Selectively Aggregated Fisher Vectors of Query Video for Mobile Visual Search

Xiaohe Zhang
[1]SECE of Shenzhen Graduate School
Peking University
Shenzhen, China
[2]The Institute of Digital Media
School of EECS, Peking University
Beijing, China
xiaohe_zhang@pku.edu.cn

Yitong Wang, Zhaoliang Liu, Ling-Yu Duan
The Institute of Digital Media
School of EECS, Peking University
Beijing, China
wangyitong@pku.edu.cn
zhaoliang_liu@pku.edu.cn
lingyu@pku.edu.cn

***Abstract*—Mobile visual search has undergone a wide development and gained much progress in recent years thanks to the ever-growing computational power of mobile devices. Most visual search methods take a single image as query and generate an image-level representation to implement image retrieval. To form a compact and discriminative representation for the query image, Fisher vectors (FV) have shown great advantage in both discriminability and computational efficiency. However, single image based visual search sometimes has unsatisfactory performance as a number of quality degeneration situations like limited view, uneven lighting, blur, occlusion and etc. may exist in the query image, while a video clip could overcome these shortcomings and contain more sufficient visual information for better retrieval performance when serving as query. Towards a compact yet discriminative representation of the query in mobile visual search, we propose a temporal-spatial based Fisher Vector (TSFV) for the query video with an equal length to an image based FV. The TSFV introduces a selective local feature aggregation scheme that employs interframe feature matching in temporal terms combined with intraframe feature attributes in spatial terms to evaluate video features' discriminability and select only the discriminative ones for aggregation. Evaluated on a diversified dataset, our proposed TSFV for query video achieves a significant performance improvement compared to typical image based FV with no additional transmission load and query latency.**

***Index Terms*—compact video representation, Fisher vector, feature selection, local feature aggregation.**

I. Introduction

With the popularization of camera embedded mobile devices and wireless Internet services, nowadays there is an emerging potential in mobile visual search and related applications. Generally speaking, most state-of-the-art mobile visual search systems follow a client-server architecture. The remote server maintains a large-scale image database and the mobile user captures an image as query. In a typical scenario, the query image transmission from the mobile device to the remote server is often over a relatively slow, bandwidth-constrained wireless network. In such case, sending the entire image will cause a heavy burden for the wireless network and a high query latency. Coming with the ever growing computational power of mobile devices, recent works have proposed to directly extract low bit-rate visual descriptors on the mobile devices and send the descriptors to implement image search.

The mobile visual search community have made a lot of efforts towards low bit-rate image descriptors. In general, low bit rate descriptors can be broadly categorized into two groups. The first group is based on local descriptor, that includes the Compressed Histogram of Gradients (CHoG) descriptor [1] proposed by Chandrasekhar et al., compression schemes [2] for SIFT [3] or SURF [4] are also well exploited in the literature. The second group is based on global descriptor, which generate a compact image-level representation by aggregating local descriptors. BoW [5] is the most popular method, which quantize image local descriptors to BoW visual words and form a BoW histogram at the mobile end. To improve the retrieval performance and efficiency at much less memory complexity, Perronnin et al. [6] introduced Fisher Kernel [7] to image retrieval. Given an image, Fisher kernel aggregates the local descriptors to form a Fisher vectors (FV) representation of fixed-length. FV achieves stronger discriminative power by employing higher order statistics compared to BoW and reduces the computational complexity significantly. Jegou et al. [8] proposed a simplified FV, the Vector of Locally Aggregated Descriptors (VLAD) and promising results have been reported.

To guarantee the descriptor's discriminative power and compactness, Gianluca et al. [9] put forward a feature selection scheme, advocating to pack only the most informative local features into the image descriptor and discard the noisy ones. This scheme evaluates a feature's discriminability by estimating the probability that the feature will yield a correct match in retrieval tasks. And the probability estimation is based on several trained models with respect to features' spatial attributes like scale, peak value, orientation and space location. Similarly with this feature selection idea, Lin et al. [10] proposed a robust Fisher vectors (RFV) representation, implementing selective local feature aggregation, which has incorporated a feature selection mechanism into the FV aggregation stage based on feature attributes.

However, the image-based mobile visual search still can't meet the requirement of a good retrieval performance in

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.45

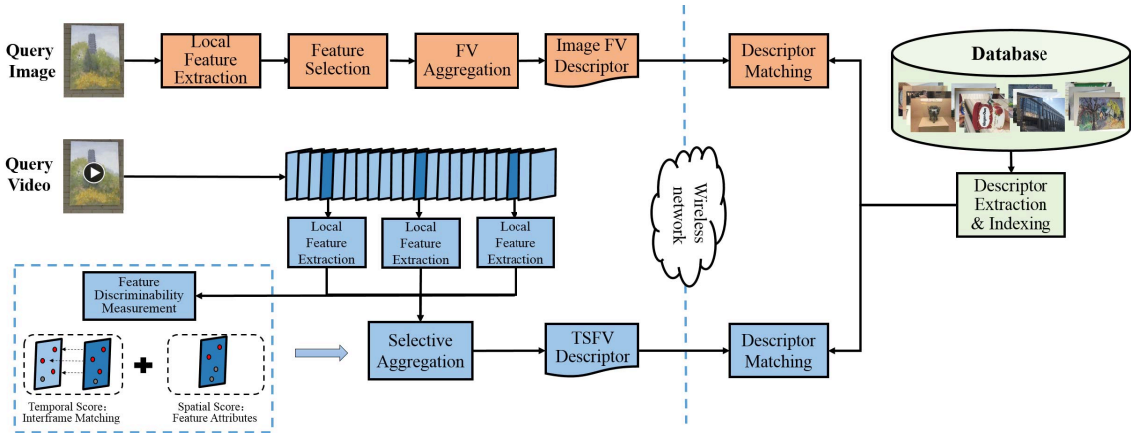

Fig. 1. The proposed TSFV generation framework and the comparison with an image based visual search scheme.

many cases, as typical quality degeneration like blur, uneven lighting, limited view, foreground occlusion as well as background clutter are unavoidable. All these factors lead to the insufficiency of discriminative and stable feature points in the query image. As a result, the image retrieval performance is unsatisfactory.

A video clip could well overcome the shortcomings mentioned above of a single image query, as multiple frames within a video can cover diversified photographing conditions and avoid just taking an image of poor quality as query. What's more, a short video clip contains much more abundant visual information about the target object by certain movement of the camera or the target object during the process of photographing. Towards video content representation, there have been a series of works. Chen [11] proposed to select a feature-rich frame from a video clip and then initiate image querying based on the selected frame. Mansour [12] has adopted a Non-negative Matrix Factorization (NMF) method to compress the features extracted from video frames. Sarkar [13] proposed to extract keyframes from a video clip and use the keyframe set to represent the video. However, there has been few work on developing a compact global descriptor including the abundant visual content of query video for efficient transmission and retrieval.

In this paper, we propose a compact yet discriminative Fisher vectors representation for query video, which is selectively aggregated by features detected from multiple frames of the query video. It includes much more abundant visual information for better image retrieval performance, but has an equal length to an image based FV descriptor. To select the discriminative features of the query video for the FV aggregation and discard the noisy ones, making the FV descriptor have a strong discriminative power, we introduce a novel feature selection scheme, which employs interframe feature matching in temporal terms combined with intraframe feature attributes in spatial terms to evaluate a feature's discriminability and decide whether the feature will be involved in the FV aggregation stage, forming a Temporal-Spatial based Fisher Vector (TSFV). Experimental results show that our feature selection approach is effective in improving the video FV representation's discriminative power and the TSFV of a query video significantly outperforms the FV descriptor of a single image in visual search performance with an equal bit rate.

The rest of this paper is organized as follows. Firstly we review related works in section 2. Then the proposed algorithm for TSFV aggregation and feature selection is presented in section 3. Section 4 analyzes experimental results, followed by the conclusions in section 5.

II. Related Work

A. Image global descriptors

Global descriptors provide compact image-level representations, usually with a fixed length, efficient for transmission and storage.

BoW. The bag-of-words (BoW) [5] is the most widely adopted method for global representation. Each local feature from an image is quantized to its closet visual word in a visual vocabulary. BoW accumulates the number (0-order statistics) of local features assigned to each visual word and form a histogram based on the statistics data.

FV. The Fisher vector (FV) [7] extends the BoW by computing higher-order statistics of the distribution of local features, e.g., Gaussian Mixture Model (GMM). Specifically, FV aggregates the gradient vector of each local feature's likelihood with respect to the GMM parameters (mean or variance) for each Gaussian. FV is finally formed by concatenating the 1-order and/or 2-order statistics of all Gaussian components. Compared to the BoW, FV achieved better retrieval performance at a much smaller visual vocabulary.

To further improve the FV representation's discriminative power, Lin et al. proposed a robust Fisher vector (RFV) [10], which incorporates a feature selection scheme in the stage of local descriptor aggregation. RFV selects the most discriminative local features to be involved in FV aggregation and discard the noisy ones. The selective aggregation significantly reduced the negative impact of noisy local descriptors on

the FV discriminability and showed superior performance in retrieval tasks.

VLAD. Jegou et al. proposed a simplified Fisher kernel representation [8] with limited dimension, namely, the Vector of Locally Aggregated Descriptors (VLAD), to aggregate residual vectors (difference between local feature and its closest visual word). VLAD has demonstrated promising results.

However, all the above global descriptors concentrated on a compact and discriminative representation for a single image, while few work has attempted to extract such a global descriptor for a video clip to include the more abundant visual information.

B. Feature selection

For a compact yet discriminative descriptor, a fundamental problem is feature selection, as the original feature set detected from an image or a video frame may include many noisy and uninformative feature points, they can cause a negative impact on the discriminative power of the descriptor and interrupt the descriptor matching process.

The state-of-the-art feature selection method for descriptor generation is based on the statistical modeling of features' spatial attributes [9][10], such as scale, orientation, peak strength and location in the image. Each of the attributes conditions the likelihood that a feature may be matched up correctly at query time by the off-line trained models. So a feature ranking list can be generated by sorting the likelihood to facilitate feature selection for descriptor generation.

Besides, Xin et al. [14] proposed a self-matching based feature selection, which randomly applies an out-of-plane rotation to the target image and match the original features to the features extracted from the out-of-plane rotated image. Then the importance of the features is ranked according to the self-matching score and the top N features are selected. Cao et al [15] put forward a feature selection method based on the entropy of the image content, entropy of extracted features and the Discrete Cosine Transformation (DCT) coefficients.

C. Video representation

A variety of video content representation methods have been proposed in recent years, for applications like video retrieval, video classification, landmark recognition, mobile augmented reality and etc.

For an efficient representation of video playing on the TV, Chen et al. [11] proposed to select a feature-rich frame from the sequence of a short user-initiated query video clip. The feature-rich frame is selected by calculating the image-level Hessian score, which can reflect the number of robust local features. But this method transforms video representation into image representation by selecting a representative frame, failing to include the numerous visual information contained within multiple frames.

A typical approach for video content representation [12][13] is to extract keyframes from the video, detect robust local features from each keyframe and gather all the local features into a total representation, followed by certain transformation or data compression to reduce the amount of data. However, this kind of methods seem not appropriate for query video representation in mobile visual search, because it is not compact enough for mobile wireless transmission.

Another kind of approaches to represent the video content are to extract features frame by frame, and to efficiently reduce the transmission bit rate there have been works on interframe predictive coding schemes [16][17][18], which exploit the temporal correlation of local descriptors and location coordinates between adjacent video frames. This kind of approaches may take any individual frame as query for image retrieval and are mainly designed for applications of mobile augmented reality and feature tracking, but have not provided a compact global signature of the whole video.

Algorithm 1 Generating a Discriminative Fisher Vectors Representation for a Query Video.

Input:

Frame Corpus V_q=$\{I_s\}_{s=1}^{S}$ of a Video Clip.

Output:

A Temporal-Spatial based Fisher Vector (TSFV).

1: Sample frames from V_q by an interval of τ, forming a subset of frames as $\hat{V_q}$=$\{I_{k\tau}\}_{k=1}^{K}$;
2: **for** each $I_{k\tau} \in \hat{V_q}$ **do**
3: Extracting local features $\{z_t^k\}_{t=1}^{T_k}$ from $I_{k\tau}$;
4: Assign a discriminability score for each feature z_t^k by Eq.11 ;
5: Sort features in descending order of discriminability value in frame $I_{k\tau}$ as Eq.9 ;
6: Decide the $sig(\cdot)$ value of each feature by Eq.10 based on the feature ranking list;
7: **end for**
8: Generate the temporal-spatial based Fisher vector (TSFV) for query video by Eq.7 ;

III. Proposed TSFV for query video

In this paper, we extend the image based FV representation to describe the visual content in a query video clip, generating a compact Fisher vectors representation of query video for mobile visual search. As shown in Fig.1, to improve the discriminative power of the video FV representation, we inject a featute selection scheme into the FV aggregation stage, which employs the interframe feature matching in temporal terms and intraframe feature attributes in spatial terms to measure a feature's discriminability and decide whether the feature will be involved in the aggregation stage, forming a temporal-spatial based Fisher vector (TSFV).

A. Brief review of FV

To formulate the problem of generating a Fisher vectors representation for a video clip, we briefly review the traditional image Fisher vectors representation. Let $I = \{z_t\}_{t=1}^{T}$ denote a set of T local features extracted from image I, $z_t \in R^d$, where d denotes the dimensionality of the local descriptor. The local features are aggregated by an offline trained GMM with

M Gaussian components: $p(z_t|\lambda) = \sum_{m=1}^{M} \omega_m p_m(z_t), \lambda = \{\omega_m, \mu_m.\sigma_m^2\}$, where ω_m, μ_m and σ_m^2 denote the weight, mean and diagonal variance matrix of Gaussian component m, respectively. The GMM parameters λ are learned through maximizing the likelihood of training images by the well-known Expectation-Maximization (EM) algorithm. Then the Fisher vectors representation is defined on the gradient vectors. The log-likelihood of image I is obtained by averaging the log-likelihood values of local descriptors:

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

Let $\mathcal{G}_m^I$ denote the d-dimensional gradient vector with respect to the m_{th} Gaussian component. The analytical form of $\mathcal{G}_m^I$ is derived as:

$$\mathcal{G}_m^I = \frac{\partial \mathcal{L}(I|\lambda)}{\partial \mu_m} = \frac{1}{\sqrt{T\omega_m}}\sum_{t=1}^{T}\gamma_t(m)\sigma_m^{-1}(z_t - \mu_m) \tag{2}$$

where $\gamma_t(m) = \frac{\omega_m p_m(z_t)}{\sum_{j=1}^{M}\omega_j p_j(z_t)}$ denotes the probability for local descriptor z_t being generated by the m_{th} Gaussian component. Finally, the Fisher vector $\mathcal{G}_I$ is formed by concatenating the aggregated gradient vectors $\mathcal{G}_m^I$ of all Gaussian components, $m = 1...M$.

B. Problem formulation

Given a query video clip of S frames $V_q = [I_1, I_2, ..., I_S]$, a large video feature set can be generated as

$$V = \{\{z_t^1\}_{t=1}^{T_1}, \{z_t^2\}_{t=1}^{T_2}, ..., \{z_t^S\}_{t=1}^{T_S}\} \tag{3}$$

where z_t^s denotes the t_{th} local feature in the s_{th} frame, T_s represents the number of local features detected from the s_{th} frame. Therefore, for the video FV aggregation, the log-likelihood of video V is obtained by averaging the log-likelihood values of all the video features:

$$\mathcal{L}(V|\lambda) = \frac{1}{T_{all}}\sum_{s=1}^{S}\sum_{t=1}^{T_s} log\ p(z_t^s|\lambda) \tag{4}$$

where T_{all} denotes the total number of features detected from every frame.

$$T_{all} = T_1 + T_2 + ... + T_S \tag{5}$$

Let $\mathcal{G}_m^V$ denote the d-dimensional gradient vector with respect to the m_{th} Gaussian component. The analytical form of $\mathcal{G}_m^V$ is derived as:

$$\mathcal{G}_m^V = \frac{\partial \mathcal{L}(V|\lambda)}{\partial \mu_m} = \frac{1}{\sqrt{T_{all}\omega_m}}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\gamma_t(m)\sigma_m^{-1}(z_t^s - \mu_m) \tag{6}$$

Then the video Fisher vector $\mathcal{G}_V$ is formed by concatenating the aggregated gradient vectors $\mathcal{G}_m^V$ of all Gaussian components, $m = 1...M$.

However, there is heavy temporal redundancy as well as many noisy feature points in the video feature set V. For a discriminative FV representation of the video clip, also considering the computational efficiency, we inject a local feature selection function $sig(\cdot)$ into the FV aggregation framework, making the video FV only aggregate discriminative features in V and discard the uninformative ones:

$$\hat{\mathcal{G}}_m^V = \frac{1}{\sqrt{\hat{T}\omega_m}}\sum_{s=1}^{S}\sum_{t=1}^{T_S} sig(z_t^s)\gamma_t(m)\sigma_m^{-1}(z_t^s - \mu_m) \tag{7}$$

$\hat{\mathcal{G}}_m^V$ is the selectively aggregated gradient vector with respect to the m_{th} Gaussian component. $sig(z_t^s)$ is a binary feature selection function, which indicates z_t^s will be involved in the FV aggregation when it has the value of 1, otherwise z_t^s will be discarded when it has the value of 0. $\hat{T}$ denotes the number of features selected for FV aggregation from the originally detected video feature set V.

C. Video feature selection

We adopt a two-step feature selection method to decide which features to be aggregated for the video FV, the first step is frame selection and the second step is feature selection of the selected frames.

In the first place, for the purpose of removing heavy temporal redundancy between adjacent frames, a subset of frames, namely keyframes, are extracted from the original video sequence. As the simplest way, we pick out a frame from every fixed-length interval τ, and the step τ is a predefined parameter and can be varied. As a result, the sampled frame subset is represented as:

$$\hat{V}_q = [I_\tau, I_{2\tau}, ..., I_{K\tau}] \tag{8}$$

Then feature selection will be implemented on the selected K frames. For each $I_{k\tau}$ of the K frames, a set of local features are detected as $I_{k\tau} = \{z_t^{k\tau}\}_{t=1}^{T_{k\tau}}$, and each feature will be assigned with a discriminability value $d(z_t^{k\tau})$, which is measured by the approach we will introduce in the next subsection. Then features in frame $I_{k\tau}$ are ranked in descending order of discriminability value:

$$d(z_1^{k\tau}) \geq d(z_2^{k\tau}) \geq d(z_3^{k\tau}) \geq ... \geq d(z_{T_{k\tau}}^{k\tau}) \tag{9}$$

the first N features in the ranking list of frame $I_{k\tau}$ will be selected to be involved in the video FV aggregation. The number N of selected features from each keyframe is a predefined parameter and can be varied.

Therefore, the binary feature selection function $sig(z_t^s)$ in Eq.7 can be determined as:

$$sig(z_t^s) = \begin{cases} 1, & s = k\tau \text{ and } t \leq N \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

D. Temporal-spatial based feature discriminability measurement

In this subsection, we introduce the measurement definition of a video feature's discriminability, which employs interframe feature matching in temporal terms combined with intraframe feature attributes in spatial terms, and this temporal-spatial based measurement provides a novel criterion for feature selection. For any feature z_t^i in the i_{th} frame of a video, its

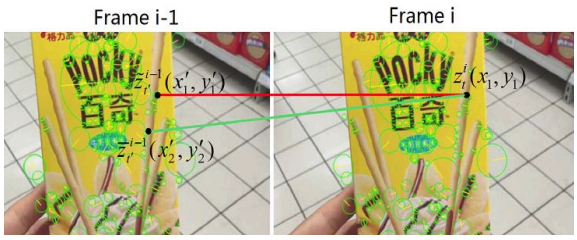

Fig. 2. An illustration of feature matching between neighboring frames. z_t^i is a feature detected from frame i, $\tilde{z}_{t'}^{i-1}$ denotes its closest neighbor in frame i-1, corresponding to the same keypoint of the target object, while $\bar{z}_{t'}^{i-1}$ denotes the second-closest neighbor in frame i-1. The closest and second-closest neighbor are linked to z_t^i respectively by a red and a green line.

discriminability value $d(z_t^i)$ is measured by a scoring function as:

$$d(z_t^i) = f(z_t^i) * h(z_t^i) \tag{11}$$

$f(z_t^i)$ is the interframe matching based score in temporal terms, and it is also the key point we want to introduce in this paper. $h(z_t^i)$ is the intraframe feature attributes based score in spatial terms.

$\mathbf{f(z_t^i)}$. As is often the case in a video clip, neighboring frames have a high degree of overlap in visual content, that is, they are very likely to depict the same object(s) with only slight difference. Therefore, unlike noisy features, the discriminative and stable features detected from the target object can find their corresponding features with very similar descriptors and locations in neighbouring frames, as illustrated in Fig.2, a pair of corresponding features of the same keypoint from two neighboring frames are linked by the red line.

For each of the detected features z_t^i (no matter discriminative or noisy features) in frame i, we find its closest and second-closest neighbor in terms of SIFT descriptor in the previous frame and calculate two quantities about it, denoted as $DDR(DescriptorDistanceRatio)$ and $NSD(NomalizedSpaceDistance)$ respectively,

$$DDR = \frac{\rho(z_t^i, \tilde{z}_{t'}^{i-1})}{\rho(z_t^i, \bar{z}_{t'}^{i-1})} \tag{12}$$

where $\tilde{z}_{t'}^{i-1}$, $\bar{z}_{t'}^{i-1}$ respectively represent the closest and second-closest neighbor of z_t^i in frame $i-1$, $\rho(\cdot,\cdot)$ represents the Euclidean distance of features' SIFT descriptors. DDR denotes the descriptor distance ratio between a feature to its closest neighbor and to its second-closest neighbor, reflecting a feature's discriminability.

$$NSD = \frac{d(z_t^i, \tilde{z}_{t'}^{i-1})}{\frac{1}{T}\sum_{t=1}^{T} d(z_t^i, \tilde{z}_{t'}^{i-1})} \tag{13}$$

where $d(\cdot,\cdot)$ represents the space distance between two coordinates, as shown in Fig.2, for the two features z_t^i and $\tilde{z}_{t'}^{i-1}$:

$$d(z_t^i, \tilde{z}_{t'}^{i-1}) = \sqrt{(x_1 - x_1')^2 + (y_1 - y_1')^2} \tag{14}$$

where (x_1, y_1), (x_1', y_1') respectively denote the coordinates of z_t^i and $\tilde{z}_{t'}^{i-1}$. NSD reflects the degree of closeness of a feature

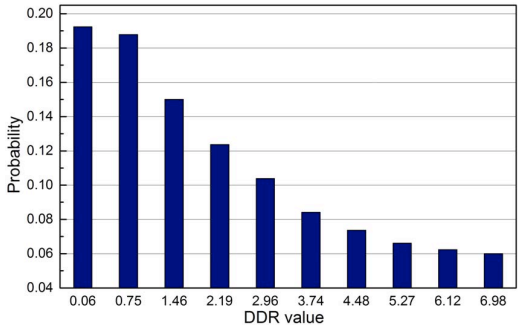

Fig. 3. Probability for being a discriminative feature as a function of quantized DDR value.

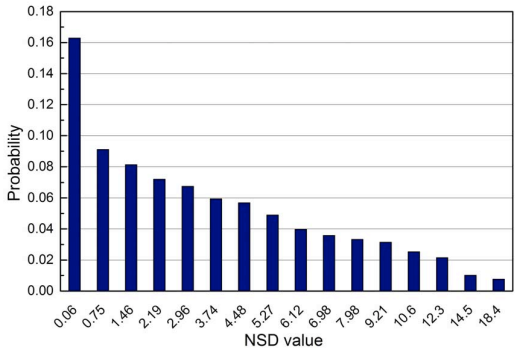

Fig. 4. Probability for being a discriminative feature as a function of quantized NSD value.

to its nearest neighbor in space location, and also indicates the feature's stability.

Obviously, discriminative and stable features detected from target object tend to have smaller values in these two quantities compared to noisy features. We want to figure out the statistical models that demonstrate the percentage of discriminative features account for in all features at different values of DDR and NSD. Hence, when given a feature, by calculating its DDR and NSD we can estimate the probability that it is a discriminative feature, in other words, its estimated discriminative power. For this goal, we have adopted a training set containing 3756 image pairs to implement image pairwise matching, each image pair including a video frame and a relevant image depicting the same object. We employ ratio test [3] followed by geometric consistency check based on RANSAC [19] to detect inliers in the video frames. The inliers will be taken as discriminative features, labeled as $l = 1$, and the other detected features are taken as noisy features labeled as $l = 0$.

Then, the percentage of discriminative features account for in all features at different clustered DDR/NSD values can be calculated as:

$$p(l = 1 | ddr \in DDR_i) = \frac{Num(l = 1 \cap ddr \in DDR_i)}{Num(ddr \in DDR_i)} \tag{15}$$

Fig. 5. Sample items from different categories of the dataset.

$$p(l=1|nsd \in NSD_i) = \frac{Num(l=1 \cap nsd \in SD_i)}{Num(nsd \in SD_i)} \quad (16)$$

where ddr and nsd respectively represent a certain value of DDR and NSD , DDR_i and NSD_i represent the i_{th} clustering center of the DDR and NSD quantities. Plots of the statistical results respectively with respect to DDR and NSD are shown in Fig.3 and Fig.4. And the interframe matching based score $f(z_t^i)$ of feature z_t^i is obtained by referring to the two statistic models based on feature's calculated ddr and nsd and multiply them.

$$f(z_t^i) = p(l=1|ddr \in DDR_i) * p(l=1|nsd \in NSD_i) \quad (17)$$

$\mathbf{h(z_t^i)}$. $h(z_t^i)$ is the intraframe feature attributes learned score for z_t^i , which is a likelihood function involving features' spatial attributes quantities. In this work, we focus on SIFT features, and a feature $z_t^i = (\eta, \theta, \upsilon, \xi)$ is of four dimensions, where η, θ, υ and ξ denote scale, orientation, peak value in scale space and the distance from a keypoint to the image center, respectively. In [9] it has been demonstrated that each of these quantities conditions the probability that a feature is a discriminative feature and will be correctly matched at query time. The probability statistical models with respect to the four attributes are also off-line trained.

Therefore, the attributes learned score $h(z_t^i)$ can be obtained by multiplying the four conditional probabilities:

$$h(z_t^i) = r_s(z_t^i) \cdot r_o(z_t^i) \cdot r_p(z_t^i) \cdot r_d(z_t^i) \quad (18)$$

where $r_s(\cdot)$, $r_o(\cdot)$,$r_p(\cdot)$, $r_d(\cdot)$ represent the individual conditional probability based on feature scale, orientation, peak value and space location, respectively.

IV. Experiments

Dataset and Evaluation metrics. To evaluate the performance of our proposed TSFV of query video, we carry out retrieval experiments on a dataset that consists of 230 query video clips depicting 115 items (2 video clips of each item),

TABLE I
STATISTICS OF THE ADOPTED DATASET IN THE EXPERIMENTS

Data category	# of query images	# of query videos	# of database images
Antiques	100	20	10
Books	50	10	5
Common objects	50	10	5
Cosmetics	100	20	10
Documents	100	20	10
Landmarks	250	50	25
Packaged cosmetics	50	10	5
Paintings	100	20	10
Signs	50	10	5
Supermarket foods	150	30	15
Wine labels	150	30	15
UKBench	-	-	10,200
In Total	1150	230	10,315

which can be divided into 11 categories, including books, cosmetics, common objects, landmarks in Peking University and etc. These videos are captured with hand-held mobile phone cameras with different amounts of camera motion, zoom, pan, rotation and perspective changes. Each video is about 150 frames long, recorded at 30fps with resolution of 640×480. Besides, each query video corresponds to an average number of 6 query images captured at the same time depicting the same object, for performance comparison of image based visual search and video based visual search. And for each item of the query set, a correspondence database image is provided and the groundtruth annotation is precisely established. The above dataset is constructed by ourselves, as those commonly used benchmark datasets like INRIA Holidays and UKBench can't provide corresponding query videos and query images of a query item at the same time for performance comparison. And to fairly evaluate the retrieval performance over a larger dataset, we use the UKBench (containing 10200 images, similar to our experiment dataset which also contain many items of small objects) as the distractor dataset to be mixed with the reference images. Details about the experimental dataset are illustrated in Table 1 and Fig.5.

The retrieval performance is measured by Recall@R, suc-

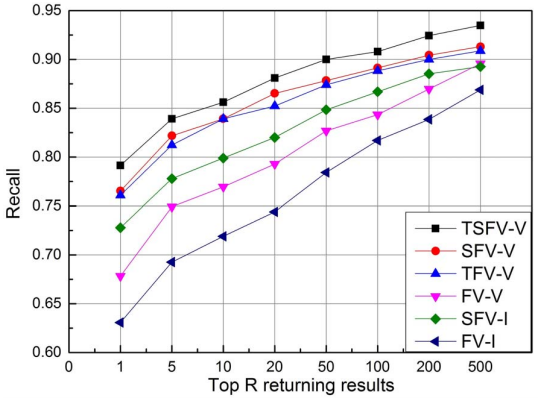

Fig. 6. The retrieval performance comparison in terms of Recall@R of the 6 baselines.

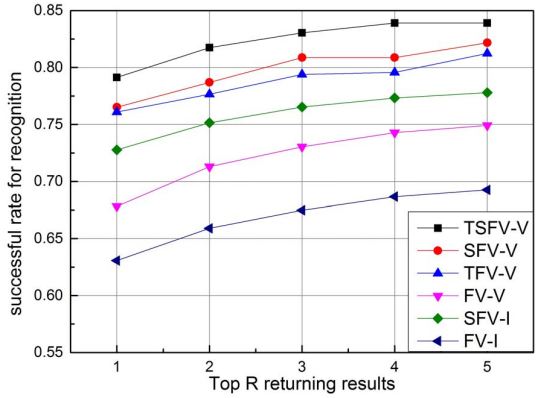

Fig. 7. The retrieval performance comparison in terms of successful rate for top R match of the 6 baselines.

cessful rate for top R match and mean Average Precision (mAP). Recall@R denotes the ratio between the number of relevant items among the identified items and the total number of relevant items at some given cut-off rank R. Top R match represents that there has relevant image been identified among the top R returning results. mAP is the mean precision of a batch of queries, each of which reveals its position-sensitive ranking precision:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{\sum_{r=1}^{M} P(r) * rel(r)}{\#of\ relevant\ images}\right) \qquad (19)$$

where Q is the number of queries, M is the size of retrieved image list for the ith query, $P(r)$ is the precision at r and $rel(r)$ is a binary function that it has the value of 1 only when the r_{th} returning result is a relevant image.

Implementation details. In our experiments the parameters are set as: frame sampling step τ=25, which is a trade-off between the coverage of video contents and computation complexity, the number of selected features from each keyframe N=300, based on the experience of MPEG-7 Compact Descriptors for Visual Search (CDVS) standard [20].

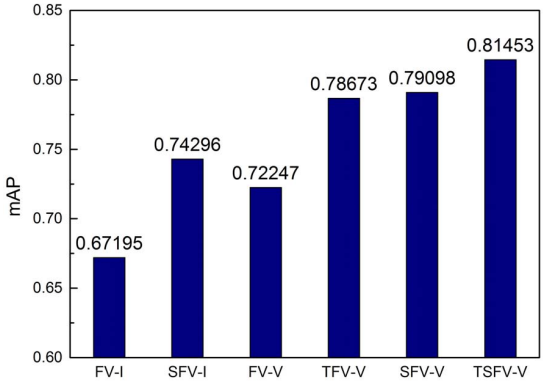

Fig. 8. The retrieval performance comparison in terms of mAP of the 6 baselines.

To improve the search efficiency and reduce descriptor storage, we compress the high-dimensional raw Fisher vectors into binary codes by an optimized Fisher vector binarization method [21], which is adopted in the MPEG CDVS standard.

Baselines. To compare the retrieval performance between the image based retrieval and video based retrieval, and also to demonstrate the feature selection's positive impact on the discriminative power of the aggregated FV representation, we implement the following baseline experiments: (1). *FV of Image(FV-I)*: Aggregating all features detected from an image to form the image Fisher vectors representation. (2). *Spatial based selective FV of Image(SFV-I)*: Based on image features' spatial attributes, implement feature selection, and aggregate the selected features to form the image Fisher vectors representation. (3). *FV of video(FV-V)*: Aggregating all features detected from keyframes to form the video Fisher vectors representation. (4). *Temporal based selective FV of video(TFV-V)*. Only based on interframe feature matching to implement video feature selection and aggregate the selected features to form the video FV representation (5). *Spatial based selective FV of video(SFV-V)*. Only based on intraframe feature attributes to implement video feature selection and aggregate the selected features to form the video FV representation. (6). *Temporal-Spatial based selective FV of video(TSFV-V)*. Based on both interframe feature matching and intraframe feature attributes to implement feature selection and aggregate the selected features into a video FV representation.

Performance comparison. The retrieval performance comparison of the 6 baselines in terms of Recall@R, successful rate for top R match and mAP are shown in Fig.6, Fig.7 and Fig.8, respectively. It can be seen that our proposed TSFV for qurey video significantly outperforms the conventional SFV of a single query image by a 7.16% promotion in mAP, under the condition that they all occupy the same descriptor length.

By comparing baseline *TSFV-V*, *TFV-V*, *SFV-V* and *FV-V*, we can see that feature selection plays an important role in enhancing the discriminative power of the video Fisher vectors representation and improving the image retrieval performance. Besides, interframe feature matching has a similar effect

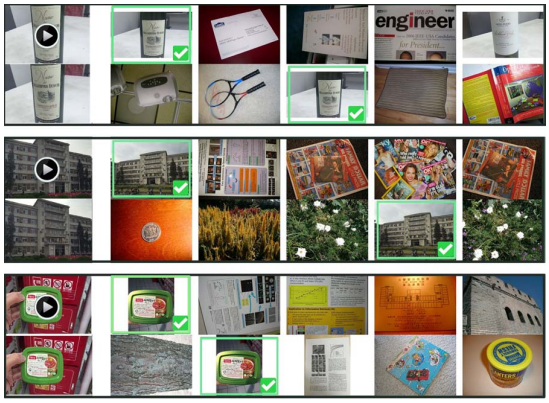

Fig. 9. Three groups of retrieval performance of TSFV-V in comparison to SFV-I. Of each group, the above: TSFV-V; the below: SFV-I. Each line corresponds to a query with top 5 dataset images returned. Green boxes indicate relevant image.

with the feature spatial attributes on video feature selection when used alone respectively. And when combining the two scoring criterion for feature selection, we obtained the best visual search performance and that is a 2.36% promotion in mAP compared to the state-of-the-art attributes based feature selection method.

Case study. Figure 9 illustrates three groups of exemplar image retrieval results, each group including two lines of results which are respectively queried by TSFV of a query video and SFV of a single query image. In these cases, the image query failed to find the relevant image at the top of the retrieved list, while a video query succeeded in recognizing the relevant image as the top return. It's obvious that our proposed TSFV of query video achieves significant retrieval performance promotion compared to conventional SFV of a singe query image.

V. Conclusion

This paper has provided a new mobile visual search solution that takes a short video clip as query and generates a compact yet discriminative temporal-spatial based Fisher vector (TSFV) for the query video. For the TSFV's generation, we have presented a novel selective local feature aggregation scheme, which employs interframe feature matching combined with intraframe feature attributes to evaluate features' discriminability and select only discriminative features for aggregation. Experimental results show that our proposed TSFV of a query video significantly outperforms traditional FV representation of a single query image for image retrieval, without incurring increase in transmission bit rate.

Acknowledgment

This work was supported by the National High-tech R&D Program of China (863 Program): 2015AA016302, and Chinese Natural Science Foundation: 61271311.

References

[1] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009, pp. 2504–2511.

[2] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," in *VCIP*, 2009.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features." *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 404–417, 2006.

[5] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[6] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010, pp. 3384–3391.

[7] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems*, 1998, pp. 487–493.

[8] H. Jgou, M. Douze, C. Schmid, and P. Prez, "Aggregating local descriptors into a compact image representation," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2010, pp. 3304–3311.

[9] G. Francini, S. Lepsy, and M. Balestri, "Selection of local features for visual search," *Signal Processing: Image Communication*, vol. 28, pp. 311–322, 2012.

[10] J. Lin, L. Y. Duan, T. Huang, and W. Gao, "Robust fisher codes for large scale image retrieval," in *International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 2013, pp. 1513–1517.

[11] D. Chen, N.-M. Cheung, V. C. Sam Tsai, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Dynamic selection of a feature-rich query frame for mobile video retrieval," in *IEEE International Conference on Image Processing(ICIP)*, 2010, pp. 1017–1020.

[12] H. Mansour, S. Rane, P. Boufounos, and A. Vetro, "Video querying via compact descriptors of visually salient objects," in *IEEE International Conference on Image Processing(ICIP)*, 2014, pp. 2789 – 2793.

[13] E. M. Anindya Sarkar, Pratim Ghosh and B.S.Manjunath, "Video fingerprinting: Features for duplicate and similar video detection and query-based video retrieval," in *SPIE*, 2008.

[14] X. Xin, Z. Li, Z. Ma, and A. K. Katsaggelos, "Robust feature selection with self-matching score," in *IEEE International Conference on Image Processing(ICIP)*, 2013, pp. 4363–4366.

[15] Y. Cao, C. Ritz, and R. Raad, "Adaptive and robust feature selection for low bitrate mobile augmented reality applications," in *International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2014, pp. 1–7.

[16] M. Makar, S. S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for mobile augmented reality," in *IEEE International Symposium on Multimedia*, 2012, pp. 50–57.

[17] M. Makar, S. S. Tsai, V. Chandrasekhar, D. Chen, and B. GIROD, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *International Journal of Semantic Computing*, vol. 7, no. 01, 2013.

[18] D. Chen, M. Makar, A. F. Araujo, and B. Girod, "Interframe coding of global image signature for mobile augmented reality," in *Data Compression Conference (DCC)*, 2014, pp. 33–42.

[19] J.Philbin, O. Chum, M.Isard, and J.Sivic, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2007, pp. 1–8.

[20] L. Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, "Overview of the mpeg-cdvs standard," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 179–194, 2015.

[21] Z. Wang, L. Y. Duan, J. Lin, J. Chen, T. Huang, and W. Gao, "Optimizing binary fisher codes for visual search," in *Data Compression Conference (DCC)*, 2015, p. 475.