CLASSIFICATION OF CORALS IN REFLECTANCE AND FLUORESCENCE IMAGES USING CONVOLUTIONAL NEURAL NETWORK REPRESENTATIONS

Lian Xu[1⋆], *Mohammed Bennamoun*[1⋆], *Senjian An*[1⋆], *Ferdous Sohel*[3†], *Farid Boussaid*[2⋆]

[1]School of Computer Science and Software Engineering
[2]School of Electrical, Electronics and Computer Engineering
[3]School of Engineering and Information Technology
[⋆]The University of Western Australia, Perth, WA 6009, Australia
[†] Murdoch University, Perth, WA 6150, Australia

ABSTRACT

Coral species, with complex morphology and ambiguous boundaries, pose a great challenge for automated classification. CNN activations, which are extracted from fully connected layers of deep networks (FC features), have been successfully used as powerful universal representations in many visual tasks. In this paper, we investigate the transferability and combined performance of FC features and CONV features (extracted from convolutional layers) in the coral classification of two image modalities (reflectance and fluorescence), using a typical deep network (e.g. VGGNet). We exploit vector of locally aggregated descriptors (VLAD) encoding and principal component analysis (PCA) to compress dense CONV features into a compact representation. Experimental results demonstrate that encoded CONV3 features achieve superior performances on reflectance and fluorescence coral images, compared to FC features. The combination of these two features further improves the overall accuracy and achieves state-of-the-art performance on the challenging EFC dataset.

Index Terms— Transfer learning, deep convolutional features, VLAD encoding, coral image classification, fluorescence

1. INTRODUCTION

Due to natural and anthropogenic factors such as ocean warming and pollution, the population of important underwater species (e.g. coral reef and kelp) has largely decreased in many areas. Such changes, which are associated with profound ecological, social and economic consequences, have drawn increased attention on marine environment protection from all sectors of society. With the rapid development of robotics and imaging techniques (e.g. AUVs and ROVs), a considerable number of underwater images are now available to study changes in marine environments. However, manual labeling this raw data for subsequent analysis is quite a laborious task for expert analysts. Thus, researchers seek automated solutions for the images annotation of underwater images.

Underwater image classification is a challenging task due to following reasons: (**i**) physical properties of the water medium (e.g. absorption and scattering), cause underwater images to suffer from color degradation which is not present in the ground images. (**ii**) water turbidity and floating particles result in underwater images exhibiting low contrast and limited visibility. (**iii**) in case of coral species, they have large variations in mophologies, size, color, shape, and texture across classes, whose boundaries are often ambiguous. (**iv**) class imbalance (i.e. non-coral species often predominate in the whole set) results in misclassified minority coral classes [1]. Moreover, due to the aforementioned challenges, common annotation techniques, such as image labels and bounding boxes, are not suitable for coral images. Instead, marine ecologists use point annotations. As shown in Figure. 1, small patches centered around each point annotation are extracted from an image for classification.

Traditional methods for coral image classification were developed based on handcrafted features, mainly relying on texture descriptors, e.g. local binary pattern (LBP) [2], gray-level co-occurrence matrix (GLCM) [3] and Gabor wavelet response [4]. After CNNs made remarkable success in ILSVRC, CNN-based methods have been used for many applications. Recent works have shown that CNN features have superior performance compared to those well-designed handcrafted

This research is partially supported by China Scholarship Council funds (CSC, 201607565016) and Australian Research Council Grants (DP150104251 and DE120102960)

features in the coral classification task [5, 6]. Generally, CNN features are extracted from fully connected layers of the network, which capture global semantic information as high-level features. However, this choice has several drawbacks: (**i**) CNN features from higher layers are more specific to their original task, while features from intermediate or lower layers are more general to other applications [7]. In case of coral images, most subcategories of corals are unseen in those large datasets (e.g. ImageNet). Moreover, corals in fluorescence images rarely appear in common datasets. Thus, coral images share less semantic information with the ImageNet task, compared to other general ground images (e.g. humans, animals, and buildings). (**ii**) Fully connected layers capture global spatial layout information. This characteristic may be effective in classifying objects with clearly defined shape or contour, but may not be useful in representing texture or edge, which are the primary features for coral classification. In contrast, deep convolutional features contain rich local information, which provides more discriminative ability in describing local regions.

The main contributions of this paper can be summerized as follows: (**i**) we propose to apply a method based on CNN and VLAD to coral image classification. (**ii**) we investigate the combined strength of two types of deep features (i.e. FC features and CONV features) in coral image classification. (**iii**) we evaluate the transferability of deep CNN features to coral classification for two image modalities (reflectance and fluorescence).

2. APPROACH

In this section, we describe our method of feature extraction and feature coding for coral image classification.

2.1. Feature extraction

In this work, we use the VGG-F pre-trained model developed by [9] for fast processing. It is provided by the MatConvNet toolbox for MATLAB [10]. This architecture is similar to AlexNet [11]. It comprises 5 convolutional layers and 3 fully connected layers. The size of input images is 224×224. The major difference between the structures of VGG-F and AlexNet is that, the former used dense connectivity between convolutional layers [9]. The output sizes of convolutional layers and fully connected layers after the rectified linear unit (ReLU) layers (thus all the activations are non-negative) are shown in Table 1.

To extract CNN features, given an input image or a patch, we resize it to 224×224, subtract the mean of images in the whole set, and feed the patch through the network. Then we take the 4096-dimensional output of the first fully connected layer as FC features, and we take activations from $13 \times 13 \times 256$ feature maps in the last three convolutional layers as different CONV features (as feature maps of first two convolutional layers are too large, they are not evaluated in this work).

Table 1: The sizes of CONV and FC layers of VGG-F

Layer	Output size ($N \times N \times D$)
CONV3	$13 \times 13 \times 256$
CONV4	$13 \times 13 \times 256$
CONV5	$13 \times 13 \times 256$
FC6	$1 \times 1 \times 4096$

2.2. VLAD encoding

Before applying VLAD encoding on CONV features, we reshape them into a group of local features. Given a matrice of CONV features from the l th (l=3,4,5) convoultional layer, $\mathbf{M} \in \mathbb{R}^{N \times N \times D}$, where the size of each feature map is $N \times N$, and the number of feature maps is D, we take activations at each location (i, j) across all feature maps as a D-dimensional local feature vector $f^l_{(i,j)}$, where $1 \leq i, j \leq$ N, thus obtaining S D-dimensional local features, where $S = N \times N$. Finally, we get a feature group $\mathbf{F}^l = \{f^l_1, f^l_2, ..., f^l_S\}$.

VLAD encoding was originally proposed for image retrieval [12]. It can be viewed as a simplified version of Fisher Vector (FV) [13]. FV uses gaussian mixture model (GMM) for clustering, while VLAD uses K-means instead. To apply VLAD, a codebook $C = \{c^l_1, ..., c^l_k\}$ is required. For this purpose, we randomly select a collection of images which were not used in the test set. K-means with k cluster centers was applied to CONV features of this subset. For each local feature in the feature group from an input image, f^l_s (where $1 \leq s \leq$ S) , it is associated with its nearest r cluster centers (We use soft assignment and set r=5 as it is in [14]). The VLAD descriptor is constructed by accumulating the differences between each local feature and their corresponding nearest centers:

$$\mathbf{v}^l_k = \sum_{s: c^l_k \in rNN(f^l_s)} w_{sk}(f^l_s - c^l_k) \quad (1)$$

$$\mathbf{v}^l = [\mathbf{v}^l_1, \mathbf{v}^l_2, ..., \mathbf{v}^l_k] \quad (2)$$

Where w_{sk} is the Gaussian kernel similarity between each local CONV feature f^l_s and each of its k nearest center c^l_k. $\mathbf{v}^l_k$ is the sum of residuals between each center c^l_k and all local CONV features which are assigned to this center. The final VLAD descriptors are normalized by L2 normalization and signed square rooting. The dimensionality of VLAD descriptor is $D \times k$. Given D=256, k=100, the VLAD descriptor of each image is 25,600 dimensional, which makes subsequent analysis

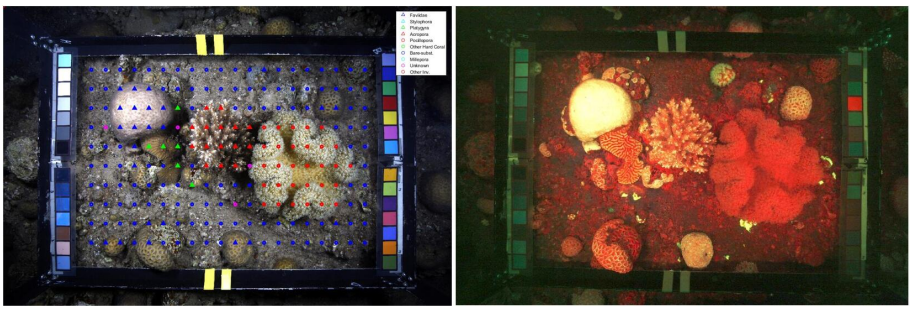

Fig. 1: An image pair of reflectance image (left) and fluorescence image (right) from the EFC dataset [8].

computationally expensive. Therefore, PCA is used to reduce the original dimension to 512-dimemsional.

2.3. Images classification

For classification, we extract FC features and CONV features in different layers. Different CONV features are encoded via VLAD, separately. Combined features are generated by concatenating multiple features. Image classification is done by using the linear support vector machine (SVM).

3. EXPERIMENTS AND ANALYSIS

We perform experiments on the EFC dataset, which consists of an expert-annotated set of registered fluorescence and reflectance image pairs captured during a nighttime reef survey in Eilat, Red Sea Israel. The whole dataset contains 212 image pairs with 200 point annotations as one of the ten dominant taxonomic categories per image. To evaluate the methods, the whole dataset of image pairs was divided into two subset. The training set consists of 142 randomly selected image pairs with 28,400 point annotations, and the test set contains 70 image pairs with 14000 point annotations.

In the following experiments, we use linear SVM as implemented in the LIBLINEAR software package [15]. Hyper-parameters are chosen by 5-fold cross-validation on the training set. The accuracy is computed as the sum of correctly classified points divided by the total number of points within each test image. The final accuracy is composed by the mean accuracy of all test images with the standard error.

In [8], end-to-end training was applied on reflectance and fluorescence images in the EFC dataset, achieving 87.8% and 85.5% accuracies, respectively. A subsequent linear SVM classifier was used to aggregate the outputs of these two networks, obtaining a joint accuracy of 90.5%.

Table 2 reports our results on reflectance images in the EFC dataset. We achieve a slight increase of 0.5% using FC6 features, while a significant increase of 1.4% using CONV3 features over reported results in [8]. The combination of FC6 and CONV3 features further improve the accuracies using these two features separately by 1.6% and 0.7%, achieving the best accuracy at 89.9%. Furthermore, as shown in Figure 2 (left), classification recalls for the five dominant corals, i.e. *Faviidae, Stylophora, Platygyra, Acropora* and *Pocillopora*, increase by 2%, 2%, 42%, 63% and 25% over reported results in [8], respectively.

Our results on fluorescence images in Table 3 follow the similar trend as those on reflectance images. FC6 features have a comparable performance 85.4% to the 85.5% reported in [8]. CONV3 features outperform their result by a significant margin of 1%. The best accuracy of our results 86.7% is achieved by concatenating FC6

Table 2: Classification accuracies on reflectance images in the EFC dataset with a comparison to [8].

Features	Feature dimension	Accuracy
CONV3	512	$89.2 \pm 0.8\%$
FC6	4096	$88.3 \pm 0.8\%$
CONV3+FC6	4608	$\mathbf{89.9} \pm 0.8\%$
[8]	-	$87.8 \pm 1.1\%$

Table 3: Classification accuracies on fluorescence images in the EFC dataset with a comparison to [8].

Features	Feature dimension	Accuracy
CONV3	512	$86.5 \pm 1.0\%$
FC6	4096	$85.4 \pm 1.0\%$
CONV3+FC6	4608	$\mathbf{86.7} \pm 1.0\%$
[8]	-	$85.5 \pm 1.2\%$

Table 4: Joint accuracies using information of both reflectance and fluorescence images in the EFC dataset with a comparison to [8].

Features	Feature dimension	Accuracy
CONV3	1024	$90.7 \pm 0.8\%$
FC6	8192	$90.4 \pm 0.8\%$
CONV3+FC6	9216	$\mathbf{91.4} \pm 0.8\%$
[8]	-	$90.5 \pm 0.8\%$

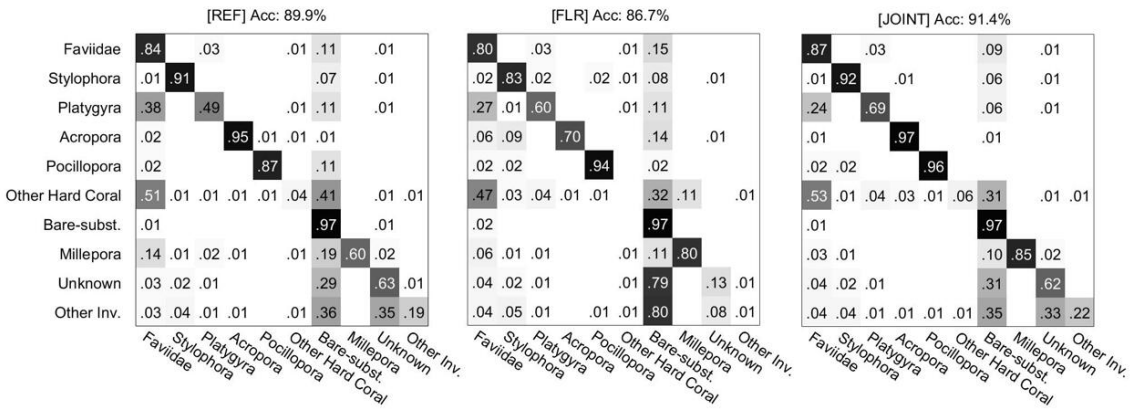

Fig. 2: Confusion matrices for our proposed methods using combined FC6 and CONV3 features from reflectance images (left), fluorescence images (middle) and both images (right).

and CONV3 features. In addition, Figure 2 (middle) show that, except *Faviidae* with the same recall 80%, for *Stylophora, Platygyra, Acropora* and *Pocillopora*, their classification recalls increase 13%, 14%, 46% and 26%, respectively, from those in [8].

Table 4 reports overall classification results by incorporating both features from reflectance and fluorescence images. CONV3 and FC6 achieve joint accuracies of 90.7% and 90.4%, which are comparable to 90.5% reported in [8]. Furthermore, by aggregating FC6 and CONV3 features, we achieve the best accuracy of 91.4%.

Experiments are also conducted to compare the performance of convolutional features from different layers. As shown in Figure 3, we can see that convolutional features from shallower layers perform better in all three cases of reflectance, fluorescence and joint images. Dash lines show that combined with FC6 features, performances of different convolutional features can be enhanced in varying degrees. Additional experiments show that combining multiple convolutional features with FC features can slightly improve the joint accuracy but at the cost of larger computational loads. From the above experiments, we find that the average accuracy of fluorescence images is much lower than that of reflectance ones. One possible reason is that, compared with reflectance coral images, fluorescence ones have more different distribution with images from ImageNet, such that CNN features generalize better on reflectance coral images than fluorescence ones. Another reason is the effect of the poor registration quality on fluorescence coral images.

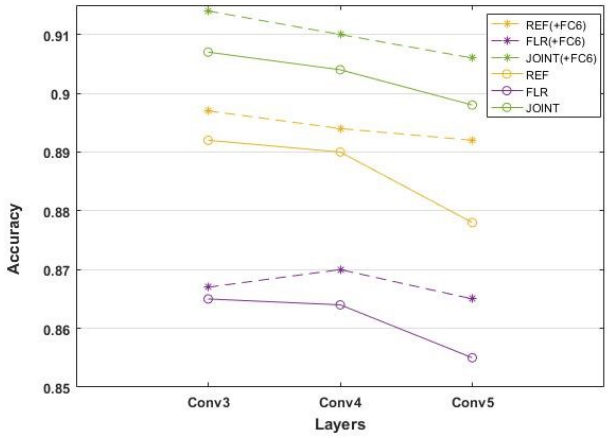

Fig. 3: Performance of convolutional features from different layers: solid and dash lines correspond to individual convolutional features and those combined with FC6 features.

VLAD is used to encode dense convolutional features into a compact feature vector. We find that in the case of lacking sufficent training data, CNN off-the-shelf features outperform training a small network from scratch. Specifically, low dimensional compact convolutional features achieve comparable and better results against fully connected features on reflectance and fluorescence coral images. We suggest, in visual tasks, deep convolutional features should be the first choice for a new dataset, which is very different from the original dataset for training deep networks. Moreover, combining features from convolutional and fully connected layers can further improve the overall accuracy.

4. CONCLUSION

In this work, we experiment with features from the fully connected layer and features from different convolutional layers. We investigate and compare their transferability to coral species in two image modalities.

5. ACKNOWLEDGEMENTS

The authors acknowledge NVIDIA for providing a Titan X GPU for the experiments involved in this research.

6. REFERENCES

[1] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.

[2] M. S. A. Marcos, L. David, E. Peñaflor, V. Ticzon, and M. Soriano, "Automated benthic counting of living and non-living components in ngedarrak reef, palau via subsurface underwater video," *Environmental monitoring and assessment*, vol. 145, no. 1, pp. 177–184, 2008.

[3] A. Gleason, R. Reid, and K. Voss, "Automated classification of underwater multispectral imagery for coral reef monitoring," in *OCEANS 2007*, pp. 1–8, IEEE, 2007.

[4] O. Pizarro, P. Rigby, M. Johnson-Roberson, S. B. Williams, and J. Colquhoun, "Towards image-based marine habitat classification," in *OCEANS 2008*, pp. 1–7, IEEE, 2008.

[5] A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. Fisher, "Automatic annotation of coral reefs using deep learning," in *OCEANS 2016 MTS/IEEE Monterey*, pp. 1–5, IEEE, 2016.

[6] A. Mahmood, M. Bennamoun, S. An, F. Sohel, F. Boussaid, R. Hovey, G. Kendrick, and R. Fisher, "Coral classification with hybrid feature representations," in *Image Processing (ICIP), 2016 IEEE International Conference on*, pp. 519–523, IEEE, 2016.

[7] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, pp. 3320–3328, 2014.

[8] O. Beijbom, T. Treibitz, D. I. Kline, G. Eyal, A. Khen, B. Neal, Y. Loya, B. G. Mitchell, and D. Kriegman, "Improving automated annotation of benthic survey images using wide-band fluorescence," *Scientific reports*, vol. 6, p. 23166, 2016.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[10] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 689–692, ACM, 2015.

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

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3304–3311, IEEE, 2010.

[13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[14] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European conference on computer vision*, pp. 392–407, Springer, 2014.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.