Overview of the MPEG-CDVS standard

Ling-Yu Duan, *Member, IEEE,* Vijay Chandrasekhar, *Member, IEEE,* Jie Chen, Jie Lin, *Member, IEEE,* Zhe Wang, Tiejun Huang, *Senior Member, IEEE,* Bernd Girod, *Fellow, IEEE,* Wen Gao, *Fellow, IEEE,*

***Abstract*—Compact Descriptors for Visual Search (CDVS) is a recently completed standard from the ISO/IEC Moving Pictures Experts Group (MPEG). The primary goal of this standard is to provide a standardized bitstream syntax to enable interoperability in the context of image retrieval applications. Over the course of the standardization process, remarkable improvements were achieved in reducing the size of image feature data and in reducing the computation and memory footprint in the feature extraction process. This article provides an overview of the technical features of the MPEG-CDVS standard and summarizes its evolution.**

***Index Terms*—compact descriptors, feature compression, MPEG-CDVS, visual search**

I. Introduction

Over the past decade, mobile phones and tablets have become devices that are suitably equipped for visual search applications. With high-resolution cameras, powerful CPUs and pervasive wireless connections, mobile devices can use images as search queries for objects observed by the user. Emerging applications include scene retrieval, landmark recognition, and product identification, among others. Examples of early commercial mobile visual-search systems include Google Goggles [1], Amazon Flow [2] and Layar [3].

The requirements for mobile visual search, such as faster searches, higher accuracy and better user experience, pose a unique set of challenges. Normally, a mobile visual search system transmits JPEG-encoded query images from the mobile end to the remote server, where a visual search is performed over a reference image database. However, image transmission could take anywhere from a few seconds to a minute or more over a slow wireless link, and wireless upload might even timeout in the case of an unstable connection. On the other hand, on-device image analysis, either for mobile image matching or for the transmission of a compact signature to the cloud, might be computationally demanding and hence slow.

In Figure 1, we present four typical client-server architectures, as follows:

- In Figure 1(a), a JPEG-encoded query image is transmitted to the server. Visual descriptor extraction and matching/retrieval are performed entirely on the server;
- In Figure 1(b), visual descriptors are extracted and compressed on the mobile client. Matching/retrieval is performed on the server using the transmitted feature data as the query;
- In Figure 1(c), a cache of the database is maintained on the mobile device, and image matching is performed locally. Only if a match is not found does the mobile device send the query to the server for a remote retrieval;
- In Figure 1(d), the mobile device performs all the image matching locally, which is feasible if the database is small and can be stored on the mobile device.

In each case, the retrieval framework must adapt to stringent mobile system requirements. First, the processing on the mobile device must be fast, lightweight and have low power consumption. Second, the size of the data transmitted over the network must be as small as possible to reduce the network latency. Finally, the algorithms used for retrieval and matching must be scalable to potentially very large databases and robust to allow reliable recognition of objects captured under a wide range of conditions, such as partial occlusions, changes in vantage point, camera parameters, and lighting.

Initial research on the topic [4], [5], [6], [7], [8], [9], [21] demonstrated that one could reduce transmission data by at least an order of magnitude by extracting compact visual features efficiently on the mobile device and sending descriptors at low bitrates to a remote server for performing the search. A significant reduction in latency could also be achieved when performing all processing on the mobile device itself.

Following initial research on the topic, an exploratory activity in the Moving Picture Experts Group (MPEG) (formal title "ISO/IEC JTC1 SC29 WG11") was initiated at the 91st meeting (Kyoto, Jan. 2010). As MPEG exploratory work progressed, it was recognized that the suite of existing MPEG technologies, such as MPEG-7 Visual, did not include tools for robust image retrieval and that a new standard would therefore be needed [10]. It was further recognized that, among several component technologies for image retrieval, such a standard should focus primarily on defining the format of descriptors and those parts of their extraction needed to ensure interoperability. Such descriptors need to be compact, image format independent, and sufficient for robust image matching. Hence, the title Compact Descriptors for Visual Search (CDVS) was coined as the name for this activity. Requirements and evaluation framework documents were subsequently produced to formulate precise criteria and evaluation methodologies to be used in the selection of technology for the standard [11],

Manuscript received July 24, 2015; revised November 2, 2015; accepted November 3, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jie Liang.

Ling-Yu Duan, Jie Chen, Zhe Wang, Tiejun Huang, and Wen Gao are with the Institute of Digital Media, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China. (e-mail: lingyu@pku.edu.cn; zhew@pku.edu.cn; cjie@pku.edu.cn; tjhuang@pku.edu.cn; wgao@pku.edu.cn)

Vijay Chandrasekhar and Jie Lin are with the Institute for Infocomm Research, Singapore. (email: vijay@i2r.a-star.edu.sg; lin-j@i2r.a-star.edu.sg)

Bernd Girod is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, USA. (email: bgirod@stanford.edu)

This work was supported by the Chinese Natural Science Foundation under Contract No. 61271311 and No. 61390515, and the National Hightech R&D Program of China (863 Program) under grant 2015AA016302. Ling-Yu Duan and Vijay Chandrasekhar are joint first authors.

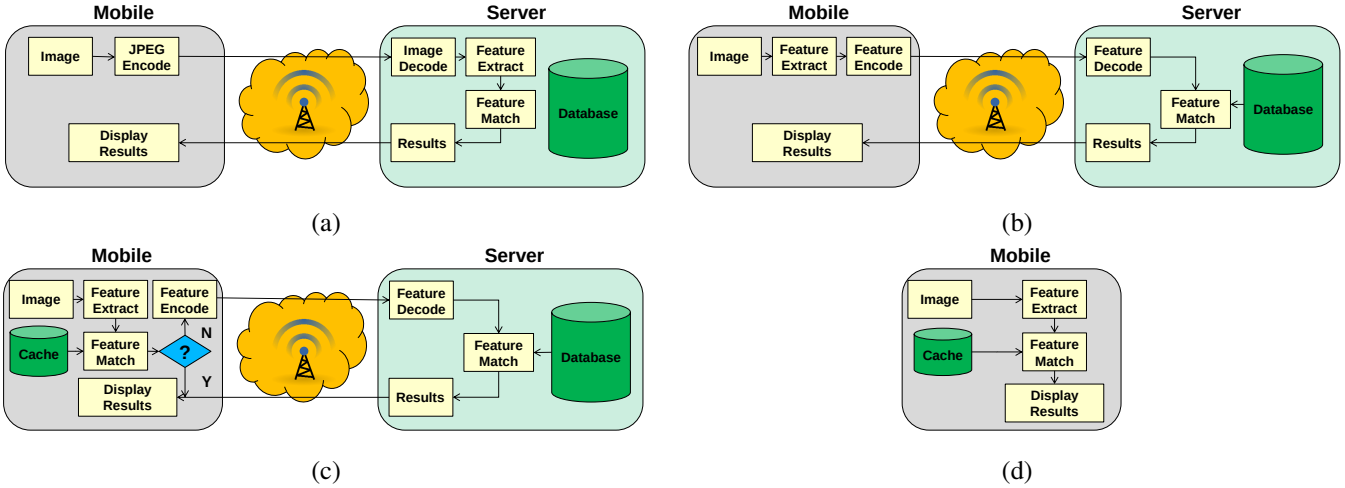

Fig. 1. In (a), the mobile device transmits a JPEG-compressed query image to the server, where all matching is performed against a database of images. In (b), the mobile device analyzes the query image, extracts features, and transmits compressed feature data. The retrieval algorithms run on the server. In (c), the mobile device maintains a cache of the database and performs image matching locally. Only if a match is not found does the mobile device send a compressed feature query request to the server. In (d), all processing is performed locally.

[12]. The envisioned MPEG-CDVS standard would have to

- ensure interoperability of visual search applications and databases,
- reduce load on wireless networks carrying visual search-related information,
- provide a basis for hardware-supported descriptor extraction and matching in mobile devices,
- enable a high level performance of implementations conformant to the standard, and
- simplify the design of descriptor extraction and matching for visual search applications.

It is envisioned that the standard might also be used in conjunction with other existing MPEG and JPEG standards, such as MPEG Query Format, HTTP, XML, JPEG, and JPSearch.

The CDVS standard (formally known as MPEG-7, Part 13) was published by ISO on August 25th, 2015 [14], and this standard evolved based on the development framework established for all MPEG standards: requirements for technology are first specified [13], technology is requested through an official "Call for Proposals" [12], and the technology proposed to MPEG is thoroughly evaluated by MPEG experts based on a previously agreed-upon methodology [11].

Over the course of the standardization process, a total of 366 input documents were received, of which there were 99 contributions to core experiments (CE). The standard witnessed active participation from a number of companies, universities and research institutes: key participants included Stanford University, Peking University, Surrey University, Telecom Italia, Qualcomm, STMicroelectronics, Huawei, Nokia, NEC, Samsung, ETRI, Visual Atoms, and others. After 14 iterations, the final software reference model, TM 14.0, was released after the 112th meeting (Warsaw, Jun. 2015) [14].

By thoroughly testing state-of-the-art visual search technology proposals and performing competitive and collaborative experiments within a rigorous evaluation framework [12], the CDVS working group has observed remarkable improvements in image retrieval performance with very compact feature data. High performance is achieved while also satisfying stringent memory and computational complexity requirements at each step of the feature extraction pipeline, making the standard ideally suited for both hardware and software implementations. This paper presents an overview of the CDVS standard, including its development and key technical contributions, and reports the most important performance results. In Section II, we present the building blocks of the standard. In Section III, we provide details of each normative block. In Section IV, we briefly discuss non-normative parts of the standard, which are also part of the reference software. In Section V, we discuss the evaluation framework, evolution of the standard, and detailed results.

II. Highlights

The MPEG-CDVS standard defines the bitstream (i.e., binary representation syntax) of descriptors and the descriptor extraction process [14]. The key building blocks are shown in Figure 2. To be compliant, the syntax of the descriptors needs to conform to the CDVS standard.

CDVS supports interoperability in two ways. First, it standardizes the bitstream syntax of descriptors. Second, it provides the framework for matching descriptors encoded at different bit rates. The latter feature allows for a compact database with compressed features and bit-rate scalability of the query.

The algorithms for retrieval and matching are not part of the standard. Video compression experts will note that the approach is the dual of what is performed for video coding standards in which the bitstream syntax and the decoder are standardized. For CDVS, the bitstream syntax and the *encoder* are standardized. The modules in Figure 2 are the minimum building blocks required to maintain interoperability.

Data Sets and Evaluation. The CDVS evaluation data set is an order of magnitude larger than other popular data sets, such as *INRIA Holidays* and *Oxford Buildings*. The data set has considerably more variety in the type of objects, scale,

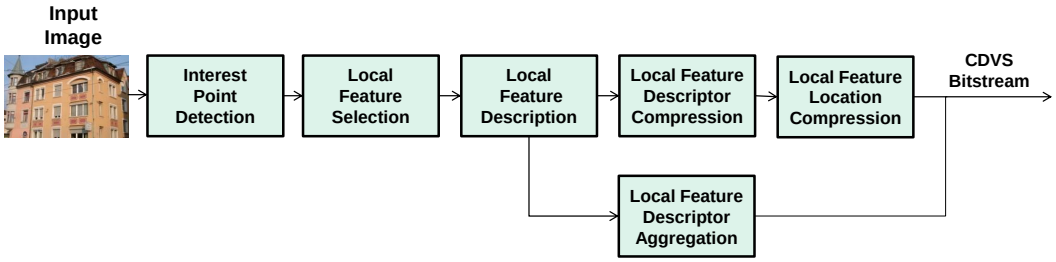

Fig. 2. Normative blocks of the CDVS standard. Compressed global and local features are extracted from the query image and combined to form the final bitstream.

rotation, occlusion and lighting conditions than the *INRIA Holidays* and *Oxford* data sets, as discussed in [15]. Both pairwise matching and retrieval experiments are included in the evaluation framework, and performance is evaluated as a function of bitrate. For pairwise matching, the ground-truth data of 10,155 matching image pairs and 112,175 non-matching image pairs are provided. For retrieval experiments, 8314 query images, 18840 reference images, and a distractor set of 1 million images from Flickr [16] are used.

Interest Point Detection and Local Feature Description. The CDVS standard adopts a low-degree polynomial (ALP) detector followed by the popular SIFT descriptor. To find interest points, ALP approximates the result of the LoG filtering using polynomials, which are used to find extrema in the scale space and to refine the spatial position of the detected points. Although not mandated by the standard, a block-based frequency domain Laplacian of Gaussian (BFLoG) approach [17] can be integrated with the ALP detector to accomplish a block-based scale-space interest point detector ALP_BFLoG [24]. ALP_BFLoG divides the original scale space into overlapping blocks, and interest point detection is performed on each block independently, thereby reducing the memory cost required by filters and scale-space buffers by an order of magnitude. The block-based interest point detection makes the entire pipeline amenable to hardware implementation with low memory cost.

Local Feature Selection. A subset of feature descriptors is selected to satisfy rate constraints at each descriptor length. A relevance measure is calculated for each local feature, which indicates the probability of a query feature matching a database feature [18]. The relevance measure is statistically learned, and it is based on the scale, peak response of the LoG, and the distance from the image center of each local feature, as well as other measures that will be discussed in the following. Features are ranked based on the relevance measure, and a fixed number of features are selected based on the total feature data budget and the number of bits per feature.

Local Feature Descriptor Compression. A low-complexity transform coding scheme is adopted in the CDVS standard [19]. The descriptor transform is followed by ternary scalar quantization and instantaneous variable-length coding. Rather than applying a transform to the entire descriptor, small linear transforms are applied to the 8 values of each individual spatial bin of the SIFT descriptor. Only a subset of transformed descriptor elements is included in the bitstream. This subset is selected according to a standardized priority table that has been optimized for the best retrieval performance. The number of transformed descriptor elements included ranges from 20 of 128 for the smallest image descriptor length (512 and 1024 bytes) to 128 of 128 for the largest image descriptor length (16384 bytes).

Local Feature Location Compression. The location coding scheme in the CDVS standard is based on the key insight that the original ordering of the features can be discarded and that one can save up to additional $\log(n!)$ bits for n features beyond the entropy-coded bitstream [39], [20]. A histogram coding scheme is adopted, which reorders feature data based on x, y location and achieves the $\log(n!)$ ordering gain. Location data are represented as a spatial histogram consisting of a binary map and a set of histogram counts. The histogram map and counts are encoded using a binary context-based arithmetic coding scheme.

Local Feature Descriptor Aggregation. A scalable compressed Fisher Vector (SCFV) representation is adopted in the CDVS standard [21]. For compressing high-dimensional Fisher vectors, a subset of Gaussian components from the Gaussian Mixture Model (GMM) are selected based on the total feature data budget, and only the information in selected components is retained. A different set of components is selected for each image based on where the energy is concentrated in the Fisher vector. A small set of header bits indicate which components are selected for each aggregated global feature. SCFV provides high matching accuracy with negligible memory requirements compared to the conventional Fisher vector compression approaches based on PCA or vector quantization.

III. Normative Blocks

A. Interest Point Detection and Local Feature Description

Local feature extraction involves detecting interest points and characterizing interest points with feature descriptors: high-dimensional representations that describe scale and rotation invariant patches [22]. The CDVS standard includes a Laplacian of Gaussian (LoG) interest point detector, followed by the popular SIFT descriptor [22]. The image scale space is represented as an image pyramid in which an image is successively filtered by a family of smoothing kernels at increasing scale factors. Normalized derivatives at each scale

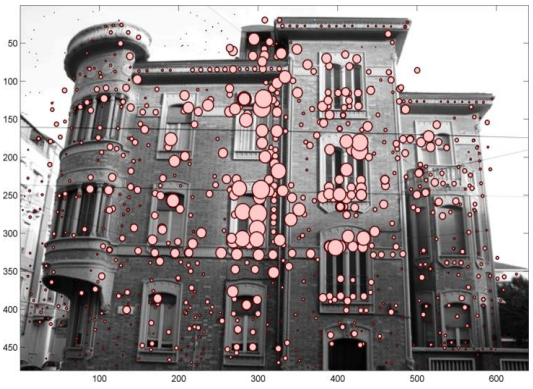

Fig. 3. The interest points of the image are plotted with circles whose diameters are proportional to the relevance measures.

in the image pyramid are computed, and interest points are computed by searching for local extrema in scale space [22].

The important novelty of the standard is a Laplacian of Gaussian (LoG) interest point detector based on polynomial approximations. The adopted low-degree polynomial (ALP) approach approximates the result of the LoG filtering [23]. Subsequently, scale-space extrema are found and refined to compute the precise spatial positions of the detected points. Specifically, to approximate the LoG scale space, ALP uses a polynomial function with regard to the scale parameter σ for each pixel (x, y) in the image:

$$p(x,y,\sigma) = \sum_{k=0}^{K-1} a_k L_k(x,y)\sigma^3 + \sum_{k=0}^{K-1} b_k L_k(x,y)\sigma^2 + \sum_{k=0}^{K-1} c_k L_k(x,y)\sigma + \sum_{k=0}^{K-1} d_k L_k(x,y) \quad (1)$$

where a_k, b_k, c_k, and d_k are the coefficients (stored in a normative table) corresponding to the $K = 4$ predefined scales σ_k and $\{L_k(.,.)|k = 0, ..., 3\}$ are K octave images produced by scale-normalized Laplacian filtering of the Gaussian-filtered images. To detect the scale-space extrema, ALP first locates the local extrema in the σ direction by setting its first derivative to zero, and then it compares the point to its 8 neighbors in the X-Y plane with respect to coordinates x and y.

Comparisons with other schemes. ALP's interest point detection is more efficient than conventional scale-space extrema detectors that compare response values at each point to $3 \times 3 \times 3 - 1 = 26$ neighbors in scale space, as ALP is built upon 4 LoG-filtered images to approximate the LoG scale space rather than 5 (or more) LoG-filtered images in the LoG detector or 6 (or more) Gaussian-filtered images in a typical Difference of Gaussians (DoG) detector [22].

B. Local Feature Selection

Based on the image content, interest point detection can result in several hundred to several thousand features, even for VGA-resolution images. For small feature data sizes (512 bytes to 4 KB), it is not feasible to include all features, even if the number of bits per descriptor is small [18]. Consequently, selecting a subset of feature descriptors becomes critical. There are also other advantages of feature selection. Local feature descriptors are aggregated to form the global feature descriptor. Incorporating noisy local features can degrade the discriminative power of the global descriptor. Finally, feature selection can also save considerable computation time in the feature extraction block: the most time-consuming module in the CDVS encoding process. Note that the feature selection problem here differs from that in supervised learning tasks, where a subset of individual dimensions of a feature are selected for improving classification performance [27].

In the CDVS standard, a relevance measure is computed for each local feature. The relevance measure indicates the a priori probability of a query feature matching a database feature. For example, query features that are closer to the center of the image are more likely to match. Similarly, features from more textured regions are more distinctive and discriminative and hence more likely to match database features. The relevance measure has been statistically learned based on five characteristics of interest points: the scale σ of the interest point, the peak response value p of the LoG, the distance d from the interest point to the image center, the ratio ρ of the squared trace to the determinant of the Hessian, and the second derivative $p_{\sigma\sigma}$ of the scale-space function with respect to scale.

By assuming that different interest point characteristics are conditionally independent given a feature match, conditional distributions of feature matching are learned for each characteristic using an independent data set [28], [18] during the standardization. Note that these parameters of learned conditional distributions are quantized within the intervals in the normative tables, and each quantization interval has an associated scalar value in the normative tables. To learn the conditional distributions, pairwise feature matching with SIFT features, ratio test, and a geometric consistency check [22], [29] were performed on a large data set of matching image pairs to obtain a set of matching and non-matching feature pairs. Note that the matching is performed with ALP as the detector and SIFT (uncompressed) as the descriptor. In the geometric verification step, the minimum number of inliers for matching image pairs was set to a stringent threshold of 30 to ensure few outliers and high-quality matching feature pairs. The statistics of the matching feature pairs were then used to estimate the conditional distributions.

The relevance score r for a feature is obtained by multiplying the conditional probabilities of each characteristic:

$$r(\sigma, p, d, \rho, p_{\sigma\sigma}) = f_1(\sigma) f_2(p) f_3(d) f_4(\rho) f_5(p_{\sigma\sigma}), \quad (2)$$

where the factors $f_1 \sim f_5$ are taken from the normative tables of the learned conditional distributions according to the interest point characteristics. Finally, features are ranked based on the relevance measure, and a fixed number of features are selected based on the total feature data budget and the number of bits per feature. Figure 3 shows an example of feature selection: the interest points of the image are plotted with circles whose diameters are proportional to the relevance measures.

Comparisons with other schemes. A naive approach for feature selection is to rank features based on the peak response from the interest point detector. The adopted approach, which takes several interest point characteristics into account,

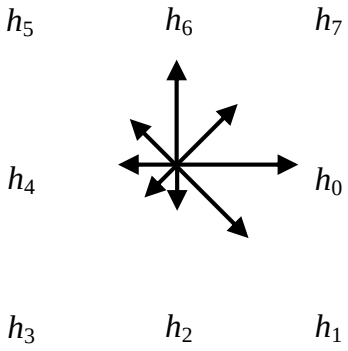

Fig. 4. Cell histogram consisting of 8 angular bins.

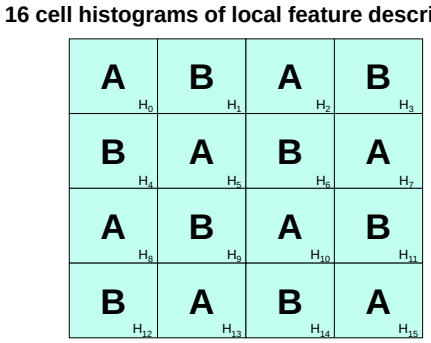

Fig. 5. Two sets of transforms (A,B) as defined in Equations 3 and 4 are applied in an alternating manner to the values in each spatial bin of the SIFT descriptor.

achieves considerably better performance, particularly at low rates [18]. A significant improvement in performance is also obtained when features are selectively aggregated based on the relevance measures in the global descriptor [30]. The initial draft of the standard also included a sixth factor in Equation 2, which made feature selection dependent on the feature orientation. However, the observed gains were small, and orientation dependence was dropped when CDVS was finalized.

C. Local Feature Descriptor Compression

The uncompressed SIFT descriptor is conventionally stored as 1024 bits per descriptor (128 dimensions and 1 byte per dimension). Even a small number of uncompressed SIFT descriptors results in tens of KBs of data; hence, local feature descriptor compression is critical for reducing the feature data size. Using novel compression schemes, the number of bits per descriptor is significantly reduced by an order of magnitude with little loss in matching performance. Several compression schemes based on Product Quantization (PQ) [31], Product Tree Structured Vector Quantization (PTSVQ) [32], [33], Multi-stage Vector Quantizer (MSVQ) [34], lattice coding [5] and transform coding [35], [36] were proposed over the course of the standardization. Eventually, a low complexity transform coding scheme was adopted after thorough evaluation.

The transform coding scheme adopted in the standard is described in [14], [35], [36]. For a local feature descriptor, each of the cell histograms $H_0, ..., H_{15}$, as shown in Figure 5, each with angular bins $h_0, ..., h_7$, as shown in Figure 4, are independently transformed.

There are two main steps: the descriptor transform, based on simple additions and subtractions of SIFT components, followed by ternary scalar quantization and entropy coding of the transformed elements. Rather than applying an order-128 transform to the entire descriptor (which can degrade performance [37]), small order-8 linear transforms are applied to individual spatial bins of the SIFT descriptor. Two sets of linear transforms are defined in Equation set 3 and Equation set 4 (also referred to as (A) and (B), respectively):

$$\begin{aligned} v_0 &= (h_2 - h_6)/2 \\ v_1 &= (h_3 - h_7)/2 \\ v_2 &= (h_0 - h_1)/2 \\ v_3 &= (h_2 - h_3)/2 \\ v_4 &= (h_4 - h_5)/2 \\ v_5 &= (h_6 - h_7)/2 \\ v_6 &= ((h_0 + h_4) - (h_2 + h_6))/4 \\ v_7 &= ((h_0 + h_2 + h_4 + h_6) - (h_1 + h_3 + h_5 + h_7))/8 \end{aligned} \quad (3)$$

$$\begin{aligned} v_0 &= (h_0 - h_4)/2 \\ v_1 &= (h_1 - h_5)/2 \\ v_2 &= (h_7 - h_0)/2 \\ v_3 &= (h_1 - h_2)/2 \\ v_4 &= (h_3 - h_4)/2 \\ v_5 &= (h_5 - h_6)/2 \\ v_6 &= ((h_1 + h_5) - (h_3 + h_7))/4 \\ v_7 &= ((h_0 + h_1 + h_2 + h_3) - (h_4 + h_5 + h_6 + h_7))/8 \end{aligned} \quad (4)$$

where $h_0 \sim h_7$ denote the bins of each cell histogram for a local feature descriptor, as shown in Figure 4.

The transforms are applied in an alternating manner, as shown in Figure 5. Adjacent spatial bins of the SIFT descriptor have similar values; hence, applying different transforms to adjacent bins improves performance, particularly at extremely low rates such as 32 bits per descriptor [38]. The subset of transform elements, the number of bits per descriptor and the number of descriptors are empirically optimized for each descriptor length, resulting in 32, 32, 65, 103, 129, and 205 bits on average at descriptor lengths of 512 bytes, 1 KB, 2 KB, 4 KB, 8 KB, and 16 KB, respectively. The alternating grid pattern shown in Figure 5 also emerges from the greedy rate allocation scheme proposed in [38].

Given two quantized local feature descriptors $V_q = \{v_i^q | i = 0, 1, ...127\}$ and $V_r = \{v_i^r | i = 0, 1, ...127\}$, their similarity

distance $Dis(.)$ is computed in the transform domain using the L_1 norm:

$$Dis(V_q, V_r) = \sum_{i=0}^{127} s_i^q s_i^r \left\| v_i^q - v_i^r \right\|_{L_1}, \quad (5)$$

where s_i^q and s_i^r denote whether the i_{th} transform element of V_q or V_r is selected. Note that Equation 5 permits the comparison of descriptors encoded at different bitrates by only considering the transformed descriptor elements that both have in common. The fixed prioritization scheme of the standard ensures that each element of a descriptor at a lower bitrate is also present in a descriptor at a higher bitrate.

Comparisons with other schemes. The transform coding scheme was selected over several other VQ and lattice coding schemes because of its simplicity, low memory and computational complexity, and excellent performance at very low rates. The memory cost of the transform coding scheme is negligible compared to the hundreds of KBs [32] or hundreds of MBs [33] required for product vector quantization schemes. The primary memory required is 128 (elements of the transform) $\times$ 2 (ternary SQ thresholds) = 256 bytes [35]. Furthermore, the scheme has lower computational complexity than VQ schemes that require nearest neighbor search over codebooks. At low rates, the transform coding scheme performs comparably or better than several VQ-based approaches, lattice coding, and binary hashing schemes, as observed from the detailed patch-level evaluation in [38]. This approach comes close to the performance of entropy constrained vector quantization and greedy rate allocation, a scheme that is close to the performance bound that can be achieved by any compression scheme [38].

In addition to the transform coding scheme, an MSVQ scheme is also available in an intermediate Test Model (4.0) [34]. The MSVQ scheme substantially reduces codebook memory requirements (38 KB [34] versus 60$\sim$150 MB for storing a large codebook containing 0.1$\sim$1 million words [32], [33]) while maintaining comparable matching performance. The MSVQ scheme uses 2-stage vector quantization: a tree structured quantizer in the first stage followed by product quantization for the residuals. Compared to the MSVQ scheme, the CDVS standard adopted transform coding scheme is superior in terms of complexity.

D. Local Feature Location Compression

A problem related to descriptor compression is the compression of the x, y location data of each feature. Each descriptor has an x, y location associated with it, which is used in the Geometric Consistency Check (GCC) step. If x, y location data are represented as floating point numbers, then the size of the location data is often comparable to the size of the compressed descriptors themselves [39].

The location histogram coding scheme adopted in the CDVS standard is described in [40], [39], [20]. The location coding scheme is based on the insight that if the original ordering of n features can be discarded, then one can achieve an additional $\log(n!)$ bits in savings over and above the original entropy-coded bitstream. The original order in which features are extracted is discarded in both steps of a visual search pipeline: matching with bag-of-words or global descriptors and the GCC step. The location histogram coding scheme is a practical scheme for achieving the $\log(n!)$ rate savings. The ordering gain normalized by the number of features is shown in Figure 6: the ordering gain increases as the number of features increases. For a few hundred features, the ordering gain is 6-8 bits per feature, which is significant considering that each feature is typically encoded with only 32 to 100 bits.

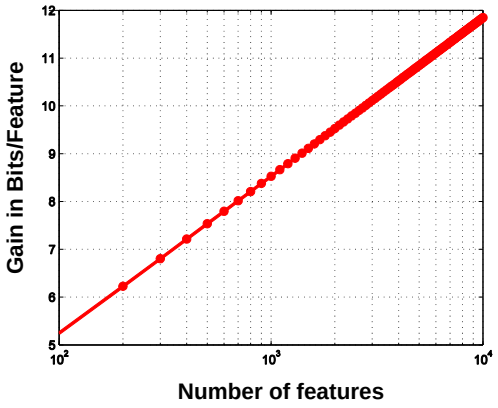

Fig. 6. Additional bitrate savings of $\log(n!)$ can be achieved if the original order of location data is discarded for n features. The ordering gain increases as the number of features increases.

The location histogram coding scheme is presented in Figure 7. Each image is subdivided into non-overlapping blocks of size 3$\times$3, and the location (x, y) data of each feature are quantized to the grid. The location data are then represented as a histogram consisting of (a) a histogram map and (b) histogram counts. The histogram map indicates which bins of the histogram are non-empty, whereas the histogram count indicates the number of features in each non-empty block. The descriptors are re-ordered based on the order of locations in the histogram map.

The histogram count is encoded using a 64 symbol, single model, static arithmetic coding scheme. The histogram map is encoded using a binary context-based static arithmetic coding scheme. As illustrated in Figure 8, the sum of features in neighboring blocks is used as context for encoding the histogram count of a given block. The sum-based context exploits the clustering of feature locations, which is typically found in images [39]. Rather than a raster-scan, a clockwise circular scanning of the histogram map is applied due to the higher density of features at the image center. Furthermore, $\sim$6 bits per feature are used for location coding compared to $\sim$12 bits per feature for lossless location coding with the default block size of 1$\times$1, which applies arithmetic coding to raw location coordinates (note that $\log_2(640 \times 640) = 18.6$ bits per feature are required to store the raw coordinates). The gain results from quantizing location data to a 3$\times$3 grid and the ordering gain discussed above. The lossy location histogram coding scheme results in a negligible decrease in matching performance for both pairwise matching and retrieval experiments [40].

Comparisons with other schemes. Several schemes have

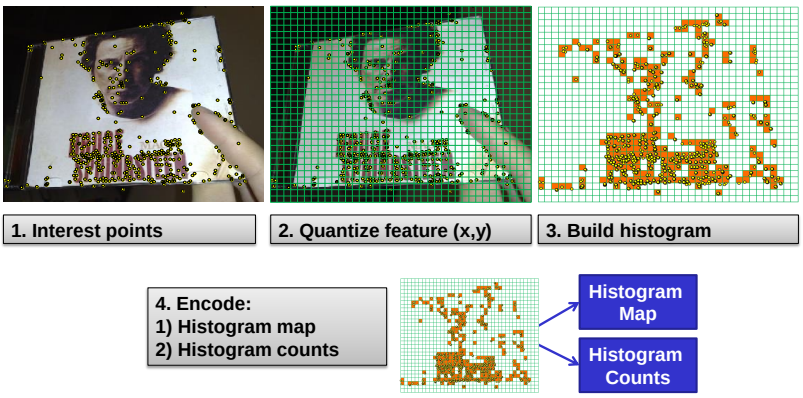

Fig. 7. Pipeline for encoding feature location data. Feature location data are quantized with a uniform grid. The histogram map and counts are subsequently encoded: the resulting scheme achieves an ordering gain of $log(n!)$ for n features.

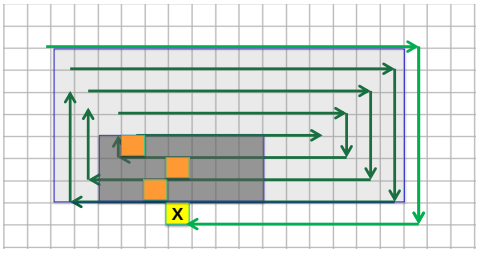

Fig. 8. The sum of features (3 in this case) in the scan neighborhood (dark gray block) is used as context for encoding the bin count of the current block (yellow). As shown, a clockwise circular scanning of the histogram map is applied, beginning with elements located at the center of the image toward elements located at a periphery of the image.

been proposed to achieve the $\log(n!)$ ordering gain. An alternative approach to reordering location data for achieving the ordering gain is to impose an ordering on the descriptor data. A scheme based on coarsely quantizing features with a vocabulary tree is proposed in [7]. Run-length encoding of the non-zero bins in the histogram results in the ordering gain. This approach is not feasible in the CDVS framework because it requires a large dictionary in the encoding step: the memory limit for the entire pipeline is set to 1 MB. Another approach is based on reordering data with binary search trees [42], [43], [44]. The Digital Search Tree (DST) coding scheme in [42], [43] can be applied to data with arbitrary long symbols, whereas the location histogram coding scheme is only applicable in situations where the histogram bins and counts can be explicitly enumerated. The location histogram coding scheme is simpler and slightly outperforms the DST coding scheme [43]. The improvement over the DST scheme is due to the context-based arithmetic coding, which exploits the statistical dependency of neighboring histogram counts.

E. Local Feature Descriptor Aggregation

State-of-the-art image retrieval systems are based on global descriptors such as Vector of Locally Aggregated Descriptors (VLAD) [45] and Fisher Vectors (FV) [48]. The Bag-of-Words (BoW) model also remains a popular choice [4], [50]. In the BoW framework, images are represented by histograms obtained by quantizing descriptors with a large vocabulary tree (e.g., 1 million visual words), and an inverted index is used for fast matching. In the global descriptor framework, images are represented by dense high-dimensional vectors (dimensions of $\sim$ 10K-100K). Finding a compact global descriptor that achieves high performance and requires little memory has been one of the main challenges of the CDVS standardization. The CDVS requirement of low memory (a maximum of 1 MB for the entire encoding process) makes BoW approaches unsuitable for this task [7].

The CDVS standard adopted the Scalable Compressed Fisher Vector (SCFV) [51], [52], [53], [54], [21] after extensive experimentation. The SCFV pipeline, illustrated in Figure 9, is built upon the baseline FV model of [46], [47], [48], [49], [30]. It uses a Gaussian Mixture Model (GMM) with 512 components to capture the distribution of up to 250 local feature descriptors. The gradient of the log-likelihood for an observed set of local feature descriptors with respect to the mean and for higher bitrates, the variances of the GMM are concatenated to form the FV representation [48], [56]. Each descriptor is assigned to multiple Gaussians (visual words) in a soft assignment step. The FV representation requires a considerably smaller vocabulary compared to the BoW model, thus satisfying the specified memory requirements.

Uncompressed FV, stored as floating point numbers, require thousands of bytes, which can be larger than the size of compressed local feature descriptors. To compress the FV, SCFV uses one-bit scalar quantizers, which allows for fast matching with the Hamming distance. At each descriptor length, the bit budget needs to be shared between the compressed global descriptor and a set of compressed local feature descriptors. For this purpose, SCFV uses rate-scalable representations (with an average size of 304, 384, 404, 1117, 1117, and 1117 bytes for the six specified bitrates) by selecting a subset of Gaussian components in the GMM based on the standard deviation of certain components (representing the gradient with respect to component mean) of the fully populated Fisher vector and retaining only the information associated with the

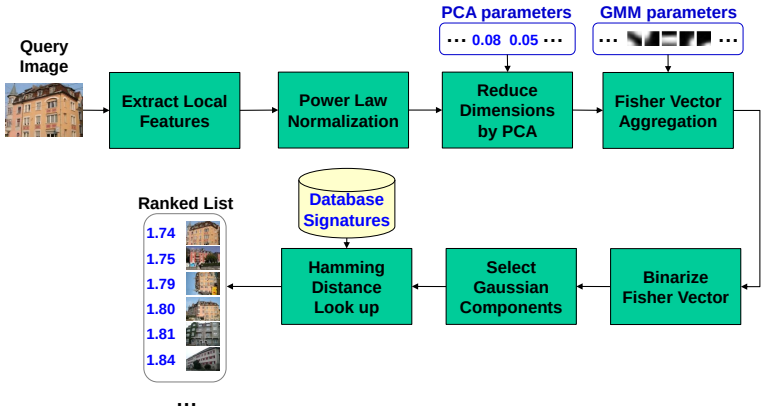

Fig. 9. The global descriptor pipeline in the CDVS standard.

selected components [54]. Extensive experiments showed that the adopted standard deviation-based approach, which excels in removing non- or less discriminative components to form a more robust FV representation, and selecting informative components to undergo less negative performance impact from sign quantization [54], has outperformed the quantization error-based selection method [21]. Specifically, for each Gaussian component i, the standard deviation $\delta(i)$ of the 32-dimensional accumulated gradient vector $g = [g_0, g_1, ..., g_{31}]$ with respect to the mean of that Gaussian component is calculated as:

$$\delta(i) = \sqrt{\frac{1}{32}\sum_{j=0}^{31}\left(g_j - \frac{1}{32}\sum_{k=0}^{31} g_k\right)^2} \tag{6}$$

Then, the Gaussian components are ranked in descending order according to $\delta(i)$. For the three lower descriptor lengths of 512 bytes, 1 KB and 2 KB, the predefined top k Gaussian components are selected; for the three higher descriptor lengths of 4 KB, 8 KB and 16 KB, the i_{th} Gaussian component is selected if $\delta(i) > \tau_\delta$, where τ_δ denotes the selection threshold. The remaining budget is filled with compressed local feature descriptors at each descriptor length. Note that a different set of GMM components are selected for each image based on which components appear to be the most informative. A small set of header bits indicate which components are selected. SCFV includes the log-likelihood gradient with respect to the GMM variance parameters for higher image descriptor lengths (4 KB, 8 KB, and 16 KB). For lower image descriptor lengths, only the gradient with respect to the mean is used.

Given two images X and Y, we can calculate the Hamming distance-based similarity score $S(.)$ of SCFV:

$$S(X,Y) = \frac{\sum_{i=0}^{511} b_i^X b_i^Y w_{Ha(u_i^X, u_i^Y)}(32 - 2Ha(u_i^X, u_i^Y))}{32\sqrt{\sum_{i=0}^{511} b_i^X}\sqrt{\sum_{i=0}^{511} b_i^Y}} \tag{7}$$

where u_i^X denotes the i_{th} binarized Gaussian component (gradient with respect to mean or gradient with respect to variance) in GMM. $b_i^X = 1$ if the i_{th} component is selected; otherwise, $b_i^X = 0$. $Ha(u_i^X, u_i^Y)$ represents the Hamming distance of the i_{th} Gaussian component between X and Y, ranging from 0 to 32. $w_{Ha(u_i^X, u_i^Y)}$ denotes the correlation weights [58] for the i_{th} Gaussian component.

Comparisons with other schemes. A watershed moment in the development of the CDVS standard was the adoption of the 512 byte Residual Enhanced Visual Vector (REVV) global descriptor into the test model TM2.0 [57]. Combining global and local feature descriptors improved the performance at each descriptor length over prior approaches based solely on local feature descriptors [57]. REVV is similar to the VLAD descriptor [45], but it incorporates several improvements to close the performance gap between VLAD and BoW representations [58], [59]. Key enhancements include more effective residual aggregation, dimensionality reduction of residuals using LDA projection matrices, and weighted distance measures for matching. An enhancement titled Robust Visual Descriptor (RVD) [60] was introduced to incorporate soft assignment to quantized words in REVV. Finally, SCFV was adopted by the CDVS standard because it outperformed other global descriptor approaches. SCFV incorporates several new ideas that REVV had introduced, such as learning correlation weights used in signature comparison. SCFV requires significantly less memory than REVV: the main memory requirement is in the descriptor PCA projection matrix (4 KB) and GMM parameters (38 KB), $\sim$ 42 KB in total. Alternate global descriptor compression approaches that require PCA projection matrices for the entire FV or large product quantization tables are prohibitive given the 1 MB memory limit of the CDVS requirements. [61], [45].

IV. Non-normative Blocks

The CDVS standard includes several useful non-normative blocks in the reference software: a block-based scale-space interest point detector called ALP_BFLoG [24], Multi-Block Index Table (MBIT) indexing structure [67], [70] for fast matching of binary global descriptors and a fast GCC algorithm called DISTRAT [66].

ALP_BFLoG employs a block-based scale-space representation [24][25][17][26]. As illustrated in Figure 10, the original scale space is sub-divided into overlapping blocks, and interest point detection is performed on each block independently, thereby significantly reducing the memory cost required by

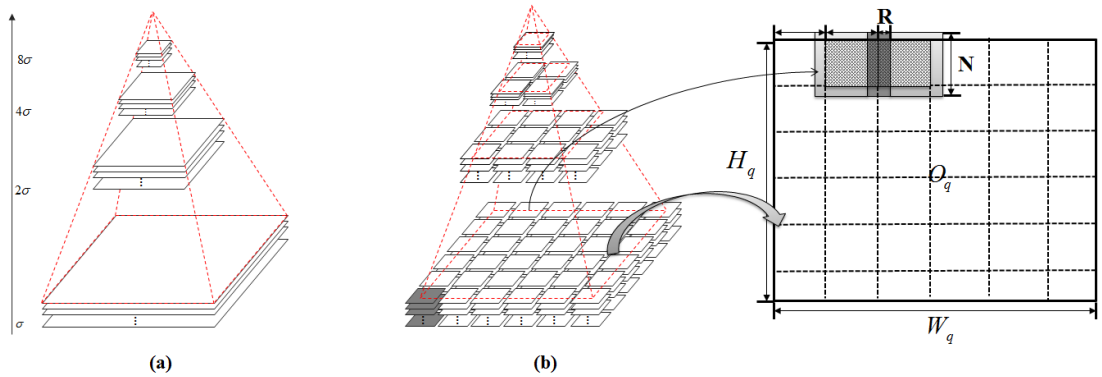

Fig. 10. The image-wise scale space in (a) is approximated by the block-wise scale space in (b) by decomposing each octave image Q_q into square R-pixel overlapped blocks; W_q and H_q are the width and height of Q_q, respectively, and N is the block size.

TABLE I
RUNTIME MEMORY, TIME COST, AND PERFORMANCE COMPARISON BETWEEN ALP_BFLOG, ALP AND DOG OVER MPEG CDVS BENCHMARK DATASETS. (TESTED ON A WINDOWS PC WITH AN INTEL CORE CPU I5 3470@3.2 GHZ). NOTE THAT THE RUNTIME MEMORY COST IS GREATER THAN THE THEORETICAL MEMORY USE IN TERMS OF FILTERS AND BUFFERS AS REPORTED ABOVE.

	Time	Memory	TPR @ FPR<0.1%	TPR @ FPR<0.5%	TPR @ FPR<1%
ALP_ BFLoG	40ms	5.3MB	87.3%	92.9%	93.9%
ALP	49ms	14.4MB	87.5%	92.4%	93.4%
DoG [22]	74ms	30.2MB	85.5%	91.3%	93.2%

filters and scale-space buffers. The block-wise scale-space implementation BFLoG allows for fast frequency domain filtering, thus reducing computational cost. In the original spatial domain (such as the spatial ALP default implementation [23] and the well-known DoG detector [22]), convolution requires that each pixel in the input image is linearly combined with its neighbors, and the neighborhood size increases with scale. The equivalent multiplication in the Discrete Fourier Transform (DFT) domain is independent of scale. Because each block has a fixed size, convolution filters are pre-computed in the DFT domain at different scales.

There is a trade-off in the block size, overlap step size, computational complexity and performance of the detector. The parameters are selected based on minimizing scale-space distortion constraints, which are subject to computational complexity constraints [17]. The optimal block size is 128×128 pixels, with an overlap of 16 pixels selected empirically to maximize BFLoG performance [24]. The BFLoG and ALP approaches were integrated to achieve the block-based scale-space interest point detector ALP_BFLoG to significantly reduce the complexity of the spatial ALP default implementation, where the block-wise processing incorporates LoG filtering, extrema detection, and orientation assignment of keypoints.

ALP_BFLoG provides two major advantages. First, ALP_BFLoG significantly reduces the footprint of filters and buffers to 850 KB, an order of magnitude smaller than the full scale-space representation of the default ALP implementation and the well-known DoG detector, which require 5.98 MB and 12.9 MB [22], respectively. Second, frequency domain filtering is used to reduce the computational complexity by ∼18% compared with the spatial ALP implementation. As listed in Table I, ALP_BFLoG has yielded comparable or slightly better performance in pairwise matching experiments, at significantly reduced runtime memory and time cost (excluding the time cost of computing local feature descriptions).

MBIT is an indexing structure for significantly improving search efficiency over large-scale image databases. For the long binary global descriptor, even though the Hamming distance can be computed very rapidly, the accumulated computational cost from exhaustive linear searches between query and database images increases linearly with the descriptor length and the scale of the image database.

MBIT reduces the exhaustive distance computing between global signatures to the problem of aligned component-to-component independent matching and constructs multiple hash tables for these components. Given a global query descriptor, its candidate neighbors can be retrieved using the query binary sub-vectors (i.e., components) as indices into their corresponding hash tables, thereby significantly reducing the required number of candidate images for subsequent linear searches. MBIT achieves a 10∼25-fold speedup over the exhaustive linear search while maintaining comparable search accuracy [70], [71].

DISTRAT is built upon a probabilistic model of wrongly matched interest points, aiming to rapidly determine whether two images contain views of the same object. DISTRAT uses the log distance ratio statistics to model outliers and inliers by assuming that the log distance ratio statistics of incorrect matches are distinct from that of correct matches. A goodness-of-fit test is performed efficiently. DISTRAT is employed to achieve very fast GCC (approximately 200∼400 times faster than RANSAC) [66].

V. EVOLUTION OF THE STANDARD

A Call for Proposals (CfP) [11], [12] was issued at the 97th MPEG meeting (Torino, July, 2011). From the 99th meeting (San Jose, Feb. 2012), the CDVS standardization entered the collaborative development phase through the definition of a

software Test Model (TM) available to all participants. A series of Core Experiments (CE) were defined to improve different software modules in the TM. The CDVS standardization entered the Committee Draft (CD) stage at the 106th meeting (Geneva, Oct. 2013), the Draft of International Standard (DIS) at the 108th meeting (Valencia, Apr. 2014), and the Final Draft of International Standard (FDIS) at the 110th meeting (Strasbourg, Oct. 2014).

A. Evaluation Framework

The MPEG-7 CDVS benchmark data set consists of 5 classes: *graphics*, *paintings*, *video frames*, *landmarks*, and *common objects*. Example images are shown in Figure 11. Both pairwise matching and large-scale image retrieval experiments are conducted in the evaluation framework, and participants are required to report performance at 6 pre-defined descriptor lengths: 512 bytes, 1K, 2K, 4K, 8K and 16K. In particular, to evaluate the interoperability, experiments of matching with different descriptor lengths (1K vs. 4K and 2K vs. 4K) are also included in the pairwise matching experiment.

The True Positive Rate (TPR) at less than 1% False Positive Rate (FPR) (a specific point on the ROC curve) is used to evaluate the pairwise matching performance. Ground-truth data of 10,155 matching image pairs and 112,175 non-matching image pairs are provided for the pairwise matching experiment. The mean Average Precision and the success rate of top match (precision at 1), are used to evaluate the retrieval performance. A total of 8314 query images, 18840 reference images and a distractor set of 1 million images from Flickr [16] are provided for the retrieval experiment. Details of the data set are provided in Table II. The *common objects* data set has the same set of images as the popular UKBench data set [62]. The data set is an order of magnitude larger than popular data sets such *INRIA Holidays* and *Oxford Buildings*. Additionally, the data set has considerably more variety in scale, rotation, occlusion and lighting conditions than the *INRIA Holidays* and *Oxford* data sets, as discussed in [15]. In addition, the query images of *graphics* data set are divided into 3 subsets called $1a$, $1b$ and $1c$. The $1a$ subset consists of original query images; the $1b$ subset consists of the down-sampled images of $1a$, such that the largest of the vertical and horizontal image dimensions is equal to 640 pixels; and the $1c$ subset consists of images obtained by applying a JPEG compression factor 20 to the $1b$ images. These 3 subsets are used to study the effects of image resolution and compression quality on performance. In addition, average values of localization accuracy parameters are produced for all pairs detected as matching from the set of annotated matching pairs for each pairwise matching experiment $1a$, $1b$, and $1c$. The localization accuracy is measured using the ratio of area of both quadrilaterals that overlap vs. the total area filled by both quadrilaterals. Readers are referred to [63] for more details of the CDVS evaluation framework. The data set is available for download at [64], [65].

B. Timeline

Table IV shows the progression of the CDVS standardization over the course of 18 MPEG meetings. Table III lists the TM milestones with the adoption of core techniques, and Table V provides the key performance improvements of TM. Remarkable performance improvements have been made over the course of the standardization. Table V shows the average mAP, TPR and success rate of Top Match of TM1(TMuC)∼TM11. Comparing TMuC and TM 11.0, CDVS has witnessed significant performance improvements, namely, in terms of the average performance over all datasets and all descriptor lengths. The mAP increased from 71.5% to 84.9% (+13.4%), the success rate of top match increased from 81.3% to 90.9% (+9.6%), and TPR increased from 90.4% to 93.4% (+2.9%). Figures 12, 13, 14 and 15 show the major (intermediate) results in the evolution of the standard from start to finish. Referring to Table III, key performance improvements result from the incorporation of a global descriptor (REVV followed by SCFV) and from the selective aggregation of features, in which statistically optimized feature selection has a large impact on performance throughout.

As illustrated in Figures 12, 13 and 14, the substantial performance improvements were achieved by the adoption of global descriptors REVV (TM2), SCFV (128 Gaussians, TM4), enhanced SCFV (with the gradient vector with respect to the variance, TM5), and improved SCFV (512 Gaussians, TM10). The detailed performance gains are listed in Table V. In particular, feature selection consistently played an important role in both BoW aggregation in TM1 and SCFV aggregation since TM4. Compared to the aggregation of randomly sampled local features, the selective aggregation led to a remarkable mAP increase of more than 25% [30].

Geometric verification is crucial for eliminating false feature

TABLE IV
THE PROGRESSION OF MPEG CDVS STANDARDIZATION.

MPEG Meeting	Place,Date	Action
93	Geneva, Jul. 2010	Requirement
97	Torino, Jul. 2011	Final Call for Proposals issued
98	Oct.28-Dec.02, 2011	Initial evaluation of proposals
101	Stockholm, Jul. 2012	Working Draft
106	Geneva, Oct. 2013	Committee Draft
108	Valencia, Apr. 2014	Draft of International Standard
110	Strasbourg, Oct. 2014	Final Draft of International Standard

TABLE V
THE AVERAGE MAP, TPR AND SUCCESS OF TOP MATCH OF TMUC∼TM14 OVER CDVS DATASETS.

Test	mAP		TPR		Top Match	
Model	value	gain	value	gain	value	gain
TMuC	71.5%	-	90.4%	-	81.3%	-
TM2	75.7%	+4.2%	90.2%	-0.2%	84.0%	+2.7%
TM3	77.1%	+1.4%	91.3%	+1.1%	85.2%	+1.2%
TM4	81.4%	+4.3%	91.5%	+0.2%	88.8%	+3.6%
TM5	83.0%	+2.6%	92.0%	+0.5%	89.9%	+1.1%
TM6	82.9%	-0.1%	91.9%	-0.1%	89.7%	-0.2%
TM7	82.9%	0%	91.8%	-0.1%	89.8%	+0.1%
TM8	82.2%	-0.7%	92.0%	+0.2%	89.1%	-0.7%
TM9	83.7%	+1.5%	93.2%	+1.2%	90.2%	+1.1%
TM10	84.6%	+0.9%	93.2%	0%	90.6%	+0.4%
TM11	84.9%	+0.3%	93.4%	+0.2%	90.9%	+0.3%
TM12	84.8%	-0.1%	93.3%	-0.1%	90.9%	0%
TM13	84.7%	-0.1%	93.3%	0%	90.9%	0%
TM14	84.7%	0%	93.3%	0%	90.9%	0%

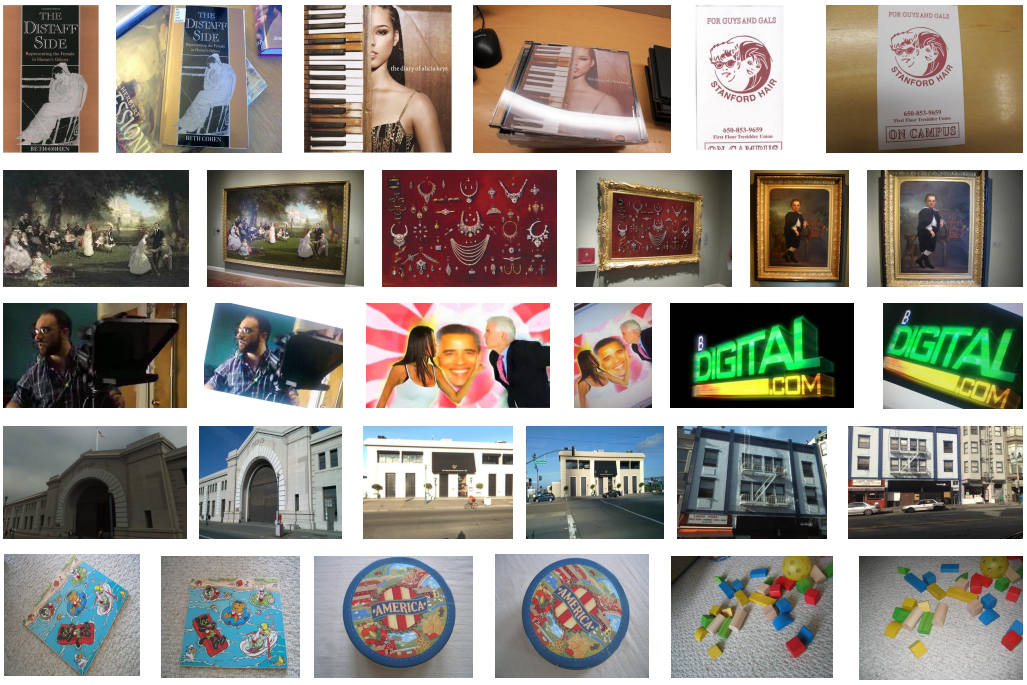

Fig. 11. Example images from the CDVS datasets (from top to bottom: graphics, paintings, video frames, buildings and objects). The data set consists of images for many different categories captured with a variety of camera phones and under widely varying lighting conditions. Database and query images alternate in each category.

TABLE II
KEY STATISTICS OF THE 5 DATA SETS USED IN THE CDVS EVALUATION FRAMEWORK

Data set	Category	# images	# matching pairs	# non-matching pairs	# retrieval queries	Mean # relevant images per query
1	Graphics	2500	3000	30000	1500	2
2	Museum Paintings	455	364	3640	364	1
3	Video Frames	500	400	4000	400	1
4	Buildings	14935	4005	48675	3499	4
5	Common Objects	10200	2550	25500	2550	3

matches and improving recognition performance. For example, in TM 11.0, geometric re-ranking yields mAP improvements of +3.76%, +4.62%, +1.73%, +3.78%, +4.73% at 4 KB for the *graphics*, *paintings*, *video frames*, *landmark*, and *common objects* data sets, respectively. GCC includes a ratio test followed by a fast geometric model estimation algorithm [66].

C. Computational and Memory Complexity

To enable efficient hardware implementation, substantial efforts were focused on reducing memory for all modules. Overall memory usage significantly decreased from over 400 MB in TM 1.0 to ∼1 MB in the final reference software. This reduction was a result of several technical breakthroughs in local feature descriptor compression, local feature descriptor aggregation, and block-based interest point detection. For local feature descriptor compression, memory cost was reduced from ∼380 MB of product tree structured vector quantization tables to the negligible ∼1 KB required for the transform coding scheme [35], [36]. For interest point detection, the memory cost significantly decreased from ∼20 MB in the original implementation, which required storing the entire scale-space stack of images, to 957 KB in the block-based approach [26], [24], [17]. For local feature descriptor aggregation, the memory cost is only ∼42 KB compared to hundreds of MBs of VQ tables required for previous approaches [61], [45].

The computational complexity of encoding also decreased drastically over the course of the standardization: from ∼500 ms to ∼150 ms (tested on a Windows PC with an Intel Core CPU i5 3470 3.2 GHz). The speed up primarily results from the block-based key point detection and the accelerated feature description step [26], [17], [24], [69], the reduction of the number of descriptors to be computed due to the feature selection stage prior to the time consuming feature description stage (even incurring up to ∼66% of the total time cost of interest point detection and local feature description) [55], and the adoption of the local descriptor scalar compression scheme rather than the tree-structured VQ [35], [36]. Efficient

TABLE III
TEST MODEL (TM) MILESTONES AND ADOPTED TECHNIQUES.

TM	Date	Place	Adopted techniques
TM1.0	Feb. 2012	San Jose, USA	Bag-of-words, feature selection, DISTRAT (m22672) [28].
TM2.0	May 2012	Geneva, Switzerland	Local feature aggregation (REVV, a global descriptor m23578 [57]).
TM3.0	Jul. 2012	Stockholm, Sweden	Scalar quantizer to compress local feature descriptors (m25929 [19]), Location coordinate coding (m25883 [40]).
TM4.0	Oct. 2012	Shanghai, China	Scalable local feature aggregation (SCFV, a scalable global descriptor, m26726 [52]), Multi-Stage Vector Quantization (MSVQ) for local feature descriptor compression (m26727 [34]), weighted matching (m25795 [41]).
TM5.0	Jan. 2013	Geneva, Switzerland	Enhanced SCFV with the addition of accumulated gradient vector with respect to the variance of the Gaussian functions for higher bit rates, i.e., 4 KB, 8 KB, and 16 KB (m28061 [51]).
TM6.0	Apr. 2013	Incheon, Korea	A block-wise frequency domain LoG filter (BFLoG, m28891 [25]), two-way key point feature matching schemes with slightly improved performance (m29359 [68]), MBIT (a fast indexing structure, m28893 [67]).
TM7.0	Jul. 2013	Vienna, Austria	Software maintenance, no technology was adopted
TM8.0	Nov. 2013	Geneva, Switzerland	A Low-degree Polynomial extrema detector (ALP, m31369 [23]).
TM9.0	Jan. 2014	San Jose, USA	Improved SCFV by increasing the number of Gaussian functions from 128 to 256 and incorporating the bit selection mechanism for the lowest descriptor length of 512 bytes (m32261 [53]).
TM10.0	Apr. 2014	Valencia, Spain	Combination of BFLoG and ALP (The block-wise processing has incorporated LoG filtering, extrema detection, and orientation assignment of keypoints, m33159 [24]), further improved SCFV by increasing the number of Gaussian functions from 256 to 512, and introducing the standard deviation-based selection method of Gaussian functions (m33189 [54]).
TM11.0	Jul. 2014	Sapporo, Japan	Software maintenance, no technology was adopted
TM12.0	Oct. 2014	Strasbourg, France	Software maintenance, no technology was adopted
TM13.0	Feb. 2015	Geneva, Switzerland	Software maintenance, no technology was adopted
TM14.0	Jun. 2015	Warsaw, Poland	Software maintenance, no technology was adopted

pairwise matching and retrieval algorithms are also made available as part of the reference software. Pairwise matching requires $\sim$0.5 ms per image pair. Retrieval takes $\sim$2.02 sec per query for highest performance and $\sim$0.2 sec per query with a negligible mAP drop of 1% for the 1 million image retrieval experiment.

VI. CONCLUSION

We have reviewed the scope and development of the MPEG CDVS standard. The CDVS standard provides a standardized bitstream of descriptors and the descriptor extraction process, which ensures the interoperability between mobile and server toward mobile image-based retrieval applications. By thoroughly testing state-of-the-art visual search technology proposals and performing competitive and collaborative experiments within a rigorous evaluation framework, the CDVS working group has made remarkable improvements in achieving high image retrieval performance with extremely compact feature data (a few KBs per image). High performance is achieved while also satisfying stringent memory and computational complexity requirements at each step of the feature extraction pipeline, thus making the standard ideally suited for both hardware and software implementations.

REFERENCES

[1] Google-Goggles, 2009, http://www.google.com/mobile/goggles/.

[2] Amazon, *Amazon Flow*, 2011, http://a9.amazon.com/-/company/flow.jsp.

[3] Layar, *Layar*, 2010, http://www.layar.com.

[4] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, "Mobile Visual Search," *IEEE Signal Processing Magazine*, vol. 28, no. 4, July 2011, pp. 61–76.

[5] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, Y. Reznik, and B. Girod, "Compressed Histogram of Gradients: A Low Bitrate Descriptor," *International Journal of Computer Vision, Special Issue on Mobile Vision*, vol. 96, no. 3, pp. 384–399, January 2012.

[6] D. Chen and B. Girod, "Memory-efficient Image Databases for Mobile Visual Search," *IEEE MultiMedia Magazine*, vol. 21, no. 3, pp. 14–23, 2014.

[7] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod, "Tree Histogram Coding for Mobile Image Matching," in *Proceedings of IEEE Data Compression Conference (DCC)*, Snowbird, Utah, March 2009, pp. 143–152.

[8] R. Ji, L.-Y. Duan, J. Chen, H. Yao, Y. Rui, S.-F. Chang, and W. Gao, "Towards Low Bit Rate Mobile Visual Search with Multiple-channel Coding," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 573–582. [Online]. Available: http://doi.acm.org/10.1145/2072298.2072372

[9] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao, "Location Discriminative Vocabulary coding for Mobile Landmark Search," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 290–314, 2012.

[10] V. Chandrasekhar, D. M. Chen, A. Lin, G. Takacs, S. S. Tsai, N. M. Cheung, Y. Reznik, R. Grzeszczuk, and B. Girod, "Comparison of Local Feature Descriptors for Mobile Visual Search," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Hong Kong, September 2010, pp. 3885–3888.

[11] "Compact Descriptors for Visual Search: Evaluation Framework," *ISO/IEC JTC1 SC29 WG11 output document N12202*, July 2011.

[12] "Compact Descriptors for Visual Search: Call for Proposals," *ISO/IEC JTC1 SC29 WG11 output document N12201*, July 2011.

[13] Y. Reznik *et al.*, "Compact Descriptors for Visual Search: Evaluation Framework and Requirements," in *ISO/IEC JTC1/SC29/WG11/N11531*.

[14] S. Paschalakis *et al.*, "Information Technology - Multimedia content descriptor interface - Part 13: Compact Descriptors for Visual Search," in *ISO/IEC 15938-13:2015* .

[15] V. Chandrasekhar, D.M.Chen, S.S.Tsai, N.M.Cheung, H.Chen, G.Takacs, Y.Reznik, R.Vedantham, R.Grzeszczuk, J.Back, and B.Girod, "Stanford Mobile Visual Search Data Set," in *Proceedings of ACM Multimedia Systems Conference (MMSys), San Jose, California, February 2011, pp. 117–122*.

[16] B. T. Mark J. Huiskes and M. S. Lew, "New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative," in

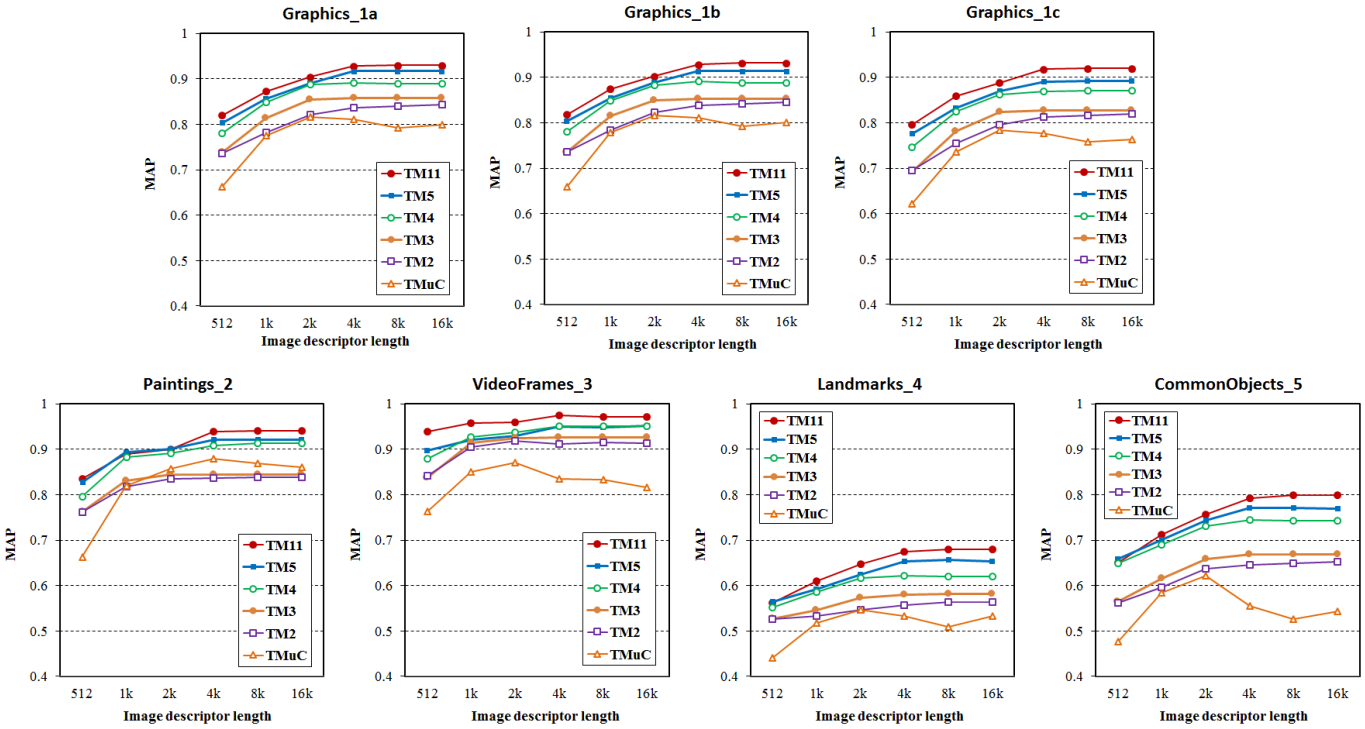

Fig. 12. Results of the Mean Average Precision (mAP) of TMuc, TM2, TM3, TM4, TM5 and TM11 over the CDVS evaluation framework.

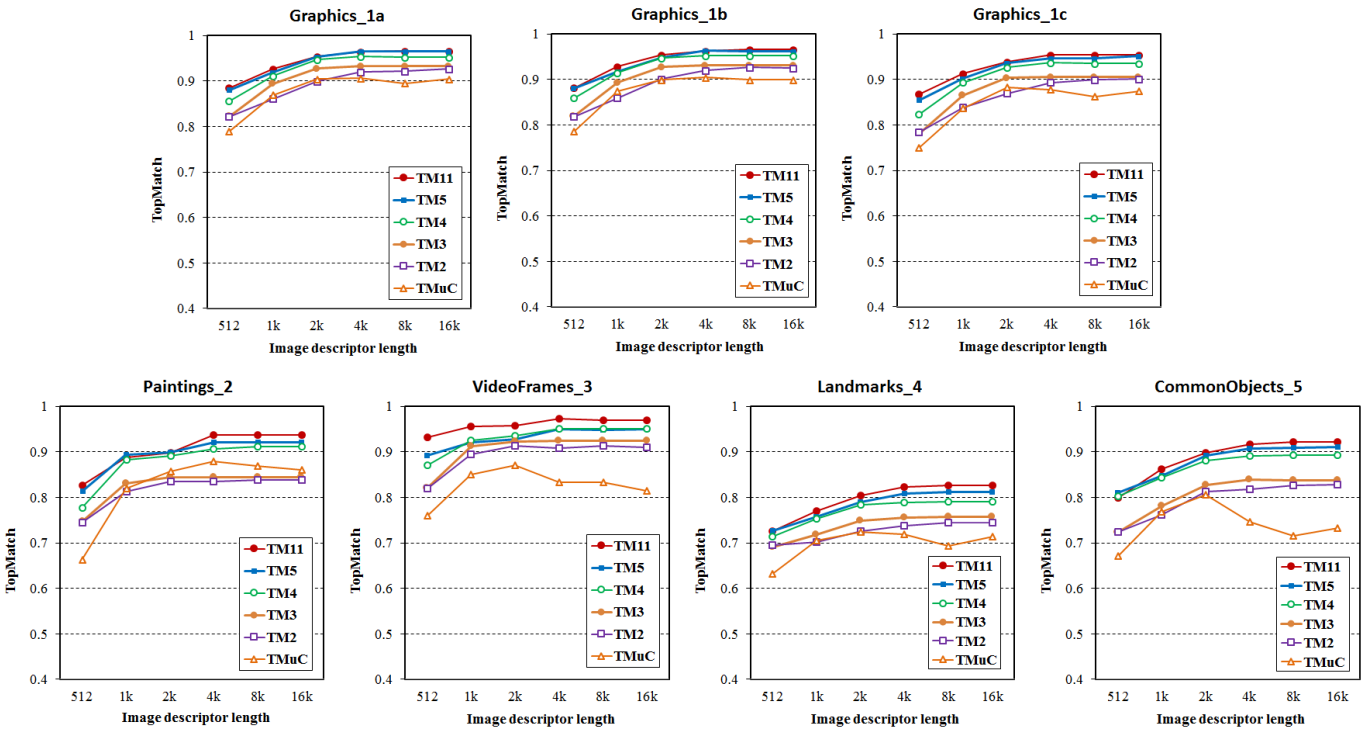

Fig. 13. Results of the success rate of top match of TMuc, TM2, TM3, TM4, TM5 and TM11 over the CDVS evaluation framework.

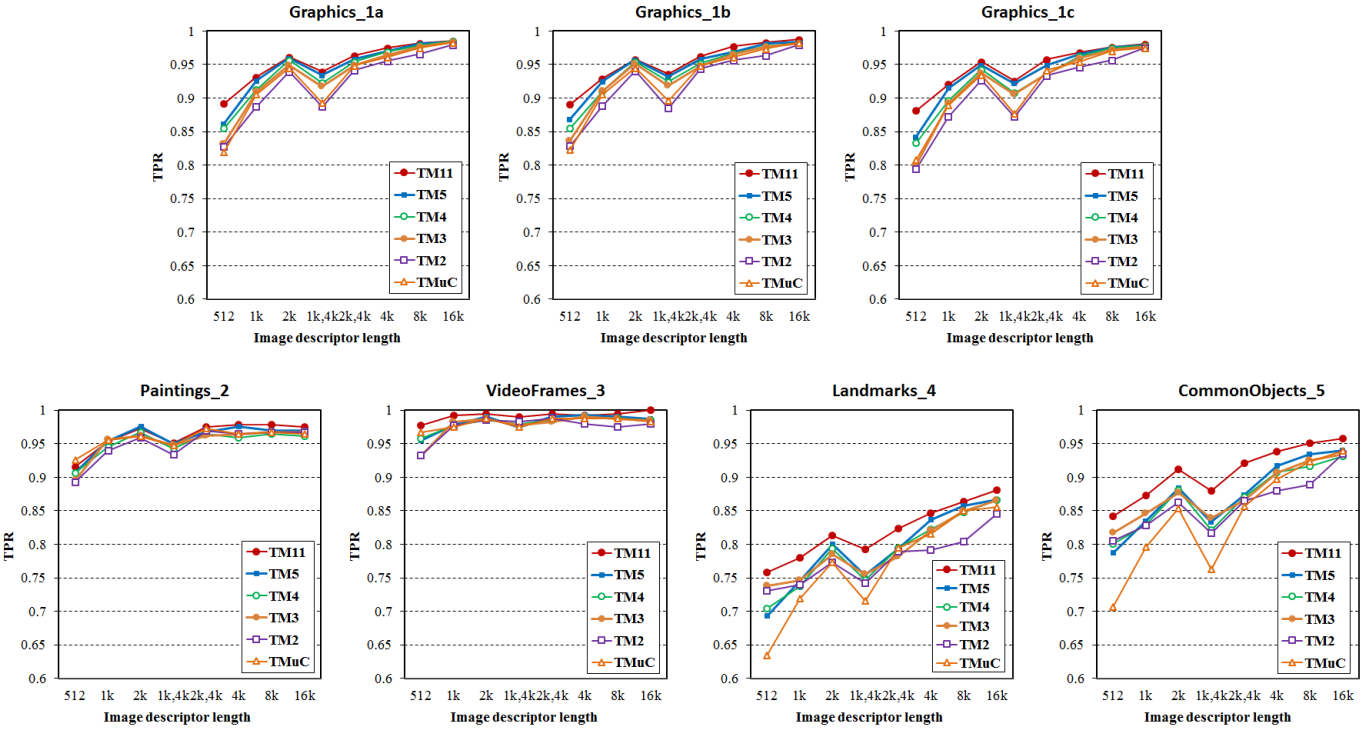

Fig. 14. Results of the True Positive Rate (TPR) of TMuc, TM2, TM3, TM4, TM5 and TM11 over the CDVS evaluation framework. Image descriptor length 1K, 4K and 2K, 4K means matching with different image descriptor lengths: 1K vs. 4K and 2K vs. 4K.

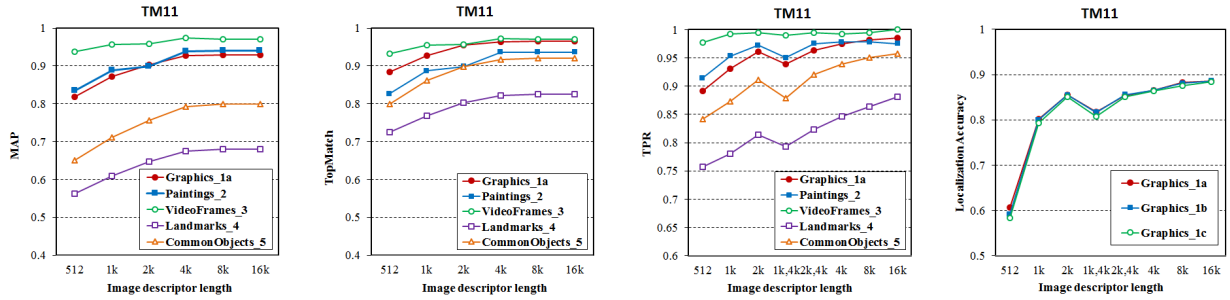

Fig. 15. The performance of TM11 at image descriptor lengths of 512 B, 1K, 2K, 4K, 8K and 16K.

Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval, New York, NY, USA, 2010, pp. 527–536.

[17] J, Chen *et al.*, "A Low Complexity Interest Point Detector," *Signal Processing Letters, IEEE*, vol. 22, no. 2, pp. 172–176, 2015.

[18] S. L. G. Francini and M. Balestri, "Selection of Local Features for Visual Search," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 311–322, April 2013.

[19] S. Paschalakis, K. Wnukowicz, M. Bober, A. Mosca, M. Mattelliano, G. Francini, S. Lepsoy, and M. Balestri, "CDVS CE2: Local Descriptor Compression Proposal," *ISO/IEC JTC1 SC29 WG11 input document M25929*, July 2012.

[20] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, M. Makar, R. Grzeszczuk, and B. Girod, "Improved Coding for Image Feature Location Information," in *Proceedings of SPIE Workshop on Applications of Digital Image Processing (ADIP) XXXV*, San Diego, California, August 2012.

[21] J, Lin *et al.*, "Rate-adaptive compact fisher codes for Mobile Visual Search," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 195–198, 2014.

[22] D. Lowe, "Distinctive Image Features from Scale-invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[23] G. Francini *et al.*, "CDVS: Telecom Italia's Response to CE1 - Interest Point Detection," in *ISO/IEC JTC1/SC29/WG11/M31369*.

[24] J. Chen *et al.*, "CDVS CE1: A Low Complexity Detector ALP_BFLoG," in *ISO/IEC JTC1/SC29/WG11/M33159*.

[25] F. Wang *et al.*, "Peking University Response to CE2: Frequency domain Interest Point Detector," in *ISO/IEC JTC1/SC29/WG11/M28891*.

[26] J. Chen *et al.*, "PKU Response to CE1: Improved BFLoG Interest Point Detector," in *ISO/IEC JTC1/SC29/WG11/M31398*.

[27] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[28] G. Francini *et al.*, "Telecom Italia's Response to the MPEG CfP for Compact Descriptor for Visual Search," in *ISO/IEC JTC1/SC29/WG11/M22672*.

[29] S. Tsai *et al.*, "Fast Geometric Re-ranking for Image-based Retrieval," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Hong Kong, September 2010, pp. 1029–1032.

[30] J, Lin *et al.*, "Robust Fisher Codes for Large Scale Image Retrieval," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 1513–1517.

[31] H. Jégou, M. Douze, and C. Schmid, "Product Quantization for Nearest Neighbor Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, January 2011.

[32] G. Francini, S. Lepsoy, and M. Balestri, "Telecom Italia Response to the CDVS Core Experiment 2," *ISO/IEC JTC1 SC29 WG11 input document M24737*, April 2012.
[33] "Description of Test Model Under Consideration," *ISO/IEC JTC1 SC29 WG11 N12367*, December 2011.
[34] J, Chen *et al.*, "Peking Univ. Response to CE2 - Local Descriptor Compression," in *ISO/IEC JTC1/SC29/WG11/M26727*.
[35] S. Paschalakis *et al.*, "CDVS CE2: Local Descriptor Compression Proposal," in *ISO/IEC JTC1/SC29/WG11/M25929*.
[36] ——, "CDVS CE2: Local Descriptor Compression," in *ISO/IEC JTC1/SC29/WG11/M28179*.
[37] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, and B. Girod, "Transform Coding of Image Feature Descriptors," in *Proceedings of Visual Communications and Image Processing Conference (VCIP)*, San Jose, California, January 2009.
[38] V. Chandrasekhar, G. Takacs, D. M. Chen, , S. S. Tsai, M. Makar, and B. Girod, "Feature Matching Performance of Compact Descriptors for Visual Search," in *Proceedings of IEEE Data Compression Conference (DCC)*, Snowbird, Utah, March 2014, pp. 3–12.
[39] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, J. P. Singh, and B. Girod, "Location Coding for Mobile Image Retrieval Systems," in *Proceedings of International Mobile Multimedia Communications Conference (MobiMedia)*, London, UK, September 2009.
[40] Z. Wang *et al.*, "Cdvs Core Experiement 3: Standford/Peking/Huawei Contribution," in *ISO/IEC JTC1/SC29/WG11/M25883*.
[41] G. Francini *et al.*, "CDVS: Improved image comparison by weighted matching," in *ISO/IEC JTC1/SC29/WG11/M25795*.
[42] V. Chandrasekhar, Y. Reznik, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "Compressing a Set of Features with Digital Search Trees," in *Proceedings of IEEE International Workshop on Mobile Vision (IWMV), International Conference on Computer Vision (ICCV)*, Barcelona, Spain, October 2011, pp. 32–39.
[43] V. Chandrasekhar, S. S. Tsai, Y. Reznik, G. Takacs, , D. M. Chen, and B. Girod, "Compressing a Set of CHoG Features," in *Proceedings of SPIE Workshop on Applications of Digital Image Processing (ADIP)*, San Diego, California, September 2011.
[44] Y. Reznik, "Coding of Sets of Words," in *Proceedings of IEEE Data Compression Conference (DCC)*, Snowbird, Utah, March 2011, pp. 43–52.
[45] H. Jégou, M. Douze, C. Schmid, and P. Perez, "Aggregating Local Descriptors into a Compact Image Representation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, June 2010, pp. 3304–3311.
[46] F. Perronnin, C. Dance, J. Sanchez, and H. Poirier, "Fisher Kernels on Visual Vocabularies for Image Categorization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, USA, June 2007, pp. 1–8.
[47] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the Fisher Kernel for Large-Scale Image Classification," in *European Conference on Computer Vision (ECCV)*, Heraklion, Crete, Greece, September 2010, pp. 143–156.
[48] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale Image Retrieval with Compressed Fisher Vectors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, June 2010, pp. 3384–3391.
[49] J. Sanchez, F. Perronnin and F. Mensink, "Image Classification with the Fisher Vector: Theory and Practice," in *International Journal of Computer Vision*, vol. 105, no. 3, pp. 222–245, January 2013.
[50] D. Nister and H. Stewenius, "Scalable Recognition with A Vocabulary Tree," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, June 2006, pp. 2161–2168.
[51] J. Lin *et al.*, "Peking Univ. Response to CE1: Performance Improvements of the Scalable Compressed Fisher Codes (SCFV)," in *ISO/IEC JTC1/SC29/WG11/M28061*.
[52] ——, "Peking Scalable Low-memory Global Descriptor SCFV," in *ISO/IEC JTC1/SC29/WG11/M26726*.
[53] ——, "Peking Univ. Response to CE2: The Improved SCFV Global Descriptor," in *ISO/IEC JTC1/SC29/WG11/M32261*.
[54] Z. Wang *et al.*, "Response to CE2: Improved SCFV," in *ISO/IEC JTC1/SC29/WG11/M33189*.
[55] M. Balestri *et al.*, "CDVS: ETRI and TI's response to CE1 - An invariant low memory implementation of the ALP detector with a simplified usage interface," in *ISO/IEC JTC1/SC29/WG11/M31987*.
[56] T. S. Jaakkola *et al.*, "Exploiting Generative Models in Discriminative Classifiers," in Proceeding of *Advances in neural information processing systems*, pp. 487–493, 1999.
[57] D. Chen *et al.*, "Compact Descriptors for Visual Search: Improvements to the Test Model under consideration with a global descriptor," in *ISO/IEC JTC1/SC29/WG11/M23578*.
[58] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual Enhanced Visual Vector as a Compact Signature for Mobile Visual Search," in *Signal Processing, Elsevier, In Press*, vol. 93, no. 8, pp 2316–2327, August 2013.
[59] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, H. Chen, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual Enhanced Visual Vectors for On-device Image Matching," in *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, Monterey, California, November 2011, pp. 850–854.
[60] M. Bober *et al.*, "Improvement to TM6 with A robust visual descriptor - proposal from university of surrey and visual atoms," in *ISO/IEC JTC1/SC29/WG11/M30311*.
[61] Y. Gong and S. Lazebnik, "Iterative Quantization: A Procrustean Approach to Learning Binary Codes," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2011, pp. 817–824.
[62] D. Nistér and H. Stewénius, "Scalable Recognition with A Vocabulary Tree," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006, pp. 2161–2168.
[63] Y. Reznik *et al.*, "Evaluation Framework for Compact Descriptors for Visual Search," in *ISO/IEC JTC1/SC29/WG11/N12202*.
[64] *MPEG CDVS (Compact Descriptors for Visual Search) Dataset*, http://pacific.tilab.com/Dataset-20120210/.
[65] *MPEG CDVS (Compact Descriptors for Visual Search) Benchmark. Stanford Digital Repository*, http://purl.stanford.edu/qy869qz5226.
[66] S. Lepsøy, G. Francini, G. Cordara, and P. P. Gusmao, "Statistical Modelling of Outliers for Fast Visual Search," in *Proceedings of IEEE Workshop on Visual Content Identification and Search (VCIDS)*, Barcelona, Spain, July 2011, pp. 1–6.
[67] Z. Wang *et al.*, "An Indexing Structure to Speed Up Retrieval," in *ISO/IEC JTC1/SC29/WG11/M28893*.
[68] X. Xin *et al.*, "CDVS: 2-Way SIFT Matching to Improve Image Matching Accuracy," in *ISO/IEC JTC1/SC29/WG11/M29359*.
[69] D. Pau *et al.*, "CDVS: STM response to Interest Point Detector CE1," in *ISO/IEC JTC1/SC29/WG11/M30233*.
[70] Z. Wang *et al.*, "Component Hashing of Variable-Length Binary Aggregated Descriptors for Fast Image Search," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Paris, September 2014, pp. 2217–2221.
[71] L.-Y. Duan *et al.*, "Weighted Component Hashing of Binary Aggregated Descriptors for Fast Visual Search," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 828–842, June 2015.

Ling-Yu Duan (M'09) received the Ph.D. degree in Information Technology from The University of Newcastle, Australia, in 2007, the M.Sc. degree in Computer Science from the National University of Singapore, Singapore, and the M.Sc. degree in Automation from the University of Science and Technology of China, Hefei, China, in 2002 and 1999, respectively. Since 2008, he has been with Peking University, Beijing, China, where he is currently a Full Professor with the School of Electrical Engineering and Computer Science. Dr. Duan is leading an image retrieval group (IMRE) in the Institute of Digital Media, Peking University. The IMRE group significantly contributed to the MPEG-CDVS (Compact Descriptors for Visual Search) standard. Since 2012, Dr. Duan is the deputy director of the Rapid-Rich Object Search (ROSE) Lab, a joint lab between Nanyang Technological University, Singapore and Peking University, China, with a vision to create the largest collection of structured domain object database in Asia and to develop rapid and rich object mobile search. Before that, he was a Research Scientist in the Institute for Infocomm Research, Singapore, from 2003 to 2008. His interests are in the areas of visual search and augmented reality, multimedia content analysis, and mobile media computing. He has authored more than 100 publications in these areas.

Vijay Chandrasekhar (M'12) is currently a scientist at the Institute for Infocomm Research. He completed his B.S and M.S. from Carnegie Mellon University (2002-2005), and Ph.D. in Electrical Engineering from Stanford University (2006-2013). His research interests include mobile audio and visual search, large-scale image and video retrieval, machine learning and data compression. He has published more than 60 papers/MPEG contributions in a wide range of top-tier journals/conferences like IJCV, ICCV, CVPR, IEEE SPM, ACM MM, IEEE TIP, DCC, ISMIR, MPEG-CDVS etc, and has filed 7 US patents (1 granted, 6 pending). His Ph.D. work on feature compression led to the MPEG-CDVS (Compact Descriptors for Visual Search) standard, which he actively contributed from 2010-2013. He was awarded the A*STAR National Science Scholarship (NSS) in Singapore in 2002.

Jie Chen is currently pursuing the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, where he is with the Institute of Digital Media. His research topics include data compression and visual search, focusing on compact visual descriptors for large scale visual search.

Jie Lin (M'14) received the Ph.D. from the School of Computer Science and Technology, Beijing Jiaotong University, in 2014. He is currently a research scientist with the Institute of Infocomm Research, Singapore. Before that, he was a research engineer in the Rapid-Rich Object Search lab at Nanyang Technological University in 2014, a visiting Ph.D. student with the Institute of Digital Media at Peking University from 2011 to 2013 and the School of EEE at Nanyang Technological University in 2010, respectively. His research interests include mobile visual search, large scale image/video retrieval and deep learning.

Zhe Wang is currently a master student with the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. He received the Bachelor degree in software engineering from Beijing Jiaotong University, Beijing, China, in 2012. His current research interests include large-scale image retrieval and fast approximate nearest neighbor search.

Tiejun Huang (M'01–SM'12) received the B.S. and M.S. degrees from the Department of Automation,Wuhan University of Technology, Wuhan, China, in 1992 and the Ph.D. degree from the School of Information Technology and Engineering, Huazhong University of Science and Technology, Wuhan, in 1999. He was a Postdoctoral Researcher from 1999 to 2001 and a Research Faculty Member with the Institute of Computing Technology, Chinese Academy of Sciences. He was also the Associated Director (from 2001 to 2003) and the Director (from 2003 to 2006) of the Research Center for Digital Media in Graduate School at the Chinese Academy of Sciences. He is currently a Professor with the National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His research interests include digital media technology, digital library, and digital rights management. Dr. Huang is a member of the Association for Computing Machinery.

Bernd Girod (M'80–SM'97–F'98) is the Robert L. and Audrey S. Hancock Professor of Electrical Engineering at Stanford University, California. Until 1999, he was a Professor in the Electrical Engineering Department of the University of Erlangen-Nuremberg. His research interests are in the area of image, video, and multimedia systems. He has published over 600 conference and journal papers and 6 books, receiving the EURASIP Signal Processing Best Paper Award in 2002, the IEEE Multimedia Communication Best Paper Award in 2007, the EURASIP Image Communication Best Paper Award in 2008, the EURASIP Signal Processing Most Cited Paper Award in 2008, as well as the EURASIP Technical Achievement Award in 2004 and the Technical Achievement Award of the IEEE Signal Processing Society in 2011. As an entrepreneur, Professor Girod has worked with numerous startup ventures, among them Polycom, Vivo Software, 8x8, and RealNetworks. He received an Engineering Doctorate from University of Hannover, Germany, and an M.S. Degree from Georgia Institute of Technology. Prof. Girod is a Fellow of the IEEE, a EURASIP Fellow, a member of the German National Academy of Sciences (Leopoldina), and a member of the National Academy of Engineering. He currently serves Stanfords School of Engineering as Senior Associate Dean at Large.

Wen Gao (S'87–M'88–SM'05–F'09) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is currently a Professor of computer science with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. Before joining Peking University, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. Dr. Gao served or serves on the editorial boards for several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, EURASIP Journal of Image Communications, and Journal of Visual Communication and Image Representation. He has chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. Prof. Gao is a Fellow of the IEEE and a member of China Engineering Academy.