Compact Descriptors for Visual Search

Ling-Yu Duan, Jie Lin, Jie Chen, Tiejun Huang, and Wen Gao

{lingyu, jlin, cjie, tjhuang, wgao}@pku.edu.cn

Institute of Digital Media,

School of Computer Science & Electronic Engineering,

Peking University,

Beijing 100871, China

Abstract: Promising visual object search technologies as well as numerous real-world applications have been developed. To ensure the application interoperability, the Moving Picture Experts Group (MPEG) has made great efforts in standardizing visual search technologies. Moreover, extraction and transmission of compact descriptors are valuable for next-generation mobile visual search applications. In this article, we review the significant progress of MPEG compact descriptors for visual search (CDVS) in standardizing technologies that will enable efficient and interoperable design of visual search applications. In addition, we report the location search and recognition oriented data collection and benchmark under MPEG CDVS evaluation framework.

Keywords: visual search, compact descriptors, standard, location recognition.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

This article has been accepted for publication in IEEE MultiMedia but has not yet been fully edited.
Some content may change prior to final publication.

Compact Descriptors for Visual Search

Ling-Yu Duan, Jie Lin, Jie Chen, Tiejun Huang, and Wen Gao

Institute of Digital Media, Peking University, Beijing 100871, China

{lingyu, jlin, cjie, tjhuang, wgao}@pku.edu.cn

Abstract: Promising visual object search technologies as well as numerous real-world applications have been developed. To ensure the application interoperability, the Moving Picture Experts Group (MPEG) has made great efforts in standardizing visual search technologies. Moreover, extraction and transmission of compact descriptors are valuable for next-generation mobile visual search applications. In this article, we review the significant progress of MPEG compact descriptors for visual search (CDVS) in standardizing technologies that will enable efficient and interoperable design of visual search applications. In addition, we report the location search and recognition oriented data collection and benchmark under MPEG CDVS evaluation framework.

Keywords: visual search, compact descriptors, standard, location recognition.

1 Introduction

Searching for a specific object in a collection of images has a variety of applications in location search, scene retrieval, landmark recognition, product search. Camera equipped mobile devices are becoming ubiquitous platforms of visual search applications such as Google Goggles, Tencent WeChat, etc.

A mobile visual search system normally transmits query images from the mobile end to the remote server, where visual search is performed over a reference image database. In wireless environment, the query response latency is subject to the network bandwidth. Sometimes it would take quite a few seconds to transmit a JPEG image (say, 30~40 KB) over a slow link. To reduce the query latency, an alternative is to extract visual features on the mobile client, and transmit the compact descriptors [1-5] to the remote server instead of images.

More recently, research efforts on compact descriptors have been done towards low bit rate mobile visual search [1-5]. Time consuming image query delivery is safely replaced by sending compact descriptors, which can significantly reduce network latency and improve user experience. Moreover, sending compact descriptors throughout 3G network may reduce power consumption of mobile devices [2] [5].

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

Existing research work in compact descriptors may be categorized into two groups. The first group works on compact local descriptors, in which vector or scalar quantization is employed to compress local descriptors. For instance, Chandrasekhar et al. proposed a Compressed Histogram of Gradient (CHoG) [1]. In geometric verification of image matching, Tsai et al. further proposed a grid-based quantization approach to code the spatial layout of local features [2].

The second group works on global descriptors. For example, research work has attempted to compress Bag-of-feature (BoF) signatures [3-5] or to aggregate the statistics of local descriptors. Chen et al. [3] proposed a Tree Histogram Coding scheme to compress the sparse BoF signature, which encodes the position difference of non-zero bins. Ji et al. proposed a multiple-channel coding based Compact Visual Descriptor (MCVD) to leverage mobile context (e.g. GPS, RFID tags) in encoding the sparse BoF signatures [4-5]. Recently, Jegou et al. [12] proposed the Vector of Locally Aggregated Descriptors (VLAD) approach to aggregate the visual word residuals. Chen et al. [10] introduced the Residual Enhanced Visual Vector (REVV), where linear discriminant analysis is employed to reduce the VLAD dimension, and then a sign binarization is applied to generate compact codes.

Both global and local descriptors contribute to the compact descriptors for visual search. The MPEG Compact Descriptor for Visual Search (CDVS) Ad-hoc group has developed leading compact descriptors as well as effective visual search pipelines. In particular, MPEG CDVS has addressed a practical issue of descriptor rate scalability. Instead of a fixed descriptor length, CDVS compact descriptors support the bit rate scalability to adapt the descriptor rate to the wireless bandwidth fluctuation.

In section 2, we introduce the progress of MPEG CDVS including important technical contributions. In section 3, we give insights of the CDVS compact descriptor, with more emphasis on the global descriptor aggregation. In section 4, we briefly discuss the pipelines of retrieval and pairwise matching. In section 5, we report the search performance of CDVS compact descriptors over the MPEG CDVS Benchmark including non-planar and planar object. Finally, we conclude this article in Section 6.

2. MPEG CDVS Standardization

2.1 Background

From the 92nd to the 96th MPEG meetings, the experts from academia and industry have widely investigated visual search applications, scope of standardization, requirements, as well as evaluation framework. At the 97th meeting, CDVS Call for Proposal was issued [9]. At the 106th meeting, CDVS entered the committee draft.

MPEG CDVS aims to define the format of compact visual descriptors as well as the pipeline of feature extraction and visual search process to enable interoperable

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

design of visual search applications. The CDVS Ad-hoc group came up with a competitive evaluation framework [9], which is in line with requirements.

The CDVS evaluation framework involves two types of experiments: retrieval and pairwise matching. The retrieval experiment is to evaluate descriptor performance in image retrieval. Mean average precision (mAP) and success rate for top match are measured. The pairwise matching experiment is to evaluate the performance in matching image pairs, which is measured by the success rate (i.e., True Positive Rate, TPR) at a given false alarm rate (False Alarm Rate, FPR) (say, 1%) as well as the localization precision. In particular, the descriptor scalability is evaluated by reporting the performance at six operating points (i.e. different descriptor lengths): 512B, 1KB, 2KB, 4KB, 8KB, and 16KB. Towards interoperability, a descriptor generated at any operating point shall allow matching with other operating points.

2.2 Test Model Development

A test model under consideration (TMuC) was determined at the 98th MPEG meeting [6]. Important stages including local descriptor extraction, feature selection, local descriptor compression, and location coding were identified [7].

After Test Model 1.0 was released, huge efforts have been made in developing or improving key technologies via core experiments including global descriptor aggregation, local descriptor compression, location coding, etc. in subsequent meeting cycles [10] [11] [14] [15] [16]. In particular, the global descriptor was introduced at the 100th meeting [10] to improve the performance of retrieval and pairwise matching, as well as search efficiency. However, the global descriptor in [10] cannot fulfill localization at the lowest operating point (512B), as no budget is allowed for local descriptors to support localization. At the 102nd meeting, a new global descriptor based on Fisher Vector aggregation, Compressed Fisher Vector (CFV), was proposed to achieve even more compactness, so that the operating point 512B allows good budget for local descriptors. Its further improvements, Scalable Compressed Fisher Vector (SCFV) [11], were adopted by Test Model.

Another important progress is made on interest point detector. Before Test Model 6.0 [26], Difference of Gaussian (DoG) [18] was employed, an approximation of Laplacian of Gaussian (LoG). To significantly reduce computational complexity and support parallel implementation, a new detector BFLoG [16] (Block based Frequency Domain Laplacian of Gaussian) was adopted by Test Model.

2.3 Core Experiments

In developing key techniques, eight core experiments were set up to investigate proposals in a competitive and collaborative platform, including global descriptor aggregation (CE1), local descriptors compression (CE2), location coding (CE3),

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

keypoint detection (CE4), local descriptor (CE5), retrieval pipeline (CE6), feature selection (CE7), and pairwise matching pipeline (CE8). CE1, CE2, CE3, and CE7 impact the normative part of MPEG CDVS standard. Table 1 lists the ever adopted key technologies in developing MEPG CDVS Test Model.

Table 1. Adopted key technologies in MPEG CDVS Test Model.

Core experiments	Adopted technologies	Technical features
CE1: Global descriptor	Scalable Compressed Fisher Vector (SCFV) [11]; Residual Enhanced Visual Vector (REVV) [10]	Fisher Vector based selective aggregation of local features; compact, discriminative, scalable in length; low computational complexity
CE2: Local descriptor compression	Transform + Scalar Quantizer [14]; Multi-stage Vector Quantizer (MSVQ) [13]	Very low memory footprint, scalable in descriptor length; both scalar and vector quantization are supported.
CE3: Location coding	Context based Location Coordinate Coding [15]	Lossy compression; the precision loss of coordinates does not impact performance.
CE4: Key point detector	Block based Frequency Domain Laplace of Gaussian Interest Point Detector [16]	Block wise processing, frequency domain filtering, low computational complexity.
CE6: Retrieval pipeline	Global Descriptor based Retrieval [10][11]; An Indexing Structure of Multi-Block Index Table (MBIT) [10] [11][17]	A scalable retrieval mechanism; the online update of database image index is supported.
CE7: Feature selection	Bayesian Learning based Feature Selection [6]	Filtering out less discriminative local features in query images.
CE8: Pairwise matching pipeline	DISTance RATio Coherence Test (DISTRAT) [6]; Global Descriptor Matching with Weighted Hamming Distance [10] [11]	Fast geometry verification; pairwise matching by combining global and local descriptors; learning based weighted Hamming Distance.

3 Compact Descriptor Extraction

A compact descriptor consists of an aggregated global descriptor and compressed local descriptors. Local descriptor compression minimizes the length of local visual descriptors, while coordinate coding minimizes the length of location coordinates of

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

each local visual descriptor. The global descriptor is generated by aggregating the statistics of local features to form a vector representation, which is subsequently compressed into compact codes to support efficient and effective matching and retrieval. Below we will introduce key technologies under CDVS Test Model.

3.1 Interest Point Detection

Compact descriptor extraction starts with interest point detection. Block-wise processing can be employed to reduce the memory complexity of the whole image based LoG filtering.. In addition, frequency domain filtering is favorable for hardware implementation. Thus, the block based frequency domain laplacian of Gaussian was adopted in Test Model 7.0 [16]. In [19], the block wise processing is further extended to extrema detection and orientation assignment.

The adopted detector BFLoG [16] works in the following steps.

Step 1). Generating the pyramid of an image with octaves.

Step 2). For each octave, do Steps (a) – (f)

a) Decomposing each octave into blocks.

b) Performing the block-wise filtering to produce the blocks of LoG response values as well as the blocks of Gaussian scaled values.

c) Recomposing the blocks of filtering results to form the LoG response images and the Gaussian scaled images.

d) Detecting scale-space extrema over the LoG response images.

e) Refining extrema output to locate interest points.

f) Computing the orientation of detected interest point.

Step 3). Outputting the union of detected interest points of all octaves, each interest point being characterized by scale, orientation, coordinates, as well as filter response values, which are applied to select local features for generating compact descriptors.

3.2 Feature Selection

The compact descriptor consists of a global descriptor and compressed local descriptors. Descriptor compactness relates to the number of local features and the size of compressed local features. On one hand, feature selection may reduce the

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

number of local features by removing noisy features. On the other hand, global descriptor aggregation may degenerate from noisy local features. The selective aggregation can significantly improve the discriminative power.

Empirical observation has shown that the characteristics of correctly matched interest points do not behave, in a statistical sense, as wrongly matched interest points. Moreover, the behavior difference pattern is consistent over heterogeneous datasets. Hence, a Bayesian learning based interest point selection strategies were developed and adopted into CDVS Test Model [6][7].

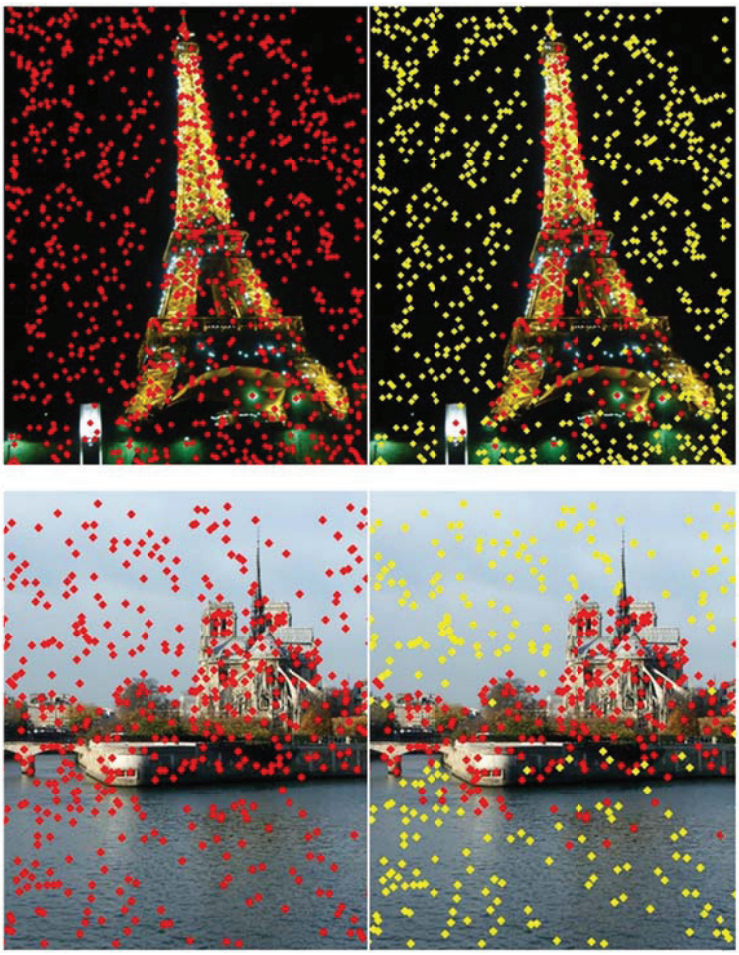


Figure 1. Selecting salient local features to generate compact descriptors (Red dots in Left Images: local features extracted from interest point detector; Yellow dots in Right images: noisy features removed by feature selection; Red dots in Right images: salient features kept after feature selection).

In training, the statistics of interest points from matching pairs of images are learnt, in terms of the characteristics of scale, orientation, coordinates and response values. From image matching pairs, training images generate a set of matched pairs of interest points. Over the training data of inliers and outliers, the statistical modelling attempts to predict the probability that a given local feature may be

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

matched correctly to a local feature in an unknown image.

In test, the characteristics of interest points in an input image are computed, and the predictor is employed to assign a probability value of correct match. Through ranking the prediction scores, a subset of local features are selected.

3.3 Local Descriptor Compression

Local descriptor compression is applied to the selected interest points. The adopted compression technique [14], namely, Transform + Scalar Quantization, first applies a predefined transform to the cell histograms of gradients, each cell histogram comprising eight bins. Based on sixteen cells, a transformed local feature descriptor shall contain all or a subset of 128 (16 x 8) transformed descriptor elements. For different target descriptor lengths, a subset of elements is formed according to a group of patterns. Recall that CDVS sets up six operating points. With the elements used in each descriptor length being the same or a subset of the elements used in the immediately higher descriptor length, simple transcoding to lower descriptor lengths by element elimination is facilitated. To fit the multiple descriptor lengths, the descriptor decimation for progressively lower descriptor lengths allows a good balance between the number of elements per descriptor and the number of descriptors that can be packed at a given descriptor length [14].

3.4 Coordinate Coding

The coordinates of local descriptors are important for geometric verification. Coordinate compression is useful for compact descriptors. Given a 640 × 480 image, 20 bits are needed to encode the location of a local feature without compression. Suppose an image contains 500 local features, to encode location needs 1250 bytes, which has been more than the lowest operating point (512 Bytes) in CDVS.

For the adopted coordinate coding techniques [15], a grid of 3×3 pixels is used for each bin, the positions of non-zero bins (“histogram map”) and the number of local features in non-zero bins (so called “histogram count”) are encoded separately. A simple arithmetic coding is employed to encode the histogram count values, while the histogram map is encoded using a context-based arithmetic coder.

3.5 Global Descriptor Aggregation

Fisher vector (FV) employs higher order statistics to achieve more discriminative power [12]. Traditional BoF approaches encode 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. Both FV and VLAD have achieved promising retrieval performance at a much smaller visual vocabulary, say a few hundreds visual words. [12].

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

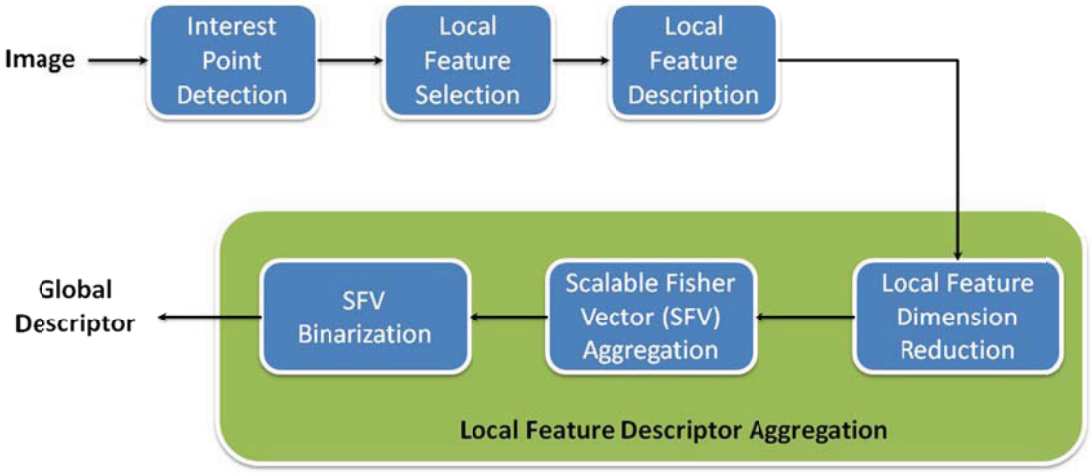


Figure 2. Selective and scalable aggregation of Fisher vector to generate global descriptor.

The adopted SCFV aggregates uncompressed local descriptors [11]. As shown in Figure 2, the pipeline of generating SCFV consists of five main stages.

a) **Feature Selection:** A subset of local features (say, SIFT [18]) are selected to perform selective aggregation. Normally, up to 300 local features per image can be selected for each descriptor length.

b) **Local Feature Dimension Reduction**: Principal Component Analysis (PCA) is employed to reduce the dimension of local features from 128-dim to 32-dim. Dimension reduction benefits the subsequent aggregation in two aspects: i) improving the feature compactness of SCFV; and ii) removing the redundancy in local feature descriptors to improve the discriminative power of SCFV.

c) **Scalable Fisher Vector Aggregation**: Offline and online phases are involved. At the offline phase, a Gaussian Mixture Model (GMM) is employed to estimate the distribution of local feature descriptors over training images. At the online phase, the GMM model is used to generate the Fisher vector [12] for each selected local feature descriptors in query/reference images. First, the gradient vector of each local feature descriptor's eigenvector w.r.t each Gaussian function is calculated. Then, the gradient vectors of selected local feature descriptors are accumulated and normalized for each Gaussian function to generate the Fisher sub-vector, and those Fisher sub-vectors of all Gaussian functions are concatenated to form the scalable Fisher vector (SFV).

Distinct from traditional FV aggregation [12], the SFV employs the rich sparseness of FV to accomplish the scalability. A subset of Gaussian functions is formed to generate informative FV sub-vectors for representing scalable FV.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

d) **Binarizing SFV:** To compress the global descriptor, the sign binarization is applied to the SFV. For each dimension of SFV, a sign function is employed to assign the value "1" to any non-negative values; otherwise, the value "0".

e) **Generating SCFV bit stream:** An overhead of 128 bits is needed to denote which Gaussian functions are used in SFV. The bit stream of binarized SFV is subsequently added to generate the SCFV bit stream.

The SCFV of different images or the SCFV at different lengths may use a distinct subset of Gaussian functions. To support the interoperability, the similarity matching of SCFV descriptors over different sets of Gaussian functions is computed based on the overlapping Gaussian functions of SCFV descriptors.

Compared to the state-of-the-art aggregation approaches, the scalability is a unique feature of the SCFV global aggregation. When bit budget is limited, the SCFV allows the code size of compact signature to be scalable with respect to different operating points. SCFV exhibits low computational complexity, and supports fast similarity matching descriptors encoded at different bit rates.

4 Visual Search Pipeline

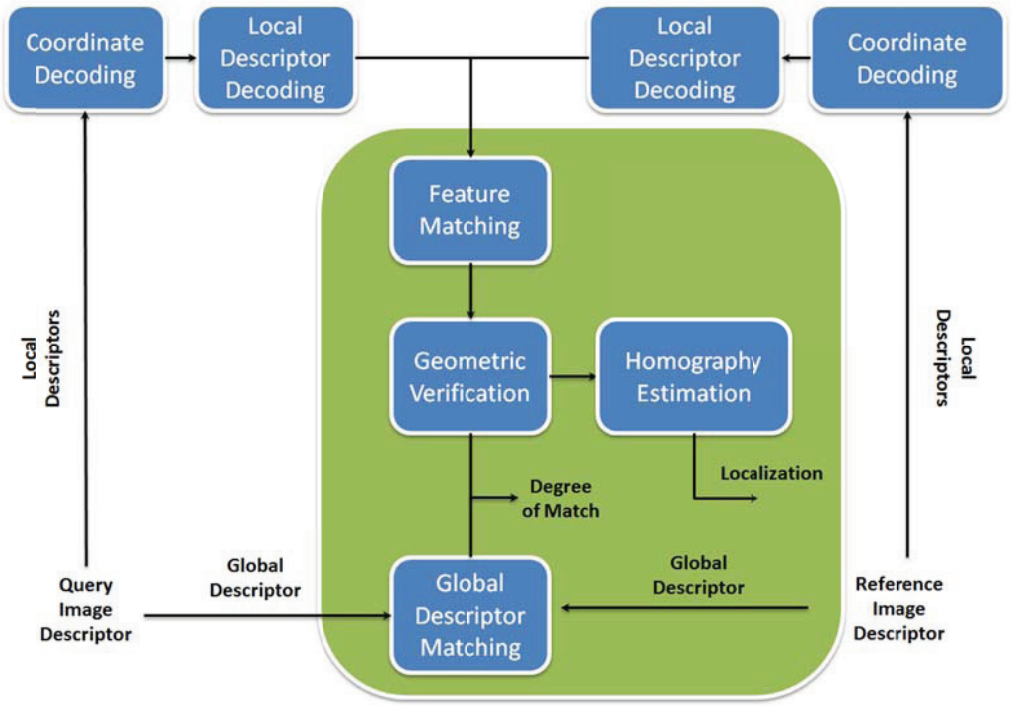
In this section, we present the visual search pipelines involving two search tasks: pairwise matching and retrieval based on the compact descriptors.

4.1 Pairwise Matching

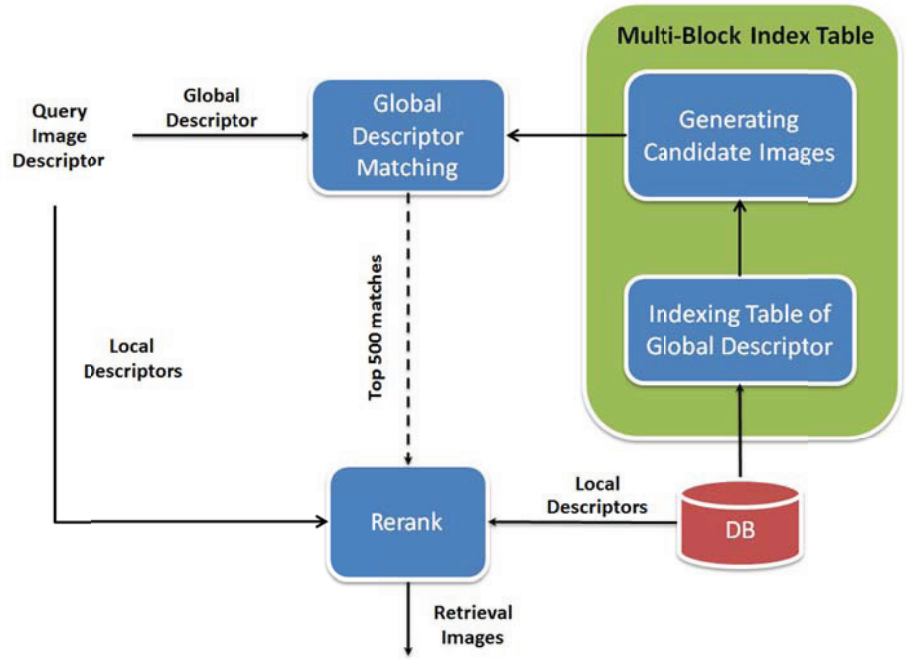
Pairwise matching first determines whether the query and the reference images depict the same objects or scene. In case of a match, it also produces localization information, i.e. the position of the matching objects in the image. Figure 3(a) shows the pairwise matching pipeline using compact visual descriptors.

Pairwise matching procedure encompasses a comparison between the global descriptors of the query image and the reference image, as well as the matching of the local descriptors present in both images. Geometric consistency check is performed to determine the number of inliers among the key point matches for the two images. If a certain hypothesis test is passed and a weighted sum of the inliers exceeds a threshold, the two images are considered as a match. In case of a match, homography estimation is conducted to produce localization information [8].

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE



(a).



(b)

Figure 3. Visual search pipelines of pairwise matching (a) and retrieval (b).

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

4.1 Retrieval

Retrieval discovers all images containing the same objects or scene as query in a large scale dataset. This retrieval pipeline is illustrated in Figure 3(b).

At the first stage, the global descriptor matching is performed to produce a shortlist of top matching images through the Hamming distance based correlations. By ranking the database images, a small set of top matches (say 500 images) are passed to the second stage, namely, geometric verification. The local descriptors are employed to re-rank the shortlisted candidate images returned from the first stage.

Moreover, a Multi-Block Index Table (MBIT) indexing structure [17] is introduced to reduce the range of Hamming distance based similarity comparison of candidate images, thereby eliminating the exhaustive search to generate the shortlist.

5 Data collection and benchmarking

5.1 Dataset

MPEG CDVS benchmark is a million-scale image dataset to evaluate the performance of compact descriptors. Figure 4 shows exemplar query/reference images. Amongst a variety of visual objects, landmark represents a typical type of non-planar objects, which often incurs ill-posed 2D photographic configurations and variances in occlusion, viewpoint, scale, illumination, and background.

As the largest dataset in CDVS benchmark, landmark contains 3,499 queries and 11,677 reference images consisting of: 1). the Zurich buildings, with 1,115 images of 200 buildings in Zurich city, 115 as queries; 2). the Turin buildings with 1,980 images of 180 landmarks in Turin city, 1,620 as queries; 3). the PKUbench with 5,574 images of 198 landmarks from PKU campus, 567 as queries; 4). the Stanford mobile visual search dataset, with 1000 images, 500 as queries; 5). the ETRI dataset, 2141 images of 14 buildings, captured with 22.5 degree changes in viewpoints by using 4 cameras, 83 as queries; 6). Paris dataset, 466 images of 29 buildings in Paris, France, 69 as queries; and 7). Huawei dataset, 2900 images of 192 buildings, 545 as queries. In addition, 3,805 matching pairs and 48,675 non-matching pairs are formed.

The second largest dataset, the Common dataset, are collected from the public available UKBench dataset containing 2,550 objects, each with 4 images taken from different viewpoints. All the 10,200 images are indexed as reference images and used as queries as well. In addition, 2,550 matching pairs and 25,500 non-matching pairs are formed. These two largest datasets constitute the most challenging test.

To study the performance over a large dataset, a FLICKR1M dataset containing 1 million Flickr images as distractor, is merged with reference datasets.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

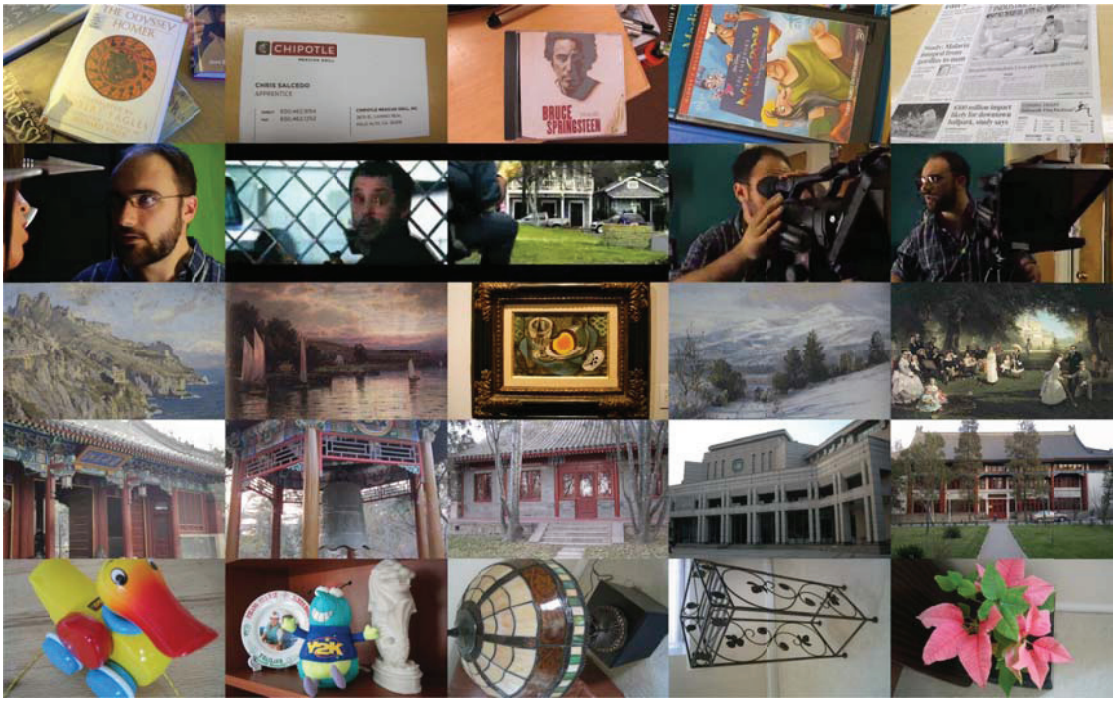


Figure 4. Exemplar images from five datasets of different categories of objects in the MPEG CDVS evaluation dataset (From top to bottom: Graphics, Video Frames, Paintings, Landmarks, and Common Objects, with the total number of query/reference images as 2,500, 500, 500, 15,176, 10,200, respectively).

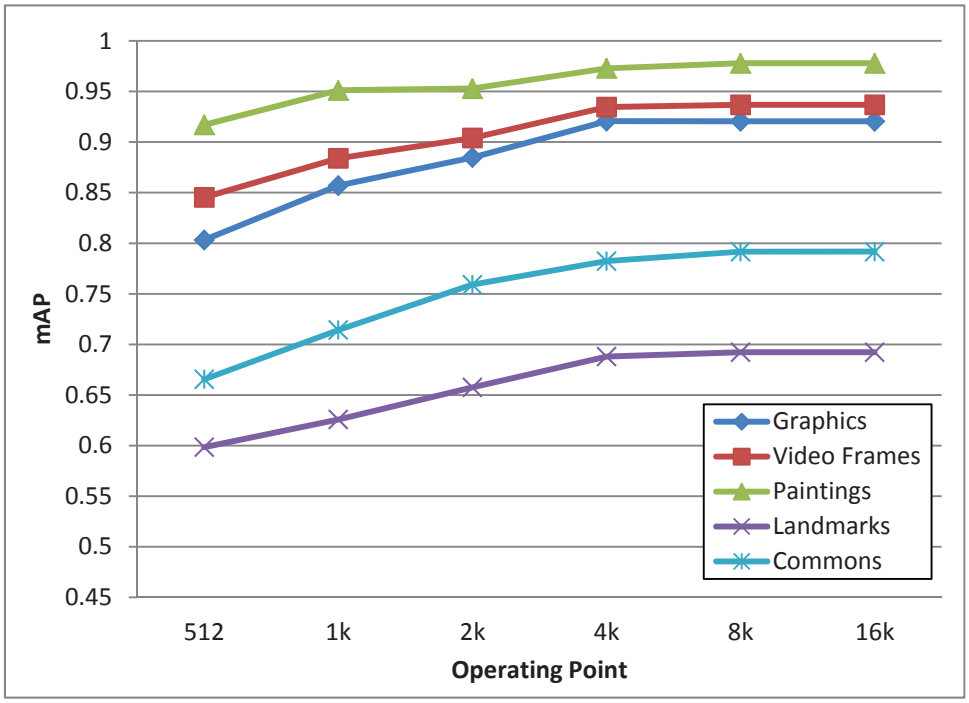
5.2 Results

Figure 5 (a) (b) report the retrieval performance over different type of datasets. With the increase of descriptor rate, the retrieval performance (mAP and Top Match) can be improved progressively, especially at lower operating points. However, the retrieval performance remains stable after 4KB. In other words, the budget of 4KB suffices for encoding discriminative information. From Figure 5 (a) (b), compared to planar objects, searching non-planar objects is more challenging. Large photometric and geometric distortions challenge correct match of query images with their closest matches in the reference databases. In particular, as landmark objects are located outdoors, more challenges may include clutter, shadow cast on buildings, reflections on windows, and severe perspective with extreme angles. See Figure 7.

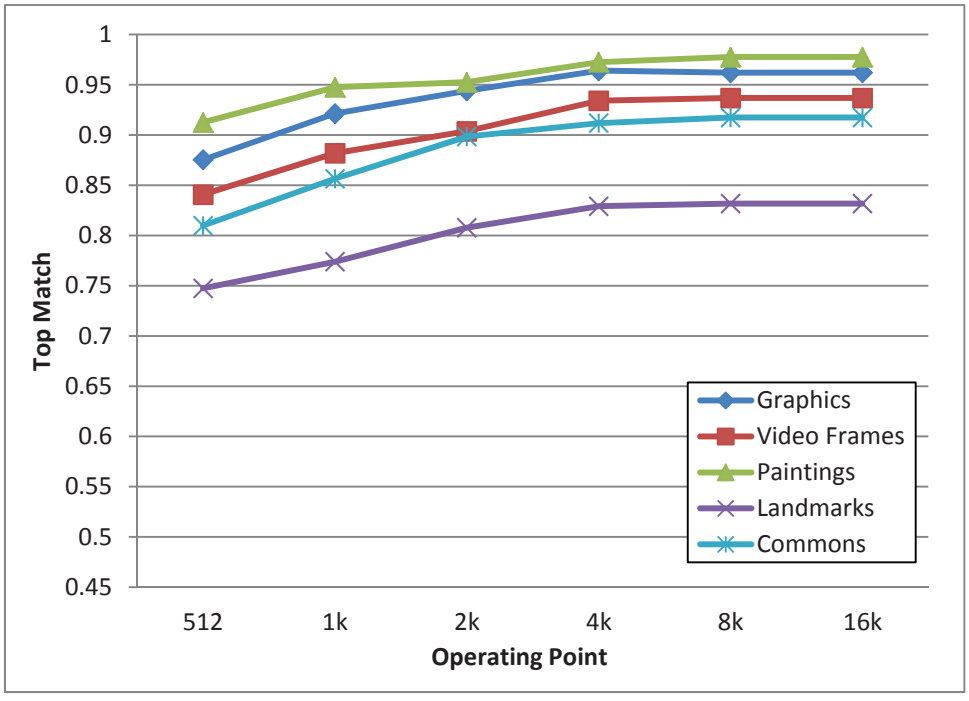
Figure 5 (c) gives the pairwise matching performance. Likewise, the matching performance increases with respect to the descriptor code rate. Non-planar objects are still the most challenging. In addition, the interoperability is shown by pairwise matching between different operating points, say 1kB vs. 4kB and 2kB vs. 4kB.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

Furthermore, the breakdown performance of landmark datasets is reported over Turin Buildings, Zurich Buildings, and PKUBench. The retrieval performance on PKUBench is much lower than Turin buildings and Zurich buildings. See Figure 6. The PKUBench dataset is the most challenging, due to dramatic changes in view angles or shooting distance, as demonstrated in Figure 7.

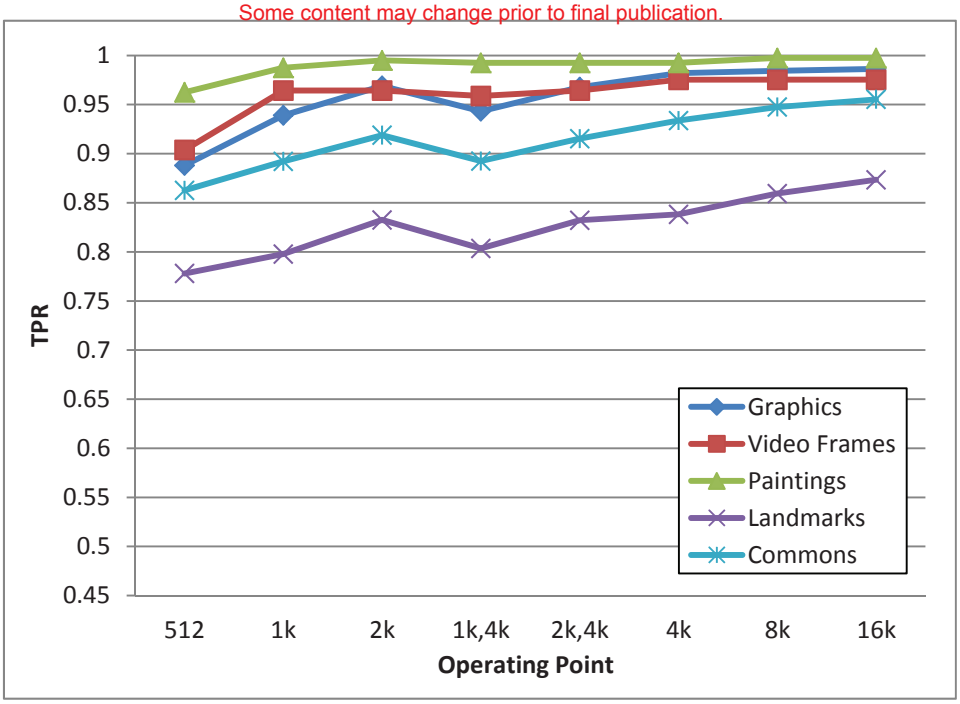


(a)



(b)

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE



(c)

Figure 5: Search performance over the datasets of different object categories with respect to different operating points: Mean Average Precision (a) and Top Match (b); and the pairwise matching success rate (at less than 1% FPR) (c). Note: these results were reported on the MPEG 106th meeting, Geneva, Switzerland, Oct. 2013.

In practice, to deal with a large-scale landmark image repository, GPS or other context information can be employed to filter out irrelevant images [4][5], prior to landmark search, so that the search performance can be further improved towards real world landmark recognition and mining, tourism recommendation from geo-tagged multimedia on the web, and other location based services.

5.3 Descriptor size and composition

The CDVS compact descriptor contains a global descriptor and local descriptors. Figure 8 presents the descriptor composition per target rate. As shown in Figure 8, the global descriptor SCFV supports bit rate scalability. Neither REVV [10] nor VLAD [12] has addressed the scalability issue; that is, their code size is not adaptive to the bit budget. When moderately increasing the SCFV code size at higher operating points, search performance can be significantly improved, while, at lower operating points, more room for local descriptors are allowed to fulfill localization and geometric verification at lower operating points.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

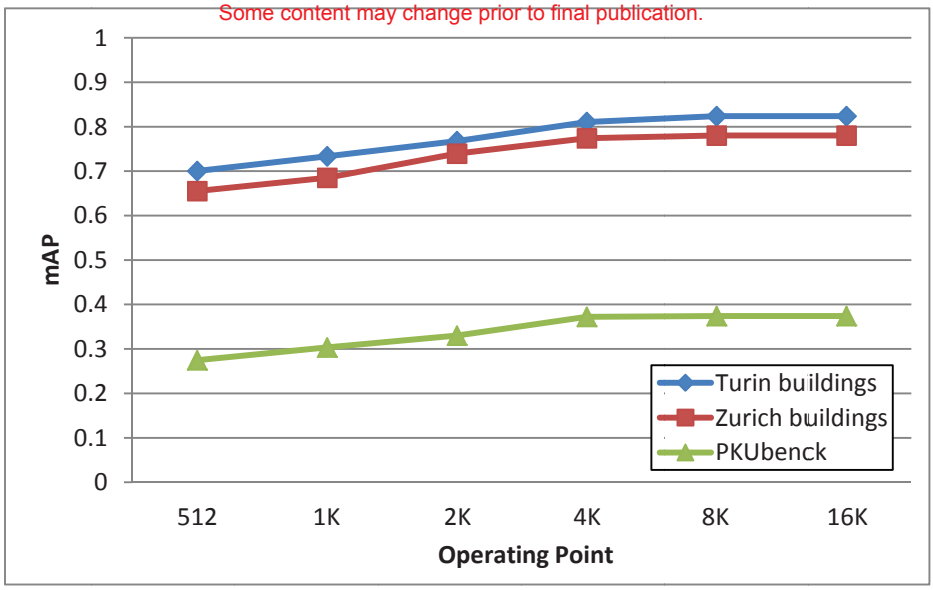


Figure 6: Retrieval performance of landmark objects over different subsets.

Figure 7: Exemplar query/reference images of landmark objects over different subsets (From top to bottom: Turin buildings, Zurich buildings, PKUBench; for each line, the left one is query image, and the rest are reference images.)

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

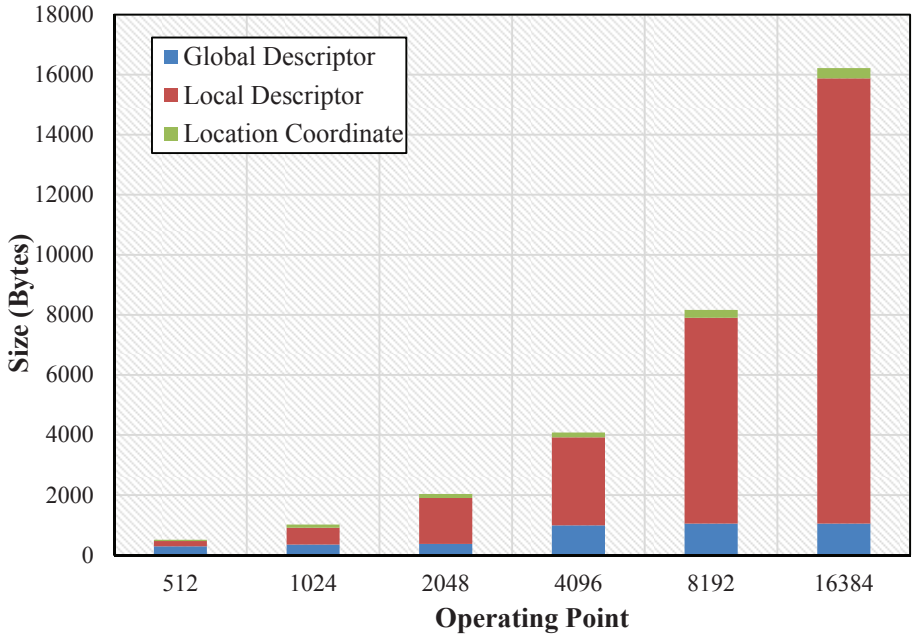


Figure 8: MPEG CDVS compact descriptor composition per target rate.

6 Conclusions

The MPEG CDVS standardization efforts have been reviewed. A competitive and collaborative platform to evaluate the state-of-the-art visual search technologies and solutions has been introduced as well.

Sending an image as a query can be safely replaced, in future systems, by extracting the compact features on the terminal instead and sending these features as the query to the server. This will also have a positive impact on privacy because the features are anonymous compared to the pictures that are not.

In addition, a small scale image database could be stored locally and matching could be performed directly on the device. Extension to other application domains could be possible like Automotive, Entertainment, Digital TV and Surveillance.

In the next few years, we can expect an increasing set of client-server visual search applications exploiting a standard approach, which offers huge advantages in term of interoperability between heterogeneous terminals and servers.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

Acknowledgement

This work was supported by the Chinese Natural Science Foundation under Contract No. 61271311, No. 61121002, 61210005, and in part by the MDA Research Fund of the ROSE LAB (A joint project between PKU and NTU).

References

[1] V. Chandrasekhar, etc. "CHoG: Compressed histogram of gradients: a low bit-rate feature descriptor," In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 2009, pp.2504-2511.

[2] B. Girod, et al. "Mobile visual search," IEEE Signal Processing Magazine, vol. 28, no. 4, pp. 61-76, 2011.

[3] D. Chen, etc. "Tree histogram coding for mobile image matching," In Proc. of Data Compression Conference, 2009, pp. 143-152

[4] R. Ji, L. Y. Duan, etc. "Location Discriminative Vocabulary Coding for Mobile Landmark Search," Int. Journal of Computer Vision, vol. 96, no. 3, pp. 290–314, 2012.

[5] R. Ji, L. Y. Duan , etc., "Towards low bit rate mobile visual search with multiple-channel coding," in Proc. *19th* ACM Int. Conf. on Multimedia, 2011, pp. 573–582.

[6] F. Gianluca, S. Lepsoy, et. al. Telecom Italia's response to the MPEG CfP for Compact Descriptors for Visual Search. m22672, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Dec. 2011.

[7] MPEG CDVS Ad-hoc Group. CDVS Test Model under Consideration: Compact descriptor for visual search. w12367, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Dec. 2011.

[8] MPEG CDVS Ad-hoc Group. CDVS Test Model 6: Compact descriptor for visual search. w13564, ISO/IEC JTC1/SC29/WG11, Inchon, Korea, April 2013.

[9] MPEG CDVS Ad-Hoc Group, "Compact Descriptors for Visual Search: Evaluation Framework," ISO/IEC JTC1/SC29/WG11/N12202, 2011/07.

[10] D. Chen, V. Chandrasekhar, et al. CE1: Improvements to the Test Model with a Low-Memory Global Descriptor. m24757, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, April 2012.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE

[11] J. Lin, L. Y. Duan, et al. Peking Univ. Response to CE1: A scalable low-memory global descriptor. m26726, ISO/IEC JTC1/SC29/WG11, Shanghai, China, Oct. 2012.

[12] Herv´e J´egou, Florent Perronnin, et. al., “Aggregating local images descriptors into compact codes,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 34, no. 9, pp. 1704-1716, 2012.

[13] J. Chen, L. Y. Duan, et al. Peking Univ. Response to CE 2 – Local descriptor compression. m26727, ISO/IEC JTC1/SC29/WG11, Shanghai, China, Oct. 2012.

[14] S. Paschalakis, K. Wnukowicz, et al. CDVS CE2: Local Descriptor Compression Proposal. m24737, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Jul. 2012.

[15] Z. Wang, S. Tsai, et al. CDVS Core Experiment 3: Stanford/Peking/Huawei contribution. m25883, ISO/IEC JTC1/SC29/WG11, Stockholm, Sweden, July 2012.

[16] F. K. Wang, L. Y. Duan, et al. Peking University Response to CE2: Frequency Domain Interest Point Detector. m28991, ISO/IEC JTC1/SC29/WG11, Inchon, Korea, April 2013.

[17] Z. Wang, L. Y. Duan, et al. An indexing structure to speed up retrieval. m28993, ISO/IEC JTC1/SC29/WG11, Inchon, Korea, April 2013.

[18] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, vol. 60, no.2, pp. 91-110, 2004.

[19] J. Chen, L. Y. Duan, et al. Peking University Response to CE 1: Improved BFLoG Interest Point Detector. M30241, ISO/IEC JTC1/SC29/WG11, Vienna, Austria, Jul. 2013.

Ling-Yu Duan (lingyu@pku.edu.cn) received his Ph.D. degree from the University of Newcastle, Australia, in 2007. From 2003 to 2008, he was a research scientist at the Institute for Infocomm Research, Singapore. Since 2008, he has been an associate professor with Peking University. Since 2012, Dr. Duan has been the deputy director of the Rapid-Rich Object Search Lab. His research interests include visual search, media content analysis, and mobile media computing.

Jie Lin (jlin@pku.edu.cn) is a Ph.D. candidate at the School of Electrical Engineering and Computer Science, Beijing Jiaotong University. Since 2011, he works as a research assistant in the Institute of Digital Media, Peking University. His research works on scalable visual descriptors and mobile visual search.

Jie Chen (cjie@pku.edu.cn) is a Ph.D. candidate at the School of Electrical Engineering and Computer Science, Peking University. His research works on compact visual descriptors and their interoperability in mobile visual search.

Tiejun Huang (tjhuang@pku.edu.cn) received his Ph.D. degree from Huazhong University of Science and Technology, China, in 1998. He is a professor in the School of Electrical Engineering and Computer Science, Peking University. He is the director of the Institute of Digital Media, Peking University. His research interests include video coding, image understanding, digital rights management, and digital library.

Wen Gao (wgao@pku.edu.cn) received his Ph.D. degree in electronics engineering from the University of Tokyo in 1991. He is a professor in the School of Electronics Engineering and Computer Science, Peking University. He led research efforts in video coding, face recognition, and multimedia retrieval. Prof. Gao was admitted as an Academician of the China Engineering Academy in 2011 and became an IEEE Fellow in 2010 for his contribution to video coding technology.

Digital Object Indentifier 10.1109/MMUL.2013.66 1070-986X/$26.00 2013 IEEE