Abstract

The paper presents a novel concept for collaborative descriptors between deeply learned and hand-crafted features. To achieve this concept, we apply convolutional maps for pre-processing, namely the convolutional maps are used as input of hand-crafted features. We recorded an increase in the performance rate of +17.06% (multi-class object recognition) and +24.71% (car detection) from grayscale input to convolutional maps. Although the framework is straightforward, the concept should be inherited for an improved representation.

1. Introduction

Neural networks computationally mirror the architecture of the human brain. Recent architectures have been made deeper in order to provide higher-level representations of object. Especially in computer vision, deep convolutional neural networks (DCNN) are becoming increasingly widely used for image recognition and object detection. Representative deep models such as VGGNet [5] have been proposed in the field of computer vision.

The use of neural net and hand-crafted feature is repeated in the computer vision field (see Figure 1). The suggestive knowledge motivates us to study an advanced hand-crafted feature after the DCNN in the recent 3rd AI. We anticipate that the hand-crafted feature should collaborate with deeply learned parameters since an outstanding performance is achieved with the automatic feature learning.

In the paper, we present a novel concept for collaborative descriptors which jointly work both DCNN and hand-crafted feature. We apply convolutional maps for pre-processing as a first implementation. The framework is simple, but it is effective for recognition tasks. We directly access the convolutional maps in order to improve the performance rate of hand-crafted features. We try to evaluate several hand-crafted features such as scale invariant feature transform (SIFT) and bag-of-words (BoW) [2] (object recognition) and higher-order local autocorrelation (HLAC) [4] (car detection). Then, the recognition rates for all hand-crafted features are compared using convolutional maps and grayscale images that have the same settings for the image description.

2. Convolutional maps for preprocessing

Here, we introduce a framework for combining convolutional maps and hand-crafted features. Figure 2 shows the process for using convolutional maps as a preprocessor. We employ hand-crafted features, namely SIFT+BoW [2] and HLAC [4].

Given a convolutional map $M_k$ with VGG [5] for kernel $k$, our goal is to extract a feature vector $V$ from the map and define a classification label $L$ as a function of SVM, $L = f_{svm}(V)$. We use VGGNet to create convolutional maps.

The 2nd max-pooling layer is used as a convolutional map that contains the $56 \times 56 \times 128$ feature map. We can handle 128 maps with dimensions of $56 \times 56$ [pixel]. The patch and kernel sizes are suitable for balanced feature ex-
We can obtain a convolutional map \( M^{\text{conv}} \) and max-pooling map \( M^{\text{mp}} \) for kernel \( k \) from input map \( M \) as below:

\[
M^{\text{conv}}_k = \sum_x \sum_y f_{kxy} M_{xy} + b_{kxy}
\]

\[
M^{\text{mp}}_k = \max_{ij} M^{\text{conv}}_{ijk}
\]

where \( f \) is a filter and \( b \) is a bias in CNN. \( x, y, i, j \) are the indices of the two-dimensional map. Hereafter, the 2\textsuperscript{nd} max-pooling layer \( M^{\text{mp}} \) is the input image.

### 3. Results

Comparisons of the hand-crafted features from convolutional maps and grayscale images are shown in Figure 3 and Table 1 respectively. We used a support vector machine (SVM) as a two- or multi-class classifier.

SIFT+BoW was employed for the Caltech 101 dataset \(^3\), since it is a kind of object categorization. According to Figure 3 the convolutional maps allow us to clearly distinguish between object classes. In the case of SIFT+BoW with convolutional maps, the codeword dictionary was shared across all 128 kernels. The dimension of the vector was 1,000, even when using convolutional maps. SIFT+BoW with convolutional maps (rate: 58.28\%) was +17.06\% better than the SIFT+BoW for the grayscale image (rate: 41.22\%) on the Caltech 101 dataset. The results show that the use of preprocessing gives a perspective for the learning-based preprocessing with CNN. Although the CNN architecture is based on the ImageNet pre-trained model, a more sophisticated model such as fine tuning with objective data may improve preprocessing.

HLAC was applied to the UIUC cars dataset \(^4\). The HLAC setting in convolutional maps was used to extract 25 dimension features per learned kernel, therefore we can extract a 25 [dim] \( \times \) 128 [kernel] = 3,200 [dim] HLAC feature from an image. Table 1 shows the proposed setting (98.82\%) performed better than the conventional grayscale setting (74.11\%). We get a +24.71\% improvement through using the convolutional maps. The preprocessing of HLAC usually requires binarization; however, we can extend this to a combination of convolution and binarization with convolutional maps.

### 4. Conclusion

The paper presented a novel concept of collaborative descriptors which combine deep convolutional neural network (DCNN) into hand-crafted features. We assigned convolutional maps as a preprocessing. We recorded an increase in the performance rate of +17.06\% and +24.71\% for the Caltech101 and UIUC cars datasets, respectively.

In the future, we must adjust the parameters of the convolutional maps, hand-crafted features and their integration. This paper only considered old-fashioned features and datasets; however, the use of convolutional maps will likely result in even greater improvements in the performance rate when used in conjunction with more sophisticated features.

### References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 286(11), 2004.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. ECCVW, 2004.

[3] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. CVPR, 2004.

[4] N. Otsu and T. Kurita. A new scheme for practical flexible and intelligent vision systems. IAPR Workshop on Computer Vision, 1988.

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representation (ICLR), 2015.