Image Classification by Codebook Updating via Joint i-Pat Topic Model Feedback

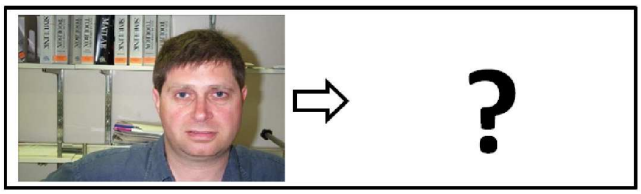
Wai Lam Hoo*, Tae-Kyun Kim†, Yuru Pei‡ and Chee Seng Chan*
*University of Malaya, 50603 Kuala Lumpur, Malaysia
Email: wailam88@siswa.um.edu.my; cs.chan@um.edu.my
†Imperial College London, London SW7 2AZ, United Kingdom
Email: tk.kim@imperial.ac.uk
‡Peking University, Beijing 100871, China
Email: peiyuru@cis.pku.edu.cn

***Abstract*—Image classification is an important task in computer vision as it plays a major role in content-based image retrieval. For an image classification system that uses the Bag-of-Words model representation, visual codebook is an essential part. Randomized forest (RF) as a tree-structure discriminative codebook has been proven highly time-efficient for real-world applications, by utilizing the image class labels which the detected features are associated with (i.e. weakly supervised learning). However, the RF codebook can be degraded if the image class labels are poorly labeled. In this paper, we tackle this image class label problem by proposing a feedback mechanism from the topic model. This statistical model can hypothesize the 'correct' labels for images, to enhance the effectiveness of the RF codebook. Besides, the feedback mechanism employs a joint image-patch (i-Pat) level information in predicting the 'correct' labels, which is in contrast to the state-of-the-art RF learning. Experiments on both the 15-Scene and C-Pascal datasets had shown that the proposed method outperforms existing methods by a margin. Besides, even with limited labeled training images, the proposed solution still able to achieve comparable accuracy compared to the fully labeled training data algorithm.**

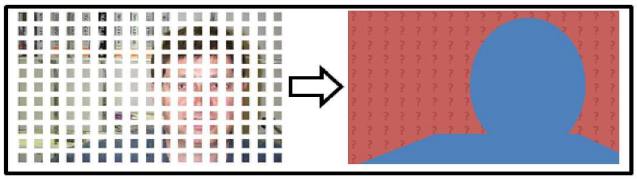
I. Introduction

Image classification from the natural images has received wide attention from the computer vision communities due to its usefulness in content-based image retrieval, video surveillance, robot localization and image understanding. Though successful algorithms have been proposed, this is still a daunting task with the presence of intra-class variation, background clutter, occlusion and changes of pose.

Recently, Bag-of-words (BoW) model has been a popular choice in the image classification task [2]–[8]. Amongst these works, [6], [7] had employed the random forest (RF) as the visual codebook. RF codebook utilizes the available image class labels that are associated with the extracted features during learning. As a result, the RF codebook will be more discriminative compare to the conventional k-means codebook [2], [3], [5]. However, such an advantage is heavily dependent on the number of accurate image class labels. As an example in Figure 1(a), from the human perspective point of view, one will most likely classify the image as a 'human face' class as our visual cognitive tends to focus on the dominated human face, and neglect the background (we will not treat this as a 'book' class, although there are books at the background). However, Figure 1(b) reflects how a machine learns. In weakly supervised settings, the machine will extract image patches from the original image source (in here, we employ dense sampling as an example), and will associate all the extracted patches to the corresponding image class label (In this case, 'human face' class). With this, the machine will also include the set of wrongly-labeled image patches (patches from the background) during training and therefore degrades the RF codebook performance.



(a) From human visual perspective, it is intuitive to label this as 'human face' class.



(b) It is clear that not all the image patches have the 'human face' characteristics (the blue region is the correct patches). However, some background patches (red region) are still associated as 'human face' class in weakly supervised setting, which is not correct and hence degrades the codebook learning performance.

Fig. 1: Image vs. patches learning: how human and machine learn over images? (image from Caltech-256 Faces class [1])

In this paper, we propose a feedback mechanism namely the **joint image-patch level (joint i-Pat) feedback mechanism** to overcome the aforementioned problem. With this strategy, we can estimate the soft class labels to each patch using the topic model. That is, the soft class labels give a probability value to each patch, describing how true that patch is belongs to the image class label as shown in Figure 2. Then, a new RF codebook can be learned from the soft class labels, enhancing the RF codebook representation. This is followed by a new pLSA topic model that is learned from the new RF codebook. This iterative learning process will continue until

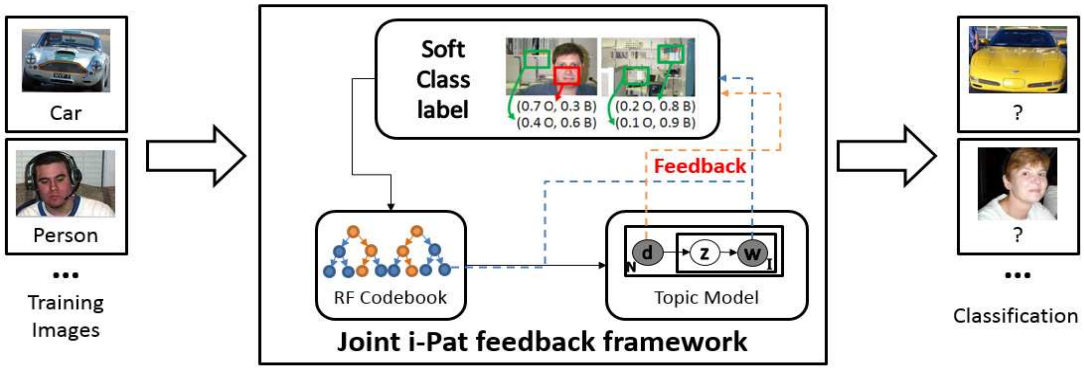


Fig. 2: An overview to the proposed Joint i-Pat feedback framework. We first learn a RF codebook from available training images. After that, we treat RF leafnodes as codebook and learn the topic model (pLSA - probabilistic latent semantic analysis, for our framework) from the BoW representation based on the learned RF codebook. To perform the joint i-Pat feedback, soft class labels for the training images are estimated from the image and patch level information from both pLSA ($p(z|w,d)$) and RF ($p(w|x)$), respectively. The blue dotted line indicates patch information from the RF and pLSA, while dark orange dotted line indicates the image level information from pLSA. For soft class label, we can notice that the red patches (correct image patch) have high object (O) score and low background (B) score. This indicates that the particular patch have high probability to be object patch. This O score are B score are resultant from unsupervised learning of pLSA, which gathers latent topics information and use a novel Dominant Topic (DT) representation to deduce the O score and B score. In contrast, green patches (the background patches) will have high B score and low O score. Then, a new RF codebook is learned and follows by, a new pLSA topic model is learned from the new RF codebook. This iterative learning process will continued until convergence.

the convergence criteria is met. Empirically, we achieve state-of-the-art performance in 2 public datasets (15-Scene and C-Pascal), even in situations that only limited training images are available.

In summary, our contributions are 2-fold: 1) we proposed **joint i-Pat feedback framework** in image classification. Such a framework includes the benefit of image level information, estimated from the topic model as well as the patch level information from both the RF codebook and topic model, respectively. This is in contrary to [6] which focused in learning the patches information and [7] that learns the image level concepts; and 2) we introduced the **soft class labels** in the joint i-Pat feedback framework to strengthen the codebook learning quality rather than conventional image class labels.

This paper is arranged as follows: Section II discusses the recent development in related topics including the codebook learning and topic model. Section III details the joint i-Pat framework. We show our results in Section IV. Finally, discussions and conclusion are drawn in Section V-VI, respectively.

II. Related Work

Visual codebook learning is an essential pipeline in the bag-of-words (BoW) representation. In order to find the optimal codewords for selected problems, unsupervised methods such as k-means [2] and kd-tree [9] had been applied to the extracted features. However, recent research work had been focused on learning the visual codebook with labeled images, in order to have a better discrimination on the codebook learning e.g. random forest sparse coding [10], supervised sparse coding [11], and ERC forest [6]). These methods extract image patches from the whole images as the input training data, which can be categorized as the patch-level learning. On the other hand, Krapac et al. [7] employed the image-level approach by minimizing image classification loss on the validation set during the RF splits. This is in order to directly maximize the image classification performance. In our method, we adopt the RF as codebook because of its discriminativeness advantage, and we tackle the RF disadvantage using the soft class label learning.

Topic model is first introduced to deal with the natural language problem by Thomas Hofmann [12] via Probabilistic Latent Semantic Analysis (pLSA) model. It has been widely studied in recent computer vision area, especially in scene understanding and object categorization [3]. Using the topic model, we can estimate image topics (or middle-level information) that reside in all image classes and find optimum combination of it to suggest the image level scores. Also, we can predict joint image-patch level scores by employing image-word-specific topic distribution $p(z|w,d)$ and codeword probability $p(w|x)$. In here, we choose to apply pLSA model [12] rather than other variants because of its simplicity (e.g. compared to Latent Dirichlet Allocation (LDA) model [13] and Hierarchical Dirichlet Process (HDP) model [14] which need to learn dirichlet prior). Note that our aim of this paper is to show the advantages of the joint i-Pat feedback mechanism for image classification.

III. Methodology

The proposed method consists of three steps as illustrated in Figure 2. First, we initiate the model by learning the image patches, that are associated with the image class labels, using the RF. Treating the RF leafnodes as codewords, we build the BoW representation from the RF codebook. Secondly, we learn the pLSA topic model from the RF codebook and estimate the

soft class labels on the training images patches for joint i-Pat feedback. Following this, a new RF codebook is learned from the patch features associated with soft class labels, and follow by a new pLSA is learned from the refined RF codebook, forming a feedback cycle. The joint i-Pat framework is iterated until the convergence criteria is satisfied. Finally, we perform the classification with the pLSA generated from the enhanced RF codebook via prediction score.

We will first discuss the initial RF codebook and topic model learning, following by the process to generate the soft class labels. Then, we will detail the joint-iPat feedback mechanism and the convergence criteria. Finally, we will explain the inference method of the proposed framework.

A. Initial RF codebook generation and pLSA learning

RF is an ensemble of the random decision trees which applied a bagging strategy on different decision trees. RF provides a very fast way of codebook learning and quantization. Moreover, when the image class labels are available, it can act as a good discriminative codebook. Each random decision tree is constructed using a random subset of the training data with replacement. The labeled training image I'_{node} at specific node, consist of x_i and l_i where x_i is the feature vectors of image patches from training images, and l_i are corresponding image class labels. These feature vectors are recursively split into left and right subsets, I'_{left} and I'_{right}, according to a set of thresholds T and a split function f, as

$$I'_{left} = \{x_i \in I'_{node} | f(x_i) < T_t\}, I'_{right} = I'_{node} \setminus I'_{left}. \quad (1)$$

Since this is a randomized decision tree, we generated a random subset of the features for split function f and T. The splits that maximize the expected information gain $\triangle ENT$ are selected. Specifically, at each split node:

$$\triangle ENT = ENT(I'_{node}) - \sum_{i=left,right} \frac{\mid I'_i \mid}{\mid I'_{node} \mid} ENT(I'_i), \quad (2)$$

$$ENT(I'_i) = p(l_i) \ \ (log_2 \ \ p(l_i)), \quad (3)$$

where $ENT(I'_i)$ is the Shannon Entropy of the probability class histogram $p(l_i)$. The leafnodes of all the trees in the forest will serves as a codebook. Then, the feature vectors are quantized by the trained RF codebook to form the BoW representation. The pLSA model is learned from the BoW whose element $BoW(w_j, d_n)$ stores the number of occurrences of a word w_j (i.e. codeword) in document d_n (i.e. image), where j is number of codewords and n is the number of images. The topics z_k of the image are selected accordingly to $p(z_k|d_n)$, where k is number of topics. The parameters are estimated by maximizing the log-likelihood algorithm:

$$p(d_n, w_j) = p(d_n) \sum_{k=1}^{K} p(w_j|z_k) p(z_k|d_n), \quad (4)$$

and we estimate the image-specific topic distribution $P(z_k|d_n)$ by

$$p(w_j|d_n) = \sum_{k=1}^{K} p(w_j|z_k) p(z_k|d_n), \quad (5)$$

B. Soft class labels

The soft class labels for each image patches are derived from the topic distributions that we learned from the pLSA. We calculate the image-codeword-specific topic distribution $p(z_k|w_j, d_n)$ as:

$$p(z_k|w_j, d_n) = \frac{p(w_j|z_k)p(z_k|d_n)}{\sum_{k=1}^{K} p(w_j|z_k)p(z_k|d_n)}. \quad (6)$$

In here, we assume that the topic distribution has a close relationship to the class-specific topic distribution, and hence we can estimate the distribution from the available labeled training images. Concretely, we define a Dominant Topic (DT) representation for each image class, i.e. there will be some topics that is representative to some classes. A mapping from the class-specific topic distribution $p(z_k|d_m)$ to the image-specific $p(c_m|d)$ class distribution can be derived as:

$$p(z_k|d_m) = \frac{\sum_{n \subset m} p(z_k|d_n)}{\sum_{m=1}^{M} p(z_k|d_m)}, \quad (7)$$

$$p(DT_m|d_n) = \frac{\sum_{k=1}^{K} p(z_k|d_m)}{\sum_{m=1}^{M} p(DT_m|d_n)}, \quad (8)$$

where $p(z_k|d_m) > 1/K$. Assume that DT_m summarizes image topics z_k that significantly represent class m, its probability distribution $p(DT_m|d_n)$ will be similar to $p(c_m|d_n)$:

$$p(c_m|d_n) \approx p(DT_m|d_n). \quad (9)$$

However, every single patch has different probability values based on the relationship between codewords w_j and patch feature vectors x_i. During the quantization process, x_i is represented by J codewords, where $J = R \times E$ while R is the number of trees used in codebook learning and E is the leafnodes per tree. Conventionally, each codeword gives a class probability based on the patches $p(c_m|x_i)$. However, by treating each codeword as an individual 'class', we can rewrite as

$$p(w_j|x_i) = \frac{1}{J} \sum_{r=1}^{R} p(w_{re}|x_i). \quad (10)$$

The image-patch-specific topic distributions $p(z_k|x_i, d_n)$ for each patches are then determined by summing the corresponding codeword probabilities of the particular feature vectors:

$$p(z_k|x_i, d_n) = \frac{p(z_k|w_j, d_n)p(w_j|x_i)}{\sum_{k=1}^{K} p(z_k|w_j, d_n)p(w_j|x_i)}. \quad (11)$$

We estimate the soft class labels for each patches, $p(c_m|x_i, d_n)$ by utilizing both the $p(z_k|x_i, d_n)$ and the $p(DT_m|d_n)$. Since $p(DT_m|d_n)$ is an image level distribution, we can define $p(DT_m|x_i, d_n)$ from it, where each x_i that belongs to similar d_n will have similar $p(DT|d_n)$ distribution. Therefore:

$$p(c_m|x_i, d_n) \approx \frac{p(DT_m|x_i, d_n)}{\sum_{m=1}^{M} p(DT_m|x_i, d_n)}. \quad (12)$$

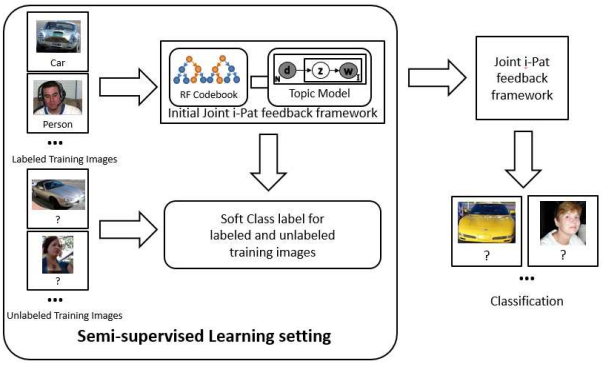


Fig. 3: Semi-supervised learning framework for joint i-Pat feedback mechanism.

These soft class labels will be used in generating a new RF codebook.

C. Feedback mechanism

The feedback mechanism continues by learning a new RF codebook using the soft class labels learned. Since the image class labels changes from discrete values to continuous, we refine the spliting criterion - Shannon Entropy ENT in Eq. 3. We compute the probability class histogram $p(l_i)_{i-Pat}$ as:

$$p(l_i)_{i-Pat} = \frac{p(c_m|x_i, d_n)}{\sum_{m=1}^{M} p(c_m|x_i, d_n)}, \tag{13}$$

where the histogram can be seen as the sum of the soft class labels of the images. Other setting of the RF codebook will remain the same. After that, a new pLSA model is learned from the new BoW based on the new RF codebook. We believe that by virtue of this close loop, the discriminative power of RF codebooks and pLSA model can be improved. Convergence is achieved when the pLSA model stop improving (in terms of classification performance). That is, we set a stopping condition where if the training results drop in a new iteration, we assume that the pLSA model has reached its convergence. Empirically, the feedback framework will converge within 2 to 5 iterations.

D. Classification

Given a testing image d_{test}, we estimate the image-specific topic distribution $p(z_k|d_{test})$:

$$p(w_j|d_{test}) = \sum_{k=1}^{K} p(w_j|z_k)p(z_k|d_{test}). \tag{14}$$

The dominant topics DT corresponding to class labels are obtained by Eq. 8. The soft class label $p(c_m|d_{test})$ of test image d_{test} are estimated:

$$p(c_m|d_{test}) \approx p(DT_m|d_{test}). \tag{15}$$

Algorithm 1 summarizes the proposed method. A set of class-specific thresholds, $thresh$ are identified from the training images $p(c_m|d_{train})$. $p(c_m|d_{test})$ that passes the respective $thresh$ will be classify as positive image.

Algorithm 1 Joint Image-Patch Level (Joint i-Pat) feedback

Require: A set of training images
Ensure: All parameters are set: number of trees, R, number of leafnodes, E, and number of topics, k
1. Initial learning of the RF codebook
2. Initial training of the pLSA using the BoW histogram based on the initial RF codebook as in step 1.
repeat
a. Infer soft class labels to associate with training image patches.
b. Re-learn RF codebook using the training image patches associated with corresponding soft class labels
c. Re-train the pLSA
until Classification result on training images are reduced
3.Classification using the final re-trained pLSA.

E. Semi-supervised Learning

Conventionally, RF cannot deal with semi-supervised setting since it needs image class labels for each features associated with it to train. Inspired by [15], we discover a possibility to extend the joint i-Pat feedback framework in semi-supervised learning (SSL) manner. Refer to Figure 3, for a set of training images that contains labeled and unlabeled images, we first learn the initial RF codebook and pLSA model from the labeled training images only. Secondly, we predict the soft class labels for both the labeled and unlabeled training images. Then, the rest of the process is same as to the original framework. With this, we can make use of the largely available unlabeled training data to further strengthen the RF codebook discriminative power.

IV. Experiments

We use the 15-Scene and C-Pascal dataset to test the difference of the proposed framework to the state-of-the-art methods. 15-Scene dataset [5] consists of both indoor and outdoor scene images. Each class consist of 200-400 images and 300 $\times$ 250 pixels respectively. We choose this dataset because this dataset has been extensively used in the research community in the image classification task that involve scene, which make it an important benchmark. C-Pascal dataset [16] is created from the bounding box annotations of the PASCAL VOC challenge 2008 training set [17] to extract the objects such that classification can be evaluated in a multi-class setting. This dataset contains 4450 images from 20 object classes but with varying object poses and background clutter.

For both datasets, we perform dense SIFT on patch size $= 8$ and step size $= 4$. We choose a small patch size and step size due to the low resolution constraint on some of the images in the C-Pascal dataset. Besides, for any image that have edge > 300 pixel, it will be resized to a maximum of 300 pixel and retain the aspect ratio. For the RF codebook settings, we use 10 random trees with 100 leafnodes, resulting in 1000 codeword histogram. Then, we use 20 topics during the pLSA learning. We use 100 training images for 15-Scene, and 30 training images for C-Pascal.

Experimental result: For the 15-Scene dataset results that depicted in Table I, it is noticed that our proposed method outperform current state-of-the-art method - ScSPM [19] which

TABLE I: Accuracy on 15-Scene Dataset compared to state-of-the-art methods

Labeled training image	10	100
Total training image	100	
ERC Forest [6]	49.95	73.84
Tree Quantizer [7]	48.14	**83.60**
Proposed method	**77.38**	82.48
KSPM [5]	81.40	
KC [18]	76.67	
ScSPM [19]	80.28	
ScSPM, base 1024, no SPM	63.13	
ScSPM, base 64, 3 level SPM	71.99	

TABLE II: Accuracy on C-Pascal dataset compared to state-of-the-art methods

Algorithm	Accuracy
multiple features + rank [20]	45.50
LP+ITML (best case) [21] (5 training)	36.40
NN with spDSIFT [22]	32.90
RALF [23]	37.30
GP-OA-Var [24] (area under AUC)	76.26
Proposed method (30 training, 5 labeled)	75.81
Proposed method (30 training, 30 labeled)	85.29

TABLE III: Convergence analysis on 15-Scene and C-Pascal Dataset in different training settings.

Dataset	labeled	Before feedback	1st iteration	convergence	best result
15 Scene	10	76.19	76.63	**77.38**	**77.38**
15 Scene	100	76.10	81.45	**82.48**	**82.48**
CPascal	5	72.51	75.56	**75.81**	**75.81**
CPascal	30	67.81	**85.20**	84.77	**85.20**

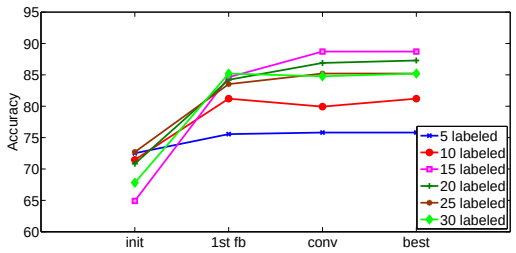


Fig. 4: Analysis on C-Pascal Dataset, specifically in convergence in semi-supervised learning (SSL) settings. init: Result by initial pLSA; 1st fb: Result after first feedback; conv: Result after convergence; best: Best result achieved out of all iteration.

is an unsupervised solution by 2.2%. However, one must note that the ScSPM that employed in this case is in the optimum settings as published in their paper. To have a fair and better understanding of the performance of ScSPM and our proposed solution, we reimplemented the ScSPM into two different settings: (ScSPM$_a$) 1024 bases with no Spatial Pyramid Matching (SPM), and (ScSPM$_b$) 64 bases with 3-level SPM, which both settings will result in 1000 bases or codewords to compare with our system with 1000 codewords. We outperform the ScSPM$_a$ and ScSPM$_b$ by 19.35% and 10.49%, respectively which is a considerable improvement. Compare to the ERC-forest [6], we also perform well as we have 8.64% improvement. This has shown that the ERC forest is affected by the wrongly labeled image patches while our proposed approach didnt. Although we are slightly weaker to Tree Quantizer [7] (a very small margin of 1.12%), in our experiment, we used a simple dense SIFT feature than their work, which sampled features on original image as well as four down-sampled version on the images. Also, in their work, they used 15 (class) $\times$ 10 (trees) $\times$ 100 (leafnodes) for their codebook representation in 15-Scene dataset experiment, which a lot larger to our current settings.

For the C-Pascal dataset results show in Table II, our proposed method yet again outperform the conventional solutions [20]–[24]. Even with less labeled training images (5 labeled in 30 training images), we still able to achieve comparable performance (rank 2 overall) with an accuracy of 75.81%. This has further justified the effectiveness of the proposed method.

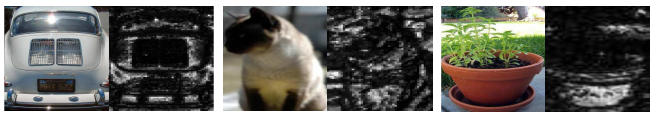
Convergence (comparison to conventional pLSA). Based on Table III, we treat each stage of the joint i-Pat feedback as an independent pLSA classifier, and report the results in iteration basis. Results on first iteration for both methods tend to be very close to the final converged results. This is expected because the first iterations and the following iterations have the same amount of features and soft class labels to learn the RF codebook, comparing to initial pLSA model which is build based on a limited labeled training images. Therefore the improvement before feedback and after first iteration is more significant. Also, the final convergence result doesn't necessary to be the best classification model, e.g. in Figure 4, the C-Pascal experiment that have 10, 20 and 30 labeled training images are fall in this category.

Semi-supervised learning: In Table I, for 15-Scene dataset, our SSL settings (10% of the total training image as labeled training image) have comparable result to the state-of-the-art solutions despite limited labeled training images are available. red Note that our proposed method in SSL is very similar to unsupervised learning using ScSPM, by a different of 10% labeled training images. Our proposed method are weaker to both the KSPM and ScSPM for 4.02% and 2.90% respectively, but we have an improvement of 14.25% and 5.39% compare to ScSPM$_a$ and ScSPM$_b$ respectively. This shows the flexibility and effectiveness of our proposed method working in the SSL environment. In the meantime, the ERC forest and Tree Quantizer methods degrade drastically to around 50% accuracy in this SSL environment, because both methods can only utilized the labeled training images during the RF learning, and the unlabeled training images cant be employed. Therefore, the number of images used during the RF learning is very limited, and results in poor performance. Bear in mind that our classifier is based on the pLSA topic model, which is a generative approach. Therefore, we believe that the classification result should be able to further increased if a hybrid approach as in [25] is applied.

The C-Pascal dataset result is explained in Figure 4, where the experiments is conducted with number of labeled training images increase gradually by 5. The experiment clearly shows the improvement from the feedback mechanism even from

(a) CALsuburb (b) Inside city (c) MIT tall building

Fig. 5: $p(c|x,d)$ visualization (right column) on selected images (left column) in 15-Scene dataset.



(a) Car (b) Cat (c) Potted plant

Fig. 6: $p(c|x,d)$ visualization (right column) on selected images (left column) in C-Pascal dataset.

1st iteration, for various settings. Also, fully labeled settings doesn't necessary to be the best result because there will be more background noise that is wrongly labeled in the initial learning, which weaken the initial RF codebook. With considerable unlabeled training images (in the experiment, roughly half of from the total training images), the RF codebook is allowed for room of improvement with the Joint i-Pat feedback mechanism and able to achieve better results.

V. Discussion

For the proposed method to work effectively, the soft class labels play a major role. In here, we visualize the image-patch-specific class distribution $p(c|x,d)$ to see the effects of soft class labels during the codebook updating process. Visualization of $p(c|x,d)$ for 15-Scene and C-Pascal dataset are illustrated in Figure 5-6 respectively. We show that $p(c|x,d)$ represent a rough silhouette to the original image itself. Besides, the high probability area (white area) normally reflects the edges of the image, which reflects the characteristic of the images especially objects in the image. $p(c|x,d)$ can be considered as a noise reduction in BoW learning technique. By assigning background patches as low probability area, we lower the chance that background patches are used in RF node splitting, to reduce RF degradation.

Computational cost of the proposed system are highly dependent on the number of iterations as one iterations will consist of RF codebook learning and pLSA learning. However, it is scalable to large dataset, as for each iterations, we just repeat the learning process by using soft class labels instead of ordinary class labels.

VI. Conclusion

We proposed a joint image-patch level (joint i-Pat) feedback framework which utilizes discriminative RF codebook learning and generative classifier learning in classification task. To achieve that, we estimate soft class labels for training images from initial pLSA model and initial RF codebook to update the RF codebook iteratively until convergence reached. We show that this framework can be applied in SSL application as well. The future work is to investigate different feature extraction parameter effect (e.g. patch size and step size) on soft class labels learning. Besides, we would like to find a more robust way for convergence decision.

References

[1] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. 1

[2] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *CVPR*, 2005. 1, 2

[3] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005. 1, 2

[4] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *ICCV*, 2005. 1

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006. 1, 4, 5

[6] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification." *T-PAMI*, vol. 30, 2008. 1, 2, 5

[7] J. Krapac, J. Verbeek, and F. Jurie, "Learning tree-structured descriptor quantizers for image categorization," in *BMVC*, 2011. 1, 2, 5

[8] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *CVPR*, 2012. 1

[9] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." in *VISAPP (1)*, 2009. 2

[10] Q. Li, C. Yao, L. Wang, and Z. Tu, "Randomness and sparsity induced codebook learning with application to cancer image classification," in *CVPR Workshop*, 2012. 2

[11] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *NIPS*, 2008. 2

[12] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning J.*, vol. 42, pp. 177–196, 2001. 2

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Jour. of Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. 2

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006. 2

[15] C. Leistner, A. Saffari, J. Santner, and H. Bischof, "Semi-supervised random forests," in *ICCV*, 2009. 4

[16] S. Ebert, M. Fritz, and B. Schiele, "Pick your neighborhood–improving labels and neighborhood structure for label propagation," in *PR*. Springer, 2011. 4

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html. 4

[18] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *T-PAMI*, vol. 32, pp. 1271 –1283, 2010. 5

[19] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009. 4, 5

[20] S. Ebert, D. Larlus, and B. Schiele, "Extracting structures in image collections for object recognition," in *ECCV*. Springer, 2010. 5

[21] S. Ebert, M. Fritz, and B. Schiele, "Pick your neighborhood–improving labels and neighborhood structure for label propagation," in *PR*. Springer, 2011. 5

[22] ——, "Semi-supervised learning on a budget: scaling up to large datasets," in *ACCV*. Springer, 2013. 5

[23] ——, "Ralf: A reinforced active learning formulation for object class recognition," in *CVPR*, 2012. 5

[24] P. Bodesheim, A. Freytag, E. Rodner, and J. Denzler, "Approximations of gaussian process uncertainties for visual recognition problems," in *Image Analysis*. Springer, 2013. 5

[25] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *T-PAMI*, vol. 30, pp. 712 –727, 2008. 5