

Learning Discriminative Mid-Level Patches for Fast Scene Classification

Angran Lin^(✉), Xuhui Jia, and Kwok Ping Chan

Department of Computer Science,
The University of Hong Kong, Hong Kong, China
{arlin, xhjia, kpchan}@cs.hku.hk
<http://www.cs.hku.hk>

Abstract. Discriminative mid-level patch based approaches have become increasingly popular in the past few years. The reason of their popularity can be attributed to the fact that discriminative patches have the ability to accumulate low level features to form high level descriptors for objects and images. Unfortunately, state-of-the-art algorithms to discover those patches heavily rely on SVM related techniques, which consume a lot of computation resources in training. To overcome this shortage and apply discriminative part based techniques to more complicated computer vision problems with larger datasets, we proposed a fast, simple yet powerful way to mine part classifiers automatically with only class labels provided. Our experiments showed that our method, the Fast Exemplar Clustering, is 20 times faster than the commonly used SVM based methods while at the same time attaining competitive accuracy on scene classification.

Keywords: Discriminative mid-level patches · Fast scene classification · Fast exemplar clustering

1 Introduction

Scene classification is not an easy task due to the various visual appearances of different scenes and the complexity in their compositions. Recently, new approaches using discriminatively trained part classifiers are applied to this problem and achieved better performance than conventional methods [1, 2]. This is not surprising since part classifiers have the ability to accumulate low level features to generate high level descriptors for each image, which carry information of the visual elements that appear frequently to better describe our real world.

Scene classification is not the only computer vision topic that benefits from part based models. As a matter of fact, in the last few years part based models have been applied to topics like object detection [3, 4], motion detection [5] and video classification [6, 7]. The reason why part based methods become so popular can be attributed to two reasons. Firstly, they focus on a key problem in computer vision. The relationship between discriminative patches and images can be described as an analogy to the relationship between words and articles.

The Bag of Words (BoW) models, including Locality-constrained Linear Coding (LLC) BoW [8] or Improved Fisher Vectors (IFV) [9] succeeded in answering this question to some extent, but the idea of training part classifiers that are visually discriminative may have pushed us one step further. In particular, [10, 11] have shown the benefits of using desired patches as visual words [12]. Secondly, most techniques in the framework are shared among different computer vision tasks, which suggests a great potential for this technique.

Given all these advantages, there are still several issues remain unsolved. The most important one is the computation consumption in the training stage. Most discriminative mid-level discovery algorithm rely on a max-margin framework which uses variants of SVMs like exemplar SVM [1] and miSVM [3]. To achieve broad coverage and better purity, thousands of training rounds are required. Moreover, in each round the classifiers/detectors are learned in an iterative manner. Thus, the complexity of using a standard procedure that involves hard negative mining for a huge amount of classifiers would be surprisingly high. It leaves us a major challenge: a simple, efficient and effective method is yet to be found.

In this paper, we proposed a fast algorithm to discover discriminative mid-level patches. It is named Fast Exemplar Clustering (FEC), which works extremely fast while at the same time, attaining competitive accuracy. As a comparison, the MIT 67 indoor scene classification problem in Sect. 5 spent only one day in training on an ordinary Core-i5 computer, while the commonly used methods today would take several days on a cluster [13].

The vastly improved efficiency of FEC method benefits from two factors. The first one is that FEC only requires spatial information of feature vectors and classifiers are trained using their distance measure rather than iteratively solving a time consuming optimization problem. The second one is that FEC uses only local information instead of global information to train classifiers. When the number of patches increases, the training time of SVM based methods for each round may increase sharply while for FEC the time consumption will rise slowly in an $O(\log N)$ manner with the help of data structures like R-tree.

The biggest challenge of FEC is the risk of over-fitting. However, we managed to solve it by using a properly designed evaluation function described in Sect. 3.3 together with a large validation set. Our experiments showed that the patches discovered by FEC were both discriminative and representative. In summary, the contributions of this paper are:

1. A novel algorithm for efficiently and effectively detecting discriminative image parts is developed, which demonstrated promising performance in the task of part-based scene classification. Besides, our approach can be seamlessly integrated into bag of visual words models to improve the results of many computer vision problems.
2. A rich training dataset for outdoor scene detection and classification (Outdoor Sight 20) is built. To our best knowledge, this is the first dataset designed for discovering meaningful mid-level patches of outdoor scenes with good in-class consistency. Our dataset consists of images covering 20 famous tourist attractions around the world.



Fig. 1. Visual elements extracted from classes (a) greenhouse (b) inside subway (c) church inside (d) video store (e) closet (f) library of MIT Indoor 67 dataset and (g) (h) Big Ben (i) (j) Mount Rushmore of our Outdoor Sight 20 dataset.

In the experiments, we evaluated our novel FEC method on the public benchmark: MIT Indoor 67 dataset, and the newly created Outdoor Sight 20 dataset, achieving extremely efficient performance (about 20x faster) while maintaining close to state-of-the-art accuracy.

Some of our results are shown in Fig. 1. (a)–(f) are discriminative visual elements extracted from MIT Indoor 67, while (g)–(j) come from our Outdoor Sight 20. As shown in the figure, our method not only captures discriminative and representative visual elements from training data with only class labels provided, but also discovers and distinguishes different visual elements of the same concept, like (g) and (h), which is naturally capable of recognizing different scenes.

2 Related Work

The practice of using parts to represent images has been adopted for quite a long time [14]. Since parts are considered more semantically meaningful compared to some low level features, the introduction of image descriptor generated by algorithms like ScSPM [15], LLC [8] and IFV [9] presented the promising future of parts. The idea of training classifiers discriminatively improved the performance of object detection [11]. However, the discovery of parts are still heavily relied on the training data. Some used the bounding box information on which several assumptions between the parts and the ground truth were based [16], while others relied on partial correspondence [17] to generate meaningful patches.

It was not until recent years that the issue of discovering discriminative mid-level patches automatically with little or no supervision was raised. Patch discovery using geometric information showed that such method has the ability to learn and extract semantically meaningful visual elements for image classification [10, 18, 19]. Unsupervised learning of patches which are frequent and discriminative in an iterative manner boosted the performance of object detection [13]. [1] summarized a simple and general framework to mine discriminative patches using exemplar SVM [20] and showed that this framework was efficient in scene classification in combination with the use of bag-of-parts and bag-of-visual-words models.

Recent works on discriminative mid-level patches can be categorized into two groups. One is to apply this method to other computer vision problems like video representation [6], 2D-3D alignment [21], movement prediction [22, 23] or learning image attributes [19, 24]. The other is to collect Internet images to enrich the visual database of discriminative mid-level patches [3, 25]. In these works the most widely used types of classifiers are mainly variants of SVM. They can achieve satisfactory accuracy but the huge time consumption really becomes a factor that must be considered if we want to apply this technique to large scale computer vision problems [26, 27].

3 Discovering Discriminative Patches: Designed for Speed

Since our purpose is to speed up the training procedure of the model, we designed it to run very fast from the very beginning. We followed the idea that discriminative

patches would be learned and discovered in a framework which had three stages: seeding, expansion and selection [1]. Generally speaking, to discover discriminative patches and the corresponding classifiers that to be able to recognize them, we first need to get a bunch of seed patches from the given images. Since the number of patches is enormous, a selection procedure is carried out. They will then be used to train classifiers using our FEC method. Subsequently the classifiers will be ranked using an evaluation function to test whether they are discriminative and representative enough. Those who have top rankings will be kept and used to represent images in the way described in Sect. 4.

3.1 Patch Selection and Feature Extraction

Commonly used ways in patch selection can be divided into two categories. One is to include all possible patches in an image or randomly select some [13], the other is to use some techniques like saliency detection [3] or superpixels [1] to reasonably remove the patches that are unlikely to contain meaningful information to reduce the problem scale and speed up the training procedure. Patch selection is an essential and indispensable part for a method which aims to run very fast as the training time can be reduced significantly with little impact on the results.

In our method, we introduced a very light-weight way by detecting the number of edges in a patch. The rationale is that we believe the most important feature that human uses to identify different objects and scenes is shape. Edge detection is able to discover the shapes of objects in patches while the number of edges inside a patch somehow suggests the importance of the patch. Intuitively, a patch containing few edges may be a part of the background which lacks discriminativeness, while a patch containing a lot of edges may involve too much details which lacks representativeness. As a result, to ensure our patch selection procedure are able to choose patches that are meaningful, we shall select those with neither too many edges nor too few edges. Figure 2 shows how this works. (a) presents the initial training image from MIT Indoor 67 and its edges detected using Canny method [28]. (b) and (c) are some patches with too few or too many edges. (d) shows the patches with modest number of edges, which contain only one or two objects and their spatial relationship. Even though edge detection is rather simple, it is very efficient and effective to find the patches that we need.

In our experiment, we selected patches with sizes of $80 * 80$, $120 * 120$, $180 * 180$, $270 * 270$. To avoid duplicates, very similar patches with close feature vectors (i.e. the city-block distance is smaller than a threshold $\delta = 0.01$) from the same image were removed. Then the percentage of the area covered by edges in each patch was calculated and a number of patches with medium number of edges among all the patches were kept for each image. We used the HOG feature [29] to represent each patch.

3.2 Classifier Training

The training procedure of the classifiers is the most time consuming section in discriminative part based techniques. Traditional approaches use SVM variants like exemplar SVM [20] and miSVM [30]. For example, Juneja introduced the exemplar SVM and the outcome was satisfactory in terms of classification results [1]. In each round, one patch from a certain class is treated as the positive input, while all patches from other classes [10] are used as negative inputs. After the SVM is trained, it is used to find the top best patches whose scores are highest among the current class. These patches are added into the positive input and the SVM is trained iteratively for several times. It is undeniable that these methods are able to mine discriminative part classifiers eventually. However, the total number of trained SVMs during the training procedure can reach millions and will take lots of time.

We managed to solve this problem by using an efficient type of classifier instead. We call it fast exemplar clustering (FEC). It follows the idea that each patch will be given a chance to see whether it is able to become a cluster [1].

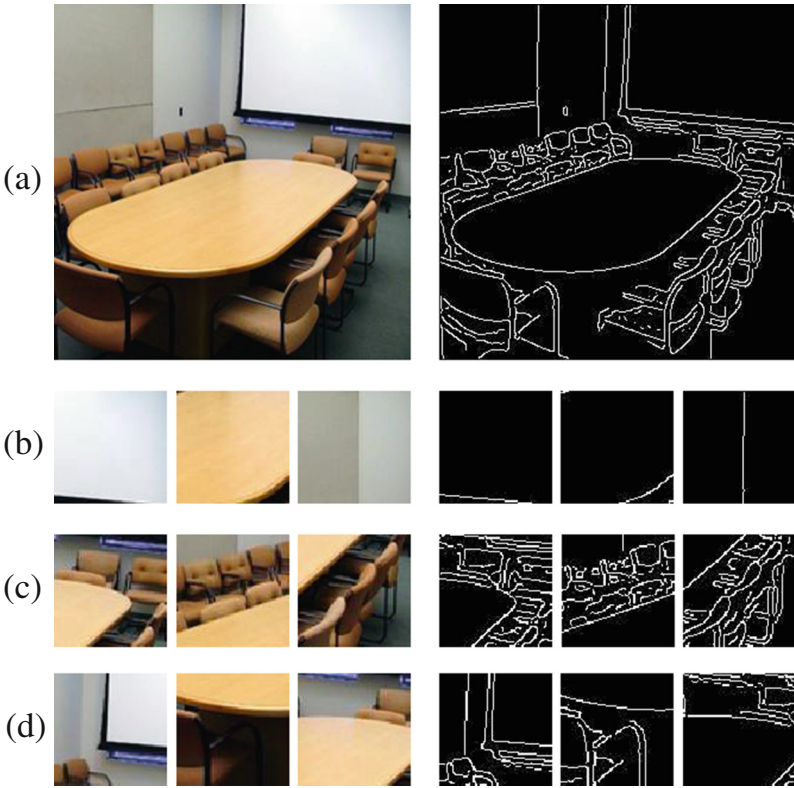


Fig. 2. Images and their ‘canny’ edges, (a) original image (b) patches with few edges (c) patches with too many edges (d) patches with modest number of edges.

Each cluster will then be tested to see if it is discriminative and representative among all the clusters.

The training procedure is shown in Fig. 3(b) and Algorithm 1. Each exemplar cluster will be trained only twice. For a specific patch, it is treated to be a cluster center at first. Then the 10 closest patches whose class labels are the same as the initial patch will be added to the cluster. The cluster center is recalculated using the mean value of these points, followed by adding the next 10 closest and non-duplicate patches with the same class label into the cluster. Each cluster C_i is represented by a clustering center P_i and a radius r_i which is equal to the largest distance between the cluster center and the patches inside the cluster. A classifier can then be built from the resultant cluster. The center and the

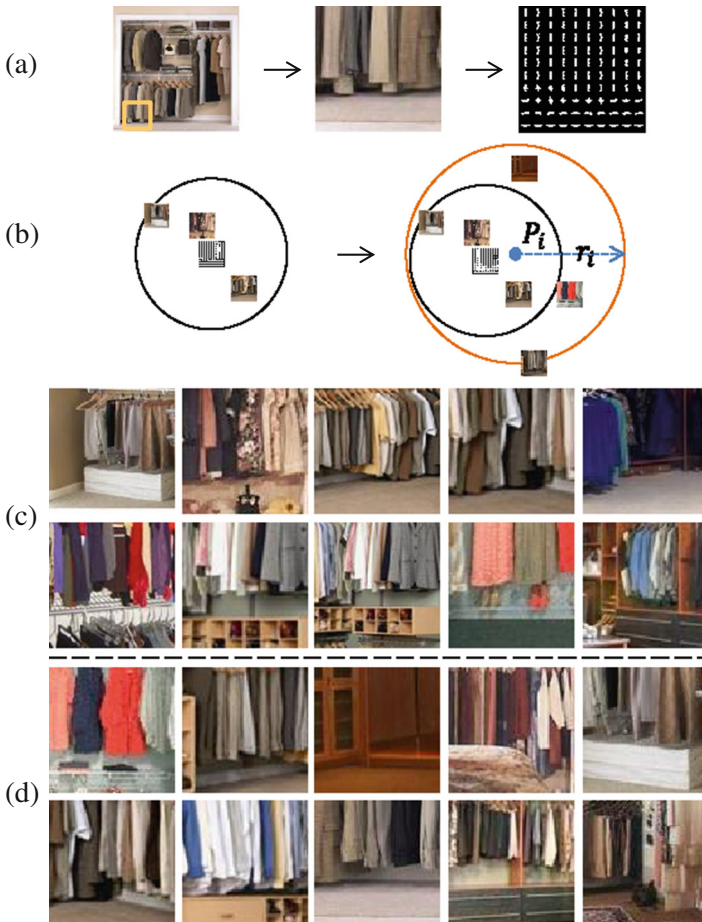


Fig. 3. Illustration of training procedure: (a) initial patch and its HOG representation (b) illustration of cluster expansion using FEC (c) example of patches added in first round of training (d) example of patches added in second round of training.

radius form an Euclidean ball which naturally divide the feature space into two parts, the inner part of the cluster and the outer part. The purpose of training is to transform the initial patch which is specific and particular to a visual concept which is generalized and meaningful.

Since we use the distance measure of feature vectors to form a cluster, the biggest challenge is the risk of over-fitting. The reason why we train only twice in clustering is that we want the clusters to be both **generalized** and **diverse** at the same time to help get rid of over-fitting. **Generalization** means that the cluster can represent not only the initial patch itself but also the patches that are visually similar to the initial patch. Generalization ensures that the patch chosen is representative and common. We want the clusters to be **diverse** since we still do not know which cluster can really represent a discriminative visual concept. If we are able to keep the diversity of the clusters, we will have more chances of obtaining the best classifier when ranking and filtering them in Sect. 3.3.

We did several experiments to decide the optimal number of training rounds, in which two results are really revealing. In one experiment we clustered until the center converges, while in the other we simply did not cluster at all, i.e., we used the initial patch as the center directly with fixed radius for all clusters. It turned out both of them worked poorly. We looked into the results and found that the first way resulted in a lot of identical classifiers which lack diversity, while the latter way resulted in serious over-fitting since one classifier is built merely on one data point. Good generalization and broad coverage are the key to find high quality classifiers.

Algorithm 1. Build exemplar cluster from a patch.

```

function BUILDCLUSTER(patch)
  cluster ← [ ]
  for  $i = 1 \rightarrow 2$  do
    euclidean(patch, patchesInSameClass)
    add 10 closestpatches → cluster
    patch ← mean(cluster)
  end for
  center ← mean(cluster)
  radius ← max(euclidean(center, cluster))
  return < center, radius >
end function

```

3.3 Classifier Selection

Though we have obtained a bunch of classifiers $C = \{C_i\}$ centered at $P = \{P_i\}$ with radius of $r = \{r_i\}$ in the training procedure, the number of classifiers is still enormous and most of them are neither representative nor discriminative. To test whether a classifier C_i is good enough, we try to find all the patches inside the Euclidean ball centered at P_i with radius r_i , and compare the class labels of these patches with the class label of the classifier. Denote n_i to be the

number of patches inside the ball and p_i to be the number of patches inside the ball with the same class label as the classifier's. Then the accuracy of each classifier is p_i/n_i .

However, if we use accuracy as the only evaluation criteria, it is very likely that the classifiers will only recognize features from very few images. It may lead to the absence of representativeness. To overcome this, we count the number of true positive patches that each image contributed and calculate the variance σ_i^2 of these numbers. A smaller σ_i^2 indicates that the true positive patches come from more training images, which suggests that the classifier is more representative than other classifiers with higher σ_i^2 values. The scoring function is then formulated as

$$F(C_i) = \frac{p_i}{n_i} \log\left(\frac{M}{\sigma_i^2 + N} + 1\right). \quad (1)$$

M, N are scaling constants to normalize the contribution of the two parts. The argument M, N are calculated by $\arg \min_{M, N} \sum_{\forall j, C_j \in C} \left(\frac{(p_i)_j}{(n_i)_j} - \log\left(\frac{M}{(\sigma_i^2)_j + N} + 1\right)\right)^2$. Actually according to our experiment results, the actual value of M, N doesn't have much impact on the results as long as it roughly balances the two parts.

Figure 4 shows the best classifier selected using different evaluation criteria. (a) shows the result of evaluating with accuracy only. The five nearest patches come from 3 different images. Even though they are visually consistent, they did not reveal the nature that really makes 'computer room' different from other classes. (b) shows the top classifier evaluated using our evaluation function. The five nearest patches come from 5 different images. The resultant classifier is more representative.

In addition to evaluating the classifiers on the training set, we introduced a large validation set to be used in the same fashion described above. A number of classifiers with top rankings will be chosen as discriminative classifiers. Figure 1 shows the results.



Fig. 4. Evaluation comparison of classifier trained on class 'computer room': (a) evaluate using only accuracy (b) evaluate using function (1) in Sect. 3.3.

4 Image Representation and Scene Classification

Since it is very hard to judge whether a patch classifier is good or not, we need to test our classifiers using a traditional computer vision task. In our experiment we introduced scene classification to compare our results with others to show that patch classifiers discovered in our method are both meaningful and useful.

For the task of scene classification, we need to first represent each image as a vector. We followed the idea of ‘bag-of-parts’ (BoP) [1] and used the discriminative classifiers learned in Sect. 3 to generate the mid-level descriptor for each image in a spatial pyramid manner [31] using 1×1 and 2×2 grids. In practice, patches are extracted using a sliding window and each patch together with its flipped mirror is evaluated using the part classifiers. As a result, each image is represented by a $5mn$ dimensional vector, in which m represents the number of classifiers kept for each class in Sect. 3.3 and n is the total number of classes.

Scene classification accuracy can be further improved if BoP representation is used in combination with Bag of Words (BoW) models like Locality-constrained Linear Coding (LLC) BoW [8] or Improved Fisher Vectors (IFV) [9]. However, to make sure our comparison is on an even base, we presented our results using only the BoP representation. We tested the union representations though in Sect. 5 as a reference.

One-vs-rest classifiers are trained to classify the scenes. Linear SVM is used for BoP representation and linear encoding. For the IFV encoding, Hellinger kernel is used.

5 Experiments and Results

The framework of FEC is simple and runs extremely fast. It is not surprising that people will question the effectiveness and correctness of these classifiers and the corresponding image descriptor generated in Sect. 4. In order to test the classifiers we obtained, we focused on the task of scene classification using two datasets. One is the MIT Indoor 67 dataset [32], the other is the Outdoor Sight 20 that we created.

MIT Indoor 67 consists of 5 main scene categories, including store, home, public places, leisure and working place. Each category contains several specific classes, making a total of 67 classes. This dataset is quite challenging thus widely used in scene classification problems.

Outdoor Sight 20 is a dataset we created which consists of outdoor views of 20 famous tourist attractions around the world such as Big Ben, The Eiffel Tower and The Great Wall of China. To test the ability of distinguishing different scenes, a 21st class which contains images of non-tourist attractions is introduced. Part of the sample images are shown in Fig. 5. We built this dataset since we wanted to test our models on both indoor and outdoor scenes. As a complementary of the MIT Indoor 67 dataset, it is specifically designed to include only outdoor images, most of which are photos taken from different angles with various lighting conditions while some are sketches or drawings. Among all the



Fig. 5. Sample images of Outdoor Sight 20 dataset of classes (a) Big Ben (b) Buckingham Palace (c) Mount Rushmore (d) Notre Dame (e) Parthenon (f) St. Paul’s Cathedral (g) St. Peter’s Basilica (h) Sydney Opera House (i) The Eiffel Tower (j) The Great Wall of China. The rest are: The Brandenburg Gate, The Colosseum, The Golden Gate Bridge, The Kremlin, The Leaning Tower of Pisa, The Pyramids of Giza, The Statue of Liberty, The Taj Mahal, The White House, Tower Bridge with an additional class of none attraction images.

images, the majority have a good within-class consistency since they are portrayals of the same object while some are even difficult for human to classify due to a lot of shared characteristics, like (g) and (f) of Fig. 5.

In our experiment on MIT Indoor 67 dataset, we draw 100 random images from each class. They are partitioned into training set containing 80 images and test set from the remaining 20 images. The training set is further split equally into two parts to be used as training part and validation part, each with 40 images. 50 classifiers for each class are kept to recognize the visual words (Fig. 6).

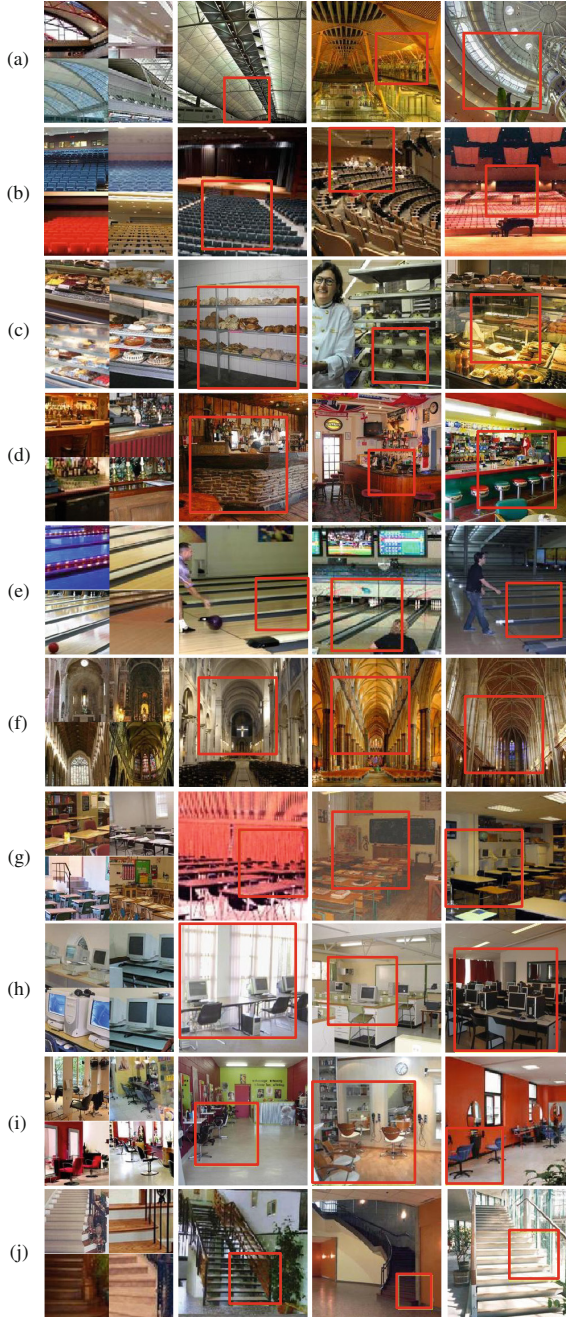


Fig. 6. Classifiers trained on classes: (a) airport inside (b) auditorium (c) bakery (d) bar (e) bowling (f) church inside (g) classroom (h) computer room (i) hair salon (j) staircase of MIT Indoor 67 dataset. The left four patches of each part show how this classifier is trained and the three images on the right show their detections on the testing image.

To test the discriminatively trained mid-level patches, we compared our results (FEC + BoP) with ROI [32], MM-scene [33], DPM [34], CENTRIST [35], Object Bank [36], RBoW [37], Patches [13], Hybrid-Parts [38], LPR [39], exemplar SVM + BoP [1] and IVC [3]. The results are shown in Table 1. Even though our method did not achieve the highest accuracy, it should be clarified that we did not mean to produce best scene classification result. We presented these numbers to show that the patches we obtained in the way described in Sect. 3 are indeed meaningful and could be used as discriminative classifiers in various computer vision problems.

Table 1. Test results on MIT Indoor 67 dataset.

Method	Accuracy (%)
ROI	26.05
MM-scene	28.00
DPM	30.04
CENTRIST	36.90
Object Bank	37.60
RBoW	37.93
Patches	38.10
Hybrid-Parts	39.80
LPR	44.84
IVC (miSVM)	47.60
Exemplar SVM + BoP	46.10
FEC + BoP (Ours)	40.30

We compared the training time required to obtain discriminative mid-level patches with exemplar SVM [20] and ours. On an ordinary Quad-core i5-3570 computer with 16 GB RAM installed using Matlab 2013b, the exemplar SVM took around 3 weeks to train while ours took only 1 day (20x faster). This is an impressive result as the accuracy did not show an enormous drop compared to the exemplar SVM + BoP method.

As is mentioned in Sect. 4, the accuracy can be further improved if BoP representation is used in combination with BoW features. In our experiment,

Table 2. Test results on Outdoor Sight 20 dataset. Comparison between accuracy and training time for part classifier is presented.

Method	Acc. (%)	Time (\approx)
Exemplar SVM + BoP	85.75	5 days
FEC + BoP (Ours)	79.25	7 h

the FEC + BoP + LLC and FEC + BoP + IFV achieved the accuracy of 49.55 % and 53.81 % respectively using parameters suggested in [40].

For the Outdoor Sight 20 dataset, we followed the exact same procedure as MIT Indoor 67 dataset on the same computers with the same number of images used in training, testing and validation for each class. We compared our results with exemplar SVM + BoP [1] in Table 2 to show that our FEC could train discriminative mid-level patches as well as the exemplar SVM with much less time.

6 Conclusion

In this paper a novel approach to learn discriminative mid-level patches from training data with only class labels provided is presented. The motivation is that current discriminative patch learning methods are too time-consuming and can hardly be applied to complicated computer vision problems with large dataset. To begin with, we trained part classifiers using the FEC algorithm. Under proper validation settings and appropriately designed evaluation function, we obtained classifiers whose accuracy could compete with state-of-the-art SVM based classifiers. We tested our classifiers on scene classification using MIT Indoor 67 and our Outdoor Sight 20. Both results revealed they were as good as classifiers generated by the contemporary methods. Our classifiers could be further applied to other computer vision problems like scene classification, video classification, object detection, 2D-3D matching.

Acknowledgement. This work is supported by the Hong Kong RGC General Research Fund GRF HKU/710412E.

References

1. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: distinctive parts for scene classification. In: 2013 IEEE Conference on CVPR, pp. 923–930. IEEE (2013)
2. Sun, J., Ponce, J.: Learning discriminative part detectors for image classification and cosegmentation. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3400–3407. IEEE (2013)
3. Li, Q., Wu, J., Tu, Z.: Harvesting mid-level visual concepts from large-scale internet images. In: 2013 IEEE Conference on CVPR, pp. 851–858. IEEE (2013)
4. Rios-Cabrera, R., Tuytelaars, T.: Discriminatively trained templates for 3D object detection: a real time scalable approach. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2048–2055. IEEE (2013)
5. Wang, L., Qiao, Y., Tang, X.: Motionlets: mid-level 3D parts for human motion recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2674–2681. IEEE (2013)
6. Jain, A., Gupta, A., Rodriguez, M., Davis, L.S.: Representing videos using mid-level discriminative patches. In: 2013 IEEE Conference on CVPR, pp. 2571–2578. IEEE (2013)

7. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2483–2490. IEEE (2013)
8. Shabou, A., LeBorgne, H.: Locality-constrained and spatially regularized coding for scene categorization. In: 2012 IEEE Conference on CVPR, pp. 3618–3625. IEEE (2012)
9. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: 2010 IEEE Conference on CVPR, pp. 3384–3391. IEEE (2010)
10. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? *ACM Trans. Graph. (TOG)* **31**, 101 (2012)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **32**, 1627–1645 (2010)
12. Mittelman, R., Lee, H., Kuipers, B., Savarese, S.: Weakly supervised learning of mid-level features with beta-bernoulli process restricted boltzmann machines. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 476–483 (2013)
13. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part II. LNCS*, vol. 7573, pp. 73–86. Springer, Heidelberg (2012)
14. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **26**, 1475–1490 (2004)
15. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: 2009 IEEE Conference on CVPR, pp. 1794–1801. IEEE (2009)
16. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multi-scale, deformable part model. In: 2008 IEEE Conference on CVPR, pp. 1–8. IEEE (2008)
17. Maji, S., Shakhnarovich, G.: Part discovery from partial correspondence. In: 2013 IEEE Conference on CVPR, pp. 931–938. IEEE (2013)
18. Shen, L., Wang, S., Sun, G., Jiang, S., Huang, Q.: Multi-level discriminative dictionary learning towards hierarchical visual categorization. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 383–390. IEEE (2013)
19. Lee, Y.J., Efros, A.A., Hebert, M.: Style-aware mid-level representation for discovering visual connections in space and time. In: 2013 IEEE International Conference on ICCV, pp. 1857–1864. IEEE (2013)
20. Malisiewicz, T., Gupta, A., Efros, A. A.: Ensemble of exemplar-svms for object detection and beyond. In: 2011 IEEE International Conference on ECCV, pp. 89–96. IEEE (2011)
21. Aubry, M., Maturana, D., Efros, A. A., Russell, B.C., Sivic, J.: Seeing 3D chairs: exemplar part-based 2D–3D alignment using a large dataset of cad models. In: 2014 IEEE Conference on CVPR. IEEE (2014)
22. Walker, J., Gupta, A., Hebert, M.: Patch to the future: unsupervised visual prediction. In: 2014 IEEE Conference on CVPR. IEEE (2014)
23. Lim, J. J., Zitnick, C. L., Dollár, P.: Sketch tokens: a learned mid-level representation for contour and object detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3158–3165. IEEE (2013)
24. Sandeep, R.N., Verma, Y., Jawahar, C.: Relative parts: distinctive parts for learning relative attributes. In: 2014 IEEE Conference on CVPR. IEEE (2014)

25. Chen, X., Shrivastava, A., Gupta, A.: Neil: extracting visual knowledge from web data. In: 2013 IEEE International Conference on ICCV, pp. 1409–1416. IEEE (2013)
26. Jia, X., Zhu, X., Lin, A., Chan, K.P.: Face alignment using structured random regressors combined with statistical shape model fitting. In: 28th International Conference on Image and Vision Computing New Zealand, IVCNZ 2013, Wellington, New Zealand, 27–29 November 2013, pp. 424–429 (2013)
27. Jia, X., Yang, H., Lin, A., Chan, K.P., Patras, I.: Structured semi-supervised forest for facial landmarks localization with face mask reasoning (2014)
28. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **8**, 679–698 (1986)
29. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Conference on CVPR, vol. 1, pp. 886–893. IEEE (2005)
30. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, pp. 561–568 (2002)
31. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Conference on CVPR, vol. 2, pp. 2169–2178. IEEE (2006)
32. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Conference on CVPR. IEEE (2009)
33. Zhu, J., Li, L.J., Fei-Fei, L., Xing, E.P.: Large margin learning of upstream scene understanding models. In: *Advances in Neural Information Processing Systems*, pp. 2586–2594 (2010)
34. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: 2011 IEEE International Conference on ECCV, pp. 1307–1314. IEEE (2011)
35. Wu, J., Rehg, J.M.: Centrist: a visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **33**, 1489–1501 (2011)
36. Li, L.J., Su, H., Fei-Fei, L., Xing, E.P.: Object bank: a high-level image representation for scene classification and semantic feature sparsification. In: *Advances in Neural Information Processing Systems*, pp. 1378–1386 (2010)
37. Parizi, S.N., Oberlin, J.G., Felzenszwalb, P.F.: Reconfigurable models for scene recognition. In: 2012 IEEE Conference on CVPR, pp. 2775–2782. IEEE (2012)
38. Zheng, Y., Jiang, Y.-G., Xue, X.: Learning hybrid part filters for scene recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part V. LNCS*, vol. 7576, pp. 172–185. Springer, Heidelberg (2012)
39. Sadeghi, F., Tappen, M.F.: Latent pyramidal regions for recognizing scenes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) *ECCV 2012, Part V. LNCS*, vol. 7576, pp. 228–241. Springer, Heidelberg (2012)
40. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods, pp. 1–12 (2011)