Overview of the MPEG CDVS standard

Ling-Yu Duan, Tiejun Huang, and Wen Gao

The Institute of Digital Media, Peking University, Beijing, China
{lingyu,tjhuang,wgao}@pku.edu.cn

Abstract

Towards mobile visual search, compact visual descriptors have been well advocated in both academic and industry endeavors. Moving Picture Experts Group (MPEG) initiated the remarkable Compact Descriptors for Visual Search (CDVS) standard activity in Jan. 2010 to push forward the frontiers of compact descriptors in mobile internet industry. In Oct. 2014, MPEG CDVS successfully entered the Final Draft of International Standard. CDVS made a series of significant breakthroughs in high performance and low complexity compact descriptors. In this paper, we give an overview of the MPEG CDVS standard, with emphasis on the development of the core techniques and their technical merits.

1. Introduction

Handheld devices like smart phone have great potentials for mobile visual search and augmented reality applications. However, online visual querying involves transferring query images from a mobile device to a remote server. The time consuming query delivery over a slow wireless link may significantly degenerate the quality of user experience. Distinct from processing a textual query, conventional visual search systems have to spend non-negligible computation on extracting features prior to the retrieval. Moreover, frequently sending entire images throughout a 3G or 4G network may incur serious power consumption. Even for on-device local search, the state-of-the-art image representations like the Bag-of-Words (BoW) [1][2] and the aggregated descriptor of local features [3][4] are unsuitable in terms of computational complexity.

To address these challenges, recent research works [5][6][7][8][9][2] have proposed to extract compact visual descriptors directly at the mobile end, and to send the descriptors at low bitrates to the remote server for scalable visual search. The descriptors shall be compact, discriminative, and meanwhile efficient in extraction to reduce the overall query delivery latency. Compact descriptors benefit three typical mobile visual search architectures as follows: 1) Send descriptors as queries; 2) Perform local search first; 3) Send images as queries. For architecture (3), although the query is not in the form of descriptors, the negative impact on search performance from noisy local features extracted from a highly compressed query image can be significantly reduced by the selective aggregation of local feature descriptors [10][11].

In order to push forward the frontiers of compact descriptors in mobile internet industry, Moving Picture Experts Group (MPEG), the formal title “ISO/IEC JTC1 SC29 WG11”, initiated the CDVS standard activity at the 91st MPEG meeting (Kyoto, Jan. 2010)[12]. By thoroughly analyzing, comparing, and testing the state-of-the-art visual search technologies, and performing competitive collaboration experiments to improve the core techniques within a rigorous evaluation framework [13], CDVS made a remarkable breakthrough in high performance and low complexity compact descriptors. This paper presents an overview of MPEG-7 CDVS standard including development progress, technical merits, and performance evaluation.

1068-0314/15 $31.00 © 2015 IEEE
DOI 10.1109/DCC.2015.72

IEEE computer society

2. Scope of MPEG-7 CDVS Standard

The objective of CDVS standard is to provide standardized description of stored or streamed images that helps to design efficient and interoperable visual search applications, allowing visual content matching in images. Visual content matching includes matching of views of textured rigid objects, landmarks, and printed documents. The requirements include robustness, sufficiency, compactness, scalability, image format independence, extraction complexity, matching complexity, and localization [14].

The CDVS standard as MPEG-7 Part 13 specifies the bitstream (i.e., binary representation syntax) of descriptors and the descriptor extraction process [11]. As a normative part, the syntax of descriptors shall conform to the MPEG-7 CDVS standard. Extraction of the CDVS descriptor is normative as well to ensure interoperability. See Figure 1. How the CDVS descriptor is used for search and filtering of content, like CDVS adopted Multi-Block Index Table (MBIT) indexing structure [15] towards a large scale search, or two-way pairwise matching algorithm [16] to improve matching accuracy, is not specified by MPEG-7 CDVS standard [11].

3. Development of MPEG-7 CDVS Standard

MPEG-7 CDVS standard activity develops specification based on a well-defined M-PEG standard development framework, following the steps: requirements for technology is specified [14], technology is requested through an official "Call for Proposals" [17], and the technology proposed to MPEG is evaluated by MPEG experts based on pre-defined performance criteria [13]. Key participants include Stanford Univ., Peking Univ., Surrey Univ., Telecom Italia, Visual Atom, Qualcomm, STMicroelectronics, Huawei, Nokia, NEC, Samsung, ETRI, etc. CDVS standard entered the committee draft (CD) on the 106th meeting (Geneva, Oct. 2013), the Draft of International Standard (DIS) on the 108th meeting (Valencia, Apr. 2014), and the Final Draft of International Standard (FDIS) on the 110th meeting (Strasbourg, Oct. 2014).

Between the 91st and 108th MPEG meeting, the MPEG-7 CDVS received in total 366 input contributions, in which there were 99 proposals on core experiments. From the 99th meeting (San Jose, Feb. 2012), MPEG-7 CDVS standard entered the collaborative development through the definition of a Test Model (TM). A series of core experiments including CE1-Global Descriptor, CE2-Local Descriptor Compression, CE3-Feature Point Location Coding, CE4-Key Point Detection, CE5-Local Descriptor, CE6-Retrieval, CE7-Feature Selection, CE8-Combining Local and Global Descriptors for Pairwise Matching were carried out to specify and implement feature extraction and encoding algorithms. The latest software reference model TM 11.0 was released after the 109th meeting (Sapporo, Jul. 2014) [11].

Significant performance improvements have been made in developing CDVS, i.e., mean Average Precision (mAP) 0.72 vs. 0.85, the success rate of Top Match 0.81 vs. 0.91, True Positive Rate (TPR) 0.90 vs. 0.93, between TM 1.0 and TM 11.0. To meet the requirement of minimum memory consumption, CDVS made substantial efforts in memory reduction with regard to CE1-Global Descriptor, CE2-Local Descriptor Compression, CE4-Key Point Detection. For CE2, the memory cost was reduced from $\sim$380MB to $\sim$1KB [18][19]; for CE4, the memory cost from $\sim$20MB down to 957KB [20]. For CE1, compared with the state-of-the-art [3][4], CDVS made a breakthrough

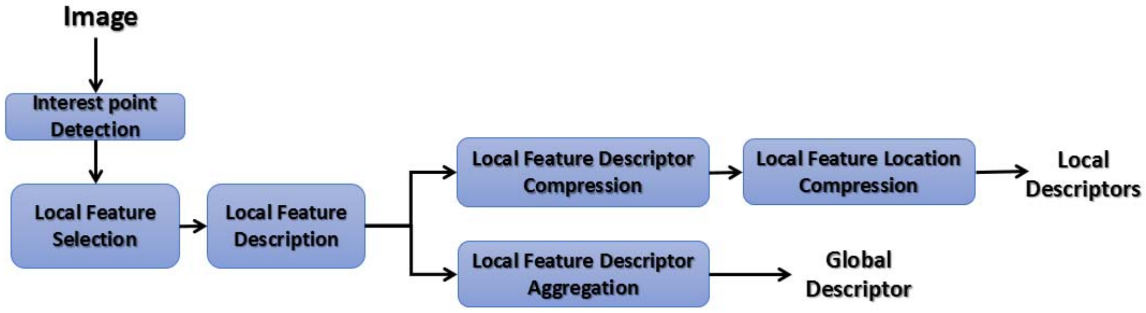
Image
Interest point Detection
Local Feature Selection
Local Feature Description
Local Feature Descriptor Compression
Local Feature Location Compression
Local Descriptors
Local Feature Descriptor Aggregation
Global Descriptor

Figure 1: The MPEG-7 CDVS encoding process.

in memory reduction from hundreds of MB to 42KB [21]. The time cost of the CDVS encoding process was reduced from ∼500ms to ∼150ms (tested on a Windows PC with Intel Core CPU i5 3470 3.2GHz), which was mainly attributed to the speedup of key point detection in CE4 [20][22] as well as local feature description. Aside from the normative encoder part, CDVS has developed efficient and effective pairwise matching and retrieval pipelines as an informative part. Pairwise matching costs ∼0.5 ms per image pair, and retrieval costs ∼2.02 sec. per query (with the best performance) and ∼0.2 sec. per query (with an mAP drop of 0.01) over 1 million image dataset.

4. CDVS Encoding

4.1 Interest Point Detection

Local feature extraction involves detecting interest points and describing the invariant feature of each interest point [23]. CDVS encoding starts with interest point detection. A 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. The normalized derivatives of each scale in an image pyramid are generated, where extrema detection is performed by searching for local extrema to identify interest points[23].

CDVS contributed to the state-of-the-art scale-space representation from two aspects: (1) how to construct a scale-space at low complexity, and (2) how to identify interest points within a scale-space. CDVS proposed a novel block-wise scale-space representation and applied Laplacian of Gaussian (LoG) filtering to implement a block based frequency domain LoG (BFLoG) detector [24][20][25][22]. BFLoG has significantly reduced the footprint of filters and buffers to 956KB, which is much smaller than a normal image-wise scale-space with 12.9MB in the baseline DoG [23]. In addition, the frequency domain filtering brought about substantial time cost reduction by a ratio of 53% on average [22]. To identify interest points, CDVS proposed a low-degree polynomial (ALP) approach [26] to approximate the LoG filtering results by polynomials, used to find extrema in the scale space and to refine the spatial position of the detected points. ALP provided an alternative interest point detector that was distinct from the prior art solutions like the canonical local extrema detection, i.e., comparing the response values of a point to its scale and spatial neighbors (say, 3×3×3). Finally, BFLoG and ALP [26] were integrated, where the block-wise processing involved LoG filtering, extrema detection, and orientation assignment.

BFLoG [20][22] subdivides the original scale space into overlapping blocks and performs interest point detection for each block independently, thereby significantly reducing the memory cost from filters and buffers. Meanwhile, the block-wise scale-space allows for the frequency domain filtering mechanism, thereby reducing time cost. In the spatial domain, each point of an input image is smoothed by its neighbor region where the region radius increases with the scale factor. By contrast, the equivalent of convolution in frequency domain is independent of the scale factor, in which the convolution involves just one dot product operation per point, a Discrete Fourier Transform (DFT), and an Inverse Discrete Fourier Transform (IDFT). As each block is of a uniform size, convolution filters can be pre-computed with respect to different scale factors. BFLoG proposed to derive the optimal block-wise scale-space decomposition by resolving the problem of minimizing the distortion of scale-space, subject to the complexity constraint. CDVS figured out the optimal block setting with the block size of 128×128 pixels, and the overlap of 16 pixels [22][24].

4.2 Local Feature Selection

Interest point detection produces a number of local features which may be greater than the number of local features that is possible to store at a given CDVS image descriptor length, for example, 512 bytes, 1KB, 2KB, 4KB, etc [13]. Therefore, a subset of local features shall be selected on the basis of a relevance measure that is computed for each of the detected local features. Only the most relevant features are packed into the compressed local feature descriptors, or aggregated to form the global feature descriptor. Besides the issue of limited descriptor length, incorporating distracting or unnecessary local features may degrade the discriminating power of compact descriptors. For example, a substantial performance improvement on Holiday dataset from 59.5% up to 67.1% mAP by the selective aggregation of local features was reported in [10]. In addition, subset selection may save considerable computation of local feature description, which is the most time consuming part in the CDVS encoding process.

CDVS has shown that the characteristics of interest points, including the scale, the orientation, the output (also called peak) of the LoG, the distance to the image centre, the ratio of the squared trace of the Hessian, and the second derivative of the scale space function with respect to the scale, embody the probabilistic relevance of correct feature match [27]. By assuming these characteristics are conditionally independent, CDVS learnt the conditional distribution, with respect to each interest point characteristic, to estimate the probability of a feature being matched correctly to some feature in some unknown image. The relevancy for a given feature in an image was obtained by multiplying its conditional probabilities. In order to model the conditional probabilistic distributions, CDVS produced an external training set of matched and non-matched features from matching pairs of images (both images in the pair depict the same object or scene), where a distance ratio test, followed by a geometric consistence check like RANSAC or faster algorithms [28], was applied to determine the correct or incorrect feature matches. For geometric verification, the minimum number of supporting inliers to consider a match pair of images as valid was set to 30, which was much higher than the minimum number of supporting inliers (say, 5 inliers empirically set by CDVS TM) in pairwise matching or retrieval tasks, to obtain the positive samples of feature matches as accurate as possible.

The issue of local feature selection is distinct from the problem of feature subset selection in classic supervised learning tasks [29]. CDVS did not explicitly classify an interest point instance as belonging to irrelevant or redundant features in the context of predicting the class of a pair of images (say, matching versus non-matching) or other class labels. This approach made an empirical assumption that the characteristics of correctly matched features behave, in a statistical sense, consistently across various datasets. By modeling feature relevance values with the conditional distribution, CDVS actually introduced a low complexity approximation of the pairwise matching algorithm to determine valid inliers from the statistical point of view.

4.3 Local Feature Descriptor Compression

Compression can significantly alleviate the storage requirement of local feature descriptors. In contrast to 1024 bits per uncompressed local feature descriptor, CDVS has obtained the compactness of 32, 32, 65, 103, 129, 205 bits per local feature descriptor on average at the descriptor length of 512 bytes, 1KB, 2KB, 4KB, 8KB, 16KB, respectively. CDVS evaluation performed comprehensive performance and complexity study from the perspectives of both vector quantization (VQ) and scalar quantization (SQ). For the CDVS standard specification, scalar quantization was finally adopted.

Since TM 3.0, CDVS adopted a novel scalar quantization based local descriptor compression scheme [18][19]. There are two major stages, namely, the descriptor transform, which entails additions and subtractions of Histogram of Gradient (HoG) bin values, and the ternary scalar quantization of the transformed elements, which entails L1-norm based comparison operations. This scheme is hardware favorable due to fast processing (in terms of transform, quantization and distance calculation), nice scalability (in terms of bitrate transcoding) directly in the compressed domain, as well as the minimum memory cost. The primary memory resource is 128 (elements from the transformed descriptor) $\times$ 2 (ternary SQ thresholds) = 256 bytes [18].

The descriptor transform may impact the recognition performance, as resolving a set of linearly uncorrelated elements (e.g., principle components) and then allocating more bits to the elements with higher variance may increase the discriminative power of the quantized descriptor, especially at lower bitrates. Instead of directly applying a PCA transform to the original descriptor, CDVS employed the subspace decomposition idea of product quantization, and proposed two sets of linear bin combination patterns (akin to an approximated PCA transform) to derive uncorrelated components (also physically meaningful) of the HoG histograms for the 16 non-overlapping subregions, respectively. The memory cost of descriptor transform is nearly zero. In order to further compress the transformed descriptor, CDVS performed element wise ternary quantization and prefix coding. To fulfill the rate scalability, a set of rules were abstracted to progressively select a subset of elements for different bitrates in the compressed domain based on some statistical importance criteria.

However, as vector quantization has been widely used in the state-of-the-art Bag-of-Words models, CDVS performed extensive experiments to investigate technical merits and limitation till TM 5.0. In particular, CDVS presented a Multi-stage Vector Quantizer (MSVQ) [30] to substantially reduce the memory requirement of Look-Up tables (i.e., 38KB [30], versus 60$\sim$150 MB for storing a large codebook containing 0.1$\sim$1 million words) while maintaining comparable matching performance. MSVQ involves the first stage tree structured quantizer for the original feature vector and the second stage product quantizer for the residuals. Searching for the nearest codeword is time consuming, as the L2 norm involves complex operations like subtraction and multiplication. However, as VQ is simple, efficient and widely used in Bag-of-Words models, MSVQ was adopted by TM 4.0 as a reference software implementation.

4.4 Local Feature Descriptor Aggregation

State-of-the-art image retrieval systems are built upon a visual vocabulary model with an inverted indexing structure, which quantizes the local features of query and database images into visual words [1]. Each database image is then represented as a BoW histogram (i.e., a compact image-level representation) and is invert indexed by visual words in the image. To improve search performance, existing BoW approaches introduce a large vocabulary tree (VT) (say, 1 million visual words) [31] to perform fine-grained quantization of local features, soft word assignment to reduce quantization errors, as well as hamming-embedding to allow for a smaller vocabulary [32]. However, to yield higher search performance usually incurs very high memory usage from storing a larger codebook and an inverted index file, given the large number of visual words and images in which the words may appear. The requirement of low memory complexity (say, CEs encourage a minimum level of 1 MB for the whole encoding process) of the CDVS standard makes conventional BoW approaches no longer suitable for generating a compact image descriptor by simply composing the quantized local descriptors of a query image. In particular, frequently updating an inverted index may consume enormous resources. This prohibits the application of BoW approaches in real-time indexing scenarios where newly arriving content must be indexed and made available for search immediately upon arrival, or local search scenarios on resource-constrained devices like mobile phones, set-top boxes, etc.

Since TM 2.0, CDVS has been dedicated to high performance and low complexity aggregation techniques to derive a global descriptor from local feature descriptors. A novel “Global + Local” strategy was established at the MPEG 100th meeting, which combining an image signature obtained by aggregating uncompressed local descriptors and a subset of compressed local descriptors to form compact descriptors at different bitrates. A novel compact signature Residual Enhanced Visual Vector (REVV) [7][33] was adopted by TM as a reference aggregation technique. Towards low memory complexity, REVV elegantly improved the Vector of Locally Aggregated Descriptor (VLAD) [3] with key enhancements in residual aggregation, dimensionality reduction, and signature comparison. Another competitive contribution robust visual descriptor (RVD) [34] introduced the soft assignment of quantized words to REVV. The ”Global + Local” solution is cost-efficient in processing and storing the index in limited amount of memory, in addition to high matching and retrieval performance.

In the standard specification, Fisher kernel [4][35] was employed to aggregate local feature descriptors by deriving the Fisher vector (FV) representation. VLAD may be considered as a non-probabilistic simplified version of FV. Finally, the Scalable Compressed Fisher Vector (SCFV) [36][37][38][21][39] was adopted by the CDVS standard. SCFV incorporated valuable elements of REVV like learning correlation weights for signature comparison. The primary memory resource is 4,224 bytes (PCA projection matrix) + 37,888 bytes (GMM parameters) = 42, 112 bytes [21].

FV employs a Gaussian Mixture Model (GMM) to estimate the distribution of local feature descriptors over a training feature set. The gradient vector [4][35] of the set of local descriptors in a query image with respect to the mean (1-order) and variance (2-order) parameters of the GMM are computed to form the FV representation. Actually, both the BoW histogram and the Fisher Vector belong to the representation of aggregating local feature descriptors. BoW encodes the 0-order statistics by counting word occurrence, while FV encodes the 1-order and 2-order statistics in addition to the 0-order statistics. FV incorporates the soft assignment of a local feature descriptor to multiple Gaussians (words). Accordingly, FV yields better performance than BoW by using a much smaller vocabulary (containing a few hundreds of words or Gaussian centroids). Memory complexity is much reduced as well.

SCFV further proposed to compress the raw FV by an one-bit scalar quantizer. The binary code supports fast Hamming distance computing. To make economic use of bit budget between local and global descriptors, SCFV offered a rate scalable representation (with an average size of 304, 384, 404, 1117, 1117, 1117 bytes for six bitrates, respectively) by selecting a subset of Gaussian centroids in the GMM and retaining the gradient vector for these selected centroids. When additional bit budget is available at higher bitrates (4KB and more), the discriminative power of SCFV can be enhanced by adding the gradient vector with respect to the variance [36].

4.5 Local Feature Location Compression

Geometric verification is crucial to eliminate false feature matches and to improve the recognition performance. For example, in TM 11.0, the geometric re-ranking yields remarkable mAP improvements of +3.76%, +4.62%, +1.73%, +3.78%, +4.73% at 4KB bitrate for all the CDVS datasets Graphics, Paintings, Video Frames, Landmark, and Common Object. Geometric consistency check consists of the ratio test and the geometric model verification. The local feature location information is required to estimate a geometric transformation model by RANSAC or faster algorithms [28].

CDVS proposed a novel local feature location lossy compression scheme [40][41]. An image is subdivided into a matrix of non-overlapping blocks of size 3×3. The location information is converted into a location histogram containing the histogram map and the histogram count. The histogram map is formed by the map of empty and non-empty histogram blocks, while the histogram count is the number of features in the non-empty block. The histogram count is encoded using a 64-symbols, single model, and static arithmetic coding scheme. The histogram map is encoded using a binary context-based arithmetic coding scheme. CDVS has yielded the rate of ∼6 bits per feature versus ∼12 bits per feature with lossless coding (i.e., block size 1×1). However, for bitrates 2KB or more, the addition of 5% more local features from lossy coding does not impact matching accuracy (say, TPR gain < +0.25%). Though lossless coding has good potentials to improve the localization accuracy in some promising applications like robotic navigation, CDVS evaluation framework limits further study due to the issues of dataset annotation and performance measurements.

5. CDVS Performance Evaluation

The MPEG-7 CDVS benchmark involves a million-scale image dataset including 8,314 query images vs. 18,840 reference images categorized into Graphics, Painting, Frame,

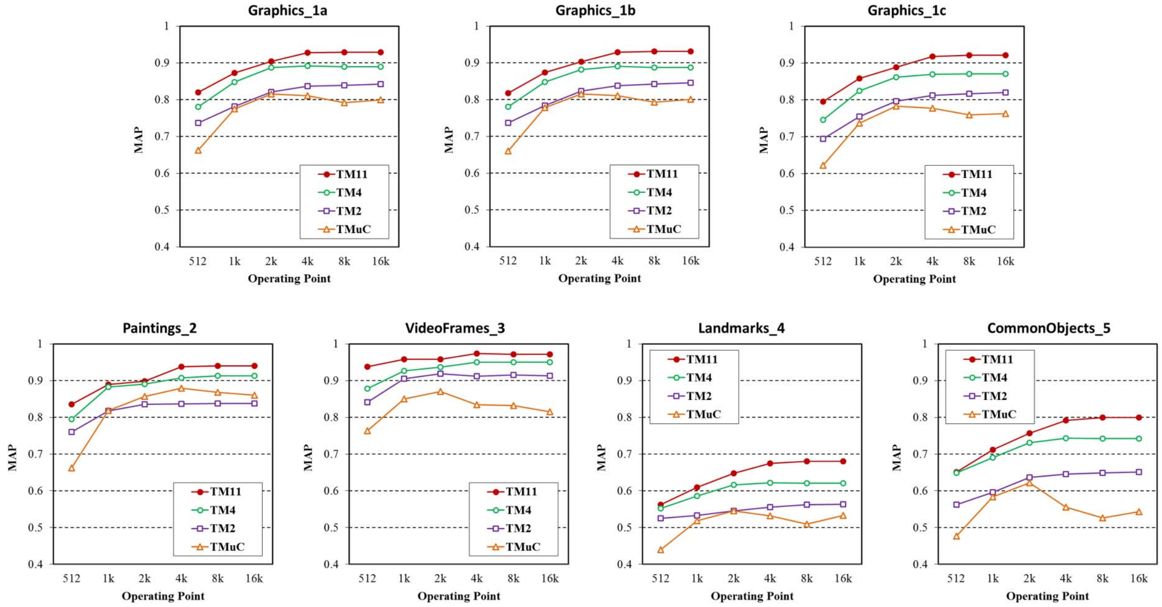

Figure 2: The retrieval performance improvements with milestone technology adoption.

Landmark, Common Objects, 10,155 matching pairs vs. 112,175 non-matching pairs, and a distractor set FLICKR1M of 1 million images collected from Flickr. The mean Average Precision and the success rate of Top Match, are used to evaluate the retrieval performance, while the True Positive Rate at less than 1% False Positive Rate is applied to evaluate the pairwise matching performance. Readers are referred to [13] for more details on the CDVS evaluation framework.

Remarkable performance improvements have been made in developing the CDVS standard. Comparing TM 1.0 (TMuC) to TM 11.0, the average performance gains over all bitrates and datasets are mAP 0.72 vs. 0.85, the success rate of Top Match 0.81 vs. 0.91, TPR 0.90 vs. 0.93. Figure 2 illustrates the retrieval performance comparison, in which the substantial performance improvements were mainly attributed to the adoption of the milestone global descriptors REVV, SCFV (128 Gaussians), improved SCFV (512 Gaussians). In particular, feature selection consistently played an important role. Compared to the aggregation of randomly sampled local features, the selective aggregation led to a remarkable mAP increase of over 0.25 [10]. In addition, memory usage substantially decreased from over 400MB down to ∼1MB . Several technical breakthroughs in local feature descriptor compression, local feature descriptor aggregation SCFV, as well as interest point detection BFLoG_ALP, contributed to the minimum primary memory use of ∼1KB, ∼40KB, ∼957KB in these modules, respectively. Overall, descriptor transform, scalar quantization, as well as subdividing (or tiling) process played a significant role in complexity reduction.

6. Summary

We have reviewed the scope and development of the MPEG-7 CDVS standard. CDVS made remarkable progress in high performance and low complexity compact descriptors. Several issues remain open, including but not limited to the "true" minimum

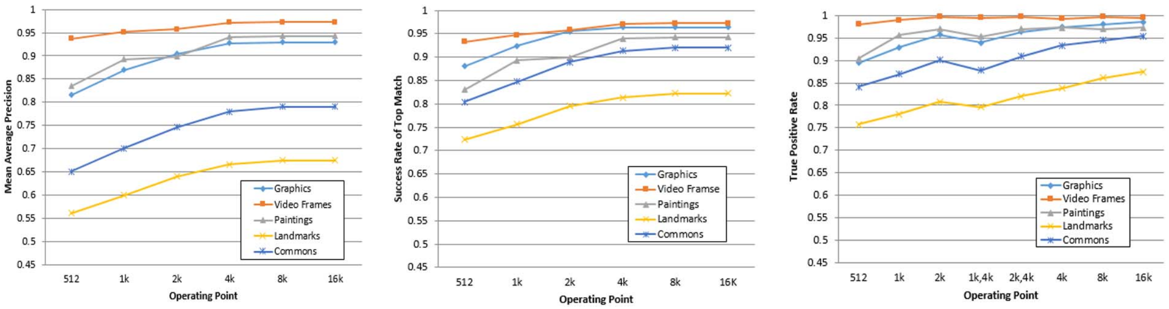

Figure 3: The performance of MPEG-7 CDVS Test Model 11.0.

normative part with sufficient interoperability, the flexibility of feature selection to maximize search performance with a toleration of minor interoperability loss, the scalability in dealing with a very large scale image dataset (say, 100 millions), etc.

Acknowledgements

The authors would like to thank MPEG CDVS experts for their dedicated efforts in collaboratively developing competitive feature descriptor coding and visual search techniques. This work was supported by the National Natural Science Foundation of China under grant 61271311, 61390515, 61421062, in part by the National Basic Research Program of China (973 Program): 2015CB351806.

References

[1] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *ICCV*, 2003.

[2] D. Chen *et al.*, “Tree histogram coding for mobile image matching,” in *DCC*, 2009.

[3] H. Jégou *et al.*, “Aggregating local descriptors into a compact image representation,” in *CVPR*, 2010.

[4] F. Perronnin *et al.*, “Large-scale image retrieval with compressed fisher vectors,” in *CVPR*, 2010.

[5] B. Girod *et al.*, “Mobile visual search,” *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 61–76, 2011.

[6] V. Chandrasekhar *et al.*, “Compressed histogram of gradients: A low-bitrate descriptor,” *International journal of computer vision*, vol. 96, no. 3, pp. 384–399, 2012.

[7] D. Chen and B. Girod, “Memory-efficient image databases for mobile visual search,” *IEEE MultiMedia Magazine*, vol. 21, no. 3, pp. 14–23, 2014.

[8] R. Ji *et al.*, “Towards low bit rate mobile visual search with multiple-channel coding,” in *ACM Multimedia*, 2011.

[9] ——, “Location discriminative vocabulary coding for mobile landmark search,” *International Journal of Computer Vision*, vol. 96, no. 3, pp. 290–314, 2012.

[10] J, Lin *et al.*, “Robust fisher codes for large scale image retrieval,” in *ICASSP*, 2013.

[11] S. Paschalakis *et al.*, “Information technology - multimedia content descriptor interface - part 13: Compact descriptors for visual search,” in *ISO/IEC DIS 15938-13.*

[12] Y. Reznik, “On mpeg work towards a standard for visual search,” in *SPIE*, vol. 8135, 2011.

[13] Y. Reznik *et al.*, “Evaluation framework for compact descriptors for visual search,” in *ISO/IEC JTC1/SC29/WG11/N12202.*

[14] ——, “Compact descriptors for visual search: Evaluation framework: Requirements,” in *ISO/IEC JTC1/SC29/WG11/N11531*.

[15] Z. Wang *et al.*, “An indexing structure to speed up retrieval,” in *ISO/IEC JTC1/SC29/WG11/M28893*.

[16] X. Xin *et al.*, “Cdvs: 2-way sift matching to improve image matching accuracy,” in *ISO/IEC JTC1/SC29/WG11/M29359*.

[17] Y. Reznik *et al.*, “Compact descriptors for visual search: Evaluation framework: Call for proposals,” in *ISO/IEC JTC1/SC29/WG11/N12201*.

[18] S. Paschalakis *et al.*, “Cdvs ce2: local descriptor compression proposal,” in *ISO/IEC JTC1/SC29/WG11/M25929*.

[19] ——, “Cdvs ce2: Local descriptor compression,” in *ISO/IEC JTC1/SC29/WG11/M28179*.

[20] J. Chen *et al.*, “Pku response to ce1: Improved bflog interest point detector,” in *ISO/IEC JTC1/SC29/WG11/M31398*.

[21] Z. Wang *et al.*, “Response to ce2: improved scfv,” in *ISO/IEC JTC1/SC29/WG11/M33189*.

[22] J, Chen *et al.*, “A low complexity interest point detector,” *Signal Processing Letters, IEEE*, vol. 22, no. 2, pp. 172–176, 2015.

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

[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] G. Francini *et al.*, “Cdvs: Telecom italia’s response to ce1 - interest point detection,” in *ISO/IEC JTC1/SC29/WG11/M31369*.

[27] ——, “Telecom italia’s response to the mpeg cfp for compact descriptor for visual search,” in *ISO/IEC JTC1/SC29/WG11/M22672*.

[28] S. Tsai *et al.*, “Fast geometric re-ranking for image-based retrieval,” in *ICIP*, 2010.

[29] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[30] J, Chen *et al.*, “Peking univ. response to ce2 - local descriptor compression,” in *ISO/IEC JTC1/SC29/WG11/M26727*.

[31] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *CVPR*, 2006.

[32] H. Jégou *et al.*, “Improving bag-of-features for large scale image search,” *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[33] 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*.

[34] 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*.

[35] T. S. Jaakkola *et al.*, “Exploiting generative models in discriminative classifiers,” *Advances in neural information processing systems*, pp. 487–493, 1999.

[36] 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*.

[37] ——, “Peking scalable low-memory global descriptor scfv,” in *ISO/IEC JTC1/SC29/WG11/M26726*.

[38] ——, “Peking univ. response to ce2: The improved scfv global descriptor,” in *ISO/IEC JTC1/SC29/WG11/M32261*.

[39] 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.

[40] Z. Wang *et al.*, “Cdvs core experiement 3: Standford/peking/huawei contribution,” in *ISO/IEC JTC1/SC29/WG11/M25883*.

[41] S. Tsai *et al.*, “Location coding for mobile image retrieval,” in *International Conference Mobile Multimedia Communication*, 2009, p. 8.