning meaningful representational units, which are useful for various vision tasks.

Acknowledgments. The work is supported by NSF DMS

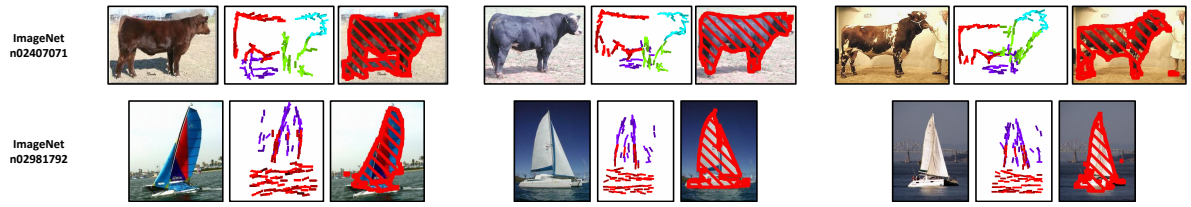

Figure 5: Examples of image encoding and cosegmentation by the hierarchical compositional templates on ImageNet.

1310391, ONR MURI N00014-10-1-0933, DARPA MSEE FA8650-11-1-7149.

References

[1] N. Ahuja and S. Todorovic. Extracting texels in 2.1D natural textures. In *ICCV*, 2007. 2

[2] J. Dai, Y. N. Wu, J. Zhou, and S.-C. Zhu. Cosegmentation and cosketch by unsupervised learning. In *ICCV*, 2013. 7

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 7

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 2

[5] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, 2007. 2

[6] S. Geman, D. F. Potter, and Z. Chi. Composition systems. *Q. Appl. Math.*, 60(4):707–736, 2002. 2

[7] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013. 7

[8] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 6, 7

[9] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 7

[10] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *ICLR*, 2013. 7

[11] Y. Hong, Z. Si, W. Hu, S.-C. Zhu, and Y. N. Wu. Unsupervised learning of compositional sparse code for natural image representation. *Q. Appl. Math.*, in press. 2

[12] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, 2012. 7

[13] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*, 2012. 7

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 7

[15] S. Lee and Y. Liu. Skewed rotation symmetry group detection. *PAMI*, 32(9):1659–1672, 2010. 2

[16] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013. 2

[17] L. Lin, K. Zeng, X. Liu, and S.-C. Zhu. Layered graph matching by composite cluster sampling with collaborative and competitive interactions. In *CVPR*, 2009. 2

[18] J. Liu and Y. Liu. Grasp recurring patterns from a single view. In *CVPR*, 2013. 2

[19] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. 1

[20] B. A. Olshausen, P. Sallee, and M. S. Lewicki. Learning sparse image codes using a wavelet pyramid architecture. In *NIPS*, 2001. 1

[21] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *TOG*, 2004. 7

[22] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2, 6, 7

[23] J. Schlecht, K. Barnard, E. Spriggs, and B. Pryor. Inferring grammar-based structure models from 3d microscopy data. In *CVPR*, 2007. 2

[24] S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, 2013. 7

[25] Z. Si and S.-C. Zhu. Learning and-or templates for object modeling and recognition. *PAMI*, 35(9):2189–2205, 2013. 2

[26] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[27] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *PAMI*, 30(12):2158–2174, 2008. 2

[28] Y. N. Wu, Z. Si, H. Gong, and S.-C. Zhu. Learning active basis model for object detection and recognition. *IJCV*, 90(2):198–235, 2010. 3, 4

[29] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011. 7

[30] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010. 2

[31] L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *ECCV*, 2008. 2