ROBUST FISHER CODES FOR LARGE SCALE IMAGE RETRIEVAL

Jie Lin[†‡] *Ling-Yu Duan*[†⋆] *Tiejun Huang*[†] *Wen Gao*[†]

[†] The Institute of Digital Media,School of EE&CS, Peking University, Beijing, 100871, China
[‡] The School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China
{jielin, lingyu, tjhuang, wgao}@pku.edu.cn

ABSTRACT

Fisher vectors (FV) have shown great advantages in large scale visual search. However, traditional FV suffers from noisy local descriptors, which may deteriorate the FV discriminative power. In this paper, we propose a robust Fisher vectors (RFV). To fulfill fast search and light storage over a large scale image dataset, we employ a simple binarization method to compress RFV to generate compact robust Fisher codes (RFC). Extensive comparison experiments on benchmark datasets have shown that both RFV and RFC outperforms the state-of-the-art performance. The scalability of RFC has been validated on a dataset of over 1 million images as well.

Index Terms— Fisher kernel, local descriptors aggregation, large scale visual search

1. INTRODUCTION

The problem of large-scale image retrieval regards the search and discovery of images contained within a large collection that depict the same objects or scenes as those depicted by a query image. This requires the database images to be processed for the creation of a descriptor database which may be indexed. Search is performed by the descriptors extracted from the query. The bag-of-features (BoF) [1] is the most popular method. Given an image, the keypoints are detected and their local descriptors (e.g. SIFT [2]) are extracted. Each local descriptor is quantized to a visual word. Inverted index file is build up to implement visual words based indexing and search. However, inverted index file brings about heavy memory consumption, e.g., 1.1 million images may incur a memory usage of 4.3GB [3], which make it difficult to scale up to a large scale image dataset.

Many research efforts have been attempted to improve the performance of traditional BoF. For example, a large vocabulary (1 million visual words) [4][3] to make fine-grained quantization of the descriptor space; soft assignment of descriptors to multiple visual words to reduce quantization error [5]; query expansion [6] and geometric verification by RANSAC [3] to improve the initial retrieval results. Unfortunately, the intensive computation renders the BoF based retrieval less efficient, despite performance improvements.

To improve the retrieval performance and efficiency at much less memory complexity, Perronnin et al. [7] introduced Fisher kernel [8] to image retrieval. Fisher kernel is a powerful tool, which exhibits the strength of both generative and discriminative models. Given an image, Fisher kernel aggregates the local descriptors to form a Fisher vectors (FV) representation of fixed-length. Jegou et al. [9] proposed a simplified FV, the Vector of Locally Aggregated Descriptors (VLAD). Promising results have been reported [7] [9].

⋆ Corresponding Author

In this paper, we propose a robust Fisher vectors (RFV) representation. The selective aggregation can significantly reduce the negative impact of noisy local descriptors on FV discriminative power. Referring to Eq.1, the aggregation of traditional FV assumes that local descriptors contribute to the FV representation equally. This could be biased from the modeling perspective. As illustrated in Figure 1, it is easy to imagine that FV mainly benefits from local descriptors extracted from the regions of *query object*, while the descriptors from *background* would deteriorate its discriminative power. Hence, it is crucial to consider a model that incorporate the selection of local descriptors into the FV aggregation stage. In this work, we propose to model the characteristics of keypoints for selecting "healthy" local descriptors to aggregate FV. Correct matching keypoints are statistically associated with useful local descriptors. The selection function may be modeled by matching pairs.

Our extensive experiments on two benchmark datasets (the UKbench and Holidays) and a Graphics dataset taken by a mobile phone, have shown the consistent superior performances of the proposed RFV over the state-of-the-arts [10][9][7]. For example, mean average precision (mAP) is improved from 59.5% to 67.1% on Holidays dataset compared to the FV [10].

More importantly, to evaluate the RFV scalability over a large-scale dataset, we employ a sign binarization approach [7] to compress RFV into small robust Fisher codes (RFC) for fast search and much less storage. Our RFC significantly outperforms [7][10], e.g., recall@1 achieves around 43% (versus 31% reported in [7]) on the Holidays dataset combined with 1 million distractor images.

2. RELATED WORK

Rather than BoF, our proposed RFC focuses on the FV based retrieval. Below we compare the BoF and FV based retrieval pipeline. Related work on FV is reviewed to distinguish the proposed RFC.

BoF versus FV Both BoF and FV generate a compact image-level representation by aggregating local descriptors. However, FV employs higher order statistics to achieve more discriminative power. BoF encodes the zero-order statistics by counting the occurrences of quantized local descriptors (visual words). Beyond the occurrence statistics, FV extends BoF by encoding higher-order statistics of local descriptors[10]. FV employs a Gaussian Mixture model (GMM) to estimate the distribution of local descriptors over a training feature set. Given an image, the gradient vectors of all local descriptors w.r.t. the parameters of each Gaussian component, are aggregated by computing their mean (1-order) and/or variance (2-order). FV is finally formed by concatenating the 1-order and/or 2-order statistics of all Gaussian components.

In particular, FV significantly reduce the computational complexity. The reasons are two-folds. Firstly, compared to BoF, FV

achieves much better performance with a few hundred visual words (Gaussian). Secondly, raw FV can be further compressed to generate compact Fisher codes (FC), e.g., binarization (0/1 bits) [7] or product vector quantization [10]. Compared to original FV, FC yields comparable performance, while FC greatly improves the retrieval efficiency due to very fast Hamming space computing (XOR operation and bit count), and reduces the storage of features. [7]. Distinct from BoF, the FV based search don't require any indexing files.

Fisher Kernel Representation Fisher kernel is a generic representation[8], which has been successfully applied to a few vision related tasks, such as image classification [11] and image retrieval [7]. Recently, further improvements have been proposed, such as incorporating spatial cues into FV [12][13], modeling the relationship of local descriptors [14], combining FV with other features to make efficient retrieval [15]. Distinct from previous work, the proposed RFV employs the keypoint characteristics to perform selective aggregation of local descriptors, which has significantly improved the discriminative power of FV representation.

3. PROBLEM FORMULATION

3.1. Brief review of Fisher vectors

To formulate the problem of robust Fisher vector, we briefly review the Fisher kernel principle. Let $\mathbf{I} = \{\mathbf{x}_t\}_{t=1}^{T}$ denote a set of T local descriptors $\mathbf{x}_t$ extracted from image $\mathbf{I}$, $\mathbf{x}_t \in \mathbb{R}^d$, where d denotes the dimensionality of raw local descriptors or the processed local descriptors after dimensionality reduction. We may assume that each local descriptor $\mathbf{x}_t$ is generated independently by an offline trained GMM with K Gaussian components: $p(\mathbf{x}_t|\lambda) = \sum_{k=1}^{K} \omega_k p_k(\mathbf{x}_t)$, $\lambda = \{\omega_k, \mu_k, \sigma_k^2\}_{k=1}^{K}$, where ω_k, μ_k and σ_k^2 denote the weight, mean and diagonal variance matrix of Gaussian component k, respectively. The GMM parameters λ are learned through maximizing the likelihood of training images with the well-known Expectation-Maximization (EM) algorithm. The log-likelihood of image $\mathbf{I}$ is obtained by averaging the log-likelihood values of local descriptors as:

$$\mathcal{L}(\mathbf{I}|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda) \tag{1}$$

Accordingly, image $\mathbf{I}$ can be represented by the aggregated gradient vectors of local descriptors with respect to the likelihood function of image $\mathbf{I}$ based on the learned parameter setting λ [8]:

$$\mathbf{G}_\lambda^{\mathbf{I}} = \frac{1}{T} \bigtriangledown_\lambda \mathcal{L}(\mathbf{I}|\lambda) \tag{2}$$

Fisher kernel [8] is elegantly defined on the gradient vector representation. Fisher kernel function $\mathbf{K}(\mathbf{I}^x, \mathbf{I}^y) = \mathbf{G}_\lambda^{\mathbf{I}^x T} \mathbf{F}_\lambda^{-1} \mathbf{G}_\lambda^{\mathbf{I}^y}$,where $\mathbf{F}_\lambda$ is the Fisher information matrix. $\mathbf{F}_\lambda = \mathbf{E}_{\mathbf{x}\sim p}[\bigtriangledown_\lambda \mathcal{L}(\mathbf{I}|\lambda) \bigtriangledown_\lambda \mathcal{L}(\mathbf{I}|\lambda)^T]$ $\mathbf{F}_\lambda$ is symmetric and positive definite, which can be decomposed as $\mathbf{F}_\lambda^{-1} = \mathbf{L}_\lambda^T \mathbf{L}_\lambda$. $\mathbf{L}_\lambda$ can be considered as a normalization matrix. Hence, we may rewrite $\mathbf{K}(\mathbf{I}^x, \mathbf{I}^y)$ as the form of dot product between normalized gradient vectors $\mathcal{G}_\lambda^{\mathbf{I}}$ with:

$$\mathcal{G}_\lambda^{\mathbf{I}} = \mathbf{L}_\lambda \mathbf{G}_\lambda^{\mathbf{I}} \tag{3}$$

, where $\mathcal{G}_\lambda^{\mathbf{I}}$ is referred to as appearance FV [11] of image $\mathbf{I}$.

Let $\mathcal{G}_k^{\mathbf{I}}$ denote the d-dimensional gradient vector w.r.t. the parameter of Gaussian k. The analytical form of $\mathcal{G}_k^{\mathbf{I}}$ is derived by:

$$\mathcal{G}_k^{\mathbf{I}} = \frac{\partial \mathcal{L}(\mathbf{I}|\lambda)}{\partial \mu_k} = \frac{1}{\sqrt{T\omega_k}} \sum_{t=1}^{T} \beta_t(k) \sigma_k^{-1}(\mathbf{x}_t - \mu_k), \tag{4}$$

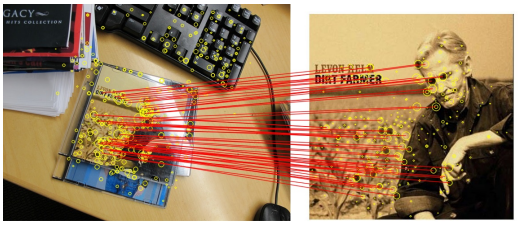

Fig. 1. Keypoint matching between a query image with background clutter (left) and the reference image (right). The bigger circles indicate more effective local descriptors for FV aggregation.

, where $\beta_t(k) = \frac{\omega_k p_k(\mathbf{x}_t)}{\sum_{j=1}^{K} \omega_j p_j(\mathbf{x}_t)}$ denotes the probability for local descriptor $\mathbf{x}_t$ being generated by Gaussian component k.

The appearance FV of image $\mathbf{I}$ $\mathcal{G}_\lambda^{\mathbf{I}}$ is finally generated by concatenating the aggregated gradient vectors $\mathcal{G}_k^{\mathbf{I}}$ of all Gaussian components $k = 1...K$. Thus, the total length of appearance FV is Kd-dimensional. Subsequently, we employ the power law ($\alpha = 0.6$) to normalize each dimension of $\mathcal{G}_\lambda^{\mathbf{I}}$.

3.2. Problem statement

We propose the selective gradient vector aggregation of local descriptors to make robust Fisher vectors. We inject a selection function $h(\mathbf{z}_t)$ of local descriptors $\mathbf{x}_t$ into the Fisher kernel aggregation framework. The selection function is defined over the detected keypoints' features $\mathbf{z}_t$. Note that most exiting works do not elegantly unify detection and description stages in robust representation.

Let $\mathbf{I} = \{(\mathbf{z}_t, \mathbf{x}_t)\}_{t=1}^{T}$ denote a collection of local descriptors $\mathbf{x}_t$ and their detected keypoints' features $\mathbf{z}_t$ in image $\mathbf{I}$. The average log-likelihood of image $\mathbf{I}$ in Eq. 1 is rewritten as follows:

$$\begin{aligned} \hat{\mathcal{L}}(\mathbf{I}|\lambda) &= \frac{1}{T}\sum_{t=1}^{T} \log p(\mathbf{z}_t, \mathbf{x}_t|\lambda) \\ &= \frac{1}{T}\sum_{t=1}^{T} \log h(\mathbf{z}_t) p(\mathbf{x}_t|\lambda), \end{aligned} \tag{5}$$

where

$$h(\mathbf{z}_t) = \begin{cases} 1 & \text{if } \mathbf{S}(\mathbf{z}_t) \leq \tau \\ 0 & \text{if } \mathbf{S}(\mathbf{z}_t) > \tau, \end{cases} \tag{6}$$

$h(\mathbf{z}_t)$ is a binary function of keypoint $\mathbf{z}_t$ to determine whether $\mathbf{x}_t$ is involved in appearance FV aggregation or not. $\mathbf{S}(\mathbf{z}_t)$ is a likelihood ratio test function and τ is a constant threshold. If $\mathbf{S}(\mathbf{z}_t) \leq \tau$, the corresponding descriptor $\mathbf{x}_t$ is adopted by appearance FV aggregation; otherwise, it is discarded. In this work, the keypoint feature $\mathbf{z}_t = (\eta, \theta, v, \xi)$ is of four dimensions, where η, θ, v and ξ denote scale, orientation, peak value in scale space and the distance from a keypoint to the image center, respectively.

Accordingly, the selectively aggregated gradient vector $\mathcal{G}_k^{\mathbf{I}}$ of Gaussian k in Eq. 4 is updated as:

$$\hat{\mathcal{G}}_k^{\mathbf{I}} = \frac{1}{\sqrt{T\omega_k}} \sum_{t=1}^{T} h(\mathbf{z}_t) \beta_t(k) \sigma_k^{-1}(\mathbf{x}_t - \mu_k). \tag{7}$$

To generate the robust Fisher vectors $\hat{\mathcal{G}}_\lambda^{\mathbf{I}}$, we concatenate the aggregated gradient vectors $\hat{\mathcal{G}}_k^{\mathbf{I}}$, $k = 1...K$ of all Gaussians components.

4. BAYESIAN ADAPTATION FOR KEYPOINT LEARNING

To fulfill the selective aggregation of local descriptors, we propose to learn $\mathbf{S}(\mathbf{z}_t)$ from the perspective of determining robust keypoints for patch-level descriptor matches. The learning procedure is to model the keypoints' characteristics of matching and non-matching keypoint pairs[16]. Given an image, for each keypoint $\mathbf{z}_t$, the learned model can be employed to predict the probability of $\mathbf{z}_t$ being correctly matched with a keypoint in the comparing image. With the output probability, we employ a likelihood ratio test to come up with the selection of local descriptors in aggregation:

$$\mathbf{S}(\mathbf{z}_t) = \frac{p(\mathbf{z}_t|\mathbf{H}_1)}{p(\mathbf{z}_t|\mathbf{H}_0)} \begin{cases} \leq \tau & \text{accept } \mathbf{H}_0 \\ > \tau & \text{reject } \mathbf{H}_0, \end{cases} \tag{8}$$

where hypothesis $\mathbf{H}_0$ and $\mathbf{H}_1$ represent whether keypoint $\mathbf{z}_t$ would yield a correct match or not, respectively. $p(\mathbf{z}_t|\mathbf{H}_i), i = 0, 1$, is the probability density function for hypothesis $\mathbf{H}_i$. $\tau \in (0, \infty)$ denotes the decision threshold to accept or reject $\mathbf{H}_0$. A smaller τ implies that there would be less keypoints in an image to accept $\mathbf{H}_0$.

Constructing the training keypoint set $\mathbb{B}_{H_1}$ and $\mathbb{B}_{H_0}$. Let $\Omega = \{\langle \mathbf{I}_n^l, \mathbf{I}_n^r \rangle\}_{n=1}^N$ denote N matching image pairs, $(\mathbf{Z}_n^e, \mathbf{X}_n^e) = \{(\mathbf{z}_{nm}^e, \mathbf{x}_{nm}^e)|e \in \{l, r\}, m = 1...M_n\}$ denote a collection of keypoints $\mathbf{z}_{nm}^e$ and the corresponding descriptors $\mathbf{x}_{nm}^e$ extracted from each image $\mathbf{I}_n^e$. We employ a distance ratio test [2] between keypoint sets $\mathbf{X}_n^l$ and $\mathbf{X}_n^r$ to detect matching keypoint pairs $\mathbb{D}_n = \langle \mathbf{X}_n^l, \mathbf{X}_n^r \rangle = \{\langle \mathbf{x}_{nd}^l, \mathbf{x}_{nd}^r \rangle | d = 1...D_n\}$ from an image pair $\langle \mathbf{I}_n^l, \mathbf{I}_n^r \rangle$, which may remove many false matches from background clutter. Subsequently, a geometric consistency check like RANSAC [3] is applied to divide keypoint set $\mathbb{D}_n$ into inliers $\hat{\mathbb{D}}_n = \langle \hat{\mathbf{X}}_n^l, \hat{\mathbf{X}}_n^r \rangle = \{\langle \hat{\mathbf{x}}_{nd}^l, \hat{\mathbf{x}}_{nd}^r \rangle | d = 1...\hat{D}_n\}$ and outliers $\mathbb{D}_n \setminus \hat{\mathbb{D}}_n$. The inlier $\hat{\mathbb{D}}_n$ are finally considered as correct matches.

We construct the entire keypoint set $\mathbb{B}_{H_1} = \{\mathbf{z}_t | t = 1...B_1, \mathbf{z}_t \in \mathbf{Z}_n^e, n = 1...N, e \in \{l, r\}\}$, including the matching keypoint (inlier) set $\mathbb{B}_{H_0} = \{\mathbf{z}_t | t = 1...B_0, \mathbf{z}_t \in \hat{\mathbf{Z}}_n^e, n = 1...N, e \in \{l, r\}\}$. $\hat{\mathbf{Z}}_n^e$ denotes the keypoint set of $\hat{\mathbf{X}}_n^e$. $\mathbb{B}_{H_1}$ contains both matching and non-matching keypoints, while $\mathbb{B}_{H_0}$ is a subset of $\mathbb{B}_{H_1}$.

To establish the hypothesis model $p(\mathbf{z}_t|\mathbf{H}_0)$ and $p(\mathbf{z}_t|\mathbf{H}_1)$, we first train a universal model $p(\mathbf{z}_t|\lambda_{\mathbf{H}_1})$ with parameters $\lambda_{\mathbf{H}_1}$ for hypothesis $\mathbf{H}_1$ over the entire keypoint set $\mathbb{B}_{H_1}$. Rather than independently learning the model $p(\mathbf{z}_t|\lambda_{\mathbf{H}_0})$ for hypothesis $\mathbf{H}_0$, we adopt Bayesian adaptation to derive $\lambda_{\mathbf{H}_0}$ by updating the well-trained parameters $\lambda_{\mathbf{H}_1}$ of the universal model using the incoming matching keypoint set $\mathbb{B}_{H_0}$. Bayesian adaptation is a popular modeling approach in speech and speaker recognition [17]. In this work, Bayesian adaptation is able to furnish sufficient prior knowledge about the distribution of keypoints via the universal model, and the consequent adaptation to matching keypoints may elegantly guarantee desirable discriminative power of the likelihood ratio test.

Estimating model $p(\mathbf{z}_t|\lambda_{\mathbf{H}_1})$. Given the training set $\mathbb{B}_{H_1}$, we adopt a GMM model to learn the distribution of keypoint features $\mathbf{z}_t$ as: $p(\mathbf{z}_t|\lambda_{\mathbf{H}_1}) = \sum_{c=1}^C \tilde{\omega}_c p_c(\mathbf{z}_t)$, where $\lambda_{\mathbf{H}_1} = \{\tilde{\omega}_c, \tilde{\mu}_c, \tilde{\sigma}_c^2\}_{c=1}^C$, C the number of Gaussian components. The covariance matrices are assumed to be diagonal and the variance vector is denoted as $\tilde{\sigma}_c^2$. We learn the parameters $\lambda_{\mathbf{H}_1}$ by maximizing the likelihood of $\mathbb{B}_{H_1}$.

Estimating model $p(\mathbf{z}_t|\lambda_{\mathbf{H}_0})$. Given the match keypoint set $\mathbb{B}_{H_0}$ and the learnt universal model $p(\mathbf{z}_t|\lambda_{\mathbf{H}_1})$, we perform Bayesian adaptation [17] in twin-stage iteration. The first step is identical to the expectation step of EM algorithm, which uses B_0 keypoint samples $\mathbf{z}_t$ from $\mathbb{B}_{H_0}$ to calculate the sufficient statistics about the GMM parameters of weight, mean and variance:

$$n_c = \sum_{b=1}^{B_0} \gamma_b(c) \tag{9}$$

$$E_c(\mathbf{z}_t) = \frac{1}{n_c} \sum_{t=1}^{B_0} \gamma_t(c) \mathbf{z}_t \tag{10}$$

$$E_c(\mathbf{z}_t^2) = \frac{1}{n_c} \sum_{t=1}^{B_0} \gamma_t(c) \mathbf{z}_t^2, \tag{11}$$

where $\gamma_t(c) = \frac{\tilde{\omega}_c p_c(\mathbf{z}_t)}{\sum_{\tilde{c}=1}^C \tilde{\omega}_{\tilde{c}} p_{\tilde{c}}(\mathbf{z}_t)}$ denotes the soft assignment of keypoint $\mathbf{z}_t$ to Gaussian c.

The second step is to apply the above sufficient statistics from $\mathbb{B}_{H_0}$ to update the parameters $\{\tilde{\omega}_c, \tilde{\mu}_c, \tilde{\sigma}_c^2\}$. The adapted parameters $\lambda_{\mathbf{H}_0} = \{\hat{\omega}_c, \hat{\mu}_c, \hat{\sigma}_c^2\}_{c=1}^C$ is derived as follows:

$$\hat{\omega}_c = \alpha_c^w n_c / B_0 + (1 - \alpha_c^w) \tilde{\omega}_c \tag{12}$$

$$\hat{\mu}_c = \alpha_c^s E_c(\mathbf{z}_t) + (1 - \alpha_c^s) \tilde{\mu}_c \tag{13}$$

$$\hat{\sigma}_c^2 = \alpha_c^t E_c(\mathbf{z}_t^2) + (1 - \alpha_c^t)(\tilde{\sigma}_c^2 + \tilde{\mu}_c^2) - \hat{\mu}_c^2, \tag{14}$$

where α_c^w, α_c^s and α_c^t are adaptation coefficients to control the impact of universal model on parameters updating. For example, when α_c^s is large, the statistics $E_c(\mathbf{z}_t)$ from matched keypoints tend to dominate in Eq. 13. In this work, we define the coefficients $\alpha_c^\rho, \rho \in \{w, s, t\}$ as the ratios $\alpha_c^\rho = \frac{n_c}{n_c + \pi^\rho}$, where π^ρ is a constant relevance factor for parameter ρ and n_c is defined in Eq. 9.

Discussion. When we select a threshold to make the likelihood ratio test $\tau \to \infty$, the RFV model degenerates to standard FV. Both RFV and FV have $K(1 + 2d)$ parameters for the appearance model, and the descriptor size Kd. However, RFV model introduces $2C(1+2z)$ more parameters for the hypothesis models, where $z = 4$ is the dimension of keypoint features $\mathbf{z}_t$. In our experiments, we adopt $C = 32$ Gaussian components for the hypothesis models, and we will show that a few additional parameters bring about significant performance improvements.

5. EXPERIMENTS

5.1. Datasets and experiment setup

Datasets: We conduct comparison experiments over two popular benchmark datasets and a public available dataset from MPEG CDVS evaluation framework.

The **UKbench** dataset contains the images of 2,550 objects, each with 4 images taken from varied viewpoints. All the 10,200 images are indexed as reference images. The retrieval performance is measured by mean average precision (mAP) of all 10,200 queries. Also, we report the average number N_s of relevant images in top 4 returns, which is the commonly used measure over this dataset [4].

The **Holidays** dataset is a collection of 1491 holiday photos. There are 500 image groups where the first image of each group is used as a query. The retrieval performance is measured by mAP.

The **Graphics** dataset is a subset of the Stanford mobile visual search dataset, involving 5 categories (CDs, DVDs, books, text documents and business cards). There are 1500 queries and 1000 reference images. The retrieval performance is measured by mAP.

To setup large-scale experiments, we use a **FLICKR1M** dataset containing 1 million distractor images collected from Flickr. This distractor set is merged with other testing datasets to evaluate the retrieval performance and efficiency over a large scale dataset. The

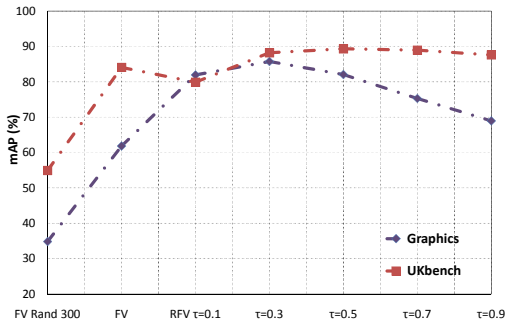

Fig. 2. Influence of the decision threshold τ on performance, and comparison of RFV and FV [7] on Graphics and UKbench datasets. $K = 64$ Gaussian components are used in all experiments.

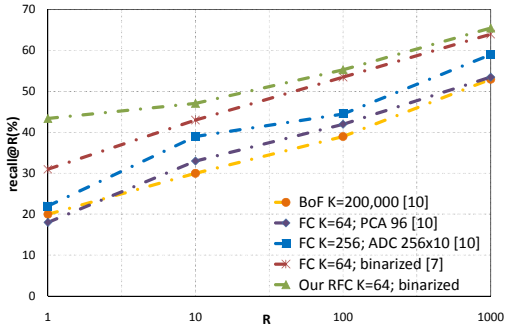

Fig. 3. Comparison of the proposed RFC and the state-of-the-arts [10][7], in terms of Recall@R over Holidays+FLICKR1M.

performance is measured by Recall@R, i.e. the rate of relevant images in top R returns. A higher recall rate indicates a better shortlist for subsequent geometric verification or other re-ranking process.

Experiment setup: All the images are resized with the longer dimension of less than 640. SIFT descriptors are extracted by the VLFeat library. For fair comparison, we reduce the dimension of SIFT feature from 128 to $d = 64$ using Principal Component Analysis (PCA), like the state-of-the-arts works [9][7][10]. Independent image datasets including the Oxford building and the Caltech building, are employed in all training stages.

5.2. Experiments on non-binarized RFV

The influence of τ. We first study the impact of varying decision thresholds τ on retrieval performance. As shown in Figure 2, the optimal τ for two datasets is slightly different. For instance, on the Graphics dataset, $\tau = 0.3$ yields the best mAP 85.7%, while $\tau = 0.5$ is optimal on the UKbench dataset. In hyphothesis ratio test, smaller τ leads to less keypoints. In Graphics, most queries depict complex scene, such as CD covers with background clutter, smaller τ may remove noise keypoints from clutter. In UKBench, the images usually consist of simple object and clean background, larger τ may filter in more keypoints of query objects to improve performance (i.e. more than 600 keypoints, versus 300 in Graphics).

RFV *vs.* FV Figure 2 shows that the proposed RFV outperforms FV significantly on both Graphics and UKbench datasets. For instance, on Graphics dataset, RFV yields the mAP 85.7% with $\tau = 0.3$ for the best and 68.9% with $\tau = 0.9$ for the worst, while FV achieves only 61.8%. To make clear the advantage of our RFV, we produce the results of FV aggregation by randomly sampling 300 SIFT descriptors from each query (FV Rand 300). We can see that the mAP of FV Rand 300 is much worse than standard FV, e.g. 34.8% *vs.* 61.8% on Graphics dataset. This demonstrates the power of keypoint learning in selective aggregation of RFV as well.

Comparison with the state-of-the-art. Table 1 compares the performance of RFV with BoF, VLAD [9] and FV [10][7] on two benchmark datasets: UKBench and Holidays. The proposed RFV outperforms BoF significantly. Compared to VLAD [9] and the standard FV [7][10], the results of RFV is much more precise when using comparable number of visual words or Gaussian components. For instance, when $K = 64$, our RFV achieves a mAP of 67.1% on Holidays dataset, while [9] reports 55.6% and [10] reports 59.5%.

5.3. Large scale retrieval experiments on binarized RFC

Figure 3 compares our RFC with the state of the arts [10][7] over Holidays dataset combined with distractor FLICKR1M, in terms of recall@R . There are two variants of FC in [10], the first one adopts PCA to reduce the dimension of FV ($K = 64$)from 4096 to 96 (PCA 96); the second one uses Product Quantization [18] to subdivide the FV ($K = 256$) into 256 subvectors where 2^{10} visual words are used to produce each subvector (256×10), the distance between query and database image is measured by Asymmetric distance computation (ADC). Both RFC and [7] employ sign binarization to compress FV ($K = 64$, binarized). As shown in Figure 3, the proposed RFC yields much better results than those reported in [7][10] for $K = 64$, especially for smaller R. For instance, for $K = 64$, our RFC yields a recall@1 of around 43% while [7] reports close to 31% .

Descriptor	dimension	UKbench	Holidays
BoF $K = 20,000$ [10]	20,000	2.87	43.7
BoF $K = 200,000$ [10]	200,000	2.81	54.0
VLAD,$K = 64$ [9]	4096	3.28	55.6
FV $K = 64$ [10]	4096	3.35	59.5
FV $K = 256$ [10]	16384	n/a	62.5
our RFV $K = 64$	**4096**	**3.44**	**67.1**
our RFV $K = 256$	**16384**	**3.58**	**72.6**

Table 1. Comparison of RFV with the state-of-the-arts [10][9], measured by mAP for Holidays and N_s score for UKbench. K denotes the number of visual words (or Gaussian components).

Timing Over 1 million image set, our RFC costs less than 0.5s per query on average for $K = 64$, due to the very fast Hamming distance computing. In contrast, the BoF with $K = 200,000$ visual words costs more than 0.6s per query on average. Moreover, the RFC based search avoids the time-consuming construction of inverted index file in the BoF retrieval framework.

6. CONCLUSION

We have proposed a robust Fisher code to improve the discriminative power of traditional Fisher vectors. The promising retrieval performance has been demonstrated over large-scale benchmark datasets. Moreover, the proposed RFC brings about the benefits of higher efficiency and lower memory complexity. Possible future work involves how to enhance the relevance between returned images and user intentions by taking the social network into consideration [19].

7. ACKNOWLEDGEMENT

This work was supported in part by the National Basic Research Program of China (2009CB320902), the Chinese Natural Science Foundation under Contract No. 61271311 and No. 61121002, and in part by the Research Fund of ZTE Corporation.

8. REFERENCES

[1] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[2] D. G. Lowe, "Distinctive image features from scale scale invariant keypoints," *IJCV*, 2004.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

[4] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006.

[5] 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*, 2008.

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007.

[7] Florent Perronnin and Christopher Dance, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*, 2010.

[8] Tommi S. Jaakkola and David Haussler, "Exploiting generative models in discriminative classifiers," in *NIPS*, 1999.

[9] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and et. al., "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[10] Hervé Jégou, Florent Perronnin, Matthijs Douze, and et. al., "Aggregating local images descriptors into compact codes," *PAMI*, 2012.

[11] Florent Perronnin and Christopher Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[12] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[13] Josip Krapac, Jakob Verbeek, and Frédéric Jurie, "Modeling spatial layout with fisher vectors for image categorization," in *ICCV*, 2011.

[14] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Image categorization using fisher kernels of non-iid image models," in *CVPR*, 2012.

[15] Matthijs Douze, Arnau Ramisa, and Cordelia Schmid, "Combining attributes and fisher vectors for efficient image retrieval," in *CVPR*, 2011.

[16] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *ECCV*, 2010.

[17] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, 2000.

[18] Hervé Jégou, Matthijs Douze, and Cordelia Schmid, "Product quantization for nearest neighbor search," *PAMI*, 2011.

[19] Shaowei Liu, Peng Cui, Huan-Bo Luan, and et. al., "Social visual image ranking for web image search.," in *MMM*, 2013.