Improving bag-of-words representation with efficient twin feature integration

Conference Paper · August 2015

DOI: 10.1145/2808492.2808543

Hanli Wang
Tongji University

Improving Bag-of-Words Representation with Efficient Twin Feature Integration

Lei Wang[1,2]
[1]Department of Computer Science and Technology, Tongji University, Shanghai, China
[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China
110_wangleixx@tongji.edu.cn

Hanli Wang[1,2]*
[1]Department of Computer Science and Technology, Tongji University, Shanghai, China
[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai, China
hanliwang@tongji.edu.cn

ABSTRACT

In recent years, the Bag-of-Words (BoW) model has been widely used in most state-of-the-art large-scale image retrieval systems. However, the standard BoW based systems suffer from low discriminative power of local features as well as quantization errors that significantly affect the retrieval performance. In this paper, twin feature is employed and well combined with two advanced techniques including Hamming Embedding (HE) and Multiple Assignment (MA) to construct a discriminative image retrieval system on BoW representation in an efficient way. Experimental results on two benchmark datasets Oxford5k and Paris6k demonstrate that the proposed technique can greatly refine the visual matching process and enhance the final performance for image retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; I.4.7 [**Feature Measurement**]: Feature representation

Keywords

Image retrieval, twin feature, Bag-of-Words, hamming embedding, multiple assignment.

1. INTRODUCTION

As we all know, image retrieval is a challenging task in computer vision area. Given a query image, its job is to return a series of relevant images from a huge database. Recently, a number of image retrieval systems are built upon the effective and efficient Bag-of-Words (BoW) [14] representation with local invariant features. In BoW, local features are extracted and encoded into compact representations by a visual vocabulary trained offline. In addition, the integration with inverted file structure makes it efficient for online retrieval in large-scale databases.

The BoW representation can be considered as an approximation method of visual feature matching between images to address the problems of high computational complexity. Two features mapped to the same visual word are deemed as a matching pair. Recently, tons of works have been proposed for extracting more discriminative features, optimizing quantization process, further investigating the attributes of features or designing better indexing strategies.

In the literature, local feature detectors such as Difference-of-Gaussian (DoG) [6] and Hessian-Affine [7] are proposed to find local image patches containing abundant structural information. These patches are often described by local invariant feature descriptors to obtain robust and powerful discrimination. Features are quantized with a pre-trained visual dictionary via clustering methods like Hierarchical K-Means (HKM) [8], Approximate K-Means (AKM) [11] as well as flat K-Means. Most of these approaches employ the inverted file structure which greatly reduces memory requirements and makes it extendable for large-scale image retrieval tasks.

However, feature locality and quantization errors significantly affect the retrieval accuracy. Local patches of different semantics may be similar and would be assigned to the same visual word. On the other hand, two features that are similar to each other may be assigned to different visual words, which will dramatically degrade visual matching accuracy. To address the problem of quantization error, Multiple Assignment (MA) [5] and soft quantization [12] have been developed. Moreover, Hamming Embedding (HE) [3] is proposed to calculate the hamming distance between two binary signatures, which improves the discriminative power of features. On the other hand, to further handle the locality of features, twin features are proposed in [15], which tend to refine the feature matching process by utilizing extra in-

[*] Corresponding author (H. Wang).

This work was supported in part by the National Natural Science Foundation of China under Grant 61472281, the "Shu Guang" project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005), and the Fundamental Research Funds for the Central Universities under Grant 0800219270.

ICIMCS'15, August 19-21, 2015, Zhangjiajie, Hunan, China
© 2015 ACM. ISBN 978-1-4503-3528-7/15/08…$15.00
DOI: http://dx.doi.org/10.1145/2808492.2808543

formation of neighboring image patch and thus improve the final retrieval performance.

Moreover, a number of works are focused on geometric consistency between features or images to filter out false matches. The RANSAC [2] based spatial re-ranking [11] achieves high precision while it is time consuming. The orderless bag-of-features are transformed to spatial-bag-of-features in [1], which takes the spatial relation of features into consideration. Analogously, spatial relationship between features are employed in [16] to further enhance the matching accuracy.

In this paper, we make a further study on the technique of twin feature. It is studied and combined with two prior techniques including HE [3] and MA [5], and their implementation details on the standard BoW model are explored and presented. Experiments on Oxford5k [10] and Paris6k [13] verify that twin features are completely complementary with HE and MA, and a more efficient and discriminative representation can be achieved by integrating these methods as compared with the original BoW. The rest of this paper is organized as follows. Section 2 gives a brief introduction to twin feature. Section 3 describes how to implement twin feature on the standard BoW model, as well as HE and MA. In Section 4, experimental results on two benchmark datasets are presented and analyzed. Finally, Section 5 concludes this paper and presents future research prospects.

2. TWIN FEATURE

As discussed in [15], twin feature is discriminative and contains extra information of local image patches which is able to enhance visual matching accuracy. For simplicity, *twin* refers to the twin feature in the rest of the paper. Twin is not detected from the image but constructed following the original feature detected by traditional methods like Hessian-Affine detector. The center of twin is generated according to the attributes of the original feature, including coordinate, scale and orientation.

Let (x, y) be the center of an original feature, with its dominant orientation θ and scale s. Then, the coordinate of twin can be calculated as

$$\begin{cases} x_{twin} = x + \alpha \cdot s \cdot \cos\theta \\ y_{twin} = y + \alpha \cdot s \cdot \sin\theta \end{cases}, \qquad (1)$$

where α is a control parameter which determines the distance between the original feature and the twin. In this work, α is set to 2.0, which is a little different from that of [15], since the original feature used here is also different. Moreover, the dominant orientation and scale of twin are defined as

$$\theta_{twin} = \theta, \; s_{twin} = s, \qquad (2)$$

which are the same as the original feature. After that, SIFT descriptor is employed for twin description to generate a 128-dimension feature vector.

3. IMPLEMENTATION DETAILS

In the following, we describe how to implement twin on standard BoW model with inverted file structure, as well as MA and HE.

3.1 BoW with Twin

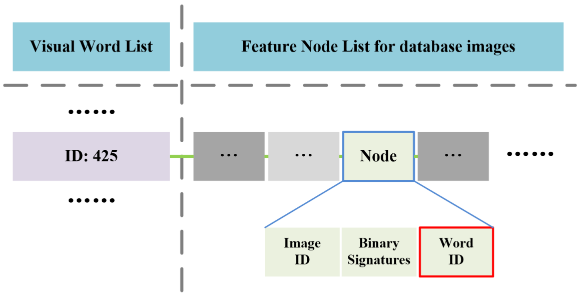

Figure 1: Illustration of inverted file structure of the proposed image retrieval framework. Binary signatures are used by HE and Word ID is used by twin.

To reduce memory cost and speed up retrieval, the inverted file structure is leveraged from text retrieval in standard BoW framework. Each visual word in the visual dictionary is followed with multiple image feature nodes quantized to it. Within each node, the image ID and some other attributes such as binary signatures used by HE are stored.

In order to retain the favorable structure of inverted files, twins are inclined to be saved as an attribute of feature nodes. As presented in [15], twin is described via SIFT-like descriptors and can be directly quantized to visual words. So an N-bit domain is added in each node to indicate the index of visual word to which twin belongs. The length of N depends on the maximum size of visual dictionary. Generally, a total of 24 bits is enough for extremely large visual dictionary. Figure 1 illustrates the inverted file structure of the proposed image retrieval framework, where each feature node contains three domains including image ID, binary signatures and Word ID.

It should be noted that some twins will be located close to or even out of the image border. On this occasion, it is unfair to compare these twins since the image patches do not exist or are not complete. In our settings, the Word ID domains of this kind of edge features are uniformly set to -1.

For offline training, features are firstly extracted, and then the corresponding twins are computed. In fact, it is computationally efficient to compute twins thanks to the simpleness of the proposed twin localization algorithm. Then, all the original features and their twins are trained into the inverted file structure. This procedure will cost twice as much time as before since the number of features (including twins) is doubled. The process of online retrieval is analogous. Specifically, two feature nodes owning the same Word ID, except -1, tend to be a good match, and should be given a higher weight ω_1. Otherwise, a lower weight ω_2 is attached, as compared to the standard BoW model which assigns a constant weight of 1 to all potential matched feature nodes. In this work, ω_1 is set to 2.0 and ω_2 is set to 0.2. The online retrieval efficiency with twin is comparable with the standard BoW framework without twin, since only several simple operations are added when using the twin technique.

3.2 MA with Twin

MA [5] is proposed to reduce quantization errors. Each feature is assigned to k nearest visual words. Following the same strategy proposed in [4], we only perform MA for query so that the inverted file structure in the search database is unchanged. Moreover, the nearest distance d_0 between a feature and a visual word is employed to remove worse assignments whose distances are above $\beta \cdot d_0$. In this work, k and β are set to 10 and 1.05, respectively.

Naturally, MA can be imitated on twins since twins also suffer from the same problem of quantization errors as the original features. The simplest way to perform MA on twin is to record the Word IDs of k_t nearest visual words which increases the probability of owning the same Word ID for two feature nodes. Similarly, the same criterion is applied to filter out bad assignments and k_t is also set to 10. Moreover, MA can be further extended both on features and twins to improve the final retrieval performance. The experimental results presented in Section 4.1 show that MA and twin are two complementary techniques that mutually reinforce.

3.3 HE with Twin

HE [3] is an effective technique based on hamming distance to improve the retrieval accuracy. As shown in Fig. 1, each node in the inverted file structure has a domain containing binary signatures, which are used to calculate the hamming distance. In this work, we mainly follow the approach in [4] to obtain weighted HE scores.

To embed twin on HE, both HE scores for original features H_f and HE scores for twins H_t should be taken into account. A naive strategy is to conduct a linear operation on H_f and H_t to obtain a more strong score. Considering the following two facts: 1) HE is a discriminative technique that can be trusted, and 2) twin is not as powerful as the original feature since it is constructed based on the original feature, we calculate the final HE score $Score_{ft}$ as

$$Score_{ft} = H_f + \lambda \cdot H_t, \tag{3}$$

$$\lambda = \mu \cdot \frac{H_f}{H_{max}} \cdot \frac{H_t}{H_{max}}, \tag{4}$$

where H_{max} represents the maximum HE score of two identical features and μ is a control parameter. As indicated in Eq. (3), λ can be regarded as a scale factor controlling the impact of H_t. Typically, H_{max} is 100,000 and μ equals 25 in this work. In addition, for those features who have no twins, H_t is assigned the same value as H_f.

4. EXPERIMENTS AND ANALYSIS

In this section, experiments on two benchmark datasets including Oxford5k [10] and Paris6k [13] are presented. The mean Average Precision (mAP) is used as the evaluation criterion. In particular, the Hessian-Affine detector [7] and SIFT descriptor [6] are applied for feature extraction and description. Different sizes of visual dictionaries are learned based on an independent Flickr100k dataset [9], with flat K-Means clustering performed.

4.1 MA-Twin Experiment

To further test the correctness and compatibility of twin, we combine the technique of MA with twin. Three kinds of MA-Twin strategies are proposed and tested, including: 1) MA on features only (SMA_f): only the original features are assigned to multiple visual words and each feature owns one Word ID; 2) MA on twins only (SMA_t): features are assigned only once but twins are assigned to multiple visual words and each feature owns more than one Word IDs; 3) MA both on features and twins (SMA_{ft}): both the original features and twins are assigned to multiple visual words and each feature owns more than one Word IDs.

Table 1: mAP of MA-Twin on Oxford5k and Paris6k with a 20k dictionary.

Methods	Baseline	MA	SMA$_f$	SMA$_t$	SMA$_{ft}$
Oxford5k	0.402	0.364	0.438	0.463	**0.474**
Paris6k	0.472	0.434	0.518	**0.556**	0.551

Table 1 presents the mAP results using both twin and MA. For performance comparison, the results of baseline BoW and standard MA without twin are also listed in the table. We obtain a consistent conclusion with [4] that MA sometimes does not improve or even deteriorates the performance of the standard BoW method due to the low discriminative power of feature nodes. More importantly, the proposed twin technique highly enhances the accuracy of visual matching so that the combination of twin with MA raises the retrieval performance to a certain degree. Specifically, SMA$_{ft}$ achieves most stable mAP with + 11.0% and + 11.7% higher than MA on the two datasets, respectively, which confirms the effectiveness of the idea of twin.

4.2 HE-Twin Experiment

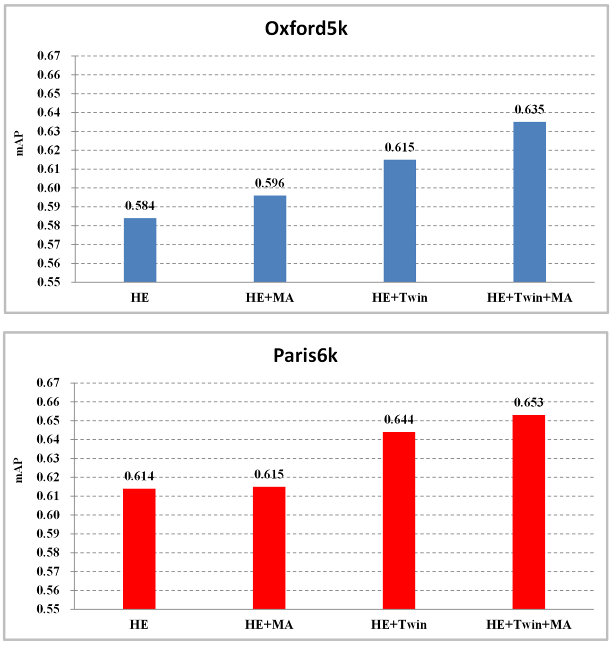

Figure 2: mAP of combination of Twin, HE and MA on Oxford5k and Paris6k with a 20k dictionary. Labels on x-axis represent the methods applied.

Besides MA, twin is also appended to the technique of HE. Figure 2 shows the experimental results with twin, HE and MA. Considering that SMA$_t$ and SMA$_{ft}$ both require much memory costs which will affect the retrieval efficiency, we only perform SMA$_f$ when twin, HE and MA are combined together. First, MA does not raise the mAP remarkably mainly because the filtering criterion we set is so strict that

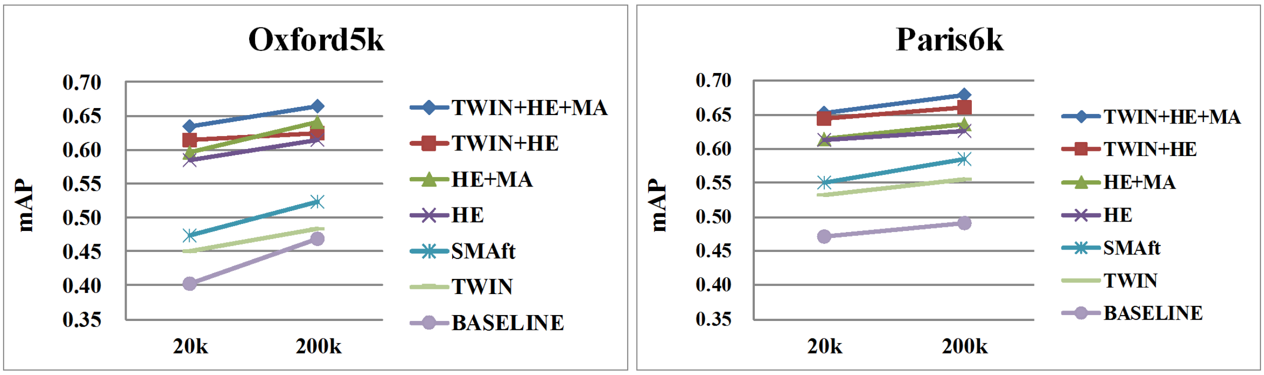

Figure 3: Performance comparison of different methods under various dictionary sizes.

few extra assignments are reserved. Second, by employing twin on HE, a + 3.1% and a + 3.0% improvements are obtained both on Oxford5k and Paris6k datasets. The twin technique complements the visual matching procedure by introducing neighboring image information while HE verifies the similarity from the feature itself. Moreover, twin also benefits from HE and provides a more precise score as shown in Fig. 2. And with the join of MA, higher performances of + 5.1% on Oxford5k and + 3.9% on Paris6k are finally achieved.

More evaluations are conducted with a larger dictionary of 200k. Consistent results can be observed in Fig. 3 and the best results of 0.665 on Oxford5k and 0.679 on Paris6k are obtained when twin is applied together with HE and MA. These two best results are also superior to [4] under almost the same settings but without twin feature.

5. CONCLUSION AND FUTURE WORK

In this paper, an in-depth study of the combination of twin feature and two advanced techniques including HE and MA is given and an efficient image retrieval framework is then realized by integrating these three techniques with the standard BoW representation. The comparative experimental results verify the effectiveness of the proposed twin feature technique on the two benchmark datasets of Oxford5k and Paris6k. In the future, we will work on more compact representation methods for twin to make the representation more discriminative to get better performances.

6. REFERENCES

[1] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *Proc. CVPR'10*, pages 3352–3359, 2010.

[2] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[3] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV'08*, pages 304–317, 2008.

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

[5] H. Jégou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Proc. CVPR'07*, pages 1–8, 2007.

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

[7] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, Oct. 2004.

[8] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR'06*, pages 2161–2168, 2006.

[9] J. Philbin, R. Arandjelovic, and A. Zisserman. Flickr100k image dataset, http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/flickr100k.html.

[10] J. Philbin, R. Arandjelovic, and A. Zisserman. Oxford5k image dataset. http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR'07*, pages 1–8, 2007.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR'08*, pages 1–8, 2008.

[13] J. Philbin and A. Zisserman. Paris6k image dataset. http://www.robots.ox.ac.uk/ vgg/data/parisbuildings/.

[14] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV'03*, pages 1470–1477, Oct. 2003.

[15] L. Wang, H. Wang, and F. Zhu. Twin feature and similarity maximal matching for image retrieval. In *Proc. ICMR'15*, 2015.

[16] P. Xu, L. Zhang, K. Yang, and H. Yao. Nested-SIFT for efficient image matching and retrieval. *MultiMedia*, 430(3):34–46, 2013.