# Performance Evaluation of SIFT and Convolutional Neural Network for Image Retrieval

Varsha Devi Sachdeva[1], Junaid Baber[2], Maheen Bakhtyar[2], Ihsan Ullah[2], Waheed Noor[2], Abdul Basit[2]
[1]Department of Computer Science, Sardar Bahadur Khan Women's University
Quetta, Pakistan
[2] Department of CS & IT, University of Balochistan, Quetta, Pakistan

***Abstract*—Convolutional Neural Network (NN) has gained a lot of attention of the researchers due to its high accuracy in classification and feature learning. In this paper, we evaluated the performance of CNN used as feature for image retrieval with the gold standard feature, aka SIFT. Experiments are conducted on famous Oxford 5k data-set. The mAP of SIFT and CNN is 0.6279 and 0.5284, respectively. The performance of CNN is also compared with bag of visual word (BoVW) model. CNN achieves better accuracy than BoVW.**

***Keywords*—Computer vision; SIFT; CNN; image retrieval; precision; recall**

## I. Introduction

In computer vision, image processing involves the information extraction from the images for the human interpretation and process it efficiently for the machine perception. Content based image retrieval (CBIR) is concerned with the retrieving images of the given subject from the expansive or huge database. Visual media content in social media channels is the most common type of the content, and it has gained the researchers attention to come up with the efficient image retrieval features which can retrieve particular object from huge databases. People always wish to retrieve the number of images for given query as much as he could be allowed by giving only the one image or object as a query. For this purpose, number of steps are included in which feature extraction is the most important one. The basic problem in image retrieval is the space amongst the high level descriptors used by human to demonstrate the descriptors of image and low level features and also the space required to save that descriptors in memory [1].

Scale Invariant Feature Transform (SIFT) has been widely used for past decade for feature extraction from the images [2]. SIFT is intrinsically robust to geometric transformation and shows better performance for different computer vision tasks such as near-duplicate image retrieval [3]. On the given image, local keypoints are detected and represented by SIFT. On average, there are $2.5 - 3.0K$ points are detected. Two images are treated similar if there are more than $T$ matched points. To match a point of one image to other image, distance is computed of given point with all the points in the second image and then closet point is considered a candidate match. Final decision that either the pair point is matched or not, is made after comparing the closest point with second nearest point, as discussed in the Section II-B. The matching process between two images takes on average 1.5 seconds on normal commodity hardware. Since, the exhaustive search of SIFT is computationally very expensive and not feasible for large databases.

There are two main problems with local keypoint based descriptors. The first problem is feature space and storage. In case of SIFT, there are $2.5$-$3.0K$ descriptors per image, each descriptor is $128$-$D$ and each value is floating point. To store the raw descriptors/image, at least 1.2 megabytes are required (given 4-bytes/float value), which is sometimes more than the image size. The second problem is the computational complexity to find the similarity between two images. As stated above, it takes around 1.5 seconds to match two images.

To overcome the the above mentioned limitations, local keypoint descriptors are quantized using BoVW. There are number of prominent techniques for quantization such as Fisher Vector [4], VLAD [5]–[8], binary quantizer [9], and BoVW model [9]–[11].

BoVW model is widely used for several computer vision and image based applications such as image retrieval [9]–[11] and image classification [7]. The idea of BoVW model is inspired from text retrieval system where the text document is represented by the frequencies of words. To normalize the size of vocabulary, stop-words and most frequent words are ignored/removed and remaining words are stemmed or lemmatization is applied, i.e., playing or played to play. To apply same idea on visual domain, descriptors are clustered and each cluster center is considered as visual word. The clustering is offline process and clusters are learned from large instances. The visual content, lets say an image, is represented by histograms of visual word present in it. The process of quantization i-e mapping each descriptor to its cluster is explained in Section II-C.

Recently, Convolution Neural Network (CNN) has achieved the state of the art performance on various different computer vision applications [12]–[14]. The main focus of CNN is on object/image classification based applications. Few papers also reported frameworks to use CNN as image features [15].

In this paper, we also used CNN as image feature for image retrieval based on visual contents. We evaluated the performance of CNN on Oxford dataset, explained in experimental section, along with SIFT and BoVW model.

## II. Related Work

In this section, we briefly discuss some recent advances and literature on SIFT, BoVW, and CNN. Later in this section we briefly explain all these three frameworks.

Image retrieval is classified in two categories: text-based search and content based search. Text based search refers

TABLE I. CNN ARCHITECTURE. THE ARCHITECTURE IN CNN CONTAINS TOTAL 8 LAYERS IN WHICH 5 LAYERS ARE CONVOLUTION LAYERS AND LAST 3 LAYERS ARE FULLY CONNECTED LAYERS [16].

| Arch. | Conv1 | Conv2 | Conv3 | Conv4 | Conv5 | Full6 | Full7 | Full8 |
|---|---|---|---|---|---|---|---|---|
| CNN-F | 64x11x11 st.4,pad 0 LRN, x2 pool | 256x5x5 st.1, pad 2 LRN, x2 pool | 256x3x3 st.1, pad 1 - | 256x3x3 st.1, pad 1 - | 256x3x3 st.1, pad 1 x2 pool | 4096 Drop-out - - | 4096 Drop-out - - | 1000 soft-max - - |

to technique where the images are first annotated manually and then text-based database management systems are used to perform retrieval tasks. Whereas, content based images search technique refers to automatically annotation of images with in their visual contents. These include colors [17]–[19], shapes [20], textures [21] or any other information that can be extracted from the image and are indexed by using indexing techniques for large scale retrieval [22].

Recently, Object search in images has also got much attention of the researchers [23], [11]. One of the most initial work on object and scene retrieval is Video Google [11] which is inspired by text based search (Google). Initially, keypoints are detected and represented by SIFT [2] which is 128 D vector against each keypoint. As described in previous section, on average there are 2.5 K to 3.0 K keypoints on single image. Each keypoint descriptor is quantized to its appropriate visual word. The process of quantization into BoVW is explained in Section II-C. The BoVW is proven to be effective and efficient for large databases [10], [24]–[26].

There are number of variations of SIFT [27], [28] where only the the robustness or distinctiveness of SIFT is improved. However, these methods are limited to small or moderate databases. To make searching computationally effective, either the descriptors are quantized to Hamming space [9] or quantized to single image feature, aka BoVW [10], [24]–[26].

CNN is also used for image representation [15], [29]–[32]. Multi-Scale Order less pooling (MOP) is introduced to represent the local feature descriptor by aggregating the CNN descriptors at three scales [29]. Different researchers first detect the subject object then extract the CNN features for each region in object [33], [34]. Pre-trained image classification neural networks have been widely used for feature extraction. The results of image retrieval can be improved by combining the FC of neural network from variant image sub-patches [14]. Images can be represented by comprising of sum of the activations of each convolution layer filter [30]. In case of R-MAC, which is a compact descriptor and contains the aggregation of the multiple regional features [15], improves the system significantly by applying the non-parametric spatial and channel-wised weighting strategies to CNN layers [35].

### A. Convolution Neural Network (CNN)

A Convolution Neural Network or feed-forward network contains number of functions and can be represented mathematically as

$$f_x = f_L(...f_2(f_1(x; w_1); w_2)...), w_L{}^{1} \quad (1)$$

Every function $f$ as shown in (1) takes a piece of information as input with a parameter vector $w_l$ and produces as output a piece of information. While the sequence of functions in CNN are handcrafted and the parameter $W = (w_1, w_2, ..., w_L)$ are the weights which are learned from the data $x$ which can be be any kind of data such as image matrix and audio/video signal. In our experiments, $x$ is color image of $m \times n \times c$, where $m \times n$ denotes the pixel in width and height, and $c$ denotes color channels.

The output of the convolution layer has filters with 3 dimensions. This is because they operate on tensor $x$ with $c$ channels. Furthermore, there are $c$' filters which are generating $c$ dimensional mapped output $y$. The convolution output $y$ has to pass from non-linear activation functions.

*1) Non-linear activation function:* CNN is composed of many functions, linear and non-filters. The major reason of having the activation function is to introduce the non-linearity into the network. Non-linear activation function has a significant importance in the network. Without activation function multi-layer neural network will behave like single layer neural network. The reason behind is that the summing of these layers would give you just another linear function. The simplest non-linearity is obtained by the non-linear activation function which is Rectified Linear Unit (ReLU) applied to each component of the feature map $y$.

*2) Pooling:* CNN has several different operators and one of them is pooling. Pooling operates on each feature channel. It combines the feature values into one suitable operator, common choices include max-pooling and sum-pooling. Max-pooling is a non-linear down sampling of the input. It divides the input image into non-overlapping rectangles and for each region it outputs the maximum value. The process of pooling reduces the computation for upper layers, facilitate translation invariance, robustness to position, and reduces the dimensionality of the input.

Convolution layer and max pooling layers are the lower layers of CNN and the fully connected layers are upper layers correspond to the traditional MLP (Multi-layer Perceptron). MLP is combination of hidden layer and logistic regression. The input to the first fully connected upper layer is the set of 4D features operated by the lower layer which is then flattened to 2D matrix of re-sized feature map.

The CNN based features are actually based on few models and each model explores different accuracy and speed trade-off. These networks are trained using same protocols and implementation. In our research, we have used pre-trained model, aka Fast CNN (CNN-F) [16], which is similar to the [36]. The CNN-F models consists of total 8 layers in which 5 layers are convolution layers and 3 layers are fully-connected layers. CNN require input image to be transformed to the fixed size which is $(224 \times 224)$. Hence the image is reduced to the $224 \times 224$. Fast processing is ensured by using the 4-pixel stride in the first convolution layer with 0 padding and max-pooling down sampling factor is $3 \times 3$. For fully connected

[1] http://www.robots.ox.ac.uk/ vgg/practicals/cnn/index.html

TABLE II. THE COLUMNS IN THE TABLE REPRESENT THE NAME OF THE LANDMARKS AND AVERAGE PRECISION IS THE AVERAGE PRECISION VALUE OF FIVE QUERIES OF EACH LANDMARK, WHEREAS MAP IS THE MEAN AVERAGE PRECISION OF OXFORD 5K DATASET

| Avg Precesion/Recall | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | L9 | L10 | L11 | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIFT | 0.449 | 0.5187 | 0.5112 | 0.6724 | 0.6588 | 0.6704 | 0.6924 | 0.8744 | 0.2758 | 0.9918 | 0.5911 | 0.6279 |
| CNN | 0.4036 | 0.4074 | 0.3964 | 0.3618 | 0.4751 | 0.4530 | 0.8065 | 0.5835 | 0.2790 | 0.7645 | 0.8813 | 0.5284 |
| BoVW | 0.3656 | 0.3425 | 0.3635 | 0.4143 | 0.3817 | 0.4051 | 0.6980 | 0.5603 | 0.2515 | 0.5658 | 0.6024 | 0.4501 |

layer from (1-3) their dimensionality is same for all types of architecture which is 4096 per image. Full6 and Full7 layers in CNN are arranged using dropout while the last Full8 layer in CNN behave as soft-max classifier, and the activation function for all layers except Full8 last layer is rectified linear unit (ReLU) [36].

Table I shows the main configuration of the pre-trained CNN network (CNN-F) which is used in our paper.

### B. Scale Invariant Feature Transform (SIFT)

The Scale Invariant Feature Transform (SIFT) is used for interest region detector using Difference of Gaussian (DoG) and feature descriptor. The SIFT feature descriptor achieves the robustness to various illumination, lighting and positional shifts by encoding in a localized set of gradient orientation histograms (HoG) [9]. In the first step, the gradient magnitude and orientation of image is examined around the key point location to select the level of Gaussian Blur using the region scale. In each examined region, sampling is performed in the $16 \times 16$ area of regular grid, which is covering the interest region. The gradient orientation is entered into the $4 \times 4$ patch of gradient histogram with 8 bins each. The main reason of this Gaussian window is to give higher weight to pixels closer to middle regions which are less position variant. Once all the histogram entries have been completed then those entries are concatenated to form a single feature vector with 128 dimensions, i-e $4 \times 4 \times 8 = 128$. Finally, all the values to normalized to the unit vector to minimize the influence of high spikes in the histogram.

### C. Bag of Visual Words

For given image, first step is to detect the keypoints. Second step is to compute the descriptors such as SIFT from each key point. In the last step, each keypoint is quantized. As stated above, the most famous quantizer is BoVW.

The BoVW, **B**, is the quantizer which quantizes the descriptor $d \in \mathbb{R}^{128}$.

$$\begin{aligned} \mathbf{B}: \quad \mathbb{R}^{128} &\rightarrow [1, K] \\ d &\rightarrow \mathbf{B}(d) \end{aligned} \tag{2}$$

**B** quantizes all the keypoint descriptors of an image into visual words by assigning each descriptor $d \in \mathbb{R}^{128}$ to any of the $K$ cluster centers, known as visual word. The set of visual words is denoted by $\mathcal{V} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_K}\}$. At the end of quantization, histogram of visual words are computed from the image. The effectiveness of model **B** is highly dependent on the number of cluster centers $K$. The BoVW is more robust if the value of $K$ is small as the voronoi cells can store more values but low value yields low distinctiveness– two different descriptors may be quantized into single cell. The BoVW is more distinctive when the value of $K$ is very large, two different descriptors which are close in feature space are quantized into two different cell as the veronoi cells are very close to each other.

Experimentally, the value of $K$ minimum quantization error is 1.0 million [10], [37]. Flat K-means [11] and hierarchical K-means [24] are extensively used for visual words.

## III. EXPERIMENTAL EVALUATION

CNN is compared with SIFT and BoVW model for image retrieval on benchmark dataset using stranded protocol. Later in this section, dataset, evaluation protocols, configuration for SIFT, BoVW, and CNN are explained.

### A. Datasets

Oxford 5k dataset is used for image retrieval [38]. This dataset contains 5062 images, denoted as $I = \{a_1, a_2, \ldots, a_n\}$, which are collected from Flickr with the name of 11 different landmarks in oxford. There are 5 queries of each landmark and total it has 55 queries with Region of Interest (ROI). Each query has *Good, OK* and *Junk* labels in ground-truth. The first two labels, *Good* and *OK*, are the treated as true positives for the query.

### B. Evaluation Metrics

Precision and recall are used as evaluation metrics and denoted by $P$ and $R$, respectively. These metrics are defined as follow

$$\begin{aligned} P &= \frac{\psi}{\tau} \\ R &= \frac{\psi}{\omega} \end{aligned} \tag{3}$$

$\psi$ denote the true positives retrieved, $\tau$ denotes the total retrieved, and $\omega$ denotes total relevant. A perfect CBIR is the one which retrieve all the true matches against the query in the database and return them in the top rank list ($P = R = 1.0$).

### C. Frameworks Configuration

Pre-trained CNN network is used [16] which is Open source distribution. Each image is represented by feature vector of 4096 dimension. The network consists of eight layers, the initial five layers of network are convolution layers and last three layers are fully connected layers. The output of the last fully connected (FC) layer of network is fed into 1000-way soft-max which produces a distribution over 1000 classes [36].

Firstly, CNN layers filters the image $i_j \in I$ where $I = (i_1, i_2 ..., i_n)$ are the images in the database (each of which is $224 \times 224 \times 3$ in size) with the 64 kernels with the stride of 4 pixels, which is the distance between the receptive

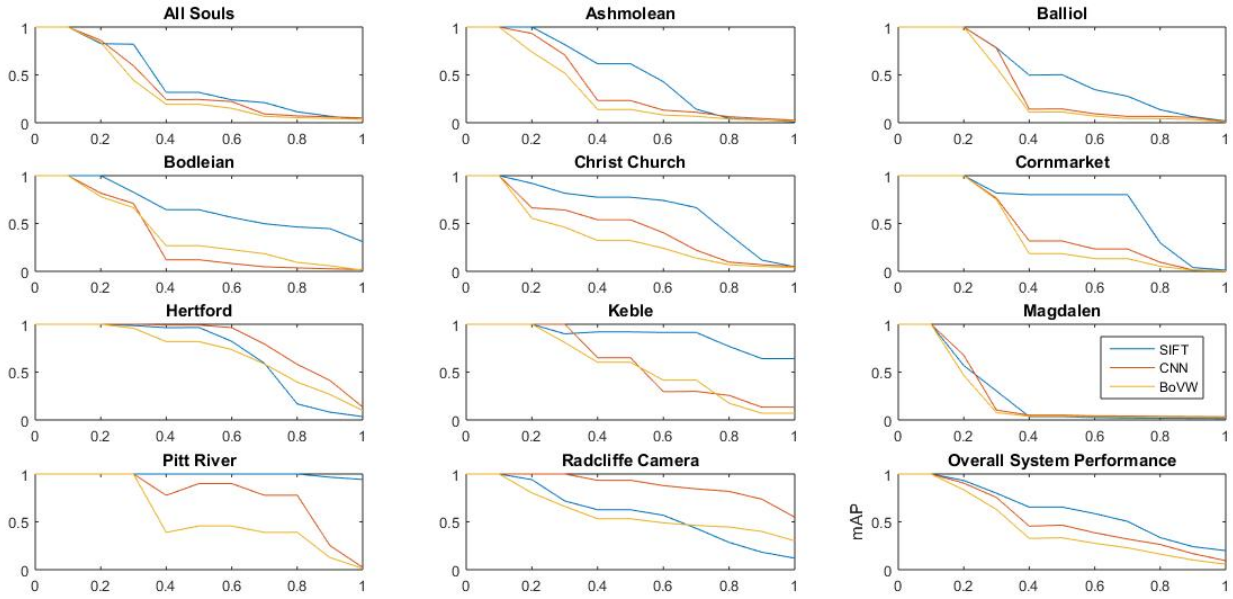


Fig. 1. Performance of SIFT, CNN, and BoVW on Oxford dataset.

center of neighboring neuron in kernel map, as shown in Table I. Then second layer in the CNN takes the input from the first CNN layer output which is response-normalized and pooled output and again filters it with 256 kernel of size $5 \times 5$. After that, the third, fourth and fifth layer of CNN are connected with one another without any intervening pooling or normalization layers. The fifth and sixth fully connected layers of CNN have 4096 neurons each, while the last layer of CNN acts as a multiway soft-max classifier. The last layer represent the number of classes in neural network and it performs the soft max classification on each of the input of the convolution neural network. Usually soft-max output is dedicated to each class in CNN, all of which are generally connected to the previous hidden layer in CNN. In the last layer of CNN, any one-to-one mapping can be used in between of the neurons which are 4096 and classes which are 1000. For any image, CNN pre-trained networks give feature vector of dimension 4096. Vocabulary of 4096 clusters are trained on the 40000 images obtained by Flickr 100K dataset[2]. For BoVW, Harris keypoints are detected and later represented by SIFT descriptors, then all descriptors are quantized into BoVW. A descriptor $\mathbf{d}$ is assigned visual word $\mathbf{w_i} \in \mathcal{V}$ provided

$$E(\mathbf{d}, \mathbf{w_i}) = \min_{\mathbf{w_j} \in \mathcal{V}} E(\mathbf{d}, \mathbf{w_j}) \tag{4}$$

Where, $E$ is the Euclidean distance defined as

$$E(\mathbf{d_1}, \mathbf{d_2}) = \sqrt{\sum_{i=1}^{m} (\mathbf{d_1}(i) - \mathbf{d_2}(i))^2} \tag{5}$$

Finally, histogram of visual word is computed. Each image is represented by histogram of visual words of dimension $K$ where $K = 4096$.

[2]http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/

To evaluate the retrieval performance of CNN and BoVW, we compute the precision for each query image. Since, there are 11 landmarks and each has 5 queries. We report the mean precision for each landmark. To compute the precision, rank-list of the each query image is obtained by computing the Euclidean distance of query feature vector with all the feature vectors in $I$, distance is then sorted in ascending order, and precision is computed on every true positive index. Same protocol cannot be applied for SIFT based retrieval. In case of SIFT, image is represented by set of features. We do exhaustive search for SIFT retrieval and rank-list is obtained by matching score of query image with all the images in $I$. The matching score $\mathcal{W}$ between two images, $a_1$ and $a_2$, is computed as follow

$$\mathcal{W}(a_1, a_2, T_m) = \frac{||S(a_1) \cap^{T_m} S(a_2)||}{||S(a_1)||} \tag{6}$$

$S(.)$ denotes the set of SIFT descriptors for given image, $||S(.) \cap^{T_m} S(.)||$ denotes the stable matched points between two sets of point, the point pair is stable if the distance between two points satisfies $T_m$, as suggested by David lowe [2], similar matching protocol is used by many researchers [9], [27].

Fig. 1 shows the performance of SIFT, BoVW, and CNN on Oxford 5K dataset. Each subplot shows the average precision over 5 queries for each landmark. Last plot shows the overall performance. It can be seen that SIFT performs better than CNN and BoVW, but at the cost of computation. SIFT image takes on average 1.5 seconds to find the matching score between pair of images whereas the distance between two images, in case of CNN and BoVW, takes 0.02 seconds.

Table II represents the mean average precision (mAP) of SIFT, CNN, and BoVW on Oxford dataset for each landmark, the last column shows the average of all landmarks. The CNN surprisingly gives better performance than BoVW despite

the fact CNN is computed as globally and BoVW represents the local features. For some frameworks BoVW gives similar performance as of SIFT provided the vocabulary size upto 1.0 millions. In our experiments, the vocubalary size is only 4096.

The feature extraction time, on average– of SIFT, CNN and BoVW are 1.6, 0.4, and 3.2 seconds, provided CNN and BoVW use pre-trained networks and clusters, respectively.

## IV. Conclusion

In this paper, we have evaluated SIFT, CNN, and BoVW for image retrieval application. CNN have been used for classification problems, in this paper, we evaluated for retrieval problem in parallel with gold standard SIFT and BoVW. Experiments show that CNN achieve comparable performance with SIFT w.r.t accuracy and outperform BoVW. SIFT matching is limited to small databases, whereas, CNN and BoVW can be used for moderate databases and easily be extended for large scale retrieval. The CNN and BoVW are faster to extract features and retrieval than SIFT, but SIFT outperforms w.r.t accuracy.

## References

[1] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, "Cnn vs. sift for image retrieval: Alternative or complementary?" in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 407–411.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] J. J. Foo and R. Sinha, "Pruning sift for scalable near-duplicate image matching," in *Proceedings of the eighteenth conference on Australasian database-Volume 63*. Australian Computer Society, Inc., 2007, pp. 63–71.

[4] 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*. IEEE, 2007, pp. 1–8.

[5] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[6] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3310–3317.

[7] A. Bergamo, S. N. Sinha, and L. Torresani, "Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 763–770.

[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] J. Baber, M. N. Dailey, S. Satoh, N. Afzulpurkar, and M. Bakhtyar, "Big-oh: Binarization of gradient orientation histograms," *Image and Vision Computing*, vol. 32, no. 11, pp. 940–953, 2014.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[12] B. Bhattarai, G. Sharma, and F. Jurie, "Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4226–4235.

[13] 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.

[14] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

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

[16] 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.

[17] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 762–768.

[18] E. Kasutani and A. Yamada, "The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval," in *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 1. IEEE, 2001, pp. 674–677.

[19] G. Pass and R. Zabih, "Histogram refinement for content-based image retrieval," in *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on*. IEEE, 1996, pp. 96–102.

[20] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[21] C. S. Won, D. K. Park, and S.-J. Park, "Efficient use of mpeg-7 edge histogram descriptor," *ETRI journal*, vol. 24, no. 1, pp. 23–30, 2002.

[22] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39–62, 1999.

[23] R. Tao, E. Gavves, C. G. Snoek, and A. W. Smeulders, "Locality in generic instance search from one example," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2091–2098.

[24] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. Ieee, 2006, pp. 2161–2168.

[25] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[26] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 511–520.

[27] J. Baber, M. Bakhtyar, W. Noor, A. Basit, and I. Ullah, "Performance enhancement of patch-based descriptors for image copy detection," *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 7, no. 3, pp. 449–456, 2016.

[28] J. Baber, E. Fida, M. Bakhtyar, and H. Ashraf, "Making patch based descriptors more distinguishable and robust for image copy retrieval," in *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*. IEEE, 2015, pp. 1–8.

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

[30] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[31] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 53–61.

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

[33] K. Reddy Mopuri and R. Venkatesh Babu, "Object level deep feature pooling for compact image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 62–70.

[34] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 3–10.

[35] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *European Conference on Computer Vision*. Springer, 2016, pp. 685–701.

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

[37] G. Amato, F. Falchi, and C. Gennaro, "On reducing the number of visual words in the bag-of-features representation," *arXiv preprint arXiv:1604.04142*, 2016.

[38] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.