A NOVEL PAIR-WISE IMAGE MATCHING STRATEGY WITH COMPACT DESCRIPTORS

Shuang Yang[⋆†], *Ling-Yu Duan*[⋆*], *Jie Lin*[⋆], *Tiejun Huang*[⋆]

[⋆] The Institute of Digital Media, School of CS & EE, Peking University, Beijing 100871, China
[†] Shenzhen Graduate School, Peking University, Shenzhen 518055, China

ABSTRACT

In this paper, we address the problem of pair-wise image matching which determines whether two images depict the same objects or scenes. SIFT-like local descriptor-based matching is the most widely adopted method for this purpose and has achieved the state-of-the-art performance. However, local descriptor-based methods usually fail when an image pair contains multiple similar local regions. This problem becomes more serious when coming to limited computational and storage resources. Although global descriptors, e.g., Fisher Vectors, can solve this issue, it is difficult for global descriptors to distinguish images containing different objects of the same class. Therefore, we propose a novel strategy to integrate local and global descriptors for better matching accuracy. To further fulfill the efficiency requirement of applications, we combine dimension reduction and product quantization to obtain compact descriptors and speed up the matching process with pre-computed lookup tables. Extensive comparisons to the state-of-the-art methods demonstrate our advantages in both matching accuracy and efficiency.

Index Terms— Image matching, SIFT, Fisher Vectors, Compact descriptors

1. INTRODUCTION

Pair-wise image matching establishes correspondence between two images and determines whether they depict the same objects or scenes. Many research efforts have been focused on image matching problem as it is a key component of many computer vision tasks[1], such as image retrieval, object recognition, 3D reconstruction, etc. Generally, image matching process is conducted as follow: descriptors are first extracted from two images. Then, similarity of the two images is computed according to the extracted descriptors. Finally, the two images are determined match or not by comparing the similarity with a pre-defined threshold.

Two factors need to be taken into account for pair-wise image matching: the matching accuracy and efficiency. Matching accuracy is usually affected by scale and viewpoint variation, deformation, occlusion, clutters, etc. Diversity of target objects or scenes also pose a great difficulty for image matching methods. Matching efficiency is required by various applications. Recent applications based on big data and mobile device bring new challenges to the efficiency of image matching methods. To take good use of the limited computational and storage resources, descriptors are compressed to low dimensional vectors for fast matching while always losing performance.

Most image matching methods are based on descriptors. Extensive research attempts on descriptors have been made to improve both matching accuracy and efficiency. The existing descriptors can be roughly categorized into two classes: local and global descriptors.

*corresponding author

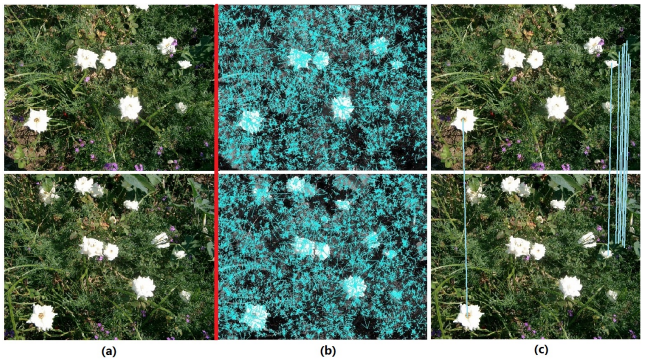

Fig. 1. Failure example of SIFT descriptor: (a) Original images; (b) SIFT descriptors extracted from the images; (c) Matching points.

Among local descriptors, Scale Invariant Feature Transform (SIFT) [2] remains the most popular local descriptor for pair-wise image matching. Other examples of local descriptors are Speeded Up Robust Features (SURF) [3] and Gradient Location Orientation Histogram (GLOH) [4]. Unfortunately, these local descriptors are high dimensional, for example 128 dimensions for SIFT, and there are usually hundreds of local descriptors per image, which costs much for storage and transmission. To save the cost, several compression schemes have been proposed to reduce the bit rate of local descriptors, e.g., hashing [5][6], transform coding [7] and vector quantization [8]. Researchers also argue to generate compact descriptors directly [9][10]. Furthermore, a subset of the extracted descriptors can be adopted for each image. For instance, a feature selection criterion [11] has been presented to choose the most promising local descriptors for subsequent image matching.

Although local descriptor-based methods have achieved the state-of-the-art performance, they fail to deal with multiple similar local regions. These local regions cannot be distinguished, since they are described by similar local descriptors. When matching, there will be a lot of mismatches of points and only few correct matches remained as illustrated in Figure1. Due to low bit-rate requirement, local descriptors are compressed to small codes and few of them are selected for matching. As a result, poor discriminative power further magnifies the issue. The performance of local descriptor-based methods is greatly degraded.

Besides local descriptors, an alternative solution for image matching is to generate global descriptors, one per image. Popular global descriptors include color histogram, shape context [12] and GIST [13]. Global descriptors can also be aggregated from local descriptors. Bag of words (BOW) [14] uses a histogram of the number of image descriptors assigned to each visual word to represent an image. Recently, Perronnin et al. [15][16] introduced Fisher kernel

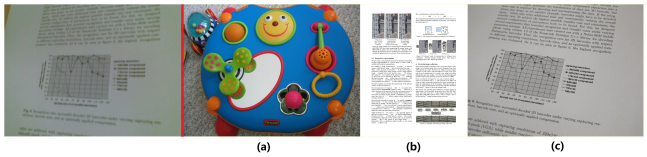

Fig. 2. Failure example of FV descriptor: The first image is a query image. (a) is the inter-class non-matching image of the query with similarity 0.299110; (b) is the intra-class non-matching image of the query with similarity 0.438452; (c) is the intra-class matching image of the query with similarity 0.401683

[17] to generate global descriptors. Given an image, Fisher kernel aggregates the local descriptors to form a Fisher Vector (FV) representation of fixed-length. Jegou et al. [18] proposed a simplified FV, the Vector of Locally Aggregated Descriptors (VLAD). Subsequently, high-dimensional global descriptors are further compressed into compact codes. Promising results have been reported [15][16][19] for classification and retrieval tasks.

The global and contextual information of global descriptors performs well when distinguishing images of objects from different classes (inter-class images). However, global descriptors may be confused by images containing different objects of the same class (intra-class images), especially for those ambiguous ones like document images. As illustrated in Figure2, we take a document image as query image and compute its similarities of FV with inter-class non-matching image(a), intra-class non-matching image(b) and intra-class matching image(c). We can observe that the similarity with inter-class non-matching image(a) is relatively low, while the similarity with intra-class non-matching image(b) is even higher than the similarity with intra-class matching image(c). Thus, it's difficult to determine whether an intra-class image pair is match or not according to similarities of global descriptors. This is mainly because attributes (such as color, edges or local features) statistics of intra-class images may be similar no matter match or not.

In this paper, we propose a novel pair-wise image matching strategy with compact descriptors, which enables local and global descriptors to complement each other at low bit-rate requirement. The proposed strategy includes two stages: first, we employ local descriptors to conduct initial matching and obtain a set of matching image pairs. Then, the unmatched image pairs of the first stage are judged again by global descriptors. Thus, local descriptors guarantee that the intra-class matching image pairs have been selected before confusing global descriptors. Meanwhile, global descriptors find out the matching image pairs with multiple similar local regions from the unmatched image pairs of local descriptors. Considering the requirement of efficiency, we compress both local and global descriptors to small codes before the matching process. To evaluate performance of the proposed strategy, we use the state-of-the-art local and global descriptors, i.e. SIFT and FV, in our implementation. Experimental results demonstrate that the proposed strategy performs better than solely using local or global descriptors at low bit rate, say hundreds of bytes. For example, the True Positive Rate of the proposed strategy over UKBench dataset is 93.73% versus 84.93% of compressed SIFT and 88.55% of compressed FV in average of different low bit rates, with 1% False Positive Rate.

This paper is organized as follows. Section 2 gives an formulated introduction of the proposed strategy. We give an exemplar implementation with SIFT and FV in Section 3. Experimental results is presented in Section 4.

2. A NOVEL STRATEGY FOR IMAGE MATCHING

To overcome the weaknesses of both local and global descriptors, we propose a novel strategy, which combines compact local descriptors with compact global descriptors.

The proposed strategy includes two stages: local descriptor-based image matching and global descriptor-based image matching. Local descriptor-based image matching is conducted first to find out matching image pairs with relatively low false positive alarm. Then, global descriptor-based image matching will further improve the matching accuracy .

Given two images X_A and X_B, local descriptors L_A, L_B and global descriptors G_A and G_B are extracted from images respectively, where $L_A = \{l_{A1}, l_{A2}, \ldots, l_{AP}\}, L_B = \{l_{B1}, l_{B2}, \ldots, l_{BQ}\}$, $l_{Ai}, l_{Bj} \in R^n$, $G_A, G_B \in R^m$, $1 \leq i \leq P$, $1 \leq j \leq Q$. $l_{Ai} \in R^n$ is the local descriptor for the i^{th} interesting point of image X_A and $l_{Bj} \in R^n$ is the local descriptor for the j^{th} interesting point of image X_B. P and Q are the numbers of local descriptors extracted from images X_A and X_B. To meet the efficiency requirement, both local and global descriptors have been compressed.

The similarity of the given images is first calculated according to L_A and L_B. For each l_{Ai}, the nearest l_{Bj} is found by ratio test. Geometric verification is subsequently applied to filter out outliers. Finally, images X_A and X_B are determined match or not by comparing the score of inliers with a predefined threshold δ.

After the matching process with L_A and L_B, we get an initial judgment about whether images X_A and X_B are match or not. If X_A and X_B are determined match, we take X_A and X_B as a matching image pair and terminate; otherwise, we judge these two images again based on G_A and G_B.

$$match\,(X_A, X_B) = \begin{cases} 1 & S_{L_A,L_B} > \delta \\ & S_{L_A,L_B} < \delta \; and \; S_{G_A,G_B} > \theta \\ 0 & S_{L_A,L_B} < \delta \; and \; S_{G_A,G_B} < \theta \end{cases}$$

where $match\,(\cdot,\cdot)$ is an indicator function, if $match\,(\cdot,\cdot) = 1$, the two images are match, otherwise not match. S_{L_A,L_B} represents the similarity between L_A and L_B, i.e. the score of inliers and S_{G_A,G_B} represents the similarity between G_A and G_B, which is depending on the distance between these two global descriptors. δ and θ are predefined thresholds respectively.

3. AN EXEMPLAR IMPLEMENTATION

Scale Invariant Feature Transform (SIFT) [2] and Fisher vector (FV) [18] are popular local and global descriptors respectively which achieve the state-of-the-art performance. In this section, we give a technical solution based on SIFT and FV employing the proposed strategy. Targeted at compact descriptors, we employ Principle Component Analysis (PCA) and Product Quantization (PQ) to compress SIFT and FV descriptors. Figure3 shows how SIFT and FV implement the proposed strategy. As seen in Figure3, in the first stage, SIFT descriptors are extracted from image pairs and compressed for fast computing. The image pairs are then matched with compressed SIFT descriptors. The unmatched image pairs of the first stage are passed to the second stage. In the second stage, original FV are generated for images pairs. PCA and PQ are applied to original FV for compact codes. Finally, the similarity is computed according to the compressed FV descriptors. Technical details are discussed below.

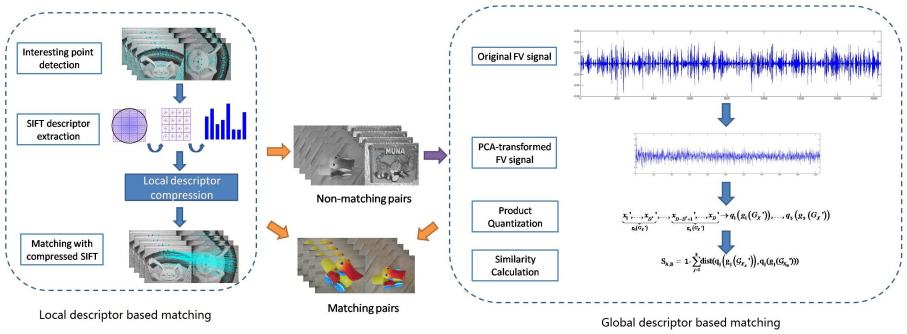

Fig. 3. Exemplar Implementation with SIFT and FV

3.1. Local descriptor based image matching

To implement the image matching process of local descriptors, SIFT descriptors are first extracted from images, for example, image X_A.

PCA is applied to eliminate the correlation among the dimensions of each SIFT descriptor. Then, a partition is given to divide the transformed descriptor ${l_{Ai}}'$ into S groups $\{g_1, \ldots, g_S\}$ consecutively and the most important yet orthogonal dimensions are grouped together.

The quantization of each group is designed in two phrases, namely Multi-Stage Vector Quantization (MSVQ)[20]. The training of the first-stage quantizer may resort to the state-of-the-art visual vocabulary techniques, such as [21][14] and their variances. We adopt Hierarchical K-Means clustering to build the initial codebook with dimension M_{1st}. Given the s-th group of transformed dimensions $g_s\left({l_{Ai}}'\right)$, we quantize it into the nearest word $w_j\ (j \in [1, M_{1st}])$. In the second stage, we adopt PQ to further quantize the residuals resulting from the codebook of the first stage. More specifically, given the transformed dimensions $g_s\left({l_{Ai}}'\right)$ and its corresponding quantization vector in the first stage w_j, a residual vector $R\left(g_s\left({l_{Ai}}'\right),\ w_j\right)$ is then formed:

$$R\left(g_s\left({l_{Ai}}'\right),\ w_j\right) = g_s\left({l_{Ai}}'\right) - w_j$$

Subsequently, PQ is applied to each residual vector.

For points matching, the distance between compressed SIFT descriptors of images X_A and X_B is calculated using a look-up table. Geometric verification is followed to filter out outliers.

Finally, images X_A and X_B are determined match or not by comparing the score of inliers with a predefined threshold δ.

3.2. Global descriptor based image matching

For global descriptor based image matching, we first generate original FV by aggregating the compressed SIFT descriptors.

Let $X = \{x_t\}_{t=1}^{T}$ denote a set of T compressed SIFT descriptors extracted from an image X, $x_t \in R^d$, where d denotes the dimensionality of the compressed SIFT descriptor. The GMM parameters λ are learned over the training set of compressed SIFT descriptors by the well-known Expectation-Maximization (EM) algorithm. Then, Fisher kernel [17] is defined on the gradient vector representation. Let $\mathcal{G}_k$ denote the d-dimensional gradient vector with respect to the k^{th} Gaussian. The analytical form of $\mathcal{G}_k$ is derived by:

$$\mathcal{G}_k = \frac{\partial L\left(X|\lambda\right)}{\partial \mu_k} = \frac{1}{\sqrt{T w_k}} \sum_{t=1}^{T} \gamma_t\left(k\right) \sigma_k^{-1}\left(x_t - \mu_k\right)$$

where $\gamma_t\left(k\right) = \frac{w_k p_k(x_t)}{\sum_{l=1}^{K} w_l p_l(x_t)}$ denotes the probability of compressed SIFT descriptor x_t being generated by the k^{th} Gaussian component. Finally, the Fisher vector $\mathcal{G}_X$ is formed by concatenating the aggregated gradient vectors $\mathcal{G}_k$ of all Gaussian components, $k = 1 \ldots K$. The dimensionality M of $\mathcal{G}_X$ is $M = Kd$. In practice, power law may be applied to normalize each element of $\mathcal{G}_X$.

For the purpose of dimensionality reduction, we employ PCA to project $\mathcal{G}_X$ to a set of orthogonal bases $\mathbf{U}$ and get a lower, D-dimensional vector ${\mathcal{G}_X}' = \mathbf{U}\mathcal{G}_X$, where $\mathcal{G}_X \in R^M$, ${\mathcal{G}_X}' \in R^D$ and $D < M$.

We employ PQ to further reduce the length of ${\mathcal{G}_X}'$. Given the PCA-transformed descriptor ${\mathcal{G}_X}'$, a partition divides it into S subvectors $g_j({\mathcal{G}_X}')$, $1 \le j \le S$, of dimension $D^* = D/S$, where D is a multiple of S. The subvectors are quantized separately using S distinct vector quantizers $q_j\left(g_j\left({\mathcal{G}_X}'\right)\right)$. Each q_j is associated with a codebook $C_j = \left\{\mathbf{c_{ji}} \in R^{D*}, i \in I\right\}$, where I is the index set. $g_j({\mathcal{G}_X}')$ is then assigned the index of its nearest center $\mathbf{c_{ji}}$.

Finally, the similarity of images X_A and X_B is calculated as follows:

$$\mathrm{S}_{\mathrm{G}_A,\mathrm{G}_B} = 1 - \sum_{j=1}^{\mathrm{S}} \mathrm{dist}(\mathrm{q_j}\left(\mathrm{g_j}\left({\mathcal{G}_{X_A}}'\right)\right), \mathrm{q_j}(\mathrm{g_j}({\mathcal{G}_{X_B}}')))$$

We determine images X_A and X_B match or not by comparing the similarity $\mathrm{S}_{\mathrm{G}_A,\mathrm{G}_B}$ with a predefined threshold θ.

4. EXPERIMENT

We evaluate the proposed strategy in the context of pair-wise image matching on publicly available MPEG Compact Descriptor Visual Search(CDVS) datasets [22][23][24][25][26]. We compare the proposed strategy with two baselines at different bit rates: (1)compressed SIFT (PCA+MSVQ) and (2)compressed FV (PCA+PQ). We use the True Positive Rate(TPR) at fixed False Positive Rate(1%) as evaluation protocol.

4.1. Experiment Datasets

Eight datasets contributed to CDVS, including ZuBud, UKBench, Stanford, ETRI, PKU, Telecom Italia, Telecom SudParis and Huawei, are used in experiments. There are 30256 images in total which are grouped to 5 categories by contents, i.e. mixed text and graphics, paintings, frames captured from video clips, landmarks and common objects. All datasets provide the annotation files of matching and non-matching images pairs. Refer to Table.1 for more details about the datasets.

Data category	Graphics	Paintings	Frame	Landmarks	Objects
♯ of images	2500	500	500	13098	10200
♯ of matching image pairs	3000	364	400	4005	2550
♯ of non-matching image pairs	30000	3640	4000	48675	25500

Table 1. Details about the datasets

4.2. Experiment Setting

We extract SIFT descriptors using the VLFeat library. For compressed SIFT, we apply different codebooks at different bit rates and the SIFT descriptors are compressed to different length. Table2 gives the number of bits per compressed SIFT descriptor at different query length.

Query Length	256	512	1K	2K	4K	8K	16K
♯ of bits	28	28	28	52	72	108	144

Table 2. Compression Details of SIFT

FV is aggregated as follow: The PCA-compressed SIFT features with dimension 32 are used. The GMM is set with K=512 Gaussian components. For compressed FV, we apply PCA to reduce the dimension of original FV from M=16384 to D=2048. We segment the transformed FV into S=256 consecutively groups of 8 dimensions, and $\log_2|C_j| = 8$. The compressed FV is at a fixed length of 256 Bytes.

In practice, we need to store the transform matrix $\mathbf{U}$, $\mathbf{U} \in R^{M\times D}$. Hence, $M \times D \times 4$ Bytes memory is necessary. Taking $M = Kd = 512 \times 32 = 16384$, $D = 2048$ as an example, we need 128MB in total to store the matrix. To meet the memory constraint of low bit-rate applications, we binarize the matrix according to the sign of each element. When projecting, if an element is 1, we multiply the corresponding dimension of $\mathcal{G}_X$ with 1, otherwise -1. Thus, we can reduce the memory cost to $M \times D \times 1$ bits (4MB), which causes little performance loss.

The Euclidean distance and cosine distance are used respectively to compute the distance between compressed local descriptors and global descriptors. The predefined parameters δ and θ are set for each dataset to make the False Positive Rate fixed.

4.3. Quantitative performance comparison

Figure4 represents a comparison of three methods with different query lengths on dataset UKBench (Objects), providing the TPR of compressed SIFT (baseline(1)), compressed FV (baseline(2)) and combination of these two descriptors with the proposed two-stage strategy, where FPR is fixed at 1%. As we can see from Figure4, compressed FV outperforms compressed SIFT a lot with the query length of 256 bytes. And the proposed strategy performs better with all query length. This is because the UKBench dataset consists of a lot of images with multiple similar local regions as illustrated in Figure1. By combining these two descriptors, compressed FV helps to find out the matching image pairs which make local descriptors fail. As the query length grows, compressed SIFT become more discriminative and helps to further improve the TPR with low FPR, reducing the FP caused by compressed FV. Figure 5 shows our approach has achieved promising results at different bit rates over different datasets.

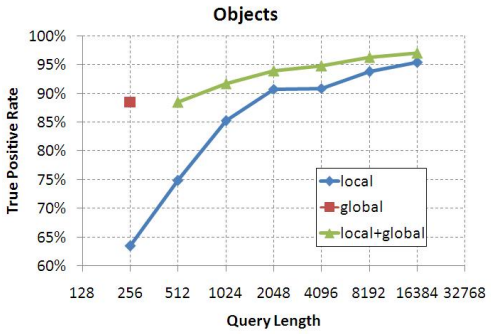

Fig. 4. Performance comparison on UKBench dataset

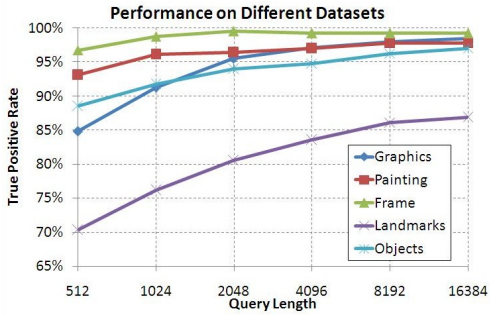

Fig. 5. Performance of compressed descriptors on different datasets

Complexity Analysis. For local descriptor SIFT, we employ MSVQ, which has significantly reduced the size of quantization tables. The total memory cost for compressing SIFT is 384KB. For global descriptor FV, our implementation uses a binarized transform matrix, which scarifies little performance but reduces the memory cost to 4MB. PQ for FV uses 2MB to store codebooks for quantization. Additional memory cost, including PCA transformation matrix for compressing raw SIFT and the GMM model, is 49KB.

5. CONCLUSION

In this paper, we propose a novel strategy to combine compact local and global descriptors, which is efficient and simple to implement. The proposed strategy takes both accuracy and efficiency into consideration to fulfill the requirements of various applications. Any two descriptors can be combined using this strategy to complement the weakness of each other.

We also provide an exemplar implementation of the proposed strategy and better accuarcy and efficiency have been achieved. To further save the memory cost, special strategy can be designed according to the characteristics of descriptors when applying the compression schemes.

6. ACKNOWLEDGEMENT

This work was supported by the Chinese Natural Science Foundation under Contract No. 61271311 and No. 61121002, and in part by the Research Fund of ZTE Corporation.

7. REFERENCES

[1] R. Ji, L.-Y. Duan, J. Chen, H. Yao, and W. Gao, "Learning compact visual descriptor for low bit rate mobile landmark search," in *Proceedings of International Joint Conference of Artificial Intelligence*, 2011.

[2] D. Lowe, "Distinctive image feature from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 10, no. 3, pp. 346–359, 2008.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1615–1630, 2005.

[5] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Proceedings of Advances in Neural Information Processing Systems*, 2009, pp. 1509–1517.

[6] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proceedings of Advances in Neural Information Processing Systems*, 2009, vol. 21, pp. 1753–1760.

[7] V. Chandrasekhar, G. Takacs, and D. Chen, "Transform coding of image feature descriptors.," in *Proceedings of Visual Communications and Image Processing*, 2009.

[8] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *Proceedings of International Conference on Computer Vision*, 2007, pp. 1–8.

[9] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proceedings of European Conference on Computer Vision*, 2010, pp. 778–792.

[10] V. Chandrasekhar, G. Takacs, D.M. Chen, S.S. Tsai, R. Grzeszczuk, and B. Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," in *Proceedings of Computer Vision and Pattern Recognition*, 2009, pp. 2501–2511.

[11] G. Francini, S. Lepsy, and M. Balestri, "Selection of local features for visual search," *Signal Processing: Image Communication*, 2012.

[12] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *Proceedings of Advances in Neural Information Processing Systems*, 2000, pp. 831–837.

[13] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," vol. 42, no. 3, pp. 145–175, 2001.

[14] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of International Conference on Computer Vision*, 2003, pp. 1470–1478.

[15] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proceedings of Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[16] F. Perronnin and C. Dance, "Large-scale image retrieval with compressed fisher vectors," in *Proceedings of Computer Vision and Pattern Recognition*, 2010, pp. 3384–3391.

[17] T.S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proceedings of Advances in Neural Information Processing Systems*, 1999, pp. 487–493.

[18] H. Jégou, M. Douze, and C. Schmid, "Aggregating local descriptors into a compact image representation," in *Proceedings of Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[19] J. Lin, L.-Y. Duan, T. Huang, and W. Gao, "Robust fisher codes for large scale image retrieval," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[20] J. Chen, L.-Y. Duan, J. Lin, R. Ji, T. Huang, and W. Gao, "On the interoperability of local descriptors compression," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[21] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

[22] "Mvs," http://mars01.stanford.edu/mvs.

[23] "Ethz," http://www.vision.ee.ethz.ch/datasets.

[24] "Cturin180," http://pacific.tilab.com/download/CTurin180.zip.

[25] "Pkubench," http://61.148.212.146:8080/landmark/benchmark.

[26] "Ukbench," http://vis.uky.edu/ stewe/ukbench.