

Enhanced Random Forest with Image/Patch-Level Learning for Image Understanding

Wai Lam Hoo*, Tae-Kyun Kim[†], Yuru Pei[‡] and Chee Seng Chan*

*Center of Image and Signal Processing, 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 understanding is an important research domain in the computer vision due to its wide real-world applications. For an image understanding framework that uses the Bag-of-Words model representation, the visual codebook is an essential part. Random forest (RF) as a tree-structure discriminative codebook has been a popular choice. However, the performance of the RF can be degraded if the local patch labels are poorly assigned. In this paper, we tackle this problem by a novel way to update the RF codebook learning for a more discriminative codebook with the introduction of the soft class labels, estimated from the pLSA model based on a feedback scheme. The feedback scheme is performed on both the image and patch levels respectively, which is in contrast to the state-of-the-art RF codebook learning that focused on either image or patch level only. Experiments on 15-Scene and C-Pascal datasets had shown the effectiveness of the proposed method in image understanding task.

I. INTRODUCTION

Recent studies on image understanding have shown to flavour part-based representation such as the Bag-of-words (BoW) model [10], [12]–[14], [17], [19], [21]. Each image given as a set of local patches is represented by a histogram of codewords. Given the codeword representations, topic discovery model such as the pLSA model [11] has been successfully applied to semantic image clustering and unsupervised learning for object categorization and scene understanding. It has been shown that the visual codebook, which is typically obtained by the k-means clustering of the local patches [10], [14], [21], is a crucial part to achieve good performance.

On the other hand, discriminative codebooks have clearly shown advantage compared to its counterpart (e.g. k-means) when the ground-truth class label of the training image is given. Random Forest (RF) [13], [17], an ensemble of decision trees with randomization, appears to be a very fast algorithm compared to the k-means codebook where the clustering and vector quantization process are highly time-demanding. Besides, the RF also shows its discriminative power as an effective codebook for object categorization and segmentation. However, this advantage is heavily relied on the accuracy of the ground-truth class label. For example, let us assume that the ground-truth class label in Figure 1a is belong to the ‘Face’ class, and hence all the local patches for the image will be associated with the ‘Face’ class label. However, we can clearly notice that not all the local patches are belong to the ‘Face’

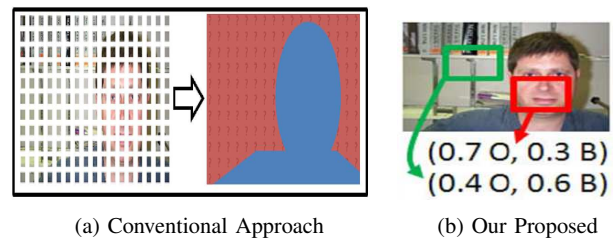


Fig. 1: (a) Weakly supervised learning by local patches. It is clear that not all the local patches (i.e. red region) in the image are belong to the ‘Face’ class label. (b) Our proposed method with the use of soft assignment where the local patches are assigned class labels comparatively independently. ‘O’ indicates the object and ‘B’ indicates the background. Best viewed in color.

class label (e.g. the background patches (red region) should not belong to the ‘Face’ class label). As such, during training the RF codebook, the background patches which are wrongly labeled in this case will greatly degrade the discriminative power of the RF codebook.

In this paper, we present a novel way of learning the RF by introducing the soft class labels, estimated from the pLSA model based on a feedback scheme. Particularly, we adapt the RF node split strategy to account for the soft class labels obtained from the initial weak pLSA classifier. The feedback scheme can be performed on the image and patch levels respectively; and we anticipated that the RF re-learning and the pLSA re-training in the close feedback loop will improve the discriminative power of the codebooks. Experiments on 15-Scene and C-Pascal datasets have shown that the proposed codebook outperforms the state-of-the-art methods such as [13], [17].

This rest of the paper is arranged as follows: Section II discusses the recent developments in related topics including the visual codebook learning and topic model. Section III details the proposed framework. We show the experimental results in Section IV. Finally, discussions and conclusion are drawn in Section V-VI, respectively.

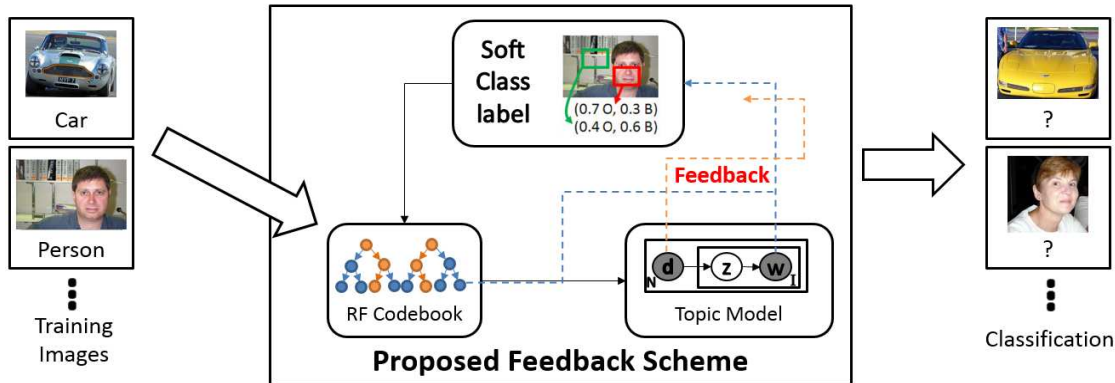


Fig. 2: An overview of the proposed feedback framework. The blue dotted line indicates patch information from the RF and pLSA, while orange dotted line indicates the image level information from the pLSA. Best viewed in color.

II. RELATED WORK

Visual codebook learning is an essential pipeline in the BoW representation. In order to find the optimal codewords, unsupervised methods such as the k-means [21] and kd-tree [18] had been employed. However, recent research work focused on learning the visual codebook using labeled images (i.e. supervised manner), in order to have a better discrimination on the codebook learning.

According to [4], RF which offers discriminative characteristics compared to the generative approaches such as the K-means is a popular choice. For instance, Moosmann et al. [17] used the image classification results from the RF as a feedback mechanism in the interest point detector to create a saliency map. The object location from the saliency map was then used in re-learning the RF for a better codebook. We denote this approach as the patch-level feedback scheme. Krapac et al. [13] proposed another variant of feedback scheme in the RF re-learning. They performed codebook learning by alternating the quantizers and classifiers to maximize the classification performance. In each split node, the data set is separated into the training and validating set to evaluate the classifier. Finally, the results from the classifier are feedback to the nodes for the optimal node splitting. Similar joint learning approaches between the dictionary and the classifier were also studied in [3], [16], [24] and we denote these approaches as the image-level feedback scheme. However, in this paper, a special feedback scheme that utilizes both the image and patch-level is proposed. In particular, we utilize the topic model for the soft assignment on image and patch labels to further improve the RF learning for a more discriminative codebook.

Topic models are widely applied in image classification [20]. The topic models are particularly effective when pairing with the BoW representation, where the models group ambiguous codewords together and generate a topic distribution over a codebook. One of the most popular topic model is the pLSA [11] which serves as a mid-level clustering method and tries to find the relationship of codewords. The codewords are grouped together for some meaningful representations. For instance, a "Face" class image as illustrated in Figure 1a where ideally the pLSA will cluster the image regions into two parts. One part represents the face by grouping eyes, hairs and mouths;

while the other represents the background (e.g. book, sky) by codewords. Similar to other clustering methods, the pLSA also requires to find the optimal number of topics to represent a particular image effectively. In this paper, we employ the pLSA to estimate the soft class label for the training data. In turn, these soft class labels are combined together to re-learn (enhance) the RF codebook and follow by re-train the pLSA model in a novel feedback scheme.

In summary, our main contribution is the introduction of **a special feedback scheme where the soft class labels** from a topic model are used to update the RF learning for a more discriminative codebook. The feedback scheme can be performed on the image and patch levels respectively for the image understanding task. Such a framework is more superior compare to the conventional solutions [13], [17] as it includes both the image-level information, estimated from the topic model, as well as the patch-level information, estimated from both the topic model and RF codebook respectively. This is in contrary to [17] that only employs the patch-level information, and [13] that only employs the image-level information.

III. METHODOLOGY

The proposed method is illustrated in Figure 2. First, we learn a weak RF using the local patches that are associated with the ground-truth image class label. Treating the RF leafnodes as codewords, we build the BoW representation from the RF codebook. Secondly, we train a weak pLSA model from the RF codebook in order to estimate the soft class labels. Thirdly, a feedback scheme where the soft class labels from the weak topic model are used to update the RF (re-learning), and follow by a new (enhanced) pLSA model is trained from the refined RF codebook. The feedback scheme will iterate until the convergence criteria is satisfied. Finally, classification is performed using the converged pLSA model.

A. Initial RF codebook learning and pLSA model training

We start with a weak classifier construction by a RF codebook and a pLSA model from a set of labeled training images. The RF as an ensemble of the random decision trees by bagging provides a very fast way of codebook learning and

quantization. Moreover, when the class labels are available, it has an advantage as a discriminative codebook. The random decision tree is constructed using a random subset of the training data with replacement. The labeled training images $I'_N = \{x_i, l_i\}$ at a specific node N , where x_i, l_i are the feature vectors of the local patches and the corresponding class labels respectively, are recursively splitted into left I'_L and right I'_R subsets, according to a set of thresholds T and a split function f , as

$$I'_L = \{x_i \in I'_N | f(x_i) < T_t\}, I'_R = I'_N \setminus I'_L. \quad (1)$$

At each split node, a random subset of the features are generated and compare to T . The ones that maximize the expected information gain ΔE are selected. Specifically, at each split node:

$$\Delta E = E(I'_N) - \sum_{i=L,R} \frac{|I'_i|}{|I'_N|} E(I'_i), \quad (2)$$

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

where $E(I'_i)$ is the Shannon Entropy of the probability class histogram $p(l_i)$ of the training images I . The leafnodes of all the trees in the forest will serve as a codebook. Then, the feature vectors x_i are quantized by the learnt RF codebook to form the BoW representation. Initial pLSA model is trained from the BoW whose element $\{w_j, d_n\}$ stores the number of occurrences of a codeword w_j in the image d_n , where j is number of codewords and n is the number of images. The image topics z_k are selected accordingly to the image-specific topic distribution $p(z_k|d_n)$, where k is number of topics. These parameters are estimated by maximizing the log-likelihood algorithm as:

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

and we can 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

In this paper, we introduce a soft class label in which the local patches are assigned class labels comparatively independently. As to one image, the local patches in the object regions are assigned the labels with respect to the object classes, while the local patches in the background regions are assigned to background labels. The feedback in the image and patch level assigns confidences to different class labels to every single patch, which produce discriminative label to patches in different regions. This is in contrast to the conventional solutions [13], [17] where all the local patches in the same image are assigned the same image label. The confusions can be caused in the codebook learning when the local patches actually in the background are assigned labels of objects as depicted in Figure 1a.

The soft class labels are estimated from $p(z_k|d_n)$ (i.e. from the initial pLSA model), and hence we can 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, it is assumed that $p(z_k|d_n)$ has a close relationship to the image-level soft class label $p(c_m|d_n)$, and so we can estimate $p(c_m|d_n)$ from the available labeled training images where c is the object class, and m is the number of classes. Concretely, we define a Dominant Topic representation \mathcal{T}_m for each c_m , i.e. a set of z_k that are representative for a particular class c_m . So, we can calculate class-specific topic distribution $p(z_k|d_m)$ as:

$$p(z_k|d_m) = \frac{\sum_{n \in \mathcal{T}_m} p(z_k|d_n)}{\sum_{m=1}^M p(z_k|d_m)}. \quad (7)$$

Then, for z_k that satisfy the condition $p(z_k|d_m) > 1/K$, we assign to \mathcal{T}_m and compute $p(c_m|d_n)$:

$$p(c_m|d_n) = \frac{\sum_{k \in \mathcal{T}_m} p(z_k|d_n)}{\sum_{m=1}^M p(c_m|d_n)}. \quad (8)$$

However, every local patch has different probability values based on the relationship between the codewords w_j and patch feature vectors x_i . During the quantization process, x_i is represented by J codewords, where $J = R \times S$ and R is the total number of trees used in codebook learning, while S is the total leafnodes per tree. Conventionally, each w_i gives a class probability based on the local patches $p(c|x_i)$. However, we treat each w_i as an individual ‘class’, and so $p(c|x_i)$ can be rewritten as codeword probability of the feature vector, $p(w_j|x_i)$:

$$p(w_j|x_i) = \frac{1}{J} \sum_{r=1}^R \sum_{s=1}^S p(w_{r,s}|x_i). \quad (9)$$

The image-patch-specific topic distributions $p(z_k|x_i, d_n)$ are then determined by summing the corresponding $p(w_j|x_i)$:

$$p(z_k|x_i, d_n) = \frac{p(z_k|w_j, d_n)p(w_j|x_i)}{\sum_{k=1}^K p(z_k|w_j, d_n)p(w_j|x_i)}. \quad (10)$$

In order to estimate the soft class labels for each local patches $p(c_m|x_i, d_n)$, we utilize $p(z_k|x_i, d_n)$ and \mathcal{T}_m . As before, the \mathcal{T}_m is a representation that is defined based on $p(z_k|d_m)$ (i.e. a set of z_k that significantly represents c_m), hence all x_i that belong to the same c_m will share the same \mathcal{T}_m . With this, we can compute $p(c_m|x_i, d_n)$ as:

$$p(c_m|x_i, d_n) = \frac{\sum_{k \in \mathcal{T}_m} p(z_k|x_i, d_n)}{\sum_{m=1}^M p(c_m|x_i, d_n)}. \quad (11)$$

C. Feedback mechanism

In order to re-learn (enhance) the weak RF codebook, we use the soft class label estimated from the topic model (Section III-B). Since x_i labels change from hard assignment $p(l_i)$ to soft assignment $p(c_m|x_i, d_n)$, we refine the splitting criterion - Shannon Entropy E in Eq. 3 and we compute the new probability class histogram $p'(l_i)$ as:

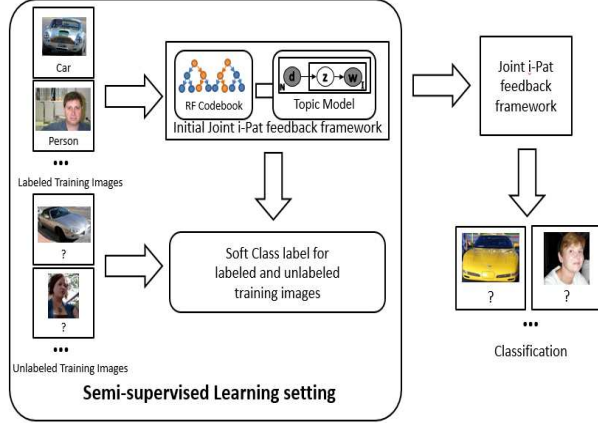


Fig. 3: Semi-supervised learning for the proposed framework.

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

Other settings of the RF codebook is remain the same. Then, an enhanced pLSA model is trained from the new BoW based on the refined RF codebook.

D. Classification

Finally, we estimate the image-specific topic distribution of the test images $p(z_k|d_{\text{test}})$ as:

$$p(w_j|d_{\text{test}}) = \sum_{k=1}^K p(w_j|z_k)p(z_k|d_{\text{test}}). \quad (13)$$

In order to estimate the image-level soft class label of test image $p(c_m|d_{\text{test}})$, we use $p(z_k|d_{\text{test}})$ and similar \mathcal{T}_m from the Section III-B:

$$p(c_m|d_{\text{test}}) = \frac{\sum_{k \in \mathcal{T}_m} p(z_k|d_{\text{test}})}{\sum_{m=1}^M p(c_m|d_{\text{test}})}. \quad (14)$$

A set of class-specific thresholds, h_m are identified from the soft class labels of a set of training images $p(c_m|d_{\text{train}})$. For images that $p(c_m|d_{\text{test}}) > h_m$, it will be denoted as the correct classification. Algorithm 1 summarizes the proposed method.

E. Semi-supervised Learning

In a real world, however, few data are labeled and obtaining exhaustive annotation is impractically expensive. As such, the semi-supervised methods have been studied. In order to show the capability of the proposed method when confronted with the unlabeled data in the training image set, we show here how the proposed framework can be extended to the semi-supervised learning paradigm. Conventionally, the RF cannot deal with semi-supervised learning since it needs image class labels for each features associated with it when training. Inspired by [15], we discover a possibility to extend the proposed framework to semi-supervised method. Refer to Figure 3, all the operations are similar as to Algorithm 1 except we will estimate the soft class labels for both the labeled and unlabeled training images.

Algorithm 1 Proposed Framework

Require: A set of labeled training image patches $\{x_i, l_i\}$.

Ensure: All parameters are set: total number of trees R , total number of leafnodes E and total number of topics, K

1. Initial learning of the RF using $\{x_i, l_i\}$.
2. Initial training of the pLSA model using the BoW histogram based on the initial RF as in step 1.

repeat

- a. Infer soft class labels $p(c_m|x_i, d_n)$ to associate with x_i .
- b. Re-learn RF using x_i that associated with corresponding $p(c_m|x_i, d_n)$.
- c. Re-train the pLSA model based on the initial RF as in step 2b.

until Convergence criteria is satisfied

3. Classification using the final pLSA model.
-

IV. EXPERIMENTS

In the experiment, we employed 15-Scene and C-Pascal datasets to test the effectiveness of the proposed framework, and as a comparison to the state-of-the-art methods. The 15-Scene dataset [14] consists of both indoor and outdoor scene images. Each class consists of 200-400 images with 300×250 pixels respectively. The C-Pascal dataset [5] is created based on the bounding box annotations for each object class from the PASCAL VOC challenge 2008 dataset [9]. As such, the classification can be evaluated in a multi-classes setting. This dataset contains 4450 images from 20 object classes with varying object poses and background clutters.

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

Experimental result: For the 15-Scene dataset results that depicted in Table I, it is noticed that our proposed method outperform the state-of-the-art method ScSPM [23] with an improvement of 2.2%. Though the improvement seems narrow, one must note that the employed ScSPM is in the optimum settings as published in their paper. To have a better understanding of the performance of the ScSPM and our proposed solution, we reimplemented the ScSPM into two different configurations: (ScSPM_a) - a 1024 bases with no Spatial Pyramid Matching (SPM), and (ScSPM_b) - a 64 bases with 3-level SPM. Both configurations will result in 1000 bases/codewords that is similar to our proposed framework for a fair comparison. Again, the proposed method outperforms the ScSPM_a and ScSPM_b with improvement of 19.35% and 10.49%, respectively. Compare to the ERC-forest [17] which is a patch-level feedback scheme solution, we also perform well as we have 8.64% improvement. This has shown that the ERC forest is affected by the wrongly labeled local patches.

Although we are slightly weaker (i.e. a very small margin

TABLE I: Accuracy (in terms of %) on the 15-Scene Dataset and comparison to the state-of-the-art methods

Labeled training image	10	100
Total training image	100	
ERC Forest [17]	49.95	73.84
Tree Quantizer [13]	48.14	83.60
Proposed method	77.38	82.48
KSPM [14]	81.40	
KC [22]	76.67	
ScSPM [23]	80.28	
ScSPM, base 1024, no SPM	63.13	
ScSPM, base 64, 3 level SPM	71.99	

TABLE II: Accuracy (in terms of %) on the C-Pascal dataset and comparison to the state-of-the-art methods

Algorithm	Accuracy (%)
Multiple features + rank [8]	45.50
LP+ITML (best case) [5] (5 training)	36.40
NN with active learning using sDSIFT [7]	32.90
RALF [6]	37.30
GP-OA-Var [1] (area under AUC)	76.26
Proposed method (30 training, 5 labeled)	75.81
Proposed method (30 training, 30 labeled)	85.29

of 1.12%) to the Tree Quantizer [13] when labeled training image = 100, in our experiment, we used a simple dense SIFT feature compared to their work, which sampled features on the original image scale, as well as at four down-sampled versions (i.e. each time, the image is re-scaled by a factor of 1.2). Also, in their work, they used 15 (class) \times 10 (trees) \times 100 (leafnodes) for their codebook representation in the 15-Scene dataset experiment, which is a lot larger to ours, and thus resulting in much higher computational cost compared to our proposed method..

For the C-Pascal dataset results show in Table II, our proposed method yet again outperform the conventional solutions [1], [5]–[8]. Even with a lower amount of labeled training images (i.e. only 5 labeled in 30 training images), the proposed method still capable to achieve comparable performance (i.e. rank 2 overall) with an accuracy of 75.81%.

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

TABLE III: Convergence analysis (accuracy) for 15-Scene and C-Pascal datasets in different training settings.

Dataset	Number of Labeled Images	Initial Learning	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%

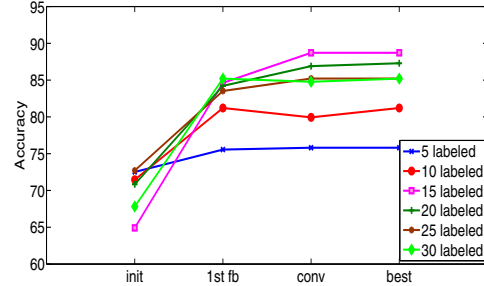


Fig. 4: Analysis on C-Pascal dataset, specifically the convergence analysis in semi-supervised learning. init: Result by initial pLSA; 1st fb: Result for first feedback; conv: Convergence result; best: Best result achieved out of all iteration. Best viewed in color.

classification model.

Semi-supervised learning: In Table I, for the 15-Scene dataset, our SSL settings (i.e. 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. Note that this amount of labeled training images are very limited, and very close to unsupervised learning using the ScSPM. 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. 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 can be better if a hybrid approach as in [2] is applied.

The C-Pascal dataset result is explained in Figure 4, where the experiments are conducted with the amount of labeled training images increase gradually by 5. The experiments clearly show the improvement from the feedback mechanism for various settings. Also, fully labeled settings doesn't necessary provide the best result because there will be more local patches that are wrongly labeled in the initial learning, which weaken the initial RF. With considerable amount of unlabeled training images (i.e. in the experiment, roughly half of from

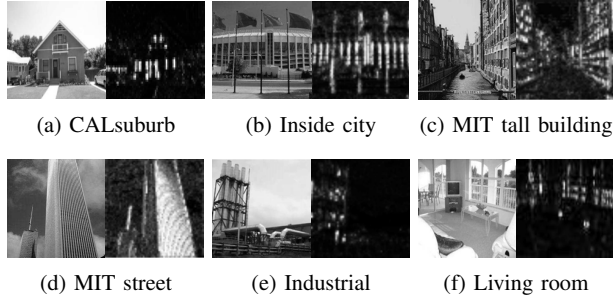


Fig. 5: $p(c|x, d)$ visualization (right column) on selected images (left column) in the 15-Scene dataset. Best viewed in color.

the total training images), the RF have high chances to be enhanced by the unlabeled training images with the 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 review the effects of soft class labels during the codebook updating process. Visualization of $p(c|x, d)$ for the 15-Scene and C-Pascal datasets are illustrated in Figure 5-6 respectively. We show that $p(c|x, d)$ represents a rough silhouette to the original image. Besides, the high probability area (white area) normally reflects the edges of the image, that is reflects the characteristic of the images especially objects in the image. $p(c|x, d)$ can be considered as an error reduction in the RF learning. By assigning background patches as low probability area, we reduce the probability that the background patches are employed in the RF node splitting, and hence improve the RF discriminative power.

On another aspect, the computational cost of the proposed method depends on the number of iterations as one iteration consists of RF codebook learning and pLSA learning. However, subsequent iterations get accelerated as we just repeat the learning process by using soft class labels instead of ordinary class labels.

VI. CONCLUSION

In this paper, we proposed a novel feedback framework which utilizes the discriminative RF codebook learning and generative classifier learning in image understanding task. To achieve that, we estimate soft class labels from the initial pLSA model and RF codebook to update the RF codebook iteratively until convergence is reached. We show that this framework can be applied in SSL paradigm as well. The future work is to investigate different feature extraction parameters effect (e.g. patch size and step size) on the soft class labels learning. Besides, we are also interested to find a more robust way for the convergence decision.

ACKNOWLEDGMENT

This research is supported by the University Malaya Research Grant (UMRG) Grant RP023-2012D, H-000000-56657-

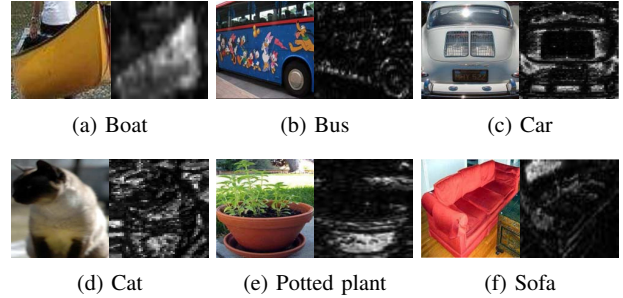


Fig. 6: $p(c|x, d)$ visualization (right column) on selected images (left column) in the C-Pascal dataset. Best viewed in color.

E13110 and the Bright Sparks Programme (BSP) from the University of Malaya.

REFERENCES

- [1] P. Bodesheim, A. Freytag, E. Rodner, and J. Denzler. Approximations of gaussian process uncertainties for visual recognition problems. In *SCIA*, 2013.
- [2] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *T-PAMI*, 30, 2008.
- [3] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.
- [4] L. Breiman. Random forest. *Mach. Learn. J.*, 45:5–32, 2001.
- [5] S. Ebert, M. Fritz, and B. Schiele. Pick your neighborhood—improving labels and neighborhood structure for label propagation. In *DAGM*, 2011.
- [6] S. Ebert, M. Fritz, and B. Schiele. Ralf: A reinforced active learning formulation for object class recognition. In *CVPR*, 2012.
- [7] S. Ebert, M. Fritz, and B. Schiele. Semi-supervised learning on a budget: scaling up to large datasets. In *ACCV*, 2013.
- [8] S. Ebert, D. Larlus, and B. Schiele. Extracting structures in image collections for object recognition. In *ECCV*, 2010.
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008).
- [10] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
- [11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning J.*, 42, 2001.
- [12] W. L. Hoo and C. S. Chan. plsa-based zero-shot learning. In *ICIP*, 2013.
- [13] J. Krapac, J. Verbeek, and F. Jurie. Learning tree-structured descriptor quantizers for image categorization. In *BMVC*, 2011.
- [14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
- [15] C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *ICCV*, 2009.
- [16] X. Lian, Z. Li, B. Lu, and L. Zhang. Max-margin dictionary learning for multiclass image categorization. In *ECCV*, 2010.
- [17] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *T-PAMI*, 30, 2008.
- [18] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009.
- [19] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR*, 2012.
- [20] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *T-PAMI*, 29:1575–1589, 2007.
- [21] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *CVPR*, 2005.
- [22] J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *T-PAMI*, 32, 2010.
- [23] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.
- [24] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *CVPR*, 2010.