Non-negative Sparse Coding Using Independent Multi-Codebooks for Near-Duplicate Image Detection

Shan Zhou[1], Jun Li[2], Junliang Xing[3], Weiming Hu[3], and Jinfeng Yang[1]

[1] College of Aviation Automation, Civil Aviation University of China, Tianjin, China
ss_zhou@yeah.net, jfyang@cauc.edu.cn
[2] School of Automation, Southeast University, Nanjing, China
lijun_automation@seu.edu.cn
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jlxing,wmhu}@nlpr.ia.ac.cn

Abstract. In this paper, we propose an efficient approach for detecting near-duplicate images and make three contributions as follows. First, for each sub-region of spatial pyramid, we learn one distinct codebook such that independent multi-codebooks (IMC) are produced. IMC is more accurate than traditional codebook because it considers the spatial information of visual words to a certain extent. Second, we adopt non-negative sparse coding (NSC) technique to encode features. This encoding scheme can effectively encourage similar features to share similar sparse representations. Third, we design an improved intersection kernel (IIK) to compute image similarity. We validate our approach on two datasets respectively, namely our 6K dataset where images are collected from three web image search engines and publicly available University of Kentucky dataset. The experimental results demonstrate our technique achieves significant performance gain compared with state-of-the-art approaches.

Keywords: Near-duplicate image detection, multi-codebooks, non-negative sparse coding, improved intersection kernel.

1 Introduction

With the rapid growth of network and multimedia technology, many near-duplicates or variants of the same image is widespread on the web. Given a query image, the results retrieved from current image search engines, such as Google, Baidu, Bing, often contains duplicate versions. It is necessary to identify these near-duplicate images and remove them for some applications, e.g. image retrieval and copyright detection. We address this problem in this paper, referred to as Near-Duplicate Image Detection (NDID), to find all the near-duplicates of an image on the web.

Zhang and Chang [1] define three categories of near-duplicate images, i.e. scene, camera and image. In this paper, we mainly concentrate on the near-duplicate images of type *image*, where we assume that the original image and its near-copy version share the same digital source. We also investigate the detection of near-duplicate images containing minor camera-type transformation. Fig. 1 shows some representtative examples of near-duplicate images.

C. Sun et al. (Eds.): IScIDE 2013, LNCS 8261, pp. 152–159, 2013.

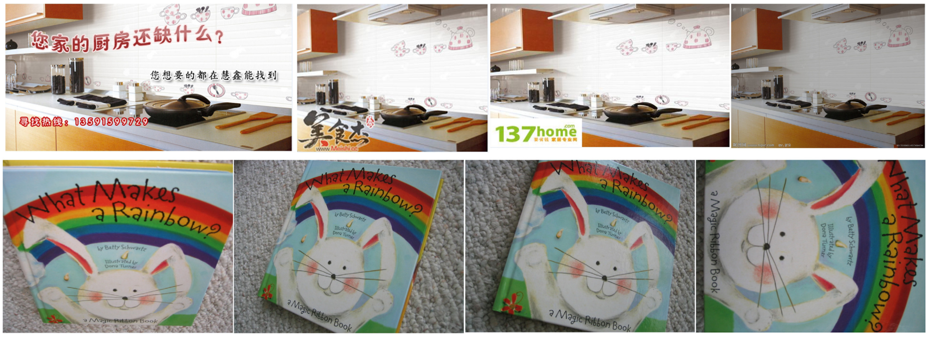

Fig. 1. Examples of duplicate images from web (the top row) and from University of Kentucky dataset [13] (the bottom row)

The remainder of the paper is organized as follows. Section 2 reviews the related work regarding NDID. Section 3 describes our framework and gives specific details of each step. Section 4 presents the experiments implemented on the two datasets and the analysis of the experimental results. Finally, in section 5, we draw a conclusion.

2 Related Work

Recently, some researchers applied the framework of image classification and retrieval to NDID for its close connection to them [2, 3]. To address the problem, bag-of-words (BoW) [4] model is considered to be an efficient method and extensively applied to NDID. The BoW-based way produces the global histogram representation of an image by encoding its local features and calculates the similarity between two global representations. It generally includes the following steps:

(1) Extraction of local features from images, e.g. SIFT and PHOG.
(2) Codebook training via unsupervised learning, e.g. K-means and GMM.
(3) Encoding and pooling of local features for generating image representation.
(4) Computation of similarity metric between images or performing classification.

However, BoW ignores the spatial information of features. According to this limitation, Lazebnik *et al* [5] proposed spatial pyramid matching (SPM) that partitioned image into increasingly finer spatial sub-regions and computed histograms of local features for each sub-region. With sparse modeling successfully applied to image and video denoising, segmentation, super-resolution, Yang *et al* [6] developed an extension of the SPM by generalizing vector quantization to sparse coding followed by multi-scale spatial max pooling. Inspired by this, Wang *et al* [7] proposed a locality-constrained linear coding (LLC) scheme that utilized locality to constrain the sparse coding process which is more computationally efficient. The two methods achieved state-of-the-art image classification performances on several benchmarks. However, both of the two schemes have the following limitations:

(1) All sub-regions of spatial pyramid share a unique codebook, which does not take into account the spatial variability of visual words. It can not describe the local details of an image with only one codebook utilized for every sub-region of spatial pyramid.

(2) The encoding scheme does not have enough constraints to coefficient. Negative coefficients are required to satisfy the sparse coding constraints and expected reconstruction error. While adopting max-pooling on spatial pyramid constructed for an image, the information loss for the negative coefficients is inevitable.

About similarity metric for comparing distributions of features, Euclidean metric is a conventional one. However, Maji *et al* [11] have shown that Euclidean distance is not the most effective way for comparing two histograms. They built a nonlinear intersection kernel (IK) based on histogram representation and reported superior results for image classification. Thus, IK attracts much attention for low computational complexity. Nevertheless, IK just considers the minimum value of pair-wise.

According to the above analysis, we make some improvement on codebook learning and encoding constraints, and design an enhanced intersection kernel function as image similarity metric. The novel framework is applied to NDID and achieves promising performances.

3 Overview of Our Framework

In this paper, we propose a NDID system where the BoW model is combined with spatial pyramid. In training phase, we learn one codebook offline corresponding to each sub-region such that independent multi-codebooks (IMC) are produced. Then we develop a spatial pyramid image representation based on non-negative sparse coding (NSC) of low-level descriptors. Next we design an improved intersection kernel function (IIK) as the similarity metric. We finally conduct comparative experiments on our 6K dataset and University of Kentucky dataset for each phase. Fig. 2 shows the flowchart of our method. The details of each part will be elaborated as follows.

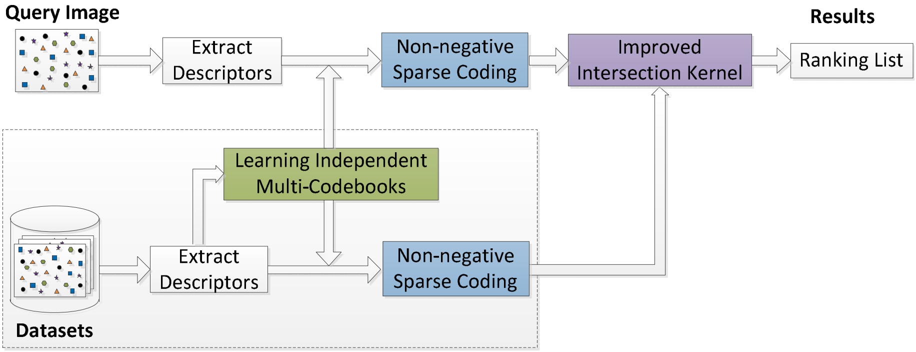
Query Image
Extract Descriptors
Non-negative Sparse Coding
Improved Intersection Kernel
Results
Ranking List
Learning Independent Multi-Codebooks
Extract Descriptors
Non-negative Sparse Coding
Datasets

Fig. 2. The flowchart of our proposed method

3.1 Learning Independent Multi-Codebooks

Spatial pyramid partitions an image into increasingly finer spatial sub-regions and computes the corresponding BoW histogram built on a pre-trained codebook for each sub-region [8]. To our knowledge, all the state-of-the-art methods to date just train a unique codebook for the spatial pyramid, which discards the local spatial variability of the codebook. Thus this unique codebook simply represents a coarse distribution of local features and not fully fitted to the descriptors from different sub-regions of spatial pyramid. So we introduce the idea of leaning independent multi-codebooks (IMC). For each sub-region per level of spatial pyramid, we learn a distinct codebook using corresponding local features. Once we have obtained a set of codebooks, we are able to encode the local features employing each independent codebook. Fig. 3 illustrates the comparison between one codebook and IMC. It is observed that IMC considers more spatial information and is more robust than one codebook.

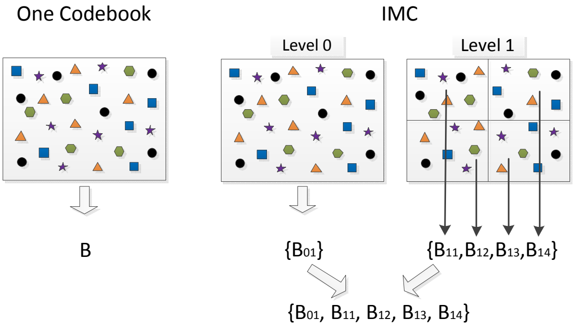

Fig. 3. Comparison between one codebook and IMC

In our experiment, we utilize the iterative online dictionary learning [9] to learn and optimize each independent codebook respectively, which construct multi-codebooks. It processes one sample at a time and sequentially minimizes a quadratic local surrogate of the expected cost. This codebook learning method can decrease the time cost in the training phase in an offline fashion to a great extent.

3.2 Non-negative Sparse Coding

Let $X = [x_1, \cdots x_K] \in R^{D\times K}$ be a set of local features in D dimensional space. Given a codebook $B = [b_1, b_2 \cdots b_M] \in R^{D\times M}$ containing M visual words of dimensionality D, the encoding technique using vector quantization (VQ) can be expressed as follows:

$$\arg\min \sum_{i=1}^{K} \left\| x_i - B\alpha_i \right\|^2 \quad s.t. \ \left\|\alpha_i\right\|_0 = 1, \left\|\alpha_i\right\|_1 = 1, \alpha_i > 0 . \tag{1}$$

From (1), we can observe VQ results in large quantization error, since each local descriptor is only assigned to its nearest neighbor of the codebook. Yang *et al* [6] relaxed the constraint on α_i, which enforced α_i to have a small number of nonzero elements. It is also referred to as sparse coding (SC):

$$\arg\min \sum_{i=1}^{K} \left\| x_i - B\alpha_i \right\|^2 + \lambda \left\|\alpha_i\right\|_1 , \tag{2}$$

where λ is a regularization parameter constraining the sparsity of α_i. SC achieves good performances on several benchmarks when only SIFT descriptors are used.

Nevertheless, SC has one significant disadvantage which consists in the influence of max-pooling on the negative encoding coefficients, namely the max-pooling of local descriptors may lead to the removal of any negative coefficients with at least zero terms preserved. Hence, this strategy will cause significant information loss and the resulting image representation is not discriminative enough.

In this work, we propose an alternative non-negative sparse coding (NSC) which can be formulated as the following optimization problem:

$$\arg\min \sum_{i=1}^{K} \|x_i - B\alpha_i\|^2 + \lambda \|\alpha_i\|_1 \qquad s.t.\ \forall i \quad \alpha_i \geq 0 . \tag{3}$$

It is can be observed that there are non-negative constraints on the coefficients. Therefore, NSC will overcome the aforementioned problem to a certain extent. Simultaneously, it can effectively encourage similar descriptors to share similar sparse representations. The problem described by (3) is generally termed the least absolute shrinkage and selection operator (LASSO). We adopt the homotopy-LARS algorithm [10] which has been demonstrated to be very efficient for solving (3).

3.3 Improved Intersection Kernel for Measuring Image Similarity

Let $H_1 = [H_{11}, H_{12}, \cdots H_{1W}] \in R_+^W$ and $H_2 = [H_{21}, H_{22}, \cdots H_{2W}] \in R_+^W$ be two histogram representations. The intersection kernel of two histograms for similarity measure is defined as follows:

$$Sim(H_1, H_2) = \sum_{i=1}^{W} \min(H_{1i}, H_{2i}) . \tag{4}$$

This expression only takes into account the minimum value of pair-wise components without considering maximum one which is also discriminative characteristics. Motivated by this, we propose an improved intersection kernel (IIK) function described as follows:

$$Sim(H_1, H_2) = \sum_{i=1}^{W} \left(\max(H_{1i}, H_{2i}) - \min(H_{1i}, H_{2i}) \right) . \tag{5}$$

If the two histograms are similar, the difference between the maximum and the minimum values of their pair-wise components is relatively small. Compared with its prototype, IIK can capture more salient differences between two histogram sequences and quantify the histogram distances more accurately. Moreover, it is robust to the noise and variance, since it is a finer expression of the distance of two histograms. Compared with Euclidean distance and Chi-square, IIK has lower computational complexity and the optimal real-time performance.

4 Experimental Implementation and Results

In order to validate the performance of our proposed NDID method, we conduct experiments on two datasets: manually collected 6K web-image dataset, and University of Kentucky dataset [13]. We also carry out the comparative study on the two datasets for other methods. In preprocessing stage, all images are transformed into gray-scale versions. We just use a single descriptor SURF [12], which approximates or even outperforms SIFT descriptor in terms of computational efficiency and robustness. Moreover, the dimensionality of a SURF descriptor is 64, which is merely half of a SIFT.

4.1 Experimental Settings

Since the setup of spatial pyramid will significantly influence the dimensionality of the final concatenated vector. In order to compromise between computational efficiency and precision, we set the level $L=1$, and choose $2^l \times 2^l$ sub-regions, i.e. the number of sub-regions overall is $N=\sum_{l=0}^{L} 4^l = \frac{1}{3}\left(4^{L+1}-1\right)=5$. Therefore, the set of multi-codebooks contain five independent codebooks learned from training local descriptors of each sub-region. We also make use of max-pooling strategy for producing the final global representation as in [7]. The pooled features from each sub-region are concatenated and normalized as the final image histogram representation. Here, we use l_2 normalization method.

4.2 6K Dataset

The 6K web-image dataset is collected from three web image search engines (Google, Baidu, and Sogou). This dataset consists of 8 class objects, i.e. animal, landmark, logo, man, musical instrument, plant, scene and transport respectively. For each class, we manually choose 25 different seed images. Therefore, there are 200 seed images overall. For every seed image, we artificially select 30 near-duplicate images from the results retrieved from the three search engines. If the search engines return less than thirty results, we transform the seed image by using ImageMagick [14], to make up to thirty near-duplicates.

Taking into account the storage and computing cost, all images are resized to be not more than 500×500 pixels. We choose 5 near-duplicate images for each seed image as training images. The low-level SURF descriptors of the training images extracted from each separate sub-region are used to train multi-codebooks. The size of each independent codebook is empirically set to 400. The precision is measured in terms of the number of relevant images in the retrieved top 30 images. We take the average precision for each class and all the query images as our evaluation criterion.

We perform three sets of comparative experiments in terms of learning different codebook, encoding methods and distance metrics on this dataset. First, we compare the effects of using IMC for our method, and the results are shown in italic in the last column of Table 1, which demonstrates the performance gain by using IMC. Next, we validate three state-of-the-art encoding methods: the baseline hard VQ, locality-constrained linear coding (LLC) [7] and our NSC. As shown in bold in Table 1, NSC outperforms others for all but one class. Last, to verify the effectiveness of IIK for our approach, we compare other three similarity metrics, i.e. Euclidean distance, Chi-square distance, intersection kernel (IK), and the results are shown in italic in Table2. It is explicitly observed that our IIK achieves the optimal performance.

Table 1. The precision (%) between different encoding schemes for our method, the last column in italic shows the result using IMC on NSC

		VQ	LLC	NSC	IMC+NSC
object class	animal	93.20	95.60	**96.13**	***96.93***
	landmark	89.20	93.60	**94.93**	***95.47***
	logo	94.13	98.27	**98.40**	***97.73***
	man	84.00	**95.07**	94.80	***95.20***
	musical	88.00	94.53	**94.80**	***95.73***
	plant	92.80	95.47	**96.66**	***97.33***
	scene	89.20	95.60	**96.80**	***97.60***
	transport	88.80	96.93	**97.60**	***98.00***
Average Precision (%)		89.92	95.63	**96.26**	***96.75***

Table 2. The precision (%) between different similarity metrics for our method

	Euclidean	Chi-square	IK	IIK
Precision (%)	93.82	94.82	83.50	***96.75***

4.3 University of Kentucky Dataset

University of Kentucky (UK) dataset includes 10, 200 images of 2550 objects where each object has 4 relevant images with different camera viewpoints or angles. The size of all the images in this database is 640×480 pixels. We resize all images to 480×360 pixels. We uniformly sample feature descriptors from all images to train multi-codebook. Every independent codebook of multi-codebooks is trained with 1000 bases. Our evaluation criterion is to calculate an average over the number of true positives of returned top four images when using a query image randomly chosen from that set of four images.

We also conduct three sets of comparative experiments on this public dataset, and the corresponding results are shown in Table 3, Table 4. It can be observed that the experimental results may be outperformed by the ones reported in some related literature, since the four relevant images of one object has different camera viewpoint whereas spatial pyramid technique is not robust enough to the variance of camera viewpoint. In spite of this, our proposed method outperforms other approaches overall.

Table 3. The average top between different encoding schemes for our method, the last column in italic shows the result using IMC on NSC

	VQ	LLC	NSC	IMC+NSC
Average Top	2.7055	2.7702	**2.9733**	***3.0012***

Table 4. The precision (%) between different similarity metrics for our method

	Euclidean	Chi-square	IK	IIK
Average Top	2.6980	2.9824	1.3212	***3.0012***

5 Conclusion

This paper presents an improved framework for near-duplicate images detection by combining the BoW model with spatial pyramid. This framework mainly consists of three elements, namely independent multi-codebooks, non-negative sparse coding and improved intersection kernel function (IMC+NSC+IIK). We learn IMC, which consider more spatial information of visual word to an extent. In addition, we adopt NSC to encode the low-level descriptors with lower information loss. We also design an IIK function to measure the similarity between two images. Experimental results on two datasets validate the proposed three improved parts effectively enhances the detection performance of near-duplicate images.

Acknowledgement. This work is partly supported by NSFC (Grant No. 60935002 and 61273023), the Ph.D. programs Foundation of Ministry of Education of China (Grant No. 20120092110024), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

1. Zhang, D., Chang, S.F.: Detecting Image Near-duplicate by Stochastic Attributed Relational Graph Matching with Learning. In: ACM Multimedia Conference, pp. 877–884 (2004)
2. Dong, W., Wang, Z., Charikar, M., Li, K.: High-Confidence Near-Duplicate Image Detection. In: ACM International Conference on Multimedia Retrieval (2012)
3. Meng, Y., Chang, E.Y., Li, B.: Enhancing DPF for near-Replica Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. II-416-23 (2003)
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In. 8th European Conference on Conference Vision, pp. 1–22 (2004)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
6. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear Spatial Pyramid Matching using Sparse Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)
7. Wang, J., Yang, J., Fu, K., Lv, F., Huang, T., Gong, Y.: Locality-Constrained Linear Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3360–3367 (2010)
8. Gauman, K., Darrell, T.: The Pyramid Match Kernels: Discriminative Classification with Sets of Image Features. In: Conf. Comput. Vision Pattern Recognit., pp. 1458–1465 (2005)
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. J. Mach. Learn. Res. 19–60 (2010)
10. SPArse Modeling Software, http://spams-devel.gforge.inria.fr/index.html
11. Maji, S., Berg, A.C., Malik, J.: Classification Using Intersection Kernel Support Vector Machines is Efficient. In: IEEE Conf. Comput. Vision Pattern Recognit., pp. 1–8 (2008)
12. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
13. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
14. ImageMagick, http://www.imagemagick.org/