RGB-D SCENE CLASSIFICATION VIA HETEROGENEOUS MODEL FUSION

Xinda Liu⋆ † *Xueming Wang*⋆ *Shuqiang Jiang*†

† Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
⋆ School of Mathematics and Computer Science, Ningxia University, Ningxia, China

ABSTRACT

We study the problem of scene classification for RGB-D images in this paper. Firstly we analyze the difference between the RGB and depth images. And then based on the difference, an efficient method is implemented to make use of the RGB and depth images and make a well fusion for the RGB and depth features. Focusing on the difference of modality between the RGB and depth images, we propose a method to learn features from color and depth separately using the heterogeneous model. Especially we use the deep ConvNet model with shallow finetuning for RGB images and the relatively shallow ConvNet model with deep finetuning, which can adequately extract different characteristics of the two modalities. After obtaining the discriminative features for each modality, a multiple fully-connected layers connected with a soft-max classifier is trained to harness the complementary relationship between the two modalities. Experimental evaluations on two publicly RGB-D datasets validate the effectiveness of the proposed method.

***Index Terms*—** RGB-D, scene classification, deep learning, finetuning, feature fusion

1. INTRODUCTION

Scene classification in the real-world environment has applications in a variety of domains including mobile robots, smart TV, and image retrieval, etc. However, scene classification is a highly challenging task due to the variability and complication of scene, the illumination conditions variations, and the wide range spatial characteristics. These problems are inherently ill-posed since a multiplicity of objects exist on any given scene and the scene is a more abstract concept. Traditionally, scene classification has been mainly studied in the 2D domain. Previous works[1, 2, 3, 4] mainly exploit the low-level visual descriptors. And recently,[5, 6, 7, 8] use a deep ConvNet model to extract the visual feature. However, despite their success, the visual-based methods are limited by the inadequate information in 2D images. Recently with the developments of RGB-D sensors, especially Microsoft Kinect series, the deep information is a useful cue to solve these scene classification problems. Motivated by the depth images possessing many useful properties that RGB images don't possess, we propose the method use the depth information to be important supplementary effectively for scene classification.

The depth is characterized directly as gray image in term of pixel intensities. Naturally, most researchers treat the depth gray image same as one of the RGB three channels, and takes RGB-D channels as input simply[9, 10]. Despite their significant efforts, the first approach(early fusion) doesn't consider the characteristic of the depth image so that they can't distinguish the difference between the depth and RGB information. In this paper we use another method to exploit an encoding method to describe the color and depth images respectively. The depth image uses the form of visual image to describe the spatial information. Thus, although the depth image is more simple than RGB image, it provides another modality discriminative information. The depth information possesses a lot of useful properties. For instance, the depth information has illumination invariance, it has a better performance in dim illumination while the color information can't even work sometimes. And the depth information also provides the stereo relationship of the object in the scene which can't be obtained by one single color image.

Owing to the depth information is represented in the form of image, we use the convolutional neural networks (CNNs) to extract the features, which is proved to be efficient for images. In view of this situation that the depth image is much different from RGB image, we choose the heterogeneous ConvNet models for them from [11, 12, 7]. The depth images use the relatively shallow AlexNet model [11] and the RGB images use the deeper VGG-16 model [12]. In consideration of independence and specificity of depth information, the chosen models are finetuned discriminatively by the RGB and depth image to be suitable for the RGB and depth images. In particular, the models are pre-trained by RGB images, the depth modality is different from the vision. Thus, we implement a deep finetuning by the depth image containing finetuning convolutional layers, on the contrary, we make a shallow finetuning by the RGB images. Finally the Neural Networks Full Connection (NNFC) feature fusion method is used to adequately exploit the complementary relationship between the two modalities and reserve the depth characteristic as much as possible.

As mentioned above, our method is illustrated in Figure 1 which has three components. Firstly the best models fit-

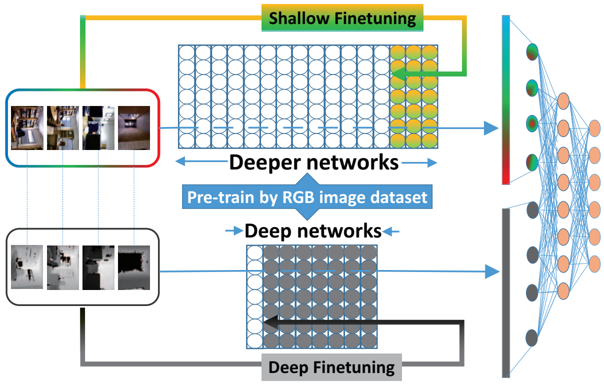


Fig. 1. Illustrating for our method.

ting the dataset are chosen. The VGG-16 model is for RGB images and the AlexNet model is for depth images. And then the chosen models are finetuned using RGB and depth images respectively. Especially we use the deeper networks with shallow finetuning for RGB images and the relatively shallow networks with deep finetuning for depth images. Finally we make a Neural Networks Full Connection (NNFC) feature fusing that was particularly suitable to deal with both RGB and depth features.

In this paper, we designed and implemented an integrated method to scene classification using a fusion of RGB features and depth features extracted from the two heterogeneous CNN models whose different layers are finetuned respectively. Being clearly different from other works which rely mainly on RGB information while depth information subsidiary, we believe depth information is equally important with visual information, and our experiment demonstrate that it makes sense to deal synchronously with RGB and depth images. Experiment on the Robot Vision dataset and SUN RGB-D dataset validate the effectiveness of our proposed method.

2. FEATURE LEARNING AND FUSION

There are many differences between RGB and depth images. For instance, the RGB image has three channels but the depth image has only one. In the meanwhile, the depth images have more presentations of spatial structure comparing to the RGB images which have more visual information. Considering of these differences, we use the different method to extract features for each kind of image respectively. After obtaining the most representative RGB and depth features, we make a NNFC feature fusion for them.

2.1. RGB feature learning

Recently, the success to extract features using convolutional neural networks (CNNs) [11] which is trained on approximately 1.2 million RGB images in the ImageNet ILSVRC 2012 raises the prospect that it can be used into scene image features extraction. Zhou et al.[7] have demonstrated the impressive results of Place-CNN, which is trained on approximately 2.5 million RGB images from Places205 dataset, for scene classification. Thus we choose the VGG-16 model[12] for RGB images. The model is pre-trained by RGB image dataset, in the other words, the modality of the pre-trained dataset is same as the finetuning dataset. It's easy to envision that their low layer parameters are much similar, thus there is no point in finetuning the convolutional layers. Based on the experience, just finetuning the fully-connected layers by RGB images appear to be sufficient for good accuracy. Thus we choose the deeper CNN models with shallow finetuing. We use this model to extract features for the RGB images directly.

2.2. Depth feature learning

Inspiring by many researches for object classification on RGB-D dataset, such as [13, 14, 15], we use ConvNet to extract the depth features. Different from them, we finetune the CNN model with color and depth images separately. Specifically, based on the characteristic of the depth modalities, we choose the relatively shallow CNN model with very deep finetuning.

We extract the features from the heterogeneous models, such as AlexNet[11], Places-CNN[7] and so on, then we train a SVM classifier using the features obtained above. Experimental verification shows that relatively shallow CNN model is adequate for the depth images. In order to make the chosen model to be more suitable for the target RGB-D dataset, we finetune the model by the depth images. The chosen model is pre-trained by the RGB images dataset, the modality is different from the depth images. Considering of this situation, we have to finetune the convolution layers to fit to the depth images. We design the experiment for the chosen models that are finetuned in different degrees from one layers until the whole eight layers by the depth images. Experimental verification shows that deeper finetuning achieves a better performance. After model selection and processing, we use the model to extract feature for the depth images directly.

2.3. Feature fusion

Feature fusion for scene classification is already generally accepted and established. Benmokhtar et al.[16] use a multi-layers perceptron scheme to fuse the low-level feature. And Shao et al. [17] use a hierarchical scheme for fusing feature. Depth and visual features are closely related but huge different from each other. Owing to the complementary properties of the feature data, in general, the combination of the depth and visual feature outperforms either of them.

Discriminative features may be extracted independently from color and depth images using the model mentioned above, both of them are 4096 dimensions. In order to fuse the two features effectively, we feed them into the last Neural Networks fully-connected layer, it's very suitable for large classes classification problem. From the final softmax layer, supervised information is back-propagated to the independent networks for both modalities. This method can not only discover the most discriminative features for each modality, but is also able to harness the complementary relationship between the two modalities. Thus, we explore this method to fuse the two kinds of features obtained from RGB and depth images. We evaluate different fusion methods, containing simple concatenation and different layers NNFC feature fusion by experiment.

3. EXPERIMENT

To evaluate the effectiveness of our proposed RGB-D feature learning and fusion method for scene classification, we perform experiments on the Robot Vision Dataset[18] and the SUN RGB-D Dataset [19, 20, 21]. The details of the experiments and the results are described in the following sections.

3.1. Dataset and Experimental setup

Robot Vision dataset has 5263 images for training and 1869 images for testing. This dataset contains 10 scene categories. Since the original dataset is unbalanced, we choose the subset by random sampling. After random sample we get the training data containing 3514 images and the testing data containing 1869 images. We also validate our approach on the Sun RGB-D dataset, a recently published indoor dataset which has 10335 images. This dataset contains 19 scene categories for scene classification.

Our work have three major aspects : original model evaluation, finetuning the convolutional neural networks for feature learning, fusing depth feature and visual feature by Neural Networks fully-connected feature fusion.

3.2. Experiment on Robot Vision dataset

3.2.1. Original model evaluation

As shown in Table 1, for the Robot vision 2013 RGB-D database, depth features have more powerful discrimination ability than the visual features getting from the same deep networks by horizontal comparison, because there are many images taken in the dim light environment, depth features are insensitive to the illumination changes. We achieve 76.03% precision classifying by the simple combination of the RGB features obtaining from the VGG-16 networks and the depth features obtaining from the AlexNet networks. The combination of the RGB and Depth features outperforms either of the two. By vertical comparison, we notice that the fewer layers networks obtain the better depth features. This result means that the depth features have a degeneration when the networks goes deeper.

Table 1. Choosing model.

Accuracy(%)	Visual	Depth
Places-CNN[7]	61.697	72.445
Places-VGG16[7]	62.868	71.268
AlexNet[11]	66.720	73.676
VGG-16[12]	67.362	73.622

3.2.2. finetuning

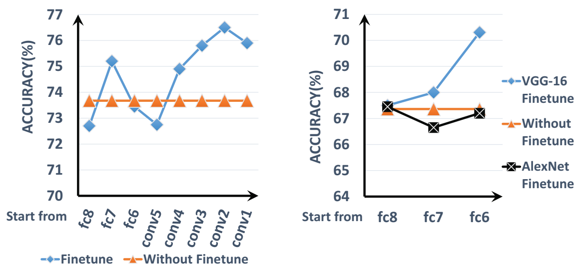
Figure 2 show the finetuning results based on the different finetuning level. For the depth images, finetuning convolutional layers could achieve better performance, and in general, finetuning deeper make the performance better. It can be indicated that finetuning convolutional layers helps the model to be more suitable for another data modality. For depth images, the finetuned model achieved 76.5%, which outperformed the model without finetuning by 2.9%. For RGB images, finetuning VGG-16 networks outperformed the AlexNet by 2.9% and enhanced the performance by 2.7% compared with the VGG-16 model without finetuning.

Experimental results shows our proposed method, the relatively deep networks with shallow finetuning for RGB images and the relatively shallow networks with deep finetuning for depth images, is effective to learn feature representations of color and depth separately.

3.2.3. feature fusion result

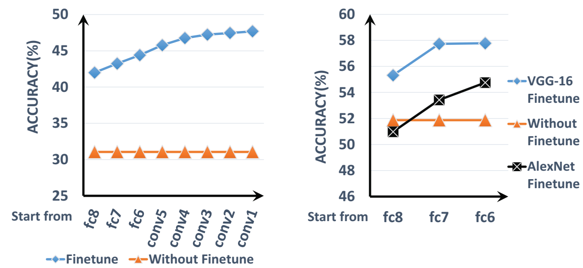
To evaluate the effectiveness of our model, we report the results of the frame-based kernel descriptor[18] model and SVM classification result based on places-CNN depth and RGB features simple concatenation[22], and they use two kinds of SVM kernels. We train our data fusion models using stochastic gradient descent with momentum of 0.9 and a 0.005 weight decay. We use a learning rate of 0.01 and 0.5 dropout ratio. We exchange the number of layer over and over again to get a better performance. As shown in table 2, feature obtaining by our proposed finetuning approach which is connected with a SVM classifier achieves a good performance. And it's obvious that our NNFC feature fusion is better than feature after finetuning simple concatenation (FTSC). Our proposed method achieves 76.7% classification accuracy for RGB-D, which outperformed Kernel descriptor[18] and Places-CNN [22] by 3.7% and 14.8%.

And the one layer NNFC feature fusion is better than the two layers NNFC feature fusion. This is because Robot Vision dataset is too small to train a large multiple layers fully-connected parameters.



(a) depth finetuning result (b) rgb finetuning result

Fig. 2. Robot Vision finetuning result.



(a) depth finetuning result (b) rgb finetuning result

Fig. 3. SUN RGB-D finetuning result.

Table 2. Robot Vision dataset experiment result.

		Accuracy(%)
Kernel descriptor[18]		76.19
Places-CNN[22]	linear kernel SVM	65.115
	RBF kernel SVM	60.942
Our FTSC	linear kernel SVM	72.45
	RBF kernel SVM	76.7255
Our NNFC	one Layer	79.8823
	two layers	78.9727

Table 3. SUN RGB-D dataset experiment result.

		Accuracy(%)
Kernel descriptor[18]		42.22
Places-CNN[22]	linear kernel SVM	56.60
	RBF kernel SVM	54.85
Our FTSC	linear kernel SVM	61.94
	RBF kernel SVM	62.84
Our NNFC	one Layer	62.93
	two layers	60.14

3.3. Experiment on SUN RGB-D dataset

We validate our approach on the SUN RGB-D dataset. For kernel descriptor method, we follow the experimental strategy in [18]. And we also give the performance of the Places-CNN features connected with a SVM classifier described in [22]. Owing to the imbalance of SUN RGB-D dataset, we pick up 2850 images for training and 2115 images for testing, different from the data partition method of [22]. In addition to the above mentioned difference, the other experimental details are following [22]. And we get 56.60% using linear kernel SVM and 54.85% using RBF kernel SVM.

To verify the our conclusions mentioned above, the AlexNet is finetuned by the depth image from fully-connected layers until convolution layers. Figure 3 shows the finetuning result on SUN RGB-D dataset, the experimental result is in line with our conclusions above. Compared with CNN without finetuning, the proposed finetuned model improved the performance by 16.4% and 5.9% for depth and RGB respectively. Consequently, the deeper ConvNet model, VGG-16, with shallow finetuning is exploited by the RGB images. And the relatively shallow ConvNet model, AlexNet, with deep finetuning is exploited by the depth images. After extracting the features of the images of the SUN RGB-D dataset, we also make a NNFC feature fusion for the RGB and depth features. The result is shown in the Table 3, our method achieved 62.93% accuracy and get 17.9% and 13.5% improvement compared with kernel descriptor and Places-CNN method.

4. CONCLUSION

In this paper, we study the scene classification problem on RGB-D dataset. We analyze the difference between the RGB and depth images. And we use the deep convolutional neural networks to extract features, especially we use the deeper networks with the shallow finetuning and the relatively shallow networks with the deep finetuning models. We make a fusion for the RGB and depth features, and archive a good performance. Experimental results on two publicly available RGB-D scene classification datasets show that the proposed approach greatly outperforms existing kernel descriptor feature methods and the baseline Places-CNN method, which indicates the effectiveness of the proposed method for RGB-D scene classification.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research 973 Program of China under Grant No. 2012CB316400, the National Natural Science Foundation of China under Grant Nos. 61532018, 61322212, the National High Technology Research and Development 863 Program of China under Grant No. 2014AA015202. This work is also funded by Lenovo Outstanding Young Scientists Program (LOYS).

6. REFERENCES

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

[2] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009.

[3] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, T. Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.

[4] J. Sanchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *Neural Comput.*, 2011.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[6] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*. April 2014, CBLS.

[7] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in Neural Information Processing Systems*, 2014, pp. 487–495.

[8] Y. Gong, L. Wang, R. Guo, and S. Lazebnik., "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.

[9] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[10] Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Yong Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," *arXiv preprint arXiv:1509.06470*, 2015.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 665–673.

[14] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014*, pp. 345–360. Springer, 2014.

[15] Xiong Lv, Xinda Liu, Xiangyang Li, Xue Li, Shuqiang Jiang, and Zhiqiang He, "Modality-specific and hierarchical feature learning for rgb-d hand-held object recognition," *Multimedia Tools and Applications*, pp. 1–18, 2016.

[16] Rachid Benmokhtar, Benoit Huet, and Sid-Ahmed Berrani, "Low-level feature fusion models for soccer scene classification," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1329–1332.

[17] Wen Shao, Wen Yang, Gui-Song Xia, and Gang Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Computer Vision Systems*, pp. 324–333. Springer, 2013.

[18] Ruihan Xu, Shuqiang Jiang, Xinhang Song, Shuang Wang, Yi Xie, Fang Wang, and Xiong Lv, "Miar ict participation at robot vision 2013," *ImageCLEF Robot Vision*, 2013.

[19] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *Computer Vision–ECCV 2012*, pp. 746–760. Springer, 2012.

[20] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell, "A category-level 3d object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*, pp. 141–165. Springer, 2013.

[21] Jianxiong Xiao, Andrew Owens, and Antonio Torralba, "Sun3d: A database of big spaces reconstructed using sfm and object labels," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1625–1632.

[22] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.