Exploiting Hierarchical Activations of Neural Network for Image Retrieval

Ying Li[1], Xiangwei Kong[1], Liang Zheng[2], Qi Tian[2]
[1]Dalian University of Technology, China
[2]Dept. of Computer Science, University of Texas at San Antonio, Texas, TX 78249
liying08@mail.dlut.edu.cn, kongxw@dlut.edu.cn, liangzheng06@gmail.com, wywqtian@gmail.com

ABSTRACT

The Convolutional Neural Networks (CNNs) have achieved breakthroughs on several image retrieval benchmarks. Most previous works re-formulate CNNs as global feature extractors used for linear scan. This paper proposes a Multi-layer Orderless Fusion (MOF) approach to integrate the activations of CNN in the Bag-of-Words (BoW) framework. Specifically, through only one forward pass in the network, we extract multi-layer CNN activations of local patches. Activations from each layer are aggregated in one BoW model, and several BoW models are combined with late fusion. Experimental results on two benchmark datasets demonstrate the effectiveness of the proposed method.

CCS Concepts

•Computing methodologies → Visual content-based indexing and retrieval;

Keywords

Image retrieval; Bag-of-Words; Feature fusion

1. INTRODUCTION

Given a certain query image, image retrieval [11, 13, 23, 25, 24, 26] aims at returning similar images from the database. To reliably measure the pairwise similarity between images, low level handcrafted descriptors (*e.g.*, SIFT [15]) are usually plunged into Bag-of-Words (BoW) model. These low level descriptors capture the local patterns of images, making BoW pipeline robust to occlusion, truncation, and rotation.

Given a visual vocabulary learned off-line, BoW counts the frequency of local descriptors in an image. As a consequence, each image is encoded using a visual histogram, which is later compared under cosine distance [16] or Hellinger kernel [1]. To reduce the quantization loss, the "soft-assignment" strategy [19] is proposed, which assigns each local descriptor to more than one visual words. Simultaneously, larger efforts are also devoted to more advanced encoding methods [12, 20] and postprocessing algorithms [7, 22, 4].

In recent years, deep learning approaches have demonstrated superior performances in various applications in multimedia community, such as image categorization [6, 14], 3D shape recognition [5], object detection [8], *etc.* In [9, 3, 2, 21], Convolutional Neutral Network (CNN) is used as a generic feature extractor for image retrieval. The extracted features are usually activations of the high layers of CNN (*e.g.*, fully connected layers), taken as holistic representations which can reveal the high level semantics of images.

In this paper, we propose an effective Multi-layer Orderless Fusion (MOF) method to simultaneously capture the low-level patterns and high-level semantics of images by using a single deep neutral network. While previous works typically focus on the discriminative ability of activations from high layers, we observe that the activations from low layers (*e.g.*, convolution layers) of deep neutral network are effective as low-level cues for image retrieval.

In order to capture the local visual patterns of images, we extract these descriptors on densely sampled patches. This is the primary difference compared with previous works [3, 2, 21] which use the global images. To estimate the matching strength between two images, we use the conventional BoW model, accelerated by inverted index [16] to reduce the computational complexity and memory usage. By doing so, we are capable of achieving satisfactory retrieval performances with acceptable indexing cost.

Meanwhile, inspired by Hamming Embedding (HE) [10], these features are also binarized to improve the matching accuracy. Furthermore, to leverage the complementarity between the low level patterns and high level semantics, we fuse the activations from different layers in the score level. As opposed to Multi-scale Orderless Pooling (MOP) [9] that aggregates activations of images at different image scales, the aggregation in this paper is applied to activations of different layers of deep neutral network. As a result, the retrieval performance is improved further due to the integration of low level cues and high level semantics.

The rest of this paper is organized as follows. The proposed MOF framework is described in Section 2. In Section 3, the experimental results are presented and discussed. Finally, we conclude in Section 4.

2. PROPOSED APPROACH

In this section, we give a formal description of the proposed Multi-layer Orderless Fusion (MOF) based image retrieval pipeline. Figure 1 illustrates the framework of the proposed method.

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2967197

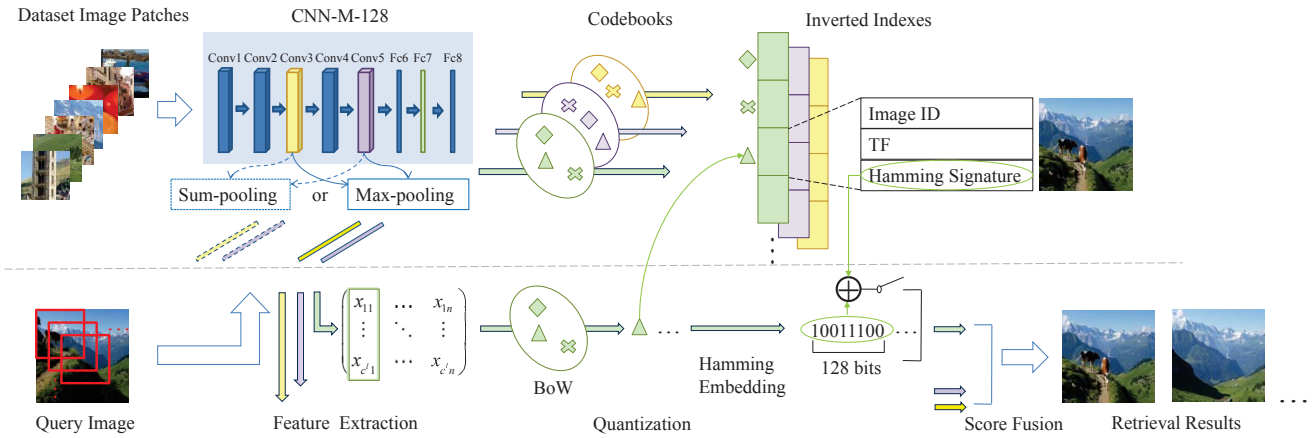


Figure 1: The pipeline of the proposed approach.

2.1 Multiple Deep Convolutional Description

Despite that several previous works [9, 3] claim that fully connected (Fc) layers give the best results compared with other layers when performing image retrieval, activations from convolutional layers should not be ignored since they preserve more instance-level details, which are especially meaningful for local area based image retrieval. In this paper, a fully connected layer, the last convolutional layer, and an intermediate convolutional layer of the CNN-M-128 [6] model are employed to construct the multiple deep convolutional description. The 7th layer (Fc7) activation of this model is compact with 128 dimensions and yields a competitive performance that will be discussed in the experiments.

For the convolutional layers, sum-pooling and max-pooling are applied separately to generate feature vectors. Given the pre-trained CNN-M model $\mathcal{C}$, deep convolutional feature maps f^l are obtained by passing image patches $\mathcal{P}$ through the network $\mathcal{C}$, with l presenting different layers. The size of f^l can be denoted as $w^l \times h^l \times c^l$, with w^l, h^l, and c^l denoting the width, height, and the number of channels of the layer l, respectively. Then sum-pooling is performed on f^l, aggregating each feature map into a single value, resulting in $p^l_{sum} \in \mathbb{R}^{1\times 1\times c^l}$. Denote the kth element of p^l_{sum} as $p_{sum}(k)$, then

$$p^l_{sum}(k) = \sum_{i=1}^{w^l}\sum_{j=1}^{h^l} f^l(i,j,k), \quad k = 1,2,...,c^l. \qquad (1)$$

The max-pooling version also generates a $1 \times 1 \times c^l$ dimensional p^l_{max} as

$$p^l_{max}(k) = \max_{1\leqslant i\leqslant w^l, 1\leqslant j\leqslant h^l} f^l(i,j,k), \quad k = 1,2,...,c^l. \qquad (2)$$

In this paper, the convolutional layers employed have 512 convolutional kernels, so the pooling result is a $1\times1\times512$ dimensional feature, which is then reshaped to a single column vector.

2.2 Two-step Quantization

In this section, we explore a feasible way to integrate deep features into the Bag-of-Words (BoW) structure, which is compatible with previous complementary techniques like Hamming Embedding (HE) in the hand-crafted low level feature based retrieval field at the same time. The procedures of assigning CNN features to visual words and generating binary signatures are two steps of quantization, with the Hamming codes acting as a finer division to boost the discriminative power of quantized features.

To get an adequate number of descriptors for constructing the BoW description of images, we densely sample image patches on a fixed size of 224×224 through a sliding window method with a stride size of 32 on the grid. We denote the set of image patches as $\mathcal{P}$ which is the input of the CNN-M-128 model $\mathcal{C}$, where $\mathcal{P} = \{P_1, P_2, ..., P_n\}$ includes n image patches in total. After the forward propagation through the pre-trained network $\mathcal{C}$, n features are obtained from each layer for a single image.

Considering activations from the three different layers separately, we train codebooks using Approximate K-Means (AKM) [18] clustering, followed by assigning features to the nearest visual word of the corresponding quantizer using the Approximate Nearest Neighbors (ANN) algorithm. Multiple Assignment (MA) [19] is applied in this quantization step to expand candidate visual centers for boosting recall. In this paper, the vocabulary size is set to be 20k and a feature is empirically assigned to three visual words.

Based on the built BoW structures, the inverted indexes are individually constructed, storing image IDs and Term Frequency (TF) scores aligned to the visual words. As complement for visual words, Hamming Embedding (HE) [10] is employed in this paper as a binary signature to refine the matching results by sub-splitting the quantized feature space. Assuming that the dimension of the original feature is c^l, and the code length after Hamming Embedding turns into h ($h \leqslant c^l$) which is set to be 128 in this paper, then a transition matrix P^l, with $P^l \in \mathbb{R}^{h*c^l}$, is trained to project both training data and target features from the c^l dimensional CNN description to another h dimensional space.

It should be paid attention to that the transition matrix P^l differs between layers. That may cause different information loss among the three layers, since convolutional layers obviously suffer more from the dimensional reduction. However, in our experiments, the two convolutional layers still gain comparable performance with the fully connected layer in this condition, proving the discriminative power of regional CNN descriptors.

2.3 Multi-layer Orderless Fusion (MOF)

Inspired by Multi-scale Orderless Pooling (MOP) [9], which extracts 4096-dimensional features from the fully connected layer at multiple scales, we proposed another Multi-layer Orderless Fusion (MOF) approach, where CNN features from convolutional layers are also taken into consideration with all the features extracted at the same single scale. Our work is different from MOP on three aspects, *i.e.*, features, quantization methods, and scales, though we both make efforts to explore local CNN based retrieval.

The two main advantages of the proposed MOF are: First, the convolutional layer activations can present low and mediate level cues apart from the high level semantics from the fully connected layer, so the fusion of them is complementary to each other. Second, all the activations can be obtained by passing a set of single scaled image patches through the deep neural network only once, without being extracted individually by several times. In addition, the activations employed in this paper are of much lower dimension than the 4096-dimensional ones utilized in previous works [9, 3] while the proposed approach still yields competitive performance.

Let x and y be two feature vectors in F_t, where $x, y \in \mathbb{R}^{c^l}$, the matching score between them is defined as

$$s = \begin{cases} idf^2_{q(x)}, & q(x) = q(y), \\ 0, & q(x) \neq q(y), \end{cases} \tag{3}$$

where $q(x)$ and $q(y)$ represent the visual words corresponding to x and y, respectively, and $idf_{q(x)}$ is the Inverse Document Frequency (IDF) of $q(x)$,

$$idf_{q(x)} = \log \frac{N}{n_{q(x)}}, \tag{4}$$

where N is the amount of images in total, and $n_{q(x)}$ counts for those ones containing $q(x)$.

Taking the binary signature into consideration, the matching results in Equation (3) are filtered as

$$s_h = \begin{cases} idf^2_{q(x)} \cdot w_h, & q(x) = q(y) \text{ and } d(b_x, b_y) < d_t, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

with w_h added as a matching strength

$$w_h = \exp\left(-\frac{d^2(b_x, b_y)}{\sigma^2}\right). \tag{6}$$

b_x and b_y are Hamming codes of feature vectors x and y respectively, and $d(b_x, b_y)$ measures the distance between them as

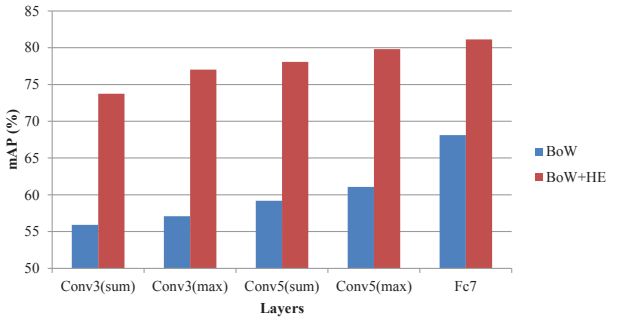
$$d(b_x, b_y) = \sum_{i=1}^{h} b_x(i) \oplus b_y(i), \tag{7}$$

where $\oplus$ means bitwise XOR operation. d_t and σ are parameters determining the acceptable intra-centroid distance and the weighting magnitude, which are set to be 52 and 30, respectively.

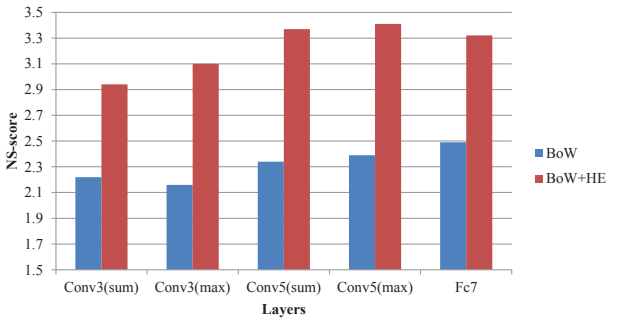
To aggregate the effects of multiple layers, score fusion scheme is adopted as

$$S_h = \sum_l \sum_x s^l_h(x) / \sqrt{tf^l_{q(x)}}, \tag{8}$$

where $tf^l_{q(x)}$ denotes the Term Frequency (TF) stored in the inverted index of node $q(x)$ corresponding to layer l.



(a) mAP on Holidays dataset



(b) NS-score on UKBench dataset

Figure 2: Image retrieval performance with activations from different layers.

3. EXPERIMENTAL RESULTS

To evaluate the effectiveness of our approach, we perform retrieval experiments on two benchmark datasets, the Holidays dataset [10] and the UKBench dataset [16]. The Holidays datset contains 1,491 holiday images, of which 500 images are used as queries. Mean Average Precision (mAP) is used to evaluate the retrieval accuracy. On the UKBench dataset, there are $10,200$ images from 2,550 object categories, with each containing 4 images. In this dataset, NS-score (average top 4 accuracy) is used to measure the retrieval accuracy.

In the following, we first compare the performance of activations from separate layers when applying the proposed local CNN based Bag-of-Words (BoW) framework with and without Hamming Embedding (HE). Then experiments on multiple layers are discussed to demonstrate the effectiveness of the Multi-layer Orderless Fusion (MOF) scheme. Finally, we compare our method with other CNN based works. In all experiments we use the pre-trained CNN-M-128 model [6].

3.1 Single Layer Evaluation

In our experiments, we choose activations from three layers in total, namely convolutional layers Conv3, Conv5, and fully-connected layer Fc7. For Conv3 and Conv5, we use sum-pooling and max-pooling to aggregate the convolutional layer feature maps, respectively. All activations undergo a square rooting for each dimension and L_2 normalization is performed to generate the final feature representation.

We summarize the image retrieval accuracies on Holidays and UKBench datasets in Figure 2 with respect to different

Table 1: MOF results of different layers

	Layers Comb.	Holidays		UKBench	
		sum	max	sum	max
BoW	Conv3+Conv5	67.23	69.80	2.67	2.68
	Conv3+Fc7	72.60	75.14	2.85	2.75
	Conv5+Fc7	74.05	75.36	2.74	2.93
	Conv3+Conv5+Fc7	75.94	76.75	2.96	3.00
BoW+HE	Conv3+Conv5	79.74	82.34	3.22	3.39
	Conv3+Fc7	81.15	84.71	3.50	3.44
	Conv5+Fc7	83.58	83.85	3.26	3.51
	Conv3+Conv5+Fc7	83.78	**85.82**	3.38	**3.53**

layers, with or without applying Hamming Embedding (denoted as BoW+HE and BoW, respectively). From Figure 2 we observe that: 1) Except for BoW+HE on UKBench dataset, the activation from Fc7 layer generally achieves the best performance among all layers, demonstrating the effectiveness of the highly abstract feature representation in the CNN architecture; 2) For convolutional layer activations, max-pooling generally has a better performance compared to sum-pooling, which is probably because max-pooling is more robust to slight geometric transformations; 3) For activations from all layers, adding Hamming Embedding in our method dramatically improves the retrieval accuracy.

The best performance is achieved by Fc7 layer features on Holidays dataset, where the mAP value reaches up to 81.12%. While on UKBench dataset, max-pooled Conv5 layer features generate the highest NS-score as 3.41.

3.2 Multi-layer Orderless Fusion Results

While the proposed CNN feature representations of separate layers achieve satisfatory results on two benchmark datasets, we further perform Multi-layer Orderless Fusion (MOF) as described in Section 2.3 to boost the discriminative power of our method.

Since in the CNN architecture, low layers can extract some low-level visual patterns such as textures and shapes, while high layers usually extract high-level semantics, the fusion of different layers is expected to generate better results than the separate counterparts. The experimental results are in Table 1. We can see that the fusion of all three layers achieves the best performance for both BoW and BoW+HE, while the fusion of one convolutional layer and one fully-connected layer generally perform better than the fusion of two convolutional features. Applying Hamming Embedding greatly improves the BoW representation, and max-pooled features outperform the sum-pooled counterparts.

The best mAP result on Holidays dataset is 85.82% by using max-pooling and fusing three layers with Hamming Embedding. On UKBench dataset, NS-score 3.53 is achieved with the same experiment settings.

We compare the proposed Multi-layer Orderless Fusion (MOF) method with other state-of-the-art CNN-based algorithms, including Neural Codes [3], Multi-scale Orderless Pooling (MOP) [9] and Patch-CKN [17]. The comparison results are shown in Table 2. On the Holidays dataset, our method gains a remarkably improvement over the compared works for more than 5% in mAP, demonstrating that the Bag-of-Words quantization with Hamming Embedding works well with deep CNN features and outperforms the vector of locally aggregated descriptors (VLAD) [12] based pooling schemes used in previous works.

Table 2: The comparison with CNN-based methods

Methods	Holidays	UKBench
Neural Codes [3]	79.30	3.56
MOP [9]	80.18	-
Patch-CKN [17]	79.30	**3.76**
MOF	**85.82**	3.53

3.3 Comparisons

On the UKBench dataset, the result is still comparable. The possible reason affecting the results may be that the patches densely cropped include too much useless information. Specifically, different from natural scenes in the Holidays dataset, most of the images in UKBench have similar plain background with a single object in the center, so patches from background are of low discriminative power while they are great in amount. However, we may apply object detection or salience detection techniques to avoid such impact in the future work. In contrast, Patch-CKN [17] extracts local descriptors on detected interest areas, which can avoid background to a certain extent. Besides, several feature processing tools like Principal Component Analysis (PCA) and whitening adopted in MOP[9] and CKN-mix[17] boost the performance in 2%. These region proposal and postprocessing methods are to be studied in the future work.

4. CONCLUSIONS

This paper proposes a method called Multi-layer Orderless Fusion (MOF) for image retrieval. MOF aggregates hierarchical activations from different layers of the Convolutional Neural Network (CNN), with the activations from each layer incorporated into the Bag-of-Words (BoW) architecture separately. With the convenience that all activations from different layers can be extracted simultaneously by passing image patches through the deep network for only once, the proposed MOF approach is more efficient than previous multi-scale and multi-category feature fusion methods. Our experimental results on two benchmark datasets suggest that the CNN activations from different layers are complementary with each other. The fusion of them under the basic BoW framework can already achieve competitive performances against other state-of-the-art algorithms. It is worth mentioning that the proposed method can be easily extended by combining other improvements (*e.g.*, spatial information and postprocessing algorithms), which would be considered in the future.

5. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (NSFC) (Grant No. 61502073 and 61429201), the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), the Open Projects Program of National Laboratory of Pattern Recognition (No. 201407349), and the China Scholarship Council. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. We thank Song Bai who provided insight and expertise that greatly assisted the research. We also thank Shaoyan Sun for assistance with the experiments, and Xiaopeng Zhang for comments that greatly improved the manuscript.

6. REFERENCES

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

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

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

[4] S. Bai and X. Bai. Sparse contextual activation for efficient visual re-ranking. *TIP*, 25(3):1056–1069, 2016.

[5] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki. Gift: A real-time and scalable 3d shape search engine. In *CVPR*, 2016.

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

[7] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, pages 1–8, 2007.

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

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

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

[11] H. Jégou, M. Douze, and C. Schmid. Improving bag-of-features for large scale image search. *IJCV*, 87(3):316–336, 2010.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010.

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

[14] L. Liu, C. Shen, and A. van den Hengel. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In *CVPR*, pages 4749–4757, 2015.

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

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

[17] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronin, and C. Schmid. Local convolutional features with unsupervised training for image retrieval. In *ICCV*, pages 91–99, 2015.

[18] 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, 2007.

[19] 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, 2008.

[20] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[21] J. Yue-Hei Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. In *CVPRW*, pages 53–61, 2015.

[22] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas. Query specific rank fusion for image retrieval. *TPAMI*, 37(4):803–815, 2015.

[23] L. Zheng, S. Wang, Z. Liu, and Q. Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR*, pages 1939–1946, 2014.

[24] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian. Query-adaptive late fusion for image search and person re-identification. In *CVPR*, pages 1741–1750, 2015.

[25] L. Zheng, S. Wang, and Q. Tian. Coupled binary embedding for large-scale image retrieval. *IEEE Transactions on Image Processing*, 23(8):3368–3380, 2014.

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