

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

Abstract—This is the supplementary material for paper 1032 titled: Image classification for codebook updating using joint i-Pat topic model feedback. Firstly, we will clarify our contribution in the paper. Secondly, we will emphasize on what are the advantages of our proposed method compare to state-of-the-art methods. Finally, we will have in-depth discussion on soft class label, $P(c|x, d)$ visualization, on top from the paper.

I. CONTRIBUTION

We propose a **joint image-patch level (joint i-Pat) feedback mechanism**. The aim is to strengthen the RF codebook utilize the pLSA topic model. To strengthen it, a new image class labelling strategy is proposed, to tackle the problem of the false-labeled background patches. In joint i-Pat feedback framework, we learn **soft class labels** from pLSA topic model, specifically image-codeword-specific topic distribution, $p(z|w, d)$. By identifying Dominant Topics, DT for each image class:

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

we can learn soft class labels $p(c|x, d)$ by relating codeword information to image patches:

$$p(c_m|x_i, d_n) \approx \frac{p(DT_m|x_i, d_n)}{\sum_{m=1}^M p(DT_m|x_i, d_n)}. \quad (2)$$

II. COMPARISON TO STATE-OF-THE-ART METHODS

In conventional BoW algorithm, unsupervised learning for codebook e.g. k-means [?], [?] and sparse coding [?] are used. However, in these works, they didn't utilize image class labels in their learning, resulting in a less discriminative codebook.

Figure 1 explains conventional methods that use RF codebook and how our method differ with them. In Moosmann et. al. [?] work, they learn patches information using RF as codebook, which replacing k-means, thus making the codebook more discriminative and time-efficient. Besides, Krapac et. al. [?] perform image level feedback learning by optimizing RF splits to directly maximize classification performance on validation set. Both methods uses RF to utilize image class labels for learning. However, both methods only focus on either patch level [?] or image level [?] information.

III. SOFT CLASS LABELS VISUALIZATION

Soft class labels $p(c|x, d)$ are estimated to enhance the RF codebook learning, so that during RF node splitting, it won't be confusing by background noises that wrongly-labeled as image class label. By visualizing the $p(c|x, d)$ of each image, we can roughly know what are the new labels that will be feedback to the RF codebook. This is a qualitative measurement to see how well the soft class labels estimation goes. Ideally, we can 'segment' the object class from the background.

The visualizations on selected images for 15-Scene and CPascal dataset are shown in Figure 2 and 3 respectively. In 15-Scene dataset case, object class would represent important elements in the class, e.g. for CALSuburb class, the important aspect should be the house itself, and may be some trees. From the Figure 2, we can see that the soft class labels assign higher probability to significant edges in the image (e.g. CALSuburb and MITtallbuilding class). However, if the image have a big grayscale color map that is smooth (e.g. MITcoast), then the soft class label might fail, and assign nearly equal distribution to all the patches. We consider this case as this image learning cannot be improved by the feedback mechanism. We believe this cause by a small patch size used in image feature extraction by DSIFT. Take note that our main contribution is to have a better image level input for image features using feedback mechanism, therefore, the problems of patch size in this experiment is not discussed.

The visualizations on selected images for CPascal dataset are shown in Figure 3. The aim is to have high probability on objects and low probability on background patches. We achieve considerable results as in the figure shown, where we manage to have high probability in some image edges (e.g. bicycle). However, there are some cases where the system mistaken the background as foreground (e.g. plane). Also, our proposed method didn't work well on extreme-low resolution image e.g. 'bottle' class with as low as 8×13 pixels.

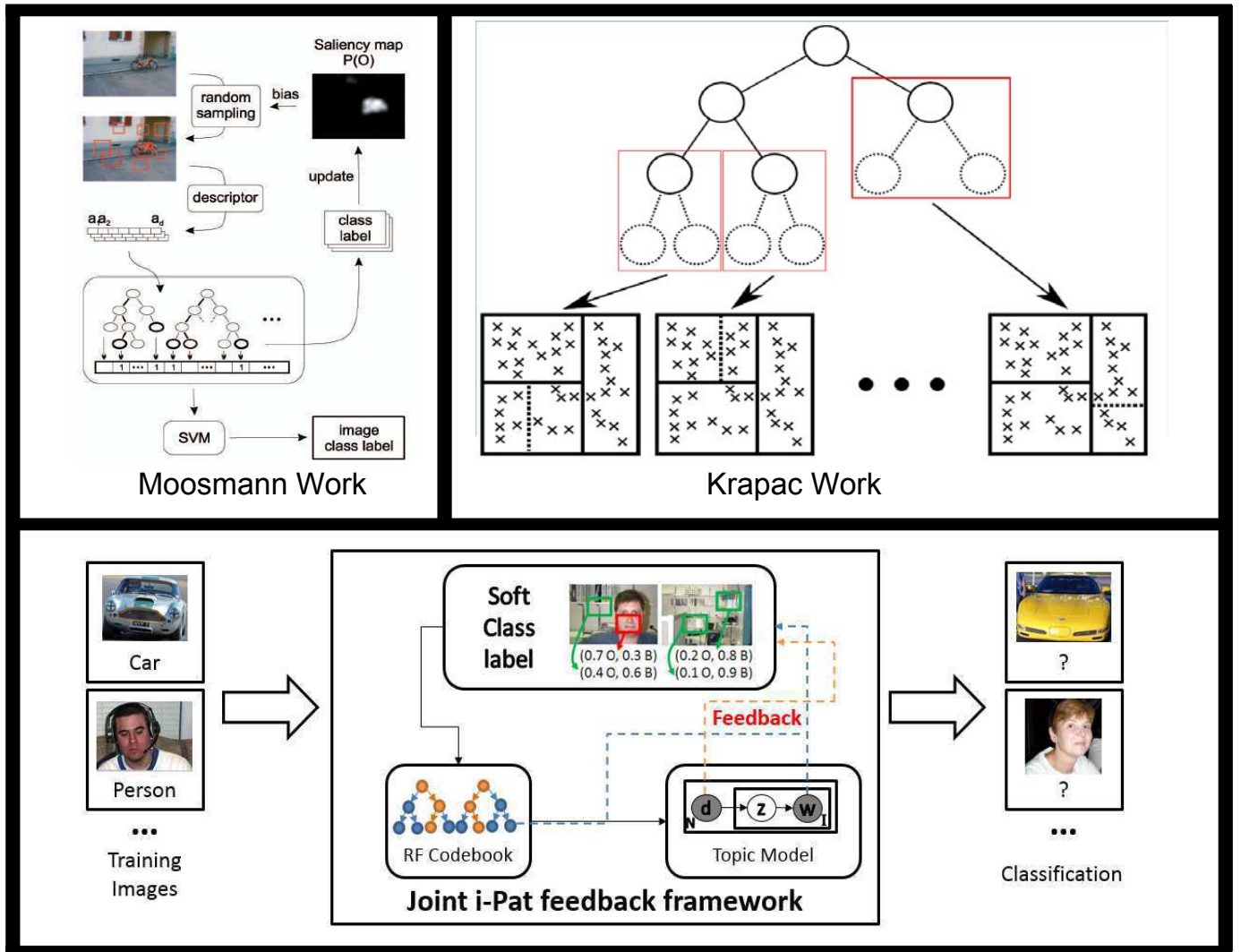


Fig. 1: Comparison of conventional method to joint i-Pat feedback framework. (Source from [?] and <http://hal.inria.fr/docs/00/61/31/18/IMG/Screenshot.png>)



























































































CALsuburb						
MITcoast						
MITforest						
MIThighway						
MITinsidacity						
MITmountain						
MITopencountry						
MITstreet						
MITtallbuilding						
PARoffice						
bedroom						
industrial						
kitchen						
livingroom						
store						

Fig. 2: $P(c|x, d)$ visualization on 15-Scene dataset. The first, third and fifth column are the sample images while the second, fourth and sixth columns are corresponding $p(c|x, d)$ visualization. Each row represents an image class.

























































































































Aeroplane	 	 	 
Bicycle	 	 	 
Bird	 	 	 
Boat	 	 	 
Bottle	 	 	 
Bus	 	 	 
Car	 	 	 
Cat	 	 	 
Cow	 	 	 
Chair	 	 	 
Diningtable	 	 	 
Dog	 	 	 
Horse	 	 	 
Motorbike	 	 	 
Person	 	 	 
Pottedplant	 	 	 
Sheep	 	 	 
Sofa	 	 	 
Train	 	 	 
Tvmonitor	 	 	 

Fig. 3: $P(c|x, d)$ visualization on CPascal dataset. The first, third and fifth column are the sample images while the second, fourth and sixth columns are corresponding $p(c|x, d)$ visualization. Each row represents an image class.