CNN vs. SIFT for Image Retrieval: Alternative or Complementary?

Ke Yan[1,2], Yaowei Wang[*,3], Dawei Liang[1,2], Tiejun Huang[1,2], Yonghong Tian[*,1,2]
[1]National Engineering Laboratory for Video Technology, School of EE&CS,
Peking University, Beijing, China
[2]Cooperative Medianet Innovation Center, China
[3]Department of Electronic Engineering, Beijing Institute of Technology, China
{keyan, dwliang, tjhuang, yhtian}@pku.edu.cn;yaoweiwang@bit.edu.cn

ABSTRACT

In the past decade, SIFT is widely used in most vision tasks such as image retrieval. While in recent several years, deep convolutional neural networks (CNN) features achieve the state-of-the-art performance in several tasks such as image classification and object detection. Thus a natural question arises: for the image retrieval task, can CNN features substitute for SIFT? In this paper, we experimentally demonstrate that the two kinds of features are highly complementary. Following this fact, we propose an image representation model, complementary CNN and SIFT (CCS), to fuse CNN and SIFT in a multi-level and complementary way. In particular, it can be used to simultaneously describe scene-level, object-level and point-level contents in images. Extensive experiments are conducted on four image retrieval benchmarks, and the experimental results show that our CCS achieves state-of-the-art retrieval results.

Keywords

Multi-level image representation; CNN; SIFT; Complementary CNN and SIFT (CCS)

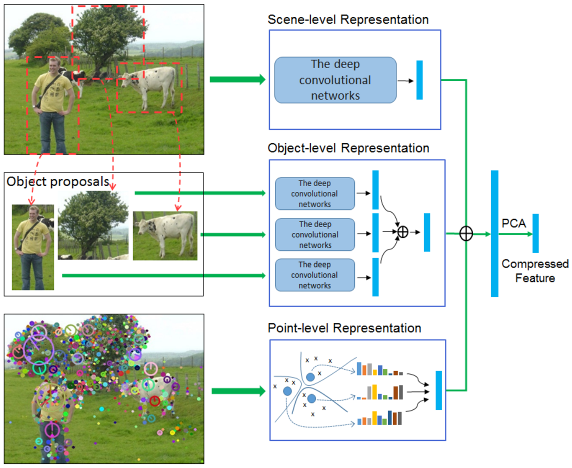
Scene-level Representation
The deep convolutional networks
Object-level Representation
Object proposals
The deep convolutional networks
The deep convolutional networks
The deep convolutional networks
PCA
Compressed Feature
Point-level Representation

Figure 1: The demonstration of our proposed complementary CNN and SIFT (CCS). The CCS aggregates three level contents, i.e., scene-level, object-level and point-level representations.

1. INTRODUCTION

Scale-invariant feature transform (SIFT) [1] has been the most widely-used hand-crafted feature for content-based image retrieval (CBIR) in the past decade. Technologically, SIFT is intrinsically robust to geometric transformations and shows good performance for near-duplicate image retrieval [2] [3]. Meanwhile, there are also many works (e.g., Fisher vector [4], VLAD [5] and their variants [6] [7] [8]), that attempt to construct semantically-richer mid-level image representations so as to improve the retrieval performance. However, in spite of significant efforts, it is still difficult to fully bridge the semantic gap between such feature representations and human's understanding of an image only with SIFT-based features.

Recently, deep convolutional neural networks (CNN) have achieved the state-of-the-art performance in several tasks, such as image classification [9] [10], object detection [11] [12] and saliency detection [13]. Compared with hand-crafted features, CNN features learned from numerous annotated data (e.g., ImageNet [14]) in a deep learning architecture, carry richer high-level semantic information. Several attempts in CBIR [15] [16] [17] showed that CNN features work well for image retrieval as a scene-level representation. Gong *et al.* [18] proposed an approach called Multi-scale Orderless Pooling (MOP) to represent local information by aggregating CNN features at three scales respectively. Konda *et al.* [19] and Xie *et.al.* [20] detected object proposals and extracted CNN features for each region at the object-level. Besides, there are also some researchers who paid attention to deep convolutional layers to derive representations [21] [22] [23] [24] for image retrieval. Although CNN features achieve good performance, we can not say that C-NN will always outperform SIFT yet. Vijay *et al.* [25] had showed that no one was better than the other consistently and the retrieval gains can be obtained by combining the two features.

[*] Corresponding author: Yaowei Wang and Yonghong Tian.

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10. . . $15.00
DOI: http://dx.doi.org/10.1145/2964284.2967252

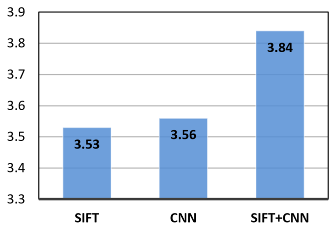

Figure 2: The overall performance comparison of SIFT, CNN, and their naive combination.

Inspired by these pioneering works and our preliminary experiments described in section 2, we demonstrate that C-NN and SIFT features are highly complementary for image retrieval tasks. Following this fact, a complementary CNN and SIFT (CCS) image representation is proposed in this paper. A similar work is Zheng *et al.* [26], which also fused CNN and SIFT from multiple levels. But [26] utilized CNN features derived from regional and global patches as auxiliary cues to BoW model to improve the matching accuracy of local feature, while our method focuses on getting a unified and powerful representation integrating CNN and SIFT coequally. As shown in Fig. 1, CNN features from scene and object levels and SIFT features from point-level are fused to form a compact image representation: First, features from the top layer of CNN are utilized as the scene-level representation. Secondly, we obtain the deep CNN features of each region generated by the object proposal method, and pool them to a fixed-length object-level representation with VLAD. Note that the pooling method really influences the power of object-level representation. It is demonstrated that VLAD pooling outperforms other pooling in our experimental results. Thirdly, a VLAD representation of SIFT features serve as the point-level representation. Finally, we concatenate all the three level representations and then leverage PCA to reduce redundancies so as to obtain the final compact representation.

Experiments are conducted on four image retrieval benchmarks, i.e., INRIA Holidays [27], Oxford5K [28], Paris6K [29] and UKBench [30]. Experimental results show that the proposed CCS outperforms the state-of-the-art methods on three datasets, and achieves comparable performance on the INRIA Holidays dataset (the smallest one).

2. PRELIMINARY EXPERIMENTS

In this section, we qualitatively discuss the complementarity between CNN and SIFT through a preliminary experiment. We conduct the experiment on the UKBench dataset [30]. As a widely-used image retrieval benchmark, this dataset consists of 10,200 images of 2,550 objects, each containing 4 images. By using every image as the query, the performance is reported as the average recall at top four results.

We obtain CNN and SIFT features by using standard settings. Specifically, we extract global activation values provided by the pool5 layer in GoogLeNet [10] as CNN features. VLAD is adopted to aggregate SIFT features. The two features hold the same dimension after adopting PCA to compress the corresponding representations.

The overall result is shown in Fig. 2. From the results, we can see that CNN and SIFT achieve nearly comparable performance. To investigate the complementarity, we conduct an experiment by simply concatenating CNN and SIFT features, and the performance is significantly boosted.

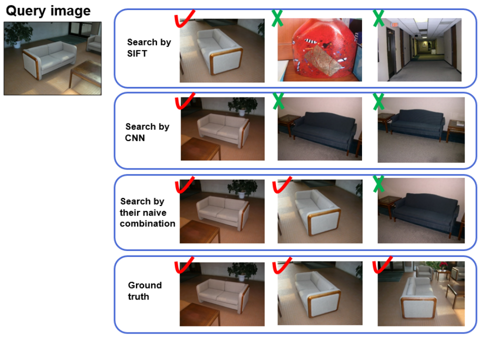

Figure 3: An image retrieval example from UK-Bench. Each image has three most similar images.

An example of results is shown in Fig. 3. CNN and SIFT return one similar image respectively. However, the returned similar images are not the same one. Their naive combination returns two similar images, demonstrating that they are indeed highly complementary. From another perspective, this example also shows that the naive combination of CNN and SIFT may not work perfectly since one similar image is also missed in this result.

3. CCS: A MULTI-LEVEL REPRESENTATION OF COMPLEMENTARY CNN AND SIFT

In this section, we describe the proposed CCS image representation. As shown in Fig. 1, it mainly consists of three level representations and a fusion method. In what follows, we introduce each component in detail respectively.

3.1 Scene-level representation

Scene-level features are used to mainly capture global and high level information for an image. It is well known that the deep learning features, especially those from the high layers, are suitable to represent semantic information [9]. In this study, we obtain CNN features from the pool5 layer in GoogLeNet [10], which is a recent deep convolutional network that ranked as the top one at the image classification task in ILSVRC 2014. Here, we denote the scene-level semantic features as $f_s \in R^{1024}$.

3.2 Object-level representation

We obtain the object-level representation using the following three steps: First, we take advantage of an object proposal approach to produce numbers of candidate object regions. Secondly, we extract deep features for each selected region. Finally, these features are pooled to a fixed-length feature vector using an appropriate pooling method.

Object proposals. Considering the trade-off between efficiency and effectiveness, we adopt edgebox [31] to produce a set of object proposals. Unlike [19] which achieved good performance by choosing the best number of proposal regions from 100 to 2,000, we choose only top 100 regions ranked by their scores considering the computational efficiency.

Features for regions. After obtaining the regions of object proposals, we also extract CNN features from the pool5 layer in GoogLeNet for them:

$$F_o = \{f_{o_1}, f_{o_2}, f_{o_3}, \cdots, f_{o_N}\}, \qquad (1)$$

where F_o is the set of object features for an image, f_{o_i} is the CNN features of the i-th proposal region, N is the number of proposals (100 in our experiments).

Pooling methods. We introduce three pooling methods *i.e.*, max pooling, sum or average pooling and VLAD pooling.

(1) Max pooling: Girshick *et al.* [11] have illuminated that at the intermediate layers of the CNN, a high activation of a node indicates a specific type of visual input. Therefore, this pooling method mainly stores the high activation for each visual content. It can be represented as

$$f_o = [max\{f_{o_n}^{(1)}\}_{n=1}^{N}, \cdots, max\{f_{o_n}^{(D)}\}_{n=1}^{N}], \qquad (2)$$

where f_o is the final representation after pooling the features of all regions, $max\{f_{o_n}^{(i)}\}_{n=1}^{N}$ represent the max value at the i-th dimension of all features, and D is the dimension of each feature, which is also the final dimension in the component (1024 in our experiments).

(2) Sum or average pooling: This pooling method retains the sum/average activation for each visual content. It pools CNN features by

$$f_o = [\sum_{n=1}^{N} f_{o_n}^{(1)}, \cdots, \sum_{n=1}^{N} f_{o_n}^{(D)}], \qquad (3)$$

where $\sum_{n=1}^{N} f_{o_n}^{(i)}$ represents the sum value at the i-th dimension of all region features. The formulation for average pooling only needs to replace $\sum_{n=1}^{N} f_{o_n}^{(i)}$ by $\frac{1}{N}\sum_{n=1}^{N} f_{o_n}^{(i)}$.

(3) VLAD pooling: As for VLAD pooling, we first utilize k-means to obtain $c_1, c_2, \cdots, c_k$ of k visual word centers. For each image, a region feature f_{o_i} is assigned to its nearest center $c_m = NC(f_{o_i})$. After subtracting the corresponding cluster center, the residuals of each region feature are retained. An VLAD description for an image can be represented as follows:

$$F_v = [\sum_{NC(f_{o_i})=c_1} (f_{o_i} - c_1), \cdots, \sum_{NC(f_{o_i})=c_k} (f_{o_i} - c_k)], \qquad (4)$$

where F_v is an aggregated feature after pooling all object-level features for an image. And we represent the VLAD features by accumulating the residuals on each visual word center. Therefore, the dimensionality of such a representation is $k * d$, which is too high to describe an image efficiently for retrieval. We subsequently apply $L2$ normalization and PCA to obtain a fixed dimension, and then derive the object-level representation f_o.

3.3 Point-level representation

Preserving geometric invariance in image representation is important for image retrieval, since some images are obtained through scaling, rotating and other geometric transformations. Additionally, considering that the hand-crafted features can keep stable performance without the need of supervised training on a large dataset, we introduce the SIFT descriptor to represent point-level information.

RootSIFT [32] is adopted to generate the descriptors surrounding interest points for an image, and then VLAD is used to obtain the point-level representation. We also apply post-processing methods on this VLAD representation, *i.e.*, $L2$ normalization and PCA, to derive a fixed dimensionality. Finally, the point-level representation is denoted by f_p.

3.4 Fusing three-level representations

To fuse the three level features, we directly concatenate them to generate the integrated representation f with

$$f = [f_s, f_o, f_p]. \qquad (5)$$

Note that this integrated representation has relatively high dimensionality (i.e., three times as much as each component). Aiming to represent an image more compactly, we conduct a series of post-processing operations on the integrated representation. First, PCA and whitening are performed on f after the $L2$ normalization, which can be represented as below:

$$f_{whiten} = diag(1./sqrt(v_1, v_2, \cdots, v_h)) * U * \frac{f}{||f||_2}, \qquad (6)$$

where U is the PCA transformation matrix, h is the number of the retained dimensions after PCA, and v_i is the ith corresponding singular value. At last, $L2$ re-normalization is performed to obtain the final compact representation $f_{final} = \frac{f_{whiten}}{||f_{whiten}||_2}$ for an image.

4. EXPERIMENTS AND RESULTS

We perform extensive experiments on commonly used benchmarks to show the effectiveness of CCS. Two sets of experiments are conducted, including the evaluation of different pooling methods for object-level and the holistic performance compared with state-of-the-art methods.

4.1 Datasets and evaluation

Four image retrieval benchmarks are used in experiments, including INRIA Holidays [27], Oxford5K [28], Paris6K [29] and UKBench [30]. Following the corresponding standard evaluation protocols, we report the performance using mean average precision (mAP) on INRIA Holidays, Oxford5K and Paris6K, and report recall at top four results on UKBench.

4.2 Experiments of object-level pooling

In Sec. 3.2, we have introduced three pooling methods. Obviously, the pooling method will remarkably influence the retrieval performance. Compared with the other two methods, the VLAD pooling method needs more parameters. We utilize k-means to generate 500 centers, and then adopt PCA to reduce the dimensionality to 1024 ultimately. In our experiments, we follow [7] which uses the soft assignment to get a more powerful representation.

We experimentally compared these pooling methods on Holidays and UKBench datasets. The results are shown in Table 1. We can see that the VLAD pooling method is the best method to aggregate object-level semantic features. A reasonable interpretation of the results is that VLAD can capture the 0^{th} and 1^{st} order statistics as advocated in [5]. Whereas, max pooling and sum/average pooling lack 1^{st} order statistics. According to the experimental results and the analysis, we adopt VLAD for aggregating object-level CNN features.

Table 1: Comparison of different pooling methods aggregating object-level contents

Method	Dimension	Holidays	UKBench
Max	1024	66.89	3.70
Sum/average	1024	62.89	3.62
VLAD	1024	**71.66**	**3.77**

Table 2: The retrieval results (mAP) on INRIA Holidays dataset.

	Dimension						
Method	32	64	128	256	512	1024	2048
VLAD [5]	48.4	52.3	55.7	-	59.8	-	62.1
TE [6]	-	-	61.7	-	-	72.0	-
NC [15]	68.3	72.9	78.9	74.9	74.9	-	-
SPoC [21]	-	-	-	81.8	-	-	-
MOP [18]	-	-	-	-	78.38	-	80.18
LC [22]	-	-	83.6	-	-	-	-
SP [17]	-	-	-	74.2	-	-	-
OC [19]	73.96	**80.67**	**85.09**	**87.77**	**88.46**	**86.58**	85.94
$f_s + f_o$:	68.87	75.09	78.57	80.39	80.56	79.34	79.46
$f_s + f_p$:	71.36	78.46	81.74	83.96	84.84	83.79	83.83
$f_o + f_p$:	67.94	75.57	81.08	83.21	84.86	82.25	82.56
f_{final}	**74.65**	79.16	84.13	86.32	87.43	85.79	**86.09**

4.3 Comparison with state-of-the-arts

To make fair comparison, we evaluate the overall retrieval performance at different dimensions from 32 to 2048. Since the original dimension of our CCS image representation is 3072, we perform PCA on the CCS representation so as to derive its low-dimension variants. The experimental results are presented in Tables 2, 3, 4, 5. The best performance at each dimension is shown in bold.

Several observations can be drawn from these results. First, after performing PCA, features at relatively low dimensions may achieve better performance since PCA eliminates the influences of redundancies. Secondly, the combination of two level features can approach or outperform the pervious state-of-the-art methods, demonstrating that multi-level representation is very powerful. However, their performance is slightly unstable. For example, $f_s + f_p$ achieves better performance on Holidays, but the other two combinations achieve better performance on Oxford5K. Finally, compared with two level features, our method almost outperforms all the possible combinations of two level features on the four datasets (apart from the Paris6K dataset at 512 and 2048 dimensionalities). Moreover, ours outperforms all these state-of-the-art methods on the Oxford5K, Paris6K and UKBench datasets. In particular, our CCS reaches 3.91 on the UKBench dataset in term of recall. We also achieve comparable performance on the Holidays dataset, which is the smallest dataset containing only 1491 images.

Table 3: The retrieval results (mAP) on Oxford5K dataset

	Dimension						
Method	32	64	128	256	512	1024	2048
VLAD [5]	-	-	28.7	-	-	-	-
TE [6]	-	-	43.3	-	-	56.0	57.1
gVLAD [8]	-	-	60	-	-	-	-
NC [15]	39.0	42.1	43.3	43.5	43.5	-	-
SPoC [21]	-	-	-	59.3	-	-	-
LC [22]	-	-	59.3	-	-	-	-
SP [17]	-	-	-	53.3	-	-	-
OC [19]	40.1	48.02	56.24	59.78	60.71	59.42	58.92
$f_s + f_o$:	45.99	53.59	58.97	62.43	62.69	62.06	61.98
$f_s + f_p$:	43.98	50.19	55.02	58.14	59.59	60.46	60.04
$f_o + f_p$:	46.87	54.67	63.78	65.13	63.47	62.17	61.04
f_{final}	**48.71**	**56.78**	**64.84**	**67.62**	**67.26**	**66.41**	**65.43**

Table 4: The retrieval results (mAP) on Paris6K dataset

	Dimension						
Method	32	64	128	256	512	1024	2048
gVLAD [8]	-	-	59.2	-	-	-	-
LC [22]	-	-	59.0	-	-	-	-
SP [17]	-	-	-	67.0	-	-	-
OC [19]	65.38	71.47	70.39	68.43	66.23	64.11	62.84
$f_s + f_o$:	74.84	76.67	75.14	72.71	70.44	69.44	69.46
$f_s + f_p$:	70.52	75.35	75.22	73.94	**72.42**	71.02	**70.99**
$f_o + f_p$:	66.17	68.54	65.76	61.76	58.55	56.49	55.91
f_{final}	**76.34**	**78.27**	**76.76**	**74.35**	72.22	**71.07**	70.72

Table 5: The retrieval results on UKBench dataset.(* *indicates the recall result transformed from precision at top 4.*)

	Dimension						
Method	32	64	128	256	512	1024	2048
VLAD [5]	-	-	3.35	-	-	-	-
TE [6]	-	-	3.4	3.45	3.49	3.51	-
SP [17]	-	-	-	3.54*	-	-	-
SPoC [21]	-	-	-	3.65	-	-	-
NC [15]	3.3	3.53	3.55	3.56	3.56	-	-
OC [19]	3.4	3.61	3.71	3.77	3.81	3.84	3.84
$f_s + f_o$	3.51	3.66	3.73	3.78	3.80	3.82	3.83
$f_s + f_p$	3.48	3.65	3.74	3.79	3.82	3.83	3.84
$f_o + f_p$	3.50	3.69	3.78	3.83	3.87	3.89	3.89
f_{final}	**3.61**	**3.75**	**3.81**	**3.86**	**3.89**	**3.91**	**3.91**

5. CONCLUSION AND FUTURE WORK

In this paper, we verify the complementarity between C-NN and SIFT features for image retrieval tasks. To benefit from the complementarity, we propose the CCS model to represent scene-level, object-level and point-level visual content simultaneously. By compressing the CCS representation with simple PCA, it outperforms or achieves comparable performance over several state-of-the-art methods on four benchmarks. This paper also provides an insightful observation that the conventional hand-crafted features can be integrated with deep learning features so as to obtain better representation of an image. In the future, we will consider a different strategy of feature combination instead of simple PCA so as to further improve the performance for image retrieval, e.g., bayesian fusion.

6. ACKNOWLEDGEMENTS

This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB-351806, the National Natural Science Foundation of China under contract No. 61390515, No. 61425025 and No. 61471042, Beijing Municipal Commission of Science and Technology under contract No. Z151100000915070 and the National Key Technology Research and Development Program of China under contract No. 2014BAK10B02. This work is also supported by Microsoft Research Asia Collaborative Research Program 2016, project ID FY16-RES-THEME-034.

7. REFERENCES

[1] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[2] J.J. Foo and R. Sinha. Pruning sift for scalable near-duplicate image matching. In *ADC*, pages 63–71. Australian Computer Society, Inc., 2007.

[3] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *CVPR*, volume 2, pages II–506. IEEE, 2004.

[4] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8. IEEE, 2007.

[5] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, 2012.

[6] H. Jégou and A. Zisserman. Triangulation embedding and democratic aggregation for image search. In *CVPR*, pages 3310–3317. IEEE, 2014.

[7] A. Bergamo, S. N. Sinha, and L. Torresani. Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In *CVPR*, pages 763–770. IEEE, 2013.

[8] Z. Wang, W. Di, A. Bhardwaj, V. Jagadeesh, and R. Piramuthu. Geometric vlad for large scale image search. *arXiv preprint arXiv:1403.3829*, 2014.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587. IEEE, 2014.

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[13] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, pages 262–270, 2015.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[15] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599. Springer, 2014.

[16] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Multimedia*, pages 157–166. ACM, 2014.

[17] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Visual instance retrieval with deep convolutional networks. *arXiv preprint arXiv:1412.6574*, 2014.

[18] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, pages 392–407. Springer, 2014.

[19] Reddy M. K. and Venkatesh B. R. Object level deep feature pooling for compact image representation. In *CVPRW*, pages 62–70, 2015.

[20] L. Xie, R. Hong, B. Zhang, and Q. Tian. Image classification and retrieval are one. In *ICMR*, pages 3–10. ACM, 2015.

[21] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *ICCV*, pages 1269–1277. IEEE, 2015.

[22] J. Y. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. *arXiv preprint arXiv:1504.05133*, 2015.

[23] L. Zheng, Y Zhao, S. Wang, J. Wang, and Q. Tian. Good practice in cnn feature transfer. *arXiv preprint arXiv:1604.00133*, 2016.

[24] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[25] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard. A practical guide to cnns and fisher vectors for image instance retrieval. *arXiv preprint arXiv:1508.02496*, 2015.

[26] L. Zheng, S. Wang, J. Wang, and Q. Tian. Accurate image search with multi-scale contextual evidences. *IJCV*, pages 1–13, 2016.

[27] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317. Springer, 2008.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8. IEEE, 2007.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8. IEEE, 2008.

[30] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, volume 2, pages 2161–2168. IEEE, 2006.

[31] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV 2014*, pages 391–405. Springer, 2014.

[32] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pages 2911–2918. IEEE, 2012.