REGION MATCHING AND SIMILARITY ENHANCING FOR IMAGE RETRIEVAL

Guixuan Zhang, Zhi Zeng, Shuwu Zhang, Hu Guan, Qinzhen Guo

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{guixuan.zhang, zhi.zeng, shuwu.zhang, hu.guan, qinzhen.guo}@ia.ac.cn

ABSTRACT

Many image retrieval systems adopt the bag-of-words model and rely on matching of local descriptors. However, these descriptors of keypoints, such as SIFT, may lead to false matches, since they do not consider the contextual information of the keypoints. In this paper, we incorporate the cues of meaningful regions where local descriptors are extracted. We describe a matching region estimation (MRE) method to find appropriate matching regions for local descriptor matching pairs. Then the region matching quality is evaluated and the true matched regions will enhance the similarity of local descriptors. Consequently, the image retrieval accuracy can be improved. Extensive experiments on benchmark datasets show the effectiveness of our method and our result compares favorably with the state-of-the-art.

Index Terms— matching region estimation, similarity enhancing, fisher vector, regional cues, image retrieval

1. INTRODUCTION

This paper considers the task of image and object retrieval. Image retrieval serves as an important basis in various applications such as personal photo search [1], location recognition [2,3], partial-duplicate search [4], 3D reconstruction [5] and product recognition [6].

Most image retrieval systems are based on matching of local descriptors, such as SIFT [7] and its variant [8]. Bag-of-words (BOW) model is adopted to achieve fast descriptor matching [9]. Local descriptors are quantized into visual words and inverted index is used for efficient retrieval. Basically, local descriptors are matched if they are assigned to the same word. Then the similarity between two images can be expressed by aggregating the similarities between the matched local descriptors. However, the coarse quantization often leads to false matches. Many efforts have been made to improve this seminal work. Hamming Embedding (HE) exploits a binary representation of local descriptors for precise matching [10]. Multiple assignment or soft assignment is adopted to alleviate quantization error [10,11]. Post-processing techniques such as re-ranking [12-14] and query expansion [15] also improve the accuracy. In this paper, we focus on improving initial results without post-processing.

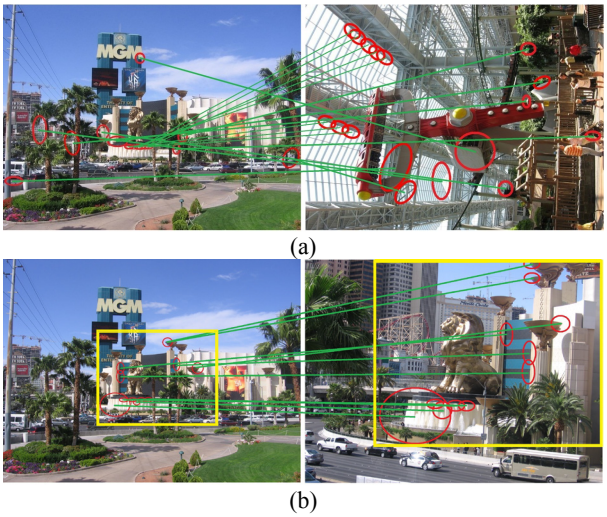

Fig.1. An example of HE based image retrieval. The left is the query image. (a) An irrelevant image with a high matching score; (b) A relevant image, but its score is lower than that of (a). All the matching SIFTs in (b) are extracted from the same objects (yellow rectangles), so they deserve higher score.

SIFT is widely used due to its discriminative power, but it may cause false matches since SIFT only describes the gradient distribution of local patches, ignoring the regional information around the keypoints (Fig.1). Some methods have been proposed to provide cues of regions. Spatial features in the nearby region are extracted and used to measure the spatial consistency of regions [16, 17]. The region size is proportional to the scale of the keypoint, without considering the integrity of a meaningful region. The corresponding region pair may introduce noise due to image occlusions (Fig.2), so the cues of this region pair are insufficient.

In this paper, we aim at finding the appropriate region pair to improve the SIFT matching accuracy, which helps to improve the image retrieval performance. Our contributions are three-fold. Firstly, a matching region estimation (MRE) algorithm is proposed to find appropriate matching regions for SIFT matching pairs. Then each SIFT pair will have a corresponding region pair. Secondly, we adopt Fisher vector (FV) [19] to describe the regions so that the similarity between regions can be measured. Specifically, we employ a binary version of FV for memory efficiency and fast compu-

ting. Finally, we introduce a similarity enhancing function to incorporate the cues of region pairs obtained from MRE. For a SIFT pair, if the corresponding region pair is a true match, e.g., the yellow rectangle pair in Fig.1(b), this SIFT pair is more likely to be a true match. So the region matching quality is evaluated and a true region match will enhance the SIFT-level similarity. After assembling some prior arts of image retrieval, we achieve state-of-the-art results.

The rest of this paper is organized as follows. The MRE algorithm is proposed in Section 2. Binarized Fisher vector is shown in Section 3. Section 4 introduces the enhancing similarity function. We discuss the experimental results in Section 5. Final conclusions are in Section 6.

2. MATCHING REGION ESTIMATION

An image usually contains multiple semantic objects. Several methods have been proposed to find a number of possible regions which may include objects for the task of image classification or object detection [20, 21]. In this paper, we employ these regions as candidates and find the appropriate region pair among them for accurate SIFT matching.

Considering the efficiency, we adopt spatial pyramid method to extracts manually defined regions. Each image I has L layers of regions. Denote the width and height of I as W and H. In the l-th layer, there are $r_l \times r_l$ regions with a fixed size $\frac{W}{s_l} \times \frac{H}{s_l}$, where r_l and s_l are the region density and scale parameters. Generally, more region proposals capture richer information. However, this is time consuming for feature extraction and region similarity measurement. We use L=4 layers, with fixed parameters $(s_1, s_2, s_3, s_4) = (1.0, 1.5, 2.0, 3.0)$ and $(r_1, r_2, r_3, r_4) = (1, 2, 3, 5)$, respectively for a good tradeoff in our experiments. Then we obtain 39 region proposals for every image. Each keypoint is located within several regions. These regions are candidates of this keypoint for the region matching step.

We adopt the HE based image retrieval method proposed in [10] as our image retrieval criterion. $\boldsymbol{x}$ and $\boldsymbol{y}$ are two descriptors (SIFT) from query image Q and database image I. $\boldsymbol{x}$ and $\boldsymbol{y}$ match if they are assigned to the same visual word and the hamming distance between their binary signatures is lower than a threshold h_t. The matching score between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by:

$$score(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} w(h(\boldsymbol{x}, \boldsymbol{y})) \cdot idf^2 & if\ q(\boldsymbol{x}) = q(\boldsymbol{y}), h(\boldsymbol{x}, \boldsymbol{y}) \le h_t \\ 0 & otherwise \end{cases} \quad (1)$$

where $h(.,.)$ is hamming distance function and idf is the inverse document frequency [9]. $w(d) = \exp(-d^2/\alpha^2)$ is a weighting function [22]. We denote as p the regional feature. For a matching pair of $\boldsymbol{x}$ and $\boldsymbol{y}$, their relevant region feature sets are $\mathcal{P}_x = \{p_t^x, t = 1, ..., \mathrm{m}\}$ and $\mathcal{P}_y = \{p_t^y, t = 1, ..., \mathrm{n}\}$, respectively, where m and n indicate the relevant region number. In order to find a region matching pair (a, c) which

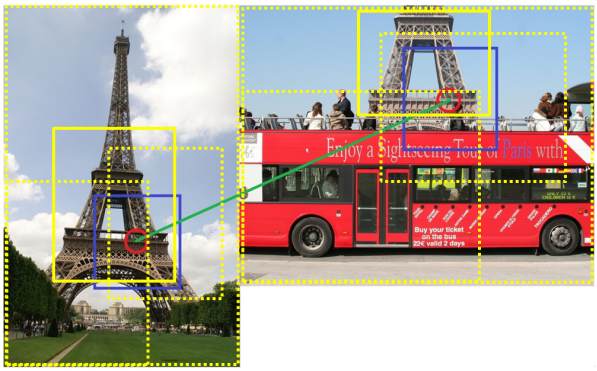

Fig.2. Illustration of the matching region estimation. Red circles are a SIFT matching pair based on hamming embedding. Some methods consider the cues of nearby regions whose size is proportional to scale of keypoints, marked in blue rectangles. They are not similar due to occlusions. In our MRE method, multiple region proposals (yellow rectangles) are extracted and only the region pair with the maximum feature similarity (solid yellow rectangles) is used for the next similarity enhancing step.

are most similar and provide sufficient cues to verify the descriptor matching, we measure the similarity between the feature sets and find the pair with the maximum similarity

$$(a, c) = \arg\max_{i,j} f(p_i^x, p_j^y), p_i^x \in \mathcal{P}_x, p_j^y \in \mathcal{P}_y, \quad (2)$$

where $f(.,.)$ is a similarity function. An example is shown in Fig.2. The SIFT match has several candidate regions marked by yellow rectangles. The two regions in solid yellow rectangles have the maximum feature similarity. These two regions are the appropriate pair for verifying the SIFT match.

3. BINARIZED FISHER VECTOR

In order to measure the similarity between regions, we should adopt a feature algorithm to describe the regions. Since we focus on an on-line retrieval system, a candidate algorithm should be robust and efficient computing. In this paper, we employ Fisher vector due to its good performance of describing global or regional information for both image classification and retrieval [19, 23].

Let $X = \{\boldsymbol{x}_t, t = 1, \ldots, T\}$ be a set of D-dimensional samples whose generation process can be modeled by an independent probability density function u_λ with parameters λ. [23] choose u_λ to be a Gaussian Mixture Model (GMM) with N centroids: $u_\lambda(\boldsymbol{x}) = \sum_{i=1}^{N} \omega_i u_i(\boldsymbol{x})$ and $\lambda = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i, i = 1, \ldots, N\}$ where ω_i, $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$ are respectively the weight, mean vector and variance matrix of Gaussian u_i. Let $\gamma_t(i)$ be the soft assignment of $\boldsymbol{x}_t$ to Gaussian i and $\boldsymbol{g}_i^X$ be the gradient with respect to the mean $\boldsymbol{\mu}_i$ of Gaussian. We obtain $\boldsymbol{g}_i^X$ after standard mathematical derivations:

$$\boldsymbol{g}_i^X = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{\boldsymbol{x}_t - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i}\right). \quad (3)$$

The final Fisher vector $\boldsymbol{g}_{\lambda}^{X}$ is formed by concatenating the $\boldsymbol{g}_{i}^{X}$ vectors for $i=1,\ldots,N$ and is therefore ND-dimensional. Here X corresponds to SIFT descriptors extracted from an image, and the dimension is reduced from 128 to 64 by PCA. Since we have extracted SIFT for the system, no extra extraction process is needed. Fisher vector further undergoes a power normalization and finally is L2-normalized.

Another reason for choosing FV is its efficient computing using an integral image of FVs. Since we partitioned the image into multiple regions, the time of generating features should be considered. FV is an aggregated representation, so we can split the image into many small grids and compute $\boldsymbol{g}_{\lambda}^{X}$ for each grid, then all the Fisher vectors of regions can be computed efficiently through the use of an integral image of unnormalized FVs.

Given the requirement of memory efficiency and low computational cost, we transform the floating-point FV into a binary signature,

$$b_f(v)=\begin{cases}1 & if\ \ r^{\mathsf{T}}v \geq thres\\ 0 & otherwise\end{cases} \qquad (4)$$

where r is a PCA projection matrix of FV and *thres* is the median value. Both the projection matrix and median value are learned on an independent dataset. In our experiment, we use a 128-bit signature. The region similarity can be reflected by the hamming distance of binarized FV. Small distance means high similarity. Then Eq. 2 can be substituted by

$$(a,c)=\arg\min_{i,j} h(b_f(p_i^x),b_f(p_j^y)),\ p_i^x\in\mathcal{P}_x,\ p_j^y\in\mathcal{P}_y\,. \qquad (5)$$

4. SIMILARITY ENHANCING FUNCTION

We have found an appropriate region pair for each SIFT match. The matching quality of this region pair should be evaluated before contributing to the SIFT-level matching score. The similarity between these two regions can be reflected by the hamming distance $d_f=h(b_f(p_a^x),b_f(p_c^y))$, where (a,c) is calculated from Eq.5. For a true-positive region match, the corresponding d_f should be small. In order to study the distribution of d_f, we extracted relevant and irrelevant regions from the Holiday dataset [10] according to the ground truth. Fig.3 depicts the distribution of the hamming distance of the relevant and irrelevant regions. It can be seen that binarized FV hamming distance separates the true matching from the false matching regions quite well.

For a SIFT matching pair based on HE, it is more likely to be a true match if the relative region pair has small hamming distance. So we consider a function that gives the SIFT pair a higher matching score when d_f is small. We have tested different kinds of functions and choose an exponential function to improve the SIFT match accuracy. The similarity enhancing function updates Eq. 1 as follows:

$$score'(\boldsymbol{x},\boldsymbol{y})=score(\boldsymbol{x},\boldsymbol{y})\times(1+\exp(-d_f{}^5\,/\,\theta^5)) \qquad (6)$$

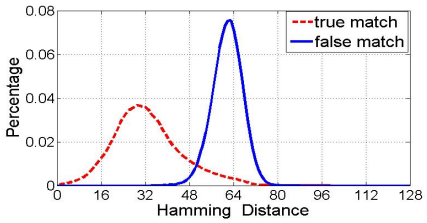

Fig.3. The distribution of hamming distance between regions.

where θ is a parameter which will be discussed in Section 5.

5. EXPERIMENTS

5.1. Datasets and Implementation Details

Datasets: We evaluate our method on the Holidays [10], Oxford5k [12], Paris [11] and Oxford105k [12] dataset. Evaluation measure is the mean Average Precision (mAP). In Oxford and Paris datasets, each query is a rectangular region delimiting the building in the image.

Features: For Holidays, keypoints are detected by Hessian-Affine detector [24]. For Oxford and Paris datasets, we use the modified Hessian-Affine detector which includes the gravity vector assumption [13]. We use SIFT descriptors and apply component-wise square rooting. The rootSIFT has proven to yield superior performance at no cost [8].

Vocabulary: We use the approximate k-means to train our visual vocabularies [25]. For Holidays, the vocabulary is trained on Flickr60k dataset [10]. Vocabulary used for Oxford is trained on Paris, and vice versa. We use a vocabulary of 65k visual words for Oxford and Paris following [26], and 20k for Holidays.

Matching region estimation: In order to mark the relative regions, 38 extra bits are required in the inverted index for each keypoint. If this keypoint is located in one region, the corresponding bit is set to 1. In order to achieve efficient retrieval, all the region similarities between the query image and the dataset image, which is first visited when traversing the inverted lists, are computed and stored in the memory. Then it can be read fast when being visited next time.

Multiple assignment and burstiness: We further combine our proposed method with multiple assignment (MA) which is applied on query side only [10]. In order to deal with the burstiness phenomenon, we also combine our approach with intra-image burstiness normalization (Burst) [22].

5.2. Parameter Analysis

We tune the parameter θ of similarity enhancing function on the Holidays and Oxford5k datasets. The baseline system is HE with 64-bit binary signature. α and h_t in Eq.1 are 16 and 22, respectively. We vary the parameter values and report the performance in Fig.4. $\theta=50$ provides good enhancing weights and no threshold is required in our similarity enhan-

Table 3. Performance comparison with state-of-the-art methods without post-processing. * denotes the case where 128-bit SIFT binary signature is used.

Methods	Ours	Ours*	[27]	[31]	[13]	[22]	[30]	[29]	[18]	[26]*	[28]*
Holiday	**82.77**	**84.27**	82.1	81.9	81.1	82.6	81.92	78.7	-	81.0	-
Oxford5k	**78.60**	81.24	78.0	70.4	72.5	64.7	65.01	77.8	71.17	80.4	**81.3**
Paris	**75.82**	**77.78**	73.6	-	-	-	-	74.1	-	77.0	77.5
Oxford105k	**73.88**	**75.33**	72.8	-	65.2	-	-	72.9	62.34	75.0	-

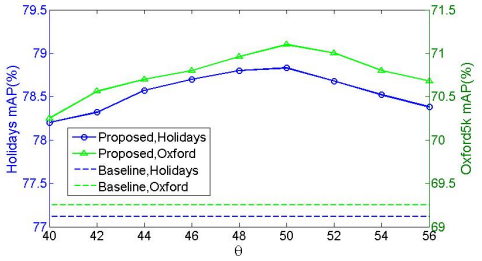

Fig.4. Impact of parameter θ on Holidays (left axis) and Oxford5k (right axis). The baseline method is based on HE.

Table 1. Impact of the number of region proposals.

Layer	$L = 1$	$L = 2$	$L = 3$	$L = 4$
Regions	1	5	14	39
Holidays	78.52	78.64	**78.81**	78.80
Oxford5k	70.01	70.51	70.87	**71.10**

Table 2. Image retrieval results for different methods. We integrate all these methods and show the accuracy in the last row.

Methods	Holidays	Oxford5k	Paris	Oxford105k
HE	77.10	69.25	68.37	56.85
HE+Proposed	78.80	71.10	70.21	62.43
HE+MA+Burst	81.00	76.83	73.75	72.06
HE+MA+Burst +Proposed	**82.77**	**78.60**	**75.82**	**73.88**

cing function. We select $\theta = 50$ for all experiments.

We also evaluate the performance with respect to the number of region proposals (Table 1). Note that each query has a specific object rectangle in Oxford and Paris, so the region proposals extraction is just processed on the database side. On the query side, only the region that has been specified is used. L=3 obtains the best performance on Holidays, but the difference between L=3 and L=4 is small. The reason is that many images in Holidays are consistent in appearance. More region proposals do not help to improve the accuracy on Holidays, but work well on Oxford5k. This is because images in Oxford dataset vary a lot in scales and viewpoint, and there are also occlusions and cluster in these images. More region proposals help to provide rich visual clues. So L=4 achieves better performance for Oxford. We use L=4 for the remaining experiments.

5.3. Evaluation

The effectiveness of our MRE algorithm and similarity enhancing function can be seen in Table 2. Our method brings improvements over the HE baseline approach for all the four datasets. Some prior arts, such as MA and Burst, have improved HE based image retrieval accuracy. Table 2 shows that our method brings consistent improvements over these two techniques. Finally, good performance is obtained.

Table 3 summarizes the performance of our method combined with MA and Burst and compares to state of the art methods without post-processing. All the reported results adopt binary version of local descriptors for efficient memory and fast computing. Most of the methods use the default detector threshold value and obtain the same number of SIFT descriptors for a fair comparison. We achieve the best performance for Holidays, Paris and Oxford105k and fall slightly behind [28] on Oxford5k when using 128-bit SIFT signature.

5.4. Time and Memory Cost

The SIFT extraction and quantization takes an average of 0.7s and 0.25s, respectively. The time spent in generating binarized FV of multiple regions is 0.05s on average, which is negligible. The average query time on Oxfor105k is 0.23s. In our MRE algorithm, extra 38 bits (L=1 bit is needless) are required for each keypoint to store the region relationship information in the inverted index. Every image needs 624 bytes for storing binarized FV. Our method consumes extra 360 Mb memory on the Oxford105k dataset when comparing to HE approach.

6. CONCLUSIONS

In this paper, we consider the contextual cues around keypoints to improve HE based image retrieval accuracy. Our MRE algorithm is proposed to find an appropriate region pair and the similarity enhancing function is used to enhance the similarity of true matched SIFTs. Experiments demonstrate the effectiveness of our methods and our results compare favorably to the state of the art approaches.

7. ACKNOWLEDGEMENTS

This work has been supported by the National Sciences & Technology Supporting Program of China under Grant No. 2015BAH49F01.

8. REFERENCES

[1] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "I Know What You Did Last Summer: Object-Level Auto-Annotation of Holiday Snaps," in *ICCV*, pp. 614-621, 2009.

[2] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual Place Recognition with Repetitive Structures," in *CVPR*, pp. 883-890, 2013.

[3] G. Schindler, M. Brown, and R. Szeliski, "City-Scale Location Recognition," in *CVPR*, pp. 1-7, 2007.

[4] W. Zhou, H. Li, Y. Lu, and Q. Tian, "SIFT Match Verification by Geometric Coding for Large-Scale Partial-Duplicate Web Image Search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, Article 4, no. 1, 2013.

[5] S. Agarwal, N. Snavely, I. Simon, S.M. Seitz, and R. Szeliski, "Building Rome in a Day," in *ICCV*, pp. 72-79, 2009.

[6] X. Shen, Z. Lin, J. Brandt, Y. Wu, "Mobile Product Image Search by Automatic Query Object Extraction," in *ECCV*, pp. 114-127, 2012.

[7] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vis*., vol. 60, no. 2, pp. 91–110, 2004.

[8] R. Arandjelović, and A. Zisserman, "Three Things Everyone Should Know to Improve Object Retrieval," in *CVPR*, pp. 2911-2918, 2012.

[9] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *ICCV*, pp. 1470–1477, 2003.

[10] H. Jégou, M. Douze, and C. Schmid, "Improving Bag-of-Features for Large Scale Image Search," *Int. J. Comput. Vis*., vol. 87, pp. 316–336, 2010.

[11] 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*, pp. 1-8, 2008.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object Retrieval with Large Vocabularies and Fast Spatial Matching," in *CVPR*, pp. 1-8, 2007.

[13] M. Perd'och, O. Chum, and J. Matas, "Efficient Representation of Local Geometry for Large Scale Object Retrieval." in *CVPR*, pp. 9-16, 2009.

[14] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object Retrieval and Localization with Spatially-Constrained Similarity Measure and K-NN Re-Ranking," in *CVPR*, pp. 3013-3020, 2012.

[15] O. Chum, A. Mikulik, M. Perdoch, J. Matas, "Total Recall II: Query Expansion Revisited," in *CVPR*, pp. 889-896, 2011.

[16] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. Han, "Contextual Weighting for Vocabulary Tree based Image Retrieval," in *ICCV*, pp. 209-216, 2011.

[17] S. Zhang, Q. Tian, Q. Huang, W. Gao, and Y. Rui, "USB: Ultrashort Binary Descriptor for Fast Visual Matching and Retrieval," *IEEE Trans. Image Process.*, vol. 23, pp. 3671-3683, 2014.

[18] R. Arandjelović, and A. Zisserman, "Visual Vocabulary with a Semantic Twist", in *ACCV*, pp. 178-195, 2014.

[19] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704-1716, 2012.

[20] M. Cheng, Z. Zhang, W. Lin, and P. Torr, "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," in *CVPR*, pp. 3286-3293, 2014.

[21] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image Classification and Retrieval are ONE," in *ICMR*, pp. 3-10, 2015.

[22] H. Jégou, M. Douze, and C. Schmid, "On the Burstiness of Visual Elements," in *CVPR*, pp. 1169-1176, 2009.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image Classification with the Fisher Vector: Theory and Practice," *Int. J. Comput. Vis*., vol 105, pp. 222-245, 2013.

[24] K. Mikolajczyk and C. Schmid, "Sclae and Affine Invariant Interest Point Detectors," *Int. J. Comput. Vis*., vol. 60, pp. 63-86, 2004.

[25] M. Muja, and D. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration," in *VISAPP*, 2009.

[26] G. Tolias, Y. Avrithis, and H. Jégou, "To Aggregate or not to Aggregate: Selective Match Kernels for Image Search," in *ICCV*, pp. 1401-1408, 2013.

[27] D. Qin, C. Wengert, and L. Van Gool, "Query Adaptive Similarity for Large Scale Object Retrieval," in *CVPR*, pp. 1610-1617. 2013.

[28] M. Shi, Y. Avrithis, and H. Jégou, "Early Burst Detection for Memory-Efficient Image Retrieval," in *CVPR*, 2015.

[29] R. Tao, E. Gavves, C. Snoek, and A. Smeulders, "Locality in Generic Instance Search from One Example," in *CVPR*, pp. 2099-2106, 2014.

[30] L. Zheng, S. Wang, W. Zhou, and Q. Tian, "Bayes Merging of Multiple Vocabularies for Scalable Image Retrieval," in *CVPR*, pp. 1963-1970, 2014.

[31] M. Jain, H. Jégou, and P. Gros, "Asymmetric Hamming Embedding: Taking the Best of Our Bits for Large Scale Image Search," in *ACM Multimedia*, pp. 1441-1444, 2011.