Multimed Tools Appl (2016) 75:12581–12595
DOI 10.1007/s11042-014-2414-3

Boosting scene understanding by hierarchical pachinko allocation

Jihong Ouyang · Ximing Li · Hongtu Li

Received: 30 July 2014 / Revised: 30 November 2014 / Accepted: 3 December 2014 /
Published online: 4 January 2015

Abstract Scene understanding is a popular research direction. In this area, many attempts focus on the problem of naming objects in the complex natural scene, and visual semantic integration model (VSIM) is the representative. This model consists of two parts: semantic level and visual level. In the first level, it uses a four-level pachinko allocation model (PAM) to capture the semantics behind images. However, this four-level PAM is inflexible and lacks of considerations of common subtopics that represent the background semantics. To address these problems, we use hierarchical PAM (hPAM) to replace PAM. Since hPAM is flexible, we investigate two variations of hPAM to boost VSIM in this paper. We derive the Gibbs sampler to learn the proposed models. Empirical results validate that our works can obtain better performance than the state-of-the-art algorithms.

Keywords Scene understanding · VSIM · hPAM · Topic modeling

1 Introduction

The scene understanding research plays a very important role in many computer vision applications [21–28]. In this area, a challenging task is the object recognition, which has been widely studied during the past decade. For example, some arts [10, 18, 19] recognize objects under the single-label setting, i.e, each image is assigned by a single label; some arts [3, 8, 20] consider the multi-label setting, i.e., each image can be assigned by one or more labels simultaneously; and the authors of [17] combine the localization task with classification task. However, nowadays recognizing objects in natural scenes is still quite difficult. That is because natural scenes are always complex, resulting in some problems such as ambiguity and occlusion. To deal with natural scenes, some works investigate more robust

J. Ouyang · X. Li · H. Li (✉)
College of Computer Science and Technology, Jilin University, Changchun, China
e-mail: lihongtu@jlu.edu.cn

J. Ouyang · X. Li · H. Li
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China

recognition algorithms by considering the semantics behind images. They commonly use topic modeling algorithms, e.g., latent Dirichlet allocation (LDA) [1], to uncover the latent semantics, or directly extend the LDA model for object recognition tasks. The representative attempts include a unified framework for total scene understanding [11], topic-supervised LDA (ts-LDA) and class-specific-simplex LDA (css-LDA) [16], and visual semantic integration model (VSIM) [4]. Different from other algorithms, VSIM is in fact a hierarchical model of latent semantic contexts and observed features. It consists of two levels: semantic level and visual level. In the semantic level, it uses pachinko allocation model (PAM) [12] to capture the scene semantics. In the visual level, it extends a nearest neighbor based LDA (nnLDA) model to represent observed visual context. This VSIM algorithm derives a joint inference process over the two levels, and empirically shows appreciated recognition performance for even complex nature images.

For VSIM, the "quality" of semantics extracted by the upstream semantic level is obviously significant for the downstream visual level and the final recognition performance. However, the original VSIM uses the simplest four-level PAM to capture the semantics. This might generate poor semantics information and further worse recognition performance. To boost VSIM, we use the enhancement PAM, i.e., hierarchical PAM (hPAM) [15], to replace the simplest four-level PAM. We develop two variations of hPAM, including an mentioned version in [15] and an additional version involving private subtopics and public subtopics. Both enhancement PAMs are preferable to VSIM for image data, and they can better uncover the semantics behind images. We have evaluated the two modifications by comparing with the original VSIM algorithm and some other state-of-the-art algorithms. The experimental results showed that our algorithms obtain better performance. For clarity, some important notations used in this paper are summarized in Table 1.

The rest of this paper is organized as follows: In Section 2, we introduce the VSIM algorithm. In Section 3, we boost VSIM using two variations of hPAM. Section 4 shows the experimental results. Finally, the conclusions are given in Section 5.

2 VSIM

VSIM [4] is a special "big" topic model used to fit the image content. It consists of two parts: semantic level and visual level. The semantic level uses the semantic topics, generated by the PAM algorithm, to represent context. Then, the visual level uses the visual topics, generated by the nnLDA model, to describe observed visual features. The full graphical model of VSIM is shown in Fig. 1, and details are provided as following.

The semantic level. In this level, the simplest four-level PAM [12] is used. This PAM contains a root node, a supertopic level, a subtopic level and a semantic label level, where adjacent levels are connected with each other. Supertopics are Dirichlet Multinomial $(\theta_{d,s}^{(t)}, \alpha^{(s)})$ distributions over subtopics, and they are used to represent general semantics. Subtopics are Dirichlet Multinomial $(\phi_t^{(l)}, \beta^{(l)})$ over semantic labels, and they are used to represent reified semantics. Formally, its generative process is given as follows:

1. For each subtopic t
   a. Sample a distribution over semantic labels: $\phi_t^{(l)} \sim Dirichlet\left(\beta^{(l)}\right)$
2. For each image d
   a. Sample a distribution over supertopics: $\theta_d^{(s)} \sim Dirichlet(\alpha^{(0)})$
   b. For each supertopic s

Table 1 Notation descriptions

Notation	Description
D	number of images
S	number of supertopics
T	number of subtopics
L	number of semantic labels
P	number of private subtopics for each supertopic in hPAM2
R	number of public subtopics in hPAM2
A	number of visual topics in nnLDA
$\alpha^{(0)}$	symmetric supertopic Dirichlet prior
$\theta^{(s)}$	the image distribution over supertopics
$\alpha^{(s)}$	asymmetric subrtopic Dirichlet prior
$\theta^{(t)}$	the supertopic distribution over subtopics
$\beta^{(l)}$	symmetric semantic label Dirichlet prior
$\phi^{(l)}$	the subtopic distribution over semantic labels
α	symmetric visual topic Dirichlet prior in nnLDA
$\theta^{(v)}$	the semantic label distribution over visual topics in nnLDA
$\beta^{(w)}$	symmetric observed label Dirichlet prior in nnLDA
$\phi^{(w)}$	the visual topic distribution over observed labels in nnLDA
$\gamma^{(l)}$	symmetric Dirichlet prior for level distribution in hPAM1
$\zeta^{(l)}$	the level distribution in hPAM1
$\phi^{(r/l)}/\phi^{(s/l)}$ $/\phi^{(t/l)}$	the root/supertopic/subtopic distribution over semantic labels in hPAM1
$\gamma^{(p)}$	Beta prior for distribution $\zeta^{(p)}$ in hPAM2
$\zeta^{(p)}$	the private/public Bernoulli distribution in hPAM2
$\alpha^{(g)}$	symmetric public subtopic Dirichlet prior in hPAM2
$\theta^{(g)}$	the image distribution over public subtopics in hPAM2
$\alpha^{(p/s)}$	asymmetric private subtopic Dirichlet prior in hPAM2
$\theta^{(p/t)}$	the supertopic distribution over private subtopics in hPAM2

i. Sample a distribution over subtopics: $\theta_{d,s}^{(t)} \sim Dirichlet(\alpha^{(s)})$

c. For each of the N_d semantic label $l_{d,n}$

i. Sample a supertopic $z_{d,n}^{(s)} \sim Multinomial\left(\theta_d^{(s)}\right)$

ii. Sample a subtopic $z_{d,n}^{(t)} \sim Multinomial\left(\theta_{d,z_{d,n}^{(s)}}^{(t)}\right)$

iii. Sample a semantic label $l_{d,n} \sim Multinomial\left(\phi_{z_{d,n}^{(t)}}^{(l)}\right)$

The visual level. In this level, the nnLDA model is provided to build a bridge between semantic labels (i.e., obtained by PAM) and observed labels based on image features, i.e., a many-to-many bipartite relation via a nearest neighbor rule. In nnLDA model, visual topics are Dirichlet Multinomial over observed labels, and they are used to describe observed visual features. Given a semantic label $l_{d,n}$ generated by PAM, the generative process of nnLDA is as follows:

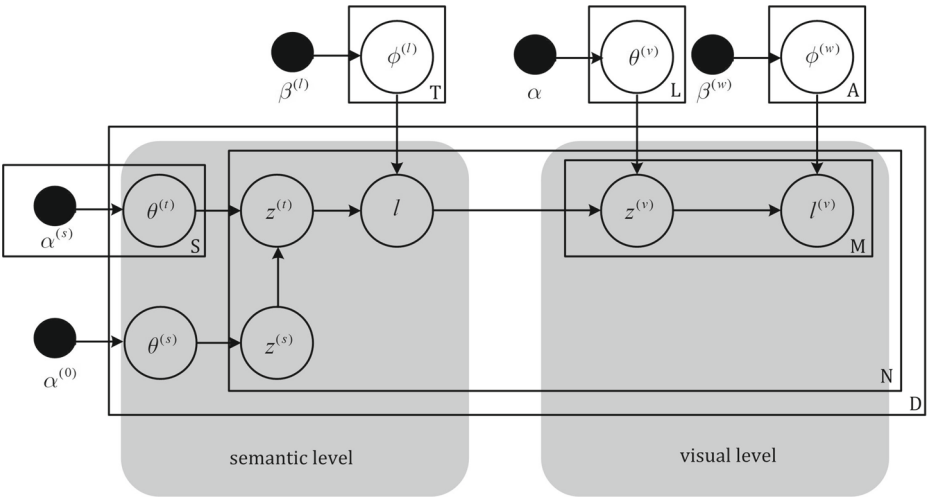


Fig. 1 The graphical model of VSIM

1. For each visual topic v
   a. Sample a distribution over observed labels: $\phi_v^{(w)} \sim Dirichlet\left(\beta^{(w)}\right)$
2. For each semantic label $l_{d,n}$
   a. Sample a distribution over visual topics: $\theta_{d,n}^{(v)} \sim Dirichlet(\alpha)$
   b. For each of the $M_{d,n}$ observed label $l_{d,n,m}^{(v)}$
      i. Sample a visual topic $z_{d,n,m}^{(v)} \sim Multinomial\left(\theta_{d,n}^{(v)}\right)$
      ii. Sample a observed label $l_{d,n,m}^{(v)} \sim Multinomial\left(\phi_{z_{d,n,m}^{(v)}}^{(w)}\right)$

3 Boosting VSIM

In the semantic level of VSIM, PAM introduces two topic levels, i.e., supertopics and subtopics, to hierarchically uncover the semantics behind images. However, PAM suffers from two problems: (1) Sometimes supertopics might prefer to make a direct connection to semantic labels, instead of a topic path; (2) Supertopics might focus on a few "private" subtopics, instead of all the subtopics. Although this PAM in VSIM performs a sparse DAG structure using an asymmetric subtopic Dirichlet prior, it is lack of considerations of "public" subtopics that describe the background semantics.

To address these problems in the semantic level of VSIM, we use hPAM [15] to replace the four-level PAM to cover semantics behind images. In contrast, hPAM is more flexible than PAM, where every node is either associated with a multinomial distribution over semantic labels, or connected with a portion of nodes in the next level. In this work, we use two variations of hPAM to point against the mentioned two problems. One is an existing version (termed hPAM1) suggested in [15], and the other one is a future version (termed

hPAM2) discussed in [15]. Since in VSIM the semantic level and visual level are conditional independent from each other, we only provide the estimation of the semantic level.

3.1 hPAM1

The hPAM1 model is mainly focusing on the first problem mentioned above. For example, suppose that a supertopic delegates "house environment", which contains a subtopic "bedroom", and this subtopic "bedroom" connects with a semantics label "bed". In PAM, for any observed semantics label "bed", it must be generated through a topic path of (house environment, bedroom). We argue that this assumption seems ramrod and redundant, because in practice, the supertopic "house environment" sometimes prefers to connect with "bed" directly. The hPAM1 model allows that all the three upper levels can generate the semantics labels, including the root node, supertopics and subtopics (e.g., "house environment" is directly connecting with "bed"). This further assumption of hPAM1 is obviously more reasonable for the natural images and can perfectly address the first problem.

To achieve this goal, hPAM1 introduces an additional variable y to show which level is generating the semantics label. Formally, hPAM1 introduces a level distribution $\zeta_{s,t}^{(l)}$ for each topic path (s, t), drawn from a symmetric Dirichlet prior $\gamma^{(l)}$. For the generation of each semantics label, we first sample a value from this level distribution to determine which level directly generates the semantics. For example, if $y_{d,n} = 1$, we will sample the semantics label from the supertopic distribution, e.g., sampling "bed" from "house environment". Under this further assumption, the generative process of hPAM1 (Fig. 2) for images is given as follows:

1. For root node, each supertopic and each subtopic i

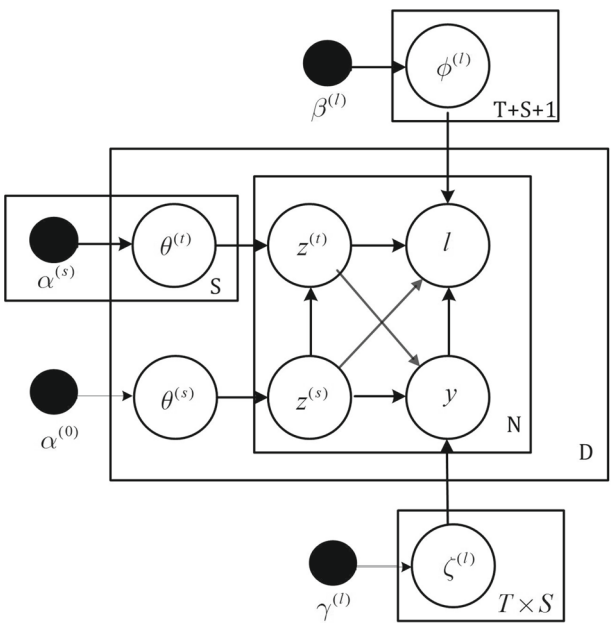


Fig. 2 The graphical model of hPAM1

a. Sample a distribution over semantic labels: $\phi_0^{(r/l)}/\phi_i^{(s/l)}/\phi_i^{(t/l)} \sim Dirichlet\left(\beta^{(l)}\right)$

2. For each topic path (s, t)
   a. Sample a level distribution: $\zeta_{s,t}^{(l)} \sim Dirichlet\left(\gamma^{(l)}\right)$
3. For each image d
   a. Sample a distribution over supertopics: $\theta_d^{(s)} \sim Dirichlet(\alpha^{(0)})$
   b. For each supertopic s
      i. Sample a distribution over subtopics: $\theta_{d,s}^{(t)} \sim Dirichlet(\alpha^{(s)})$
   c. For each of the N_d semantic label $l_{d,n}$
      i. Sample a supertopic $z_{d,n}^{(s)} \sim Multinomial\left(\theta_d^{(s)}\right)$
      ii. Sample a subtopic $z_{d,n}^{(t)} \sim Multinomial\left(\theta_{d,z_{d,n}^{(s)}}^{(t)}\right)$
      iii. Sample a level indicator $y_{d,n} \sim Multinomial\left(\zeta_{z_{d,n}^{(s)},z_{d,n}^{(t)}}^{(l)}\right)$
      iiii. if $y_{d,n} = 0/1/2$, sample a semantic label $l_{d,n} \sim Multinomial\left(\phi_0^{(r/l)}/\phi_{z_{d,n}^{(s)}}^{(s/l)}/\phi_{z_{d,n}^{(t)}}^{(t/l)}\right)$

We use collapsed Gibbs sampler [9] to train hPAM1. This is achieved by sequentially updating the supertopic assignment $z_{d,n}^{(s)}$, subtopic assignment $z_{d,n}^{(t)}$, and level indicator $y_{d,n}$ to each semantic label by the following rule:

$$P\left(z_{d,n}^{(s)}, z_{d,n}^{(t)}, y_{d,n} | I, \alpha^{(0)}, \alpha^{(s)}, \beta^{(l)}, \gamma^{(l)}\right) \propto \frac{N_{-n}^{s/d} + \alpha^{(0)}}{N_{-n}^{d} + S\alpha^{(0)}} \times \frac{N_{-n}^{st/d} + \alpha^{(s)}}{N_{-n}^{s/d} + \sum_{i=0}^{T}\alpha_i^{(s)}} \times \frac{N_{-n}^{y/st} + \gamma^{(l)}}{N_{-n}^{st} + 3\gamma^{(l)}} \times \frac{N_{-n}^{l/sty} + \beta^{(l)}}{N_{-n}^{sty} + L\beta^{(l)}} \quad (1)$$

where $N^{st/d}$ and $N^{s/d}$ are the times that a topic path (s, t) and supertopic s have occurred in image d, respectively; N^d is the total semantic labels in image d; $N^{y/st}$ and N^{st} are the number of level indicator y exists for the topic path (s, t) and the total number of (s, t) occurs; $N^{l/sty}$ and N^{sty} are the number that a semantic label l corresponds to pair (s, t, y) and the total number of pair (s, t, y) occurs; the subscript -n denotes a quantity except for the token in position n.

Finally, the distributions $\zeta_{s,t}^{(l)}$ and $\phi_0^{(r/l)}/\phi_i^{(s/l)}/\phi_i^{(t/l)}$ can be estimated as follows:

$$\zeta_{s,t,y}^{(l)} = \frac{N^{y/st} + \gamma^{(l)}}{N^{st} + 3\gamma^{(l)}} \quad (2)$$

$$\phi_{0,l}^{(r/l)}/\phi_{s,l}^{(s/l)}/\phi_{t,l}^{(t/l)} = \frac{N^{l/sty} + \beta^{(l)}}{N^{sty} + L\beta^{(l)}} \quad if\ y = 0/1/2 \quad (3)$$

Note that the asymmetric Dirichlet prior $\alpha^{(s)}$ is used to capture the sparse relations between supertopics and subtopics. So we need to optimize this prior during model training. Following [4], we use the moment matching method to estimate the approximate MLE of $\alpha^{(s)}$.

3.2 hPAM2

This hPAM2 model is mainly focusing on the second problem mentioned above. On one hand, in PAM the supertopic distribution over subtopics is in fact sparse. That is, for each supertopic, some subtopics commonly correspond to very low probabilities. For example, the supertopic "house environment" connects with two subtopics "bedroom" and "road". Obviously, "house environment" must significantly prefer to "bedroom" than "road". The hPAM2 model fully considers this situation, and introduces the concept of private subtopic, where each supertopic only samples from its own private subtopics, e.g., defining "bedroom" is the private subtopic for "house environment". In this setting, supertopics are constrained with the less relevant subtopics. On the other hand, some subtopics represent the background semantics. In contrast to private subtopics, these subtopics are widespread to all the supertopics. For example, the subtopic "weather" might be popular for almost all the supertopics. Based on this analysis, hPAM2 further introduces the concept of public subtopic, where each public subtopic is shared by all the supertopics.

Formally, hPAM2 makes two assumptions: (1) Each supertopic has P private subtopics, which are used to describe the sparse structure between supertopics and subtopics; (2) All the supertopics share R public subtopics, which are used to describe the background semantics. From the generative perspective, hPAM2 introduces a private/public Bernoulli distribution $\zeta_d^{(p)}$ for each image d, drawn from a Beta prior $\gamma^{(p)}$. This distribution is used to determine whether generating a private subtopic from its superior supertopic or generating a public subtopic directly. The generative process of hPAM2 (Fig. 3) is given as follows:

1. For each private or public subtopic t

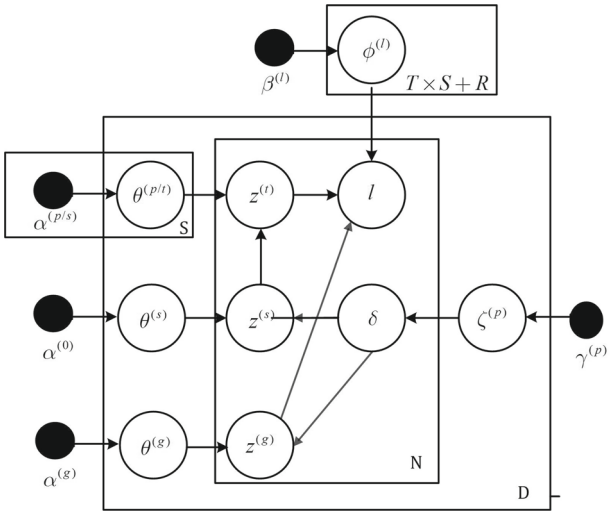


Fig. 3 The graphical model of hPAM2

a. Sample a distribution over semantic labels: $\phi_t^{(l)} \sim Dirichlet\left(\beta^{(l)}\right)$

2. For each image d

a. Sample a distribution over supertopics: $\theta_d^{(s)} \sim Dirichlet(\alpha^{(0)})$

b. Sample a distribution over public subtopics: $\theta_d^{(g)} \sim Dirichlet(\alpha^{(g)})$

c. Sample a private/public distribution: $\zeta_d^{(p)} \sim Dirichlet\left(\gamma^{(p)}\right)$

d. For each supertopic s

i. Sample a distribution over subtopics: $\theta_{d,s}^{(p/t)} \sim Dirichlet(\alpha_s^{(p/t)})$

e. For each of the N_d semantic label $l_{d,n}$

i. Sample an indicator $\delta_{d,n} \sim \text{Bernoulli}\left(\zeta_d^{(p)}\right)$

ii. If $\delta_{d,n} = 1$: Sample a supertopic $z_{d,n}^{(s)} \sim Multinomial\left(\theta_d^{(s)}\right)$, and then Sample a subtopic $z_{d,n}^{(t)} \sim Multinomial\left(\theta_{d,z_{d,n}^{(s)}}^{(p/t)}\right)$; finally, Sample a semantic label $l_{d,n} \sim Multinomial\left(\phi_{z_{d,n}^{(t)}}^{(l)}\right)$

iii. If $\delta_{d,n} = 0$: Sample a public subtopic $z_{d,n}^{(g)} \sim Multinomial\left(\theta_d^{(g)}\right)$; and then Sample a semantic label $l_{d,n} \sim Multinomial\left(\phi_{z_{d,n}^{(g)}}^{(l)}\right)$

Collapsed Gibbs sampler is also used for hPAM2. The updating rule with respect to the supertopic assignment $z_{d,n}^{(s)}$, private subtopic assignment $z_{d,n}^{(t)}$, public subtopic assignment $z_{d,n}^{(g)}$ and private/public indicator $\delta_{d,n}$ is given as follows:

$$P\left(z_{d,n}^{(s)}, z_{d,n}^{(t)}, z_{d,n}^{(g)}, \delta_{d,n} | I, \alpha^{(0)}, \alpha^{(g)}, \alpha^{(p/s)}, \beta^{(l)}, \gamma^{(p)}\right) \propto \frac{N_{-n}^{\delta/d} + \gamma^{(p)}}{N_{-n}^{d} + 2\gamma^{(p)}} \times \frac{N_{-n}^{s/d} + \alpha^{(0)}}{N_{-n}^{pd} + S\alpha^{(0)}} \times \frac{N_{-n}^{st/d} + \alpha_{s,t}^{(p/s)}}{N_{-n}^{s/d} + \sum_{i=0}^{P} \alpha_{s,i}^{(p/s)}} \times \frac{N_{-n}^{t/d} + \alpha^{(g)}}{N_{-n}^{gd} + R\alpha^{(g)}} \times \frac{N_{-n}^{l/t} + \beta^{(l)}}{N_{-n}^{t} + L\beta^{(l)}} \tag{4}$$

where $N^{\delta/d}$ and N^d are the number of semantic labels assigned by indicator δ and the total number of semantic labels in image d; N^{pd} is the total semantic labels generated by private subtopics image d; $N^{t/d}$ and N^{gd} are the times that public subtopic t has been occurred and the total semantic labels generated by public subtopics in image d.

Similar with hPAM1, the distributions $\zeta_d^{(p)}$ and $\phi_t^{(l)}$ with hPAM2 are obtained:

$$\zeta_d^{(p)} = \frac{N^{\delta/d} + \gamma^{(p)}}{N^d + 2\gamma^{(p)}} \tag{5}$$

$$\phi_{t,l}^{(l)} = \frac{N^{l/t} + \beta^{(l)}}{N^t + L\beta^{(l)}} \tag{6}$$

We also use the moment matching method to optimize the asymmetric Dirichlet prior $\alpha^{(p/s)}$.

4 Experiment

In this section, we evaluate the performance of the proposed algorithms on two diverse image datasets: Scene-15 [6] and SUN09 [5]..

4.1 Experimental setting

Datasets: The Scene-15 dataset contains totally 4,485 images grouped into fifteen scene classes, where thirteen were collected by [6] and two were collected by [10]. Each class has 200 to 400 images, and the average size is about 300 × 250 pixels.

The SUN09 dataset contains totally 8600 natural indoor and outdoor images. Each image is averagely annotated by seven various objects and each object averagely covers 5 % of the image size. Following [4], we consider the top frequent 200 classes, and use 4,367 images for training and 4,317 images for testing. During training, we consider the annotated ground-truth locations and labels pre-signed in the dataset. During testing, we use the bounding boxes detected by DPM detector[1] [7].

For Scene-15, we use two types of features to represent images. i.e., texton histograms using a codebook of 100 textons obtained by a 40-filter bank textons and visual word histograms using a codebook of 200 visual words generated by dense SIFT descriptor. For SUN09, we use three types of features, as described in [14], including the two mentioned histograms and normalized R. G. B histograms (i.e., color) with their means and variances.

Model parameters: Since the focusing of this work is on the semantic level of VSIM, we use the same number of visual topics (i.e., $A = 50$) for nnLDA as suggested in [4]. In terms of hPAM1, 20 supertopics and 50 subtopics are defined; the three symmetric Dirichlet priors are set as: $\alpha^{(0)} = 1$, $\beta^{(l)} = 0.1$ and $\gamma^{l} = 10$; the asymmetric subtopic Dirichlet prior $\alpha^{(s)}$ is learnt during model training. In terms of hPAM2, there assumes 20 supertopics with 5 private subtopics and 10 public subtopics (i.e., totally 110 subtopics); the four symmetric Dirichlet priors are set as: $\alpha^{(0)} = 1$, $\alpha^{(g)} = 0.1$, $\beta^{(l)} = 0.1$ and $\gamma^{l} = 10$; also, the private subtopic Dirichlet prior $\alpha^{(p/s)}$ is learnt during model training. In terms of Gibbs sampler, we run five independent MCMC chains with a burn-in of 500 iterations for each model, the averaged results are reported finally.

4.2 Performance

We evaluate our algorithms with the same tasks in [4]. Here, we name our algorithms as hVSIM1 and hVSIM2.

4.2.1 Semantic Scene Prediction

We evaluate the scene detection performance with respect to the ground-truth on dataset SUN09. As described in [4], we estimate the ground-truth distribution by grounding semantic labels with ground-truth labels and inferring supertopics and subtopics from the semantic level. We use the symmetric Kullback-Leibler (KL) divergence between the subtopic distributions as the performance metric. Note that the lower value of KL divergence implies better performance.

[1] downloaded at http://cs.brown.edu/pff/

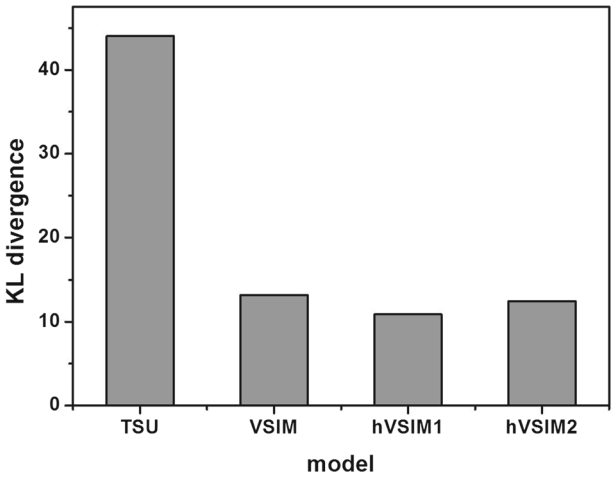


Fig. 4 KL divergence between estimated and the ground-truth on SUN09

We use two state-of-the-art models, i.e., VSIM and total scene understanding model[2] (TSU) proposed by [11], as baselines. The results are shown in Fig. 4. Clearly, the KL divergences of hVSIM1 and hVSIM2 are lower than the original VSIM and TSU. hVSIM1 shows a little better performance than hVSIM2, That is because we neglect the public subtopics in hVSIM2. To sum up, we argue that our modifications can provide closely matching with the ground-truth labels.

4.2.2 Predicting Top-N Labels

In this experiment, we evaluate the presence proportion of estimated top-N labels in the ground-truth on dataset SUN09. The VSIM and hcontext[3] [5] are used as performance baselines. For this evaluation, the higher value of proportion indicates better performance.

We perform this experiment with different values of N over the following set $\{1, 2, 3, 4, 5\}$. The results are shown in Fig. 5. The hVSIM2 performs better than other algorithms in all the settings of N. The hVSIM1 also outperforms VSIM (4/5) and hcontext (4/5) in the most settings. Note that hVSIM2 seems more robust than others. That is because, unlike VSIM, it considers the private and public subtopics so that it can capture local and background semantics. In contrast, our scheme is more reasoning for the natural scene. Empirical results further indicate this cognition.

4.2.3 Object Detection

Finally, we evaluate object detection performance of our two algorithms on both datasets Scene-15 and SUN09.

For Scene-15, we use 40, 60, 80 and 100 images per-class for training respectively, and for each case the remaining images are used for testing. Apart from VSIM, other two existing algorithms, i.e., Sparse coding (SC) [2] and localized soft-assignment coding (LSC)

[2]downloaded at http://vision.stanford.edu/projects/totalscene/

[3]downloaded at http://people.csail.mit.edu/myungjin/HContext.html

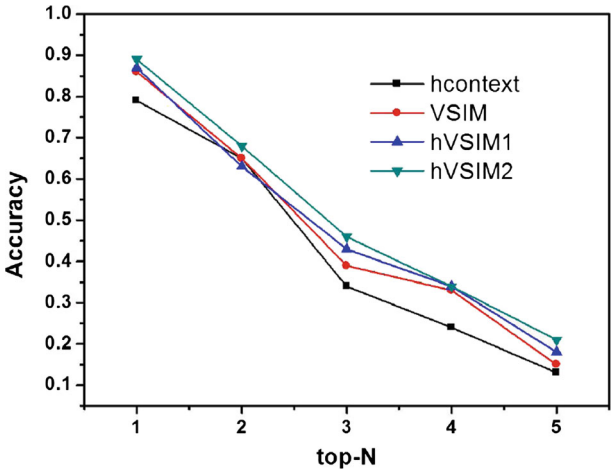


Fig. 5 Performance of top-*N* prediction on SUN09

[13], are used as baselines. Figure 6 shows the classification accuracy with different values of training image per-class. We observe that our algorithms perform better than baseline algorithms in the most cases, and hVSIM2 is even better than hVSIM1. They both outperform VSIM in all the settings. When the training images per-class is 100, hVSIM2 is almost at the same level with the two arts SC and LSC, and hVSIM1 is a little lower than the two. With the decrease of training images, our algorithms outperform SC and LSC. This indicates that our algorithms are robust, and can achieve competitive performance with fewer training images. This advantage is very helpful for small sample cases in practice.

For SUN09, as in [4], we sort the size of object classes from most to least; and report the averaged accuracy over every 25 object classes. We compare our algorithms with VSIM and the DPM detector proposed in [7]. Figure 7 illustrates the results. Clearly, hVSIM1 and hVSIM2 show similar performance, and both of them outperform VSIM in the most

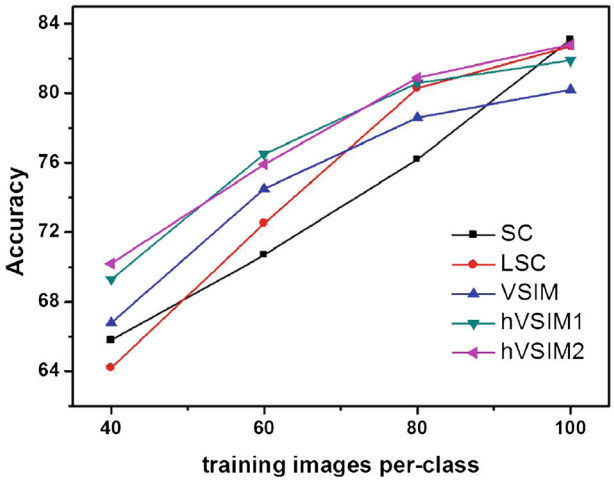


Fig. 6 The performance of object detection on Scene-15

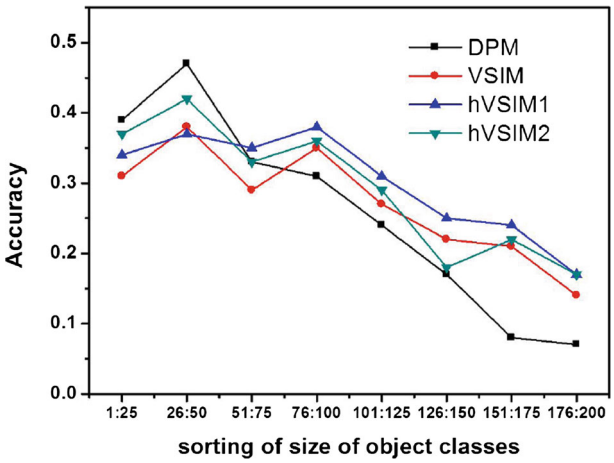


Fig. 7 The performance of object detection on SUN09

settings. DPM performs a bit better than the three VISM-based algorithms for large size classes, however, it seems too sensitive to the size of training set. For relatively small rare classes, our two algorithms significantly outperform DPM. In contrast, hVSIM1 and hVSIM2 show more smooth trend and are competitive with VSIM and DPM on the overall performance.

5 Conclusion

In this paper, we investigate the problem of naming object in the complex natural images. We attempt to boost the recent VSIM algorithm, which is composed of a semantics level, i.e., a four-level PAM, and a visual level, i.e, nnLDA. Our focus is on the semantics level, and we use two variations of hPAM to replace the simple four-level PAM. The first one allows connections among supertopics and semantic labels; the other one takes the local and background semantics into account. They are more flexible and powerful to capture semantics behind natural images.

We compare the proposed algorithms with the original VSIM and some other state-of-the-art algorithms on datasets Scene-15 and SUN09. Empirical results indicate that our modifications are more robust and achieve competitive performance on the basic tasks, e.g., top-N predication and object detection.

In the future, we will explore our algorithms on other popular and challenging image collections and tasks.

Acknowledgments This work was supported by National Nature Science Foundation of China (NSFC) under the Grant No. 61170092, 61133011, and 61103091.

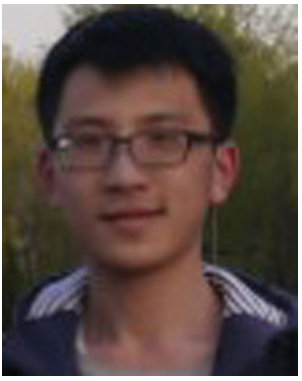
References

1. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

2. Boureau YL, Bach F, LeCun Y, Ponce J (2010) Learning mid-level features for recognitionral scene categories. In: Conference on Computer Vision and Pattern Recognition, pp 2559–2566
3. Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recog 37(9):1757–1771
4. Chakraborty I, Elgammal A (2013) Visual-semantic scene understanding by sharing labels in a context network. CoRR
5. Choi MJ, Lim JJ, Torralba A, Willsky AS (2010) Exploiting hierarchical context on a large database of object categories. In: Conference on Computer Vision and Pattern Recognition, pp 129–136
6. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: Conference on Computer Vision and Pattern Recognition, pp 524–531
7. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
8. Frnkranz J, Hllermeier E, Menca EL, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153
9. Griffiths TL, Steyvers M (2004) Finding scientific topics. In: National academy of Sciences of the United States of America, vol. 101, pp 5228–5235
10. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Conference on Computer Vision and Pattern Recognition, pp 2169–2178
11. Li LJ, Socher R, Li FF (2009) Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In: Conference on Computer Vision and Pattern Recognition, pp 2036–2043
12. Li W, McCallum A (2006) Pachinko allocation: Dag-structured mixture models of topic correlations. In: International Conference on Machine Learning, pp 577–584
13. Liu L, Wang L, Liu X (2011) In defense of soft-assignment coding. In: International Conference on Computer Vision, pp 2486–2493
14. Malisiewicz TJ, Huang JC, Efros AA (2006) Detecting objects via multiple segmentations and latent topic models. Carnegie Mellon University Tech Report
15. Mimno D, Li W, McCallum A (2007) Mixtures of hierarchical topics with pachinko allocation. In: International Conference on Machine Learning, pp 633–640
16. Rasiwasia N, Vasconcelos N (2013) Latent Dirichlet allocation models for image classification. IEEE Trans Pattern Anal Mach Intell 35(11):2665–2679
17. Russakovsky O, Lin Y, Yu K, Fei-Fei L (2012) Locality-constrained linear coding for image classification. In: European conference on Computer Vision, pp 1–15
18. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: Conference on Computer Vision and Pattern Recognition, pp 3360–3367
19. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: Conference on Computer Vision and Pattern Recognition, pp 1794–1801
20. Yang Y, Huang Z, Shen HT, Zhou X (2011) Mining multi-tag association for image tagging. World Wide Web J 14(2):133–156
21. Yang Y, Huang Z, Yang Y, Shen HT, Luo J (2013) Local image tagging via graph regularized joint group sparsity. Pattern Recog 46(5):1358–1368
22. Yang Y, Yang Y, Shen HT (2013) Effective transfer tagging from image to video. ACM Trans Multimedia Comput Commun Appl 9(2). Article No. 14
23. Yang Y, Zha ZJ, Gao Y, Zhu X, Chua TS (2014) Exploiting web images for robust semantic video indexing via sample-specific loss. IEEE Trans Multimedia 16(6):1677–1689
24. Zhang L, Gao Y, Hong C, Feng Y, Zhu J, Cai D (2014) Feature correlation hypergraph: exploiting high-order potentials for multimodal recognition. IEEE Trans Cybernetics 44(8): 1408–1419
25. Zhang L, Gao Y, Xia Y, Dai Q, Li X (2014) A fine-grained image categorization system by cellet-encoded spatial pyramid modeling. IEEE Transactions on Industrial Electronics
26. Zhang L, Han Y, Yang Y, Song M, Yan S, Tian Q (2013) Discovering discrminative graphlets for aerial image categories recognition. IEEE Trans Image Process 22(12):5071–5084
27. Zhang L, Ji R, Xia Y, Zhang Y, Li X (2014) Learning a probabilistic topology discovering model for scene categorization. IEEE Transactions on Neural Networks and Learning Systems PP(99)
28. Zhang L, Song M, Deng X, Bu J, Chen C (2011) Large-scale outdoor scene classification by boosting a set of highly discriminative and low redundant graphlets. In: IEEE International Conference on Data Mining Workshops, pp 847–852

Jihong Ouyang is a professor at the College of Computer Science and Technology, Jilin University of China. She received her Ph.D degree in Jilin university in 2005. Her main research interests include artificial intelligence and machine learning, more specifically in spatial reasoning, multi-label learning, topic modeling, and online learning.

Ximing Li received the M.S. degree in computer science from Jilin University, China, in 2011. Currently he is a Ph.D. candidate in the College of Computer Science and Technology at Jilin University. His main research interests include topic modeling and multi-label learning.



Hongtu Li was born in 1984. He received the PhD degrees from the College of Computer Science and Technology, Jilin University (JLU), Changchun, China, in 2012. He has been a lecturer with the School of Jilin University since 2012. His current research interests include Network Security, Cryptography and so on.