Encoded Deep Features for Visual Place Recognition

A.H. Abdul Hafez, Saed Alqaraleh and Ammar Tello
Computer Engineering Department, Hasan Kalyoncu University, Gaziantep, Turkey
Email: abdul.hafez@hku.edu.tr, saed.alqaraleh@hku.edu.tr, ammar.tello@std.hku.edu.tr

***Abstract*—In this work, a new VPR approach that uses the features extracted from a Convolutional Neural Network (CNN) architecture that will be encoded by the Fisher Vector (FV) is introduced. As the main aim of this work is to develop a robust approach that can meet real-life challenges, the deep features are encoded with FV, which as shown in the experiments section, can lead to getting more robust features. Our approach was evaluated using two classifiers, Dynamic Time Warping (DTW) and Support Vector Machine (SVM) in particular. Using both classifiers, the FV-based encoded features have outperformed the non-encoded features.**

***Index Terms*—Dynamic time warping, Deep features, Fisher Vector, CNN, Image sequence matching, Visual place recognition.**

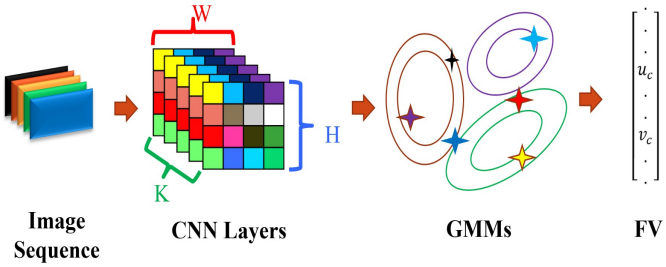

Fig. 1. The main concept of the proposed approach. A set of features are extracted from the nth CNN layer. The set of image features are clustered using GMM which are used to generate FV code of the input image.

I. Introduction

Visual Place Recognition (VPR) has an impressive effect on achieving the localization task for autonomous robots and vehicles using visual input. Due to this it has attracted researchers in the few recent years and been well studied since then, particularly the recent autonomous robots works based on CNN models. It has been shown in the literature that image retrieval and image classification tasks can be achieved using pre-trained CNN [1]. The deep features can be extracted from any layer of the used CNN.

The work presented in [2] has focused on the performance of the last layer of CNN which is actually the layer that gives the final classification decision. SVM and Softmax are the two frequently layers that can be used as the last layer, and the result of this study showed that SVM has slightly outperformed the Softmax. We also propose here to use Dynamic Time Warping (DTW) as the last layer to produce the final classification decision.

All the aforementioned works have used general pre-trained CNN architectures, i.e., these models were not trained specifically for the place recognition task. The work of [3] can be considered as the first work that attempted to solve this problem by collecting a dataset called Places that consists of 10 million images of places in different environmental situations from around the world. Then, some well-known architectures like (AlexNet [4], GoogLeNet [5], and VGG16 [6]) were trained using this dataset, and as expected, it has outperformed the previous pre-trained models on ImageNet [4] dataset.

Another dataset called Specific PlacEs Dataset (SPED) was collected using nearly 30000 outdoor cameras from all around the world. This dataset contains images from days and nights through several months and several years attempting to include all possible changes under different conditions [7]. This work presented two models named as AmosNet and HybridNet, the AmosNet was trained on the SPED, while the HybridNet was trained on the ImageNet, then fine-tuned with SPED. The results of this study showed that HybridNet has outperformed AmosNet.

Despite the fact that deep features are able to outperform the handcrafted features in most cases, it has been found that it is not efficiently handling several visual challenges [8]. To overcome this problem and as shown in [8][9], researchers have tried to add some additional steps before and after the deep feature extraction process to compensate this limitation. Encoding features to enhance the ability of deep features to be more robust to appearance changes was one of these steps. The approach of [10], uses the BOW [11] to encode the image's landmarks extracted from the output of the convolutional layers of a pre-trained VGG16 deep neural network.

A slightly different approach was followed in [9] where the encoding method was VLAD [12] and detecting the landmarks was done through only one layer of a pre-trained AlexNet365 [3]. This approach outperformed the one with BOW. Another improved approach was introduced in [9] by the same authors where more features from different layers of the CNN are gathered to improve the performance. In this contest, we investigate the Fisher vectors as encoding method to enhance the performance of place recognition algorithms using deep features.

The main contribution of this work is to introduce a new VPR algorithm that utilizes deep features encoded using the fisher vector. In addition, DTW is used as a last classification

layer and compared to SVM. In more detail, the feature maps are extracted from a selected convolution layer after applying the test and the reference image sequences as an input to the network. Then, these features are encoded using FV-based codebook. After that, the DTW algorithm (or SVM) makes a decision by classifying each input image to an image from the reference sequence of images.

II. Encoded Deep features using Fisher Vectors for visual place recognition

This section presents the proposed approach which is depicted in Figure 1. The approach starts by extracting features through convolutional layers of the deep networks. Then, these features are encoded using FV-built codebook. Finally, a classifier is to make the decision on the test image.

A. Deep features Extractions

In this work, the whole pixels of each image are fed into the deep network, then, the features are extracted through the output of one of the layers of the CNN architecture. In other words, the image is represented using the output of a specific Layer from a CNN model. In this paper, the layers of the VGG16 [13], the ResNet50 [14] and HybridNet [7] networks are considered. In fact, the convolutional layers of the three mentioned models were used in this work.

The structure of VGGNet [13] has 16 convolutional layers and it has been proved that it is one of the most suitable choices for extracting image features. Using the VGG16 model, the input I is passed through the first two convolutional layers each of which has 64 feature maps produced by using a kernel of size 3×3. Then, its output is reduced into $112 \times 112 \times 64$ with a pooling layer. This output is fed after that into the next block which consists of two layers in each of which there are 128 feature maps, followed by a pooling layer that further reduces the size into $56 \times 56 \times 128$. The third block gives an output with size $28 \times 28 \times 256$ where each of the three layers in this block consists of 256 feature maps. There are three layers in each of the fourth and fifth blocks with 512 feature maps in each layer. The output of the fourth block is $14 \times 14 \times 512$ and the output of the fifth block is $7 \times 7 \times 512$.

ResNet [14] has a novel architecture that is based on skip connections. In addition, Heavy batch normalization was introduced in ResNet, which leads to training the model using 152 layers while still having lower complexity compared to other models. The ResNet50 model consists of five blocks (like VGG16) but with different parameters for each block. The input image is sub-sampled into $112 \times 112 \times 64$ as an output of the first block. The next block contains two stages, the first one uses the max pooling, while the second stage consisted of three duplicated copies of three convolutional layers that have different parameters. The output of the second block is pooled into $56 \times 56 \times 256$ and then fed into the next block which gives an output with $28 \times 28 \times 512$ size. The fourth block gives an output with $14 \times 14 \times 1024$ size, where each convolutional layer is repeated 6 times. Similarly, the last convolutional block gives a $7 \times 7 \times 2048$ output.

The Hybrid Net model was developed for VPR applications [7]. Hybrid Net has 6 convolutional layers followed by two FC layers. This CNN architecture initialized with weights taken from a previously developed model, i.e., CaffeNet [4]. This due to the fact that both the Hybrid Net and CaffeNet models have the same dimensions of the first five layers. Then, the Hybrid Net was fine-tuned on the SPED VPR dataset. In more details, HybridNet model consists of six blocks, each has only one convolutional layer and one pooling layer, the pooling layer of the first block gives an output of $55 \times 55 \times 96$ size, this output is reduced in the next block into $27 \times 27 \times 256$. Each of the third and fourth layers produces a $13 \times 13 \times 384$ output size, Similarly, the output of the fifth and sixth blocks is $13 \times 13 \times 256$ and $6 \times 6 \times 256$ respectively.

Based on the above, when using the output of a convolutional layer, the output dimensions will be $W \times H \times K$ where W is the width, H is the height and K is the depth of the Feature maps in the selected layer. For such a layer, there are $W \times H$ feature vectors, where each one consists of K feature maps. Let N denote $W \times H$, these features are stacked together in a matrix with size $N \times K$ to be fed into the next stage.

B. Place Recognition using FV-based Encoded Deep Features

Fisher Vector (FV) encoding method is a global descriptor of an image. FV is obtained by clustering the training data using the Gaussian Mixture Model (GMM) and then representing the input samples by projecting them onto this GMM.

To build a GMM, we use the training data from the Garden point dataset and Berlin_A100. This step provides us the main components of the GMM including weight (w_c), mean (M_c) and covariance (E_c) for each cluster (c). These components can be described as $\lambda = \{w_c, M_c, E_c\}$, where $c = 1, \ldots, M$, and M is the number of clusters which set to 128 in this work.

The next phase is to represent both, the training and testing datasets using FV. For every feature vector (x_t) extracted from a convolutional layer of a CNN model, the following two components are calculated as follows:

$$u_{dc} = \frac{1}{T\sqrt{2\pi_c}} \sum_{t=1}^{T} P_r(c|x_t, \lambda) \frac{x_{dt} - \mu_{dc}}{\sigma_{dc}}, \quad (1)$$

$$v_{dc} = \frac{1}{T\sqrt{2\pi_c}} \sum_{t=1}^{T} P_r(c|x_t, \lambda) [(\frac{x_{dt} - \mu_{dc}}{\sigma_{dc}})^2 - 1], \quad (2)$$

Where $d = 1, \ldots, K$ is the components of x data vector with dimension K represents the number of feature maps which varies according to the selected CNN layer. The posterior probability $P_r(c|x_t, \lambda)$ of each cluster is given as

$$P_r(c|x_t, \lambda) = \frac{\omega_c g(x_t|\mu_c, \Sigma_c)}{\sum_{j=1}^{M} \omega_j g(x_t|\mu_j, \Sigma_j)}; \quad (3)$$

Here, $g(x_t|\mu, \Sigma)$ is the Gassuian density function. As a result, for each image, the calculated components are concatenated to formulate the final fisher vector illustrated as

$$\Phi(I) = [\ldots, u_c, \ldots, v_c, \ldots]^T, \tag{4}$$

The length of this vector equals $M * K * 2$, where, as mentioned previously, M is 128 and K is the number of feature maps according to the selected layer. An improved version of FV can be generated with a square root normalization followed by L_2 normalization applied on $\Phi(I)$.

As soon as the input image is represented using the FV representation, it is supplied to the last stage classifier to find out its label or its matching place. Two classifiers (or matching algorithms) are used in our work. They are the SVM and DTW. Finally, the system has a decision on whether it is a prior visited place or a new place.

III. Experimental Study

A. Datasets and Evaluation

In this study both well-known datasets "Garden Point" and "berlin_A100" [15] were used. The "Garden Point" is a dataset that captures the changes in the pose and lighting conditions through the Garden point campus. It consists of three sub-datasets: Day left, Day right and Night right. The first two sequences were collected during the day, but with a different viewpoint. In addition, the third one has a very close viewpoint to the second one but it differs in the illumination and the images of this series where taken in the night. Each of these series has 200 images labeled by referring to the corresponding images.

The "berlin_A00": is a dataset collected from a platform called Mapillary where images of the same route were collected by different users with a variation in viewpoint and appearance.

B. Experiments and Results analysis

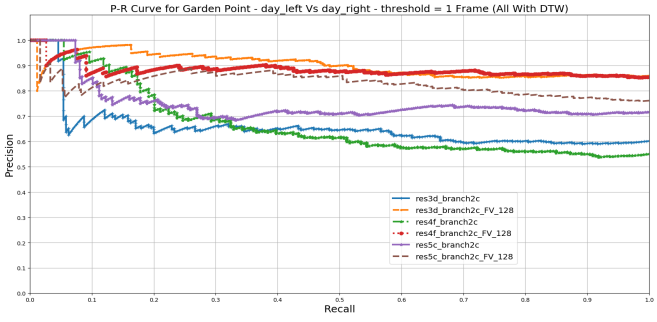

Fig. 2. PRC for convolutional layers in ResNet50 without FV against the same layers encoded with FV and all are integrated with DTW.

Related to the performance evaluation, the precision-recall curve (PRC), Area under curve (AUC), and the average precision (AP) measures were used. 1) The precision (P) is given as $P = TP/(TP + FP)$ while the recall (R) is given as $R = TP/(TP + FN)$. Every match between i and j frames is considered as a positive if the visual distance $D(i, j)$ is greater than a threshold t. Otherwise, the match is considered as negative matches. 2) AUC can be calculated using the trapezoidal rule

$$AUC = \sum_{i=1}^{n-1} \frac{p_i^{min} + p_{i+1}^{max}}{2}(r_{i+1} - r_i), \tag{5}$$

where p is the precision value and r is the recall value. Also, p_i^{min} is the minimum precision corresponding to r_i and p_i^{max} is the maximum precision corresponding to r_i and n is the considered number of recalls. 3) AP which is the weighted mean of precision was used for each threshold in the PRC, and it can be calculated as

$$AP = \sum_n (R_n - R_{n-1})P_n, \tag{6}$$

where R_n and P_n are the recall and precision respectively obtained for the threshold n in the PRC.

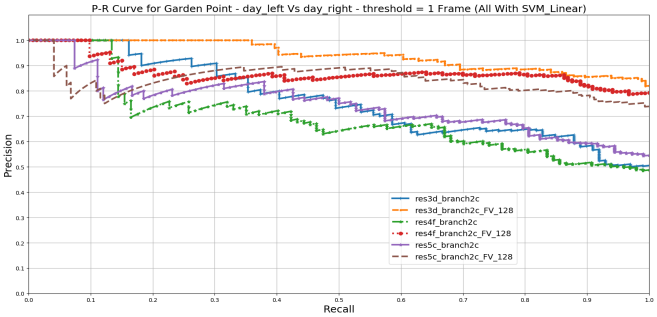

Fig. 3. PRC for convolutional layers in ResNet50 without FV against the same layers encoded with FV and all are integrated with SVM.

1) Experiment 1: Performance of DTW-based Fisher Vector: In this experiment, we have evaluated the DTW-based deep features encoded with the fisher vector against the non-encoded features. In addition, the same scenario was repeated using the SVM classifier based deep features. In addition, the Pre-trained VGG16 and ResNet50 networks were used. Furthermore, the number of GMM was set to 128 for all the scenarios of this experiment.

Overall, as shown in Table I and Figures 2, 3 the following can be observed:

- A Using the Garden Point (Day left vs Day right), the FV leads to improve the performance of a) all the used layers and b) the used architecture.
- B Using the Garden Point (Day left vs Night right), whenever the DTW was used as the classifier, the features extracted from VGG16 and encoded with FV outperforms the same deep features without FV.
- C Using the berlin_A100, when the DTW was used as the classifier, the features with FV outperformed the feature without.

2) Experiment 2: DTW against SVM: In this experiment, the performance of the DTW for place recognition is compared with the SVM based algorithm. The results are shown in Table II, and it can be summarized as follows:

TABLE I
AUC AND AP FOR THE SELECTED CONVOLUTIONAL LAYERS WITH AND WITHOUT FV ENCODING USING THE DTW, AND A THRESHOLD OF 1 FRAME.

Dataset	Model	Layers	AUC		AP	
			Without FV	With FV	Without FV	With FV
Garden Point (Day left vs Day right)	VGG16	block3_Conv3	0.664	0.883	0.663	0.883
		block4_Conv3	0.604	0.899	0.603	0.898
		block5_Conv3	0.699	0.632	0.698	0.630
	ResNet50	res3d_branch2c	0.654	0.902	0.652	0.901
		res3f_branch2c	0.667	0.884	0.666	0.883
		res5c_branch2c	0.750	0.830	0.749	0.829
berlin_A100	ResNet50	res3d_branch2c	0.247	0.397	0.240	0.372
		res3f_branch2c	0.315	0.699	0.299	0.697
		res5c_branch2c	0.333	0.627	0.319	0.624
Garden Point (Day left vs Night right)	ResNet50	res3d_branch2c	0.421	0.619	0.418	0.617
		res3f_branch2c	0.538	0.722	0.535	0.721
		res5c_branch2c	0.443	0.583	0.439	0.581

TABLE II
AUC RESULTS FOR CONVOLUTIONAL LAYERS IN RESNET50 WITH FV BASED DTW AGAINST SVM.

Dataset	Conv layer	DTW	SVM
Garden PointDay leftDay right	res3d_branch2c	0.901	0.941
	res3f_branch2c	0.883	0.872
	res5c_branch2c	0.829	0.843
Garden PointDay leftNight right	res3d_branch2c	0.619	0.331
	res3f_branch2c	0.722	0.53
	res5c_branch2c	0.583	0.431
berlin_A100	res3d_branch2c	0.397	0.332
	res3f_branch2c	0.699	0.389
	res5c_branch2c	0.627	0.317

1) Using the Garden Point (Day left vs Day right), where the challenge is only the viewpoint, the SVM was able to outperform the DTW using two of the three used layers, i.e., "res3d_branch2c" and "res5c_branch2c".
2) Using the Garden Point (Day left vs Night right) and berlin_A100 datasets, which have viewpoint, appearance and illumination challenges, the DTW can significantly outperform the SVM.

IV. CONCLUSIONS AND FUTURE WORKS

This algorithm has enhanced the features extracted from a deep convolutional neural network (CNN) by encoding them using the Improved Fisher Vector (IFV). In our experiments, the performance of the DTW and SVM are used as a classifier at the last stage of our proposed algorithm was investigated.

Using the FV encoding scheme, the experimental results show superior performance for our approach especially with the challenging datasets in terms of viewpoint and appearance. However, for the viewpoint problem, using the Garden Point (Day left vs Day Right), SVM was able to get a little bit better performance. On the other hand, SVM was not robust enough to face the challenges existed in other datasets like Garden Point (Day left vs Night Right) and Berlin_A100, and for such dataset, there is a clear advantage of our approach as shown in the related experiments.

V. ACKNOWLEDGMENT

The Titan Xp used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1] Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. Multi-attention network for one shot learning. In *Proceedings of the IEEE CVPR*, pages 2721–2729, 2017.
[2] Ke Du and Kai-Yu Cai. Comparison research on iot oriented image classification algorithms. In *ITM Web of Conferences*, volume 7, page 02006. EDP Sciences, 2016.
[3] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE PAMI*, 40(6):1452–1464, 2017.
[4] A Krizhevsky, I Sutskever, and GE Hinton. Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems*, pages 1097–1105, 2012.
[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, pages 1–9, 2015.
[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[7] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE ICRA*, pages 3223–3230. IEEE, 2017.
[8] Zhe Xin, Xiaoguang Cui, Jixiang Zhang, Yiping Yang, and Yanqing Wang. Real-time visual place recognition based on analyzing distribution of multi-scale cnn landmarks. *Journal of Intelligent & Robotic Systems*, 94(3-4):777–792, 2019.
[9] Ahmad Khaliq, Shoaib Ehsan, Michael Milford, and Klaus McDonald-Maier. Camal: Context-aware multi-scale attention framework for lightweight visual place recognition. *arXiv preprint arXiv:1909.08153*, 2019.
[10] Zetao Chen, Fabiola Maffra, Inkyu Sa, and Margarita Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *IEEE/RSJ IROS*, pages 9–16. IEEE, 2017.
[11] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.
[12] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *IEEE CVPR*, pages 3304–3311, 2010.
[13] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on PAMI*, 38(10):1943–1955, 2015.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE CVPR*, pages 770–778, 2016.
[15] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4297–4304. IEEE, 2015.