Compact Descriptors for Mobile Visual Search and MPEG CDVS Standardization

Ling-Yu Duan, Feng Gao, Jie Chen, Jie Lin, Tiejun Huang
The Institute of Digital Media, the School of EE & CS, Peking University, China
{lingyu, fgao, cjie, jieling, tjhuang}@pku.edu.cn

***Abstract*— In this paper, we present the state-of-the-art compact descriptors for mobile visual search. In particular, we introduce our MPEG contributions in global descriptor aggregation and local descriptor compression, which have been adopted by the ongoing MPEG standardization of compact descriptor for visual search (CDVS). Standardization progress will be introduced. Other issues including visual object databases and MPEG CDVS impact on visual search industry will be discussed as well.**

I. Introduction

Smart camera phones and tablet PCs have shown great potentials in mobile visual search, thanks to the integrated functionality of high resolution color camera, powerful CPU, pervasive 3G wireless connection. Emerging mobile Apps have involved rich visual objects, such as CD/book cover, poster, logo, landmark, scene, product, etc. Existing mobile visual search systems are deployed in the client-server architecture. The server end maintains a duplicate or near-duplicate visual search system, which employs approximate visual matching techniques such as Bag-of-Words (BoW) or Hashing [1][2][3]. To accelerate similarity search, reference images are offline indexed, say inverted index table. In online search, a snapped picture is sent to the server, where visual search is conducted to identify the top matched or relevant images; meanwhile, further recommended information linked with visual objects is returned to mobile users.

Undoubtedly, in 3G wireless environment, the upstream delivery of a visual query is subject to the network constraint of unstable or limited bandwidth. Latency from query delivery may degenerate user experience significantly. However, with fast growing processing power in mobile devices or the development of application-specific integrated circuit (ASIC), sending an entire image seems unnecessary, since visual feature extraction and compression can be performed on mobile devices. To reduce the visual query delivery latency, a visual descriptor is required to be compact and discriminative. Nowadays, this trend has received dedicated efforts in MPEG standardization, namely, Compact Descriptor for Visual Search (CDVS) [4].

On the other hand, although both academia and industry have made significant progress in technical components of visual search, it remains unclear how to make visual search applications compatible across a broad range of devices and platforms. To ensure application interoperability is becoming an important and practical issue, which is one essential motivation for MPEG CDVS standardization as well.

In this paper, we present the state-of-the-art compact visual descriptors, including our MPEG contributions recently adopted by CDVS Test Model. In addition, we discuss the progress of MPEG CDVS standardization. Further discussion will be given, including visual object databases, CDVS impact on visual search industry, etc.

II. Compact Visual Descriptors

Comparing to previous works in compact local descriptors, e.g. SURF [5], GLOH [6], PCA-SIFT [7], and MSR [8], recent works have attempted to address compact descriptors towards low bit rate mobile visual search [9 -12]. Readers are referred to [13] for a review of compact descriptors for visual search. In this section, we first introduce existing compact descriptors for visual search. Then we present the state-of-the-art compact descriptors under MPEG CDVS.

Most existing work may be classified into two categories. The first category employs quantization to compress local descriptors without any noticeable loss of discriminative power. For instance, Chandrasekhar et al. proposed a Compressed Histogram of Gradient (CHoG) [9], which adopts Gagie Tree coding to compress each local descriptor into some 60 bits. Tsai et al. further proposed a grid-based quantization approach to code the spatial layout of local features for geometric verification in image matching [13]. Distinct from directly compressing local descriptors, the second category attempts to compress vector quantization (VQ) based BoW signatures [10-12]. For instance, Chen et al. proposed a Tree Histogram Coding scheme [10] to compress the sparse BoW signature, which encodes the position difference of non-zero bins. Ji et al. proposed a multiple-channel coding based Compact Visual Descriptor (MCVD), which enables a context-specific coding of sparse BoW signatures [11-12].

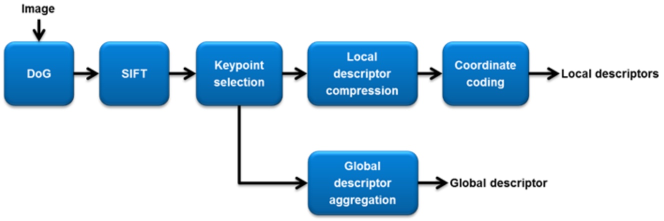


Figure 1. Compact descriptor extraction pipeline in CDVS TM [14]

Based on these recent developments, MPEG is currently pursuing the standardization of compact descriptors for visual search. Figure 1 illustrates how the current CDVS Test Model (TM) [14] produces a compact descriptor of an image in a series of processing steps. Difference of Gaussian (DoG) based keypoint detection identifies keypoints in an image based on multi-scale representation. SIFT features are extracted for the detected keypoints. Selection of a limited number of keypoints is performed to identify those that maximize a measure of expected quality for subsequent matching. Local descriptor compression and coordinate coding give rises to compressed local descriptors, by compressing the SIFT features and their coordinates on a keypoint subset. Global descriptor is formed by aggregating uncompressed SIFT features to describe the whole image.

A. *Global Descriptor Aggregation*

State-of-the-art visual search techniques usually adopt a BoW model. The retrieval pipeline involves two stages: 1) using the BoW representation to produce a shortlist of matching images, followed by 2) a geometric re-ranking which relies on the local descriptors. More recently, research evidences have shown visual search performance and efficiency can be significantly improved by leveraging global and local descriptors [15, 16, 24]. On the MPEG 102nd meeting, we have proposed a novel global descriptor, named Scalable Compressed Fisher Vector (SCFV), which has been adopted by CDVS TM [17].

Below we summarize the different processing stages of SCFV:

- SIFT dimension reduction

Principal Component Analysis (PCA) is employed to reduce the dimension of raw SIFT features from 128-dim to 32-dim, which benefits the SCFV in: i) significantly reducing the FV dimension to make SCFV much more compact; and ii) effectively removing redundant information in raw SIFT features.

- Scalable Fisher Vector (SFV) Aggregation

Offline: We train a Gaussian Mixture Model (GMM) with 128 Gaussian functions over a training set of SIFT features (the same set as SIFT dimension reduction). The GMM model is employed in the online stage to generate the fisher vector (FV) [18] for each selected local feature in an image from Keypoint Selection.

Online: We first calculate the gradient vector of each SIFT eigenvector (32-dim), w.r.t. each Gaussian function. Then we accumulate the gradient vector of all the selected keypoints in an image, w.r.t. each Gaussian function. By concatenating the accumulated gradient vectors of all Gaussian functions, we generate the aggregated FV with in total 128 x 32 = 4096 dim.

Beyond traditional FV aggregation [15, 24], we propose scalable fisher vector aggregation with FV sparseness. Through measuring the FV sub-vector sparseness of each Gaussian function, we determine the informative Gaussian functions and select their FV sub-vectors to form the scalable fisher vector (SFV). Note that the SFV based global descriptor may use distinct sets of Gaussian functions to represent different images. However, SFV enables the interoperability in pair-wise matching between different sets of Gaussian functions by computing the correlation of FV sub-vectors over the overlapped Gaussians functions of two images.

- Binarizing SFV

To further compress SFV, we employ a sign function to binarize SFV signatures. For each dimension of SFV, we apply the sign function to assign the value "1" to any non-negative values, and the value "0" to any negative values, respectively.

- Generating the SCFV bit stream

First, a "head" segment of 128 bits is formed, indicating which Gaussian functions are adopted in SFV signatures. The binarized SFV is then produced, consisting of selected Gaussian functions' binarized FV sub-vectors. Figure 2 gives an example of SCFV bit stream, in which up to 128 Gaussian functions may be involved to generate FV sub-vectors for SCFV. As a result, the maximum descriptor length is limited to 528 bytes (4096 + 128 = 4224 bits).

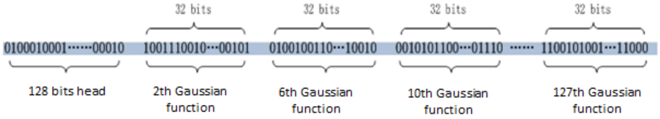


Figure 2. Example of SCFV bit stream.

In order to generate a shortlist of matching images (say 500), we compare a query's global descriptor with the pre-computed global descriptors of database images by computing the Hamming distance. It is well known that Hamming distance computing is very efficient, which is suitable for hardware implementation (say ASIC). Moreover, the global descriptor based retrieval pipeline does not rely on the time consuming indexing construction. In terms of performance, we have shown that SCFV significantly outperformed the BoW model as well as the state-of-the-art global descriptor like REVV [14, 24].

B. *Local Descriptor Compression*

Local descriptors mainly contribute to the pairwise matching between the compact descriptors extracted from the query and the reference images. The pairwise matching determines whether the query and the reference images depict the same objects or scenes or not. For match detection, geometric consistency check is performed to determine the number of inliers among the key point matches (correspondences) between the two images. In case of a match, localization information is provided, i.e. the position of the matching objects in the image, where homography estimation is conducted.

To generate compact visual descriptors, we have to compress local descriptors in addition to global descriptor aggregation. In the current CDVS TM4, both scalar and vector quantizers are given to compress the selected SIFT features (depending on the mode) [14]. As CDVS standard would generally prefer hardware-friendly tools, lower memory requirements (typically below 128KB) are specified.

On the MPEG 102nd meeting, our proposed multi-stage vector quantizer (MSVQ) [19] has been adopted by CDVS TM. With significantly reduced memory usage (from 6MB in TM3 to 38KB in TM4), we have achieved comparable performance. MSVQ starts by quantizing raw SIFT features via a flat codebook at the first stage, and the resulting quantization residual errors are subsequently quantized via a small flat codebook at the second stage. The PSNR of compressed local descriptors, together with descriptor number, jointly impact the performance at different budgets. MSVQ is an effective approach to a quantizer with desirable PSNR, but with much lower complexity. Readers are referred to [20] for design details of MSVQ.

In addition, S. Paschalakis et al. [21] proposed an effective and efficient scalar quantization on the MPEG 101st meeting. For the formation of an image descriptor, the transformed and quantised local descriptors are concatenated and undergo adaptive arithmetic coding.

In general, scalar quantization is more hardware-friendly than vector quantization in terms of computing efficiency. However, vector quantization is supposed to bring about better performance with more computational complexity. One core experiment in CDVS has been setup to investigate local descriptor compression.

C. *Location Coordinate Coding*

Location coding is important. For example, the lowest operating point is set to 512B per query in the CDVS evaluation framework. Given a VGA image, 20 bits are needed to code the coordinate of a keypoint without any compression. Given an image containing 500 keypoints, it would cost about 1250 bytes, much more than 512 bytes.

The coordinates of selected keypoints in an image are encoded by a spatial grid based quantization and arithmetic coding [13], which is employed.by CDVS TM. A spatial grid of 6 x 6 pixels is overlaid on top of the original image. The keypoints within each spatial bin are counted, forming the location histogram. In details, the positions of the nonzero bins (the histogram map), and the number of coordinates in the nonzero bins (the histogram count) are encoded separately. A simple arithmetic coding is applied to encode the histogram count values, while the histogram map is encoded using a context-based arithmetic coder in a circular scan fashion. Training models can be learned to come up with different contexts.

III. MPEG CDVS STANDARDIZATION

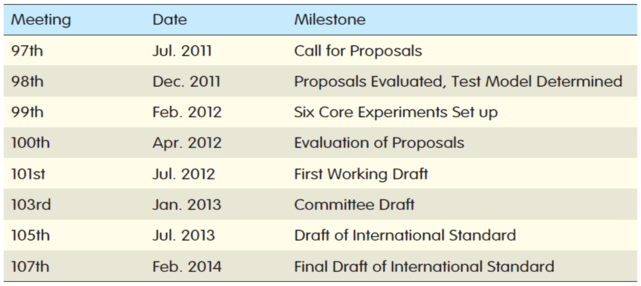
A. MPEG CDVS Standardization Progress

At the 92nd to 96th MPEG meetings, the experts from academia and industry have in depth investigated potential applications, scope of standardization, requirements, and evaluation framework. The 1st and 2nd Workshops on Mobile Visual Search hosted by Stanford Univ. and Peking Univ. made useful inputs to CDVS requirements. At the 97th MPEG meeting, the CFP was issued [22]. To fulfill interoperability, CDVS standardization aims to define the format of compact visual descriptors as well as the pipeline of feature extraction and visual search process [22].

The visual descriptors need to be robust, compact, and easy to compute on a wide range of platforms. High matching accuracy must be achieved for images of rigid, textured objects; landmarks; and documents. Matching should be accurate despite partial occlusions and changes in vantage point, camera parameters, and lighting. To reduce the amount of query information, and to alleviate query transmission latency, the descriptor length must be minimized. Descriptor extraction shall allow adaptation of descriptor length so that the required performance level can be satisfied. Extracting descriptors must not be complex in terms of memory and time [22].

Table I lists the progress of CDVS standardization. Significant achievements have been made in global descriptor aggregation, local descriptor compression and location coordinates coding. CDVS is to enter the phase of committee draft on the upcoming 103rd meeting.

TABLE I. TIMELINE AND MILESTONES OF CDVS STANDARDIZATION



Meeting	Date	Milestone
97th	Jul. 2011	Call for Proposals
98th	Dec. 2011	Proposals Evaluated, Test Model Determined
99th	Feb. 2012	Six Core Experiments Set up
100th	Apr. 2012	Evaluation of Proposals
101st	Jul. 2012	First Working Draft
103rd	Jan. 2013	Committee Draft
105th	Jul. 2013	Draft of International Standard
107th	Feb. 2014	Final Draft of International Standard

B. Core Experiments

One important contribution of CDVS Ad-hoc group lies in the evaluation framework, which exactly aligns with requirements [22]. Proposals are evaluated by two types of experiments: retrieval and pairwise matching. The retrieval experiment is carried out to evaluate descriptor performance in image retrieval. Mean average precision (mAP) and success rate for top match are measured. The pairwise matching experiment is to evaluate the performance in matching image pairs, which is measured by the success rate (True Positive Rate, TPR) at a given false alarm rate (False Alarm Rate, FPR) (say 1%) as well as the localization precision. In particular, the descriptor scalability is evaluated by reporting the performance at six operating points of different query size: 512B, 1KB, 2KB, 4KB, 8KB, and 16KB. For interoperability, a descriptor generated at any operating point shall allow matching with other operating points.

Eight datasets are collected in the evaluation framework, in which 30,256 images are categorized into *mixed text and graphics*, *paintings*, *frames captured from video clips*, *landmarks*, and *common objects*. Ground-truth annotation files of pairs of matching and non-matching images, as well as query images and relevant reference images are provided. For localization, the *mixed text and graphics* category provides bounding boxes for each matching pair. In retrieval experiment, a distractor image set containing 1 million images of varying resolutions and content (collected from Flickr) are used.

Up to seven CEs have been performed to investigate CDVS core techniques, including global descriptor (CE1), local descriptors compression (CE2), location coding (CE3), keypoint detection (CE4), better uncompressed local descriptor (CE5), etc. As introduced in Section II, CE1, CE2, CE3 have significant impact on CDVS standard normative part. CE4 is to figure out an alternative of Difference-of-Gaussian (DoG) in order to avoid SIFT patent [23].

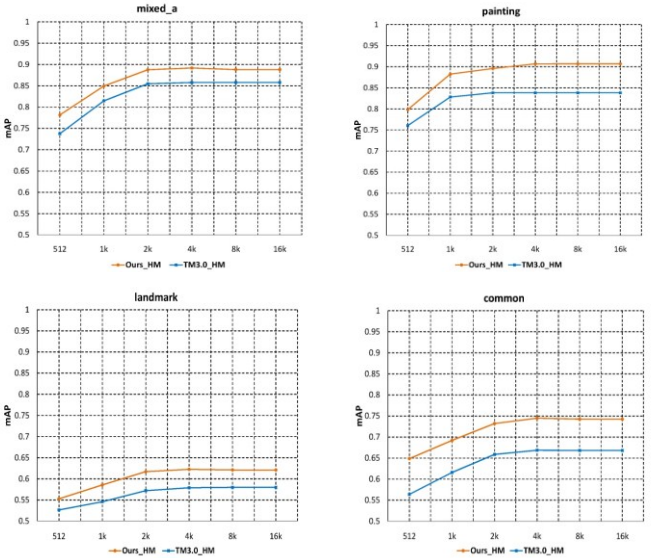


Figure 3. Comparison of retrieval performance under CDVS TM. The TM4 adopted SCFV [17] significantly outperforms REVV [24].

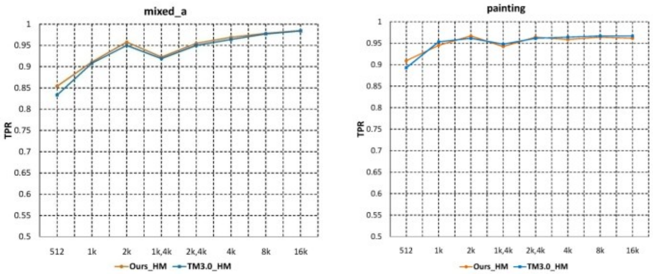


Figure 4. Comparison of pairwise matching under CDVS TM. SCFV involved matching achieves comparable performance with REVV.

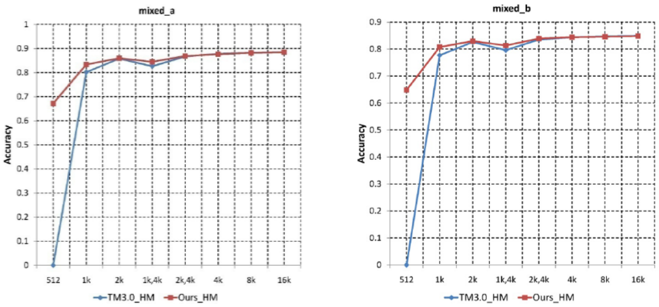


Figure 5. Comparison of localization accuracy under CDVS TM. Due to extreme compactness, SCFV enables the localization at the lowest point 512B, with extra budget for local descriptors. Moreover, the compactness brings about more budgets for local descriptors, which leads to improved localization accuracy at 1K and 1K vs.4K.

To provide quantitative analysis of evaluation framework, and to present our MPEG contribution in global descriptor, we present the retrieval and matching results of CE 1 on several datasets in Figures 3, 4, 5. In terms of performance improvements, for retrieval, with scalar quantization, we have achieved mAP 81.47% on average (vs. 77.04% in H-Mode REVV); for pair-wise matching, we have achieved the TPR of 91.47% on average (vs. 91.35% in H-Mode REVV); in localization, we have achieved the accuracy of around 67% on *mixed text and graphics* at 512B (vs. 0.00% in H-Mode REVV). These improvements led to the integration of SCFV into CDVS TM. Due to space limit, we cannot list complete results. Readers are referred to our MPEG input contribution [17] or request the document by email.

In particular, as the hardware-friendly tool favors lower memory, SCFV memory usage has been significantly reduced to 49KB, much smaller than REVV (119KB) [24]. The SCFV tables involve SIFT PCA projection matrix (16.5KB) and GMM parameters (32.5KB).

IV. FURTHER DISCUSSIONS

A. *Visual Object Databases*

In mobile visual search, the user snaps a photo of an object with a mobile device to retrieve information about the object. In addition to robust visual feature extraction and indexing, there is an equally important problem of large scale object database collection and analysis. Collecting, analyzing and managing large-scale database of real-life objects are critical to facilitate effective mobile search.

More recently, Peking University (PKU, China) and Nanyang Technological University (NTU, Singapore) have setup a joint lab **ROSE** (Rapid-Rich Object SEarch Lab) to create the largest collection of structured domain object databases in Asia and to develop a rapid and rich object mobile search. **ROSE** will build a large database with 50M images to support mobile visual search. This project proposes to intelligently collect large amount of media data through a combination of filter-out and filter-in strategies. The first target domain of this project is mobile commerce (M-commerce). The goal is to provide mobile object search and recommendation for consumer products such as apparels (clothing, dresses, shirts, and ties), handbags, shoes, etc. The second target domain is centered on tourism industries. The aim is to develop capabilities which will enable a smartphone to function as an interactive mobile city guide. The third target domain is focused on lifestyle and hobbies. The main idea is to develop mobile applications and construct relevant object databases to assist users in their pursuit of lifestyle and hobbies.

B. *Identical Non-Rigid or Non-Textured Object Search*

The ongoing MPEG CDVS are seeking technologies for visual content matching of rigid and textured objects including matching of views of objects, landmarks, and printed documents. The aim is to detect identical objects, where the algorithm shall be robust to partial occlusions as well as changes in vantage point, camera parameters and lighting conditions. However, in certain domains, say consumer product such as apparels, handbags, shoes, etc., visual objects are often non-rigid or non-textured, so that the state-of-the-art compact descriptors and corresponding search process would probably fail. An alternative solution is to address identical non-rigid or non-texture object search from the similar object search point of view. The big challenge lies in the collection of large-scale image datasets and ground-truth annotation to train classifiers or recognizers. To meet this challenge and promote mobile media applications, **ROSE** project establishes a goal to build the capability to be the largest collection of domain object database in Asia for mobile image search in five years.

C. *CDVS Impact on Visual Search Industry*

It is envisioned that MPEG CDVS standard technologies will ensure interoperability of visual search applications and databases, enable high level of performance of implementations conformant to the standard, simplify design of descriptor extraction and matching for visual search applications, enable hardware support for descriptor extraction and matching in mobile devices, reduce load on wireless networks carrying visual search-related information. To build full visual search application this standard may be used jointly with other existing standards, such as MPEG Query Format, HTTP, XML, JPEG, JPSec, and JPSearch [4].

ACKNOWLEDGEMENT

This work was supported in part by the Chinese Natural Science Foundation under Contract No. 61271311 and No. 61210005, and in part by the Rich and Rapid Object Search (ROSE) Project.

REFERENCES

[1] D. Nister and H. Stewenius , “Scalable recognition with a vocabulary tree,” in Proc. of IEEE Conf. CVPR 2006.

[2] G. Schindler and M. Brown, “City-scale location recognition,” in Proc. of IEEE Conf. CVPR 2007.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. “Object retrieval with large vocabularies and fast spatial matching,” in Proc. of IEEE Conf. CVPR 2007.

[4] CDVS1.“Call for proposals for compact descriptors for visual search,” N12201. Turin, Italy: ISO/IEC JTC1/SC29/WG11, 2011

[5] H. Bay, T, Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in Proc. of ECCV 2006.

[6] K. Mikolajczyk and C. Schmid, “Performance evaluation of local descriptors”, IEEE Trans. PAMI, vol. 27, no.10, pp. 1615-1630, 2005.

[7] Y. Ke, R. Sukthankar, “PCA-SIFT: A robust distinctive representation for local image descriptors,” In Proc. of IEEE CVPR 2004.

[8] G. Hua, M. Brown, and S. Winder, “Discrimianative embedding for local image descriptors,” In Proc. of IEEE Conf. ICCV 2007.

[9] V. Chandrasekhar, etc. “CHoG: Compressed histogram of gradients:a low bit-rate feature descriptor,” In Proc. of IEEE Conf. CVPR 2009.

[10] D. Chen, “Tree histogram coding for mobile image matching,” In Proc. of Conf. DCC 2009.

[11] R. Ji, et al.“Location discriminative vocabualry coding for mobile landmark search,” Int. Journal of Computer Vision, vol. 96, no. 3, 2012.

[12] R. Ji, et al. “Towards low bit rate mobile visual search with multiple channel coding,” In Proc. of ACM Multimedia, pp. 573-582, 2011.

[13] B. Girod, et al. “Mobile visual search,” IEEE Signal Processing Magazine, vol. 28, no. 4, 2011.

[14] CDVS2. “Test Model 4: Compact descriptor for visual search,” W12929, ISO/IEC JTC1/SC29/WG11, Shanghai, China, 2012

[15] H. Jegou, M. Douze, C. Schmid, P. Perez, “Aggregating local descriptors into a compact image representation,” in Proc. of CVPR 2010.

[16] F. Perronnin, Y. Liu, J. Sanchez, H. Poirier, “Large-scale image retrieval with compressed fisher vectors,” in Proc. Conf. CVPR 2010.

[17] J. Lin, L.-Y. Duan, J. Chen, T. Huang, W. Gao,“Peking Univ. Response to CE 1: A Scalable Low-Memory Global Descriptor,” M26726,, ISO/IEC JTC1/SC29/WG11, Shanghai, China, Oct. 2012.

[18] F. Perronnin and C. Dance, “Fisher Kernels on Visual Vocabularies for Image Categorization,” in Proc. of IEEE Conf. CVPR 2007.

[19] J. Chen, L.-Y. Duan, J. Lin, T. Huang, W. Gao, “Peking Univ. Response to CE 2: Local descriptor compression,” M26727, ISO/IEC JTC1/SC29/WG11, Shanghai, China, Oct. 2012.

[20] J. Chen, L.-Y. Duan, R. Ji, Z. Wang, “Multi-stage vector quantization towards low bit rate visual search,” in Proc. of IEEE ICIP 2012.

[21] S. Paschalakis, et al., “CDVS CE2: Local descriptor compression proposal,” M25929, ISO/IEC JTC1/SC29/WG11, Sweden: Jul. 2012.

[22] CDVS3, “Evaluation framework for compact descriptors for visual Search,” N12202. Turin, Italy: ISO/IEC JTC1/SC29/WG11, 2011.

[23] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. Journal of Computer Vision, vol. 60, pp. 91-110, 2004.

[24] D. Chen, et al. “Residual enhanced visual vector as a compact signature for mobile visual search,” Signal Processing, In press. 2012.