JOINT OPTIMIZATION OF JPEG QUANTIZATION TABLE AND COEFFICIENT THRESHOLDING FOR LOW BITRATE MOBILE VISUAL SEARCH

Yitong Wang, Ling-Yu Duan, Jie Lin, Tiejun Huang, Wen Gao*

The Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China
e-mail:{wangyitong, lingyu, jielin, tjhuang, wgao}@pku.edu.cn

ABSTRACT

Low latency query delivery over wireless network is a key problem for mobile visual search. Extracting compact descriptors directly on the mobile device is computational expensive, an alternate approach is to send highly compressed JPEG query images. As JPEG baseline optimizes the rate-distortion from a perceptual perspective rather than maintaining search performance, recent work proposed to learn a feature-preserving JPEG quantization table for improved search accuracy. However, this method is data-dependent and the quantization table cannot adapt to image blocks. To address these issues, we propose to jointly optimize the JPEG quantization table and coefficient thresholding. The matching score between uncompressed image and its compressed JPEG image is employed as the distortion measure to avoid time consuming image labeling, and coefficient thresholding eliminates the redundant coefficients. Extensive experiments on benchmark datasets show that our approach obtains superior performance than state-of-the-art at low bitrates, meanwhile, it consumes lower cost including processing time, memory and battery on mobile device.

***Index Terms*—** JPEG Compression, Mobile Visual Search, Quantization Table, Coefficient Thresholding, Optimization

1. INTRODUCTION

Camera equipped mobile devices have shown great potentials in mobile visual search [1] applications like Google Goggles. In general, a query is sent from the mobile client to the server via a wireless link, then visual search is performed to identify the relevant images from a reference image database hosted at the server end. In wireless environment, the upstream of a visual query is subject to network constraint of unstable or limited bandwidth. To reduce latency for better user experience, the upstream query data is expected to be as small as possible.

Recent works have proposed to extract compact visual descriptors of query images directly on the mobile device, and send such descriptors over a wireless link at low bitrates (See Fig.1 (b)). In particular, this topic relates to an ongoing MPEG standardization, namely, Compact Descriptors for Visual Search (CDVS) [2][3][4]. Existing compact descriptors mainly build upon local invariant features (e.g., SIFT [5], SURF [6]), which are subsequently compressed into compact codes without incurring considerable loss of discriminative power. For instance, Chandrasekhar et al. [7] proposed a Compressed Histogram of Gradient (CHoG), which adopts Huffman Tree coding to compress each local feature into approximate

∗ is corresponding author. National Natural Science Foundation of China under grant 61271311, 61121002, 61390515, and 61210005 supported this work.

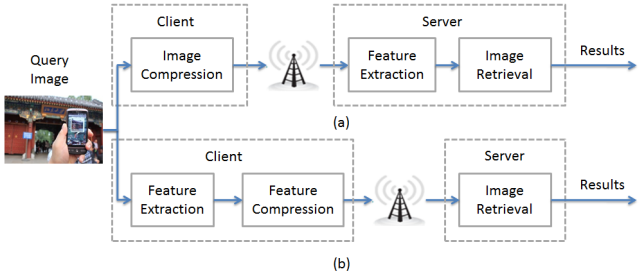


Fig. 1. Framework of low bitrate mobile visual search: (a) Transmitting highly compressed JPEG query images, and subsequent descriptors extraction and matching (retrieval) are performed on the server (Top), and (b) Extracting and compressing visual descriptors directly on the mobile client, and sending compact descriptors over a wireless link (Bottom).

60 bits. Other examples of compact descriptors are Vector of Locally Aggregated Descriptors (VLAD) [8], Compressed Fisher Vector (CFV) [9][10][11] and Residual Enhanced Visual Vector (REVV) [12]. However, the heavy feature extraction complexity poses a challenge to work with the mobile device with limited computation and memory resources. For example, a less optimized feature extraction process may cost over 2 seconds and 20MB RAM to extract SIFT features from a VGA image (640x480) on iPhone 4S.

An alternate approach is to perform fast JPEG compression [13] with low memory footprint on the mobile client, then transmit the JPEG compressed image as a query to the server (See Fig.1 (a)). The JPEG-like compression first partitions the image into 8×8 blocks, converts each block to frequency domain using discrete cosine transform (DCT), then quantizes the resulting DCT coefficients using a 8×8 quantization table, followed by entropy coding. JPEG coder largely reduces image size, however, it also degrades feature detection and description due to the compression artifacts, resulting in decreased search performance. Previous works [14][15] mainly aim to minimize the rate-constrained pixel-wise information loss from a perceptual perspective based on the human visual system, but possibly not optimal for visual search.

To improve search accuracy, the image compression scheme is required to preserve informative visual features. Makar et al. [16] employed the location information of detected features to determine the corresponding image patch for subsequent patch compression. Chao et al. [17] proposed a rate-distortion optimization method to preserve the blocks containing SIFT features, while allocating fewer bits to encode blocks without features. As previously mentioned, these methods involve resource consuming feature detection and description, which are not suitable for implementation on the

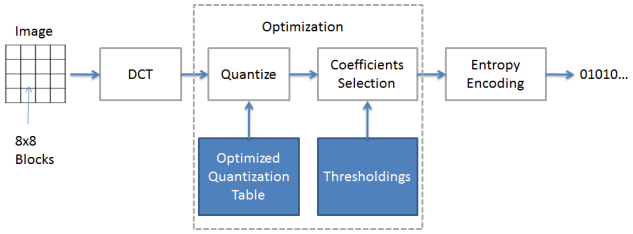


Fig. 2. The proposed image compression pipeline based on JPEG baseline. The optimized quantization table and thresholdings lead to compressing image deeply by eliminating redundant coefficients, which are suitable for low bitrate mobile visual search.

mobile device. To address this issue, Duan et al. [18] adopted pairwise matching precision as the distortion measure to optimize the JPEG quantization table for preserving informative features implicitly. However, this approach still has some drawbacks: (1) the JPEG quantization table is learned in a supervised manner, i.e., it requires manually labeled match/non-match image pairs; (2) the optimized quantization table is uniformly applied to all image blocks, which cannot adapt to individual blocks.

In this paper, we propose to jointly optimize the JPEG quantization table and coefficient thresholding for low bitrate mobile visual search. Firstly, the matching score between uncompressed image and its compressed JPEG image is employed as the distortion measure to avoid time consuming match/non-match image labeling. Secondly, we take into account the location information of image blocks to select informative DCT coefficients for block-adaptive coding, considering the hypothesis of informative visual information relating to the distance from the image center [19][20]. The joint optimization approach (see Fig.2) is able to eliminate the redundant coefficients for effective feature preserving. In addition, the output bitstream of our approach is compatible with the stream format of JPEG standard, ensuring a better interoperability for mobile visual search applications. Extensive experiments on MPEG CDVS benchmark datasets combined with 1 million distractor images [4] have shown the consistent superior search accuracy of the proposed approach over the state-of-the-art at low bitrates. Compared with extracting local features directly on the mobile phone, our approach provides prominent advantages in terms of time cost, memory cost and battery consumption.

2. PROBLEM FORMULATION

The objective is to simultaneously optimize the JPEG quantization table and coefficient thresholding for better preserving informative local features for effective and efficient mobile visual search. The quantization table Q used in this work follows the definition of the JPEG standard. Thus, each 8×8 block (64 pixels) is transformed by DCT and compressed by the 8×8 quantization table Q:

$$Q = \begin{pmatrix} Q_{00} & \cdots & Q_{70} \\ \vdots & \ddots & \vdots \\ Q_{70} & \cdots & Q_{77} \end{pmatrix}, 1 \leq Q_{ij} \leq 255 \quad (1)$$

The quantized DCT coefficients $C_n = \{C_n^i\}, i = 0, ..., 63$ (in zigzag order) are further encoded by adaptively selecting the most important elements, where C_n^i is the i^{th} quantized DCT coefficient of the n^{th} image block. Specifically, we employ the thresholding mask $T_n = \{T_n^i\}, T_n^i \in \{0, 1\}, i = 0, ..., 63$ for the n^{th} image block to perform the coefficient selection:

$$T_n^i = \begin{cases} 1 & if \quad C_n^i \quad selected \\ 0 & otherwise \end{cases} \quad (2)$$

Algorithm 1 Joint optimization of Q and T

1: **Input:** images $I_i, i = 1, ..., M$
2: Initialized Q_0 and T_0 with rate constraint Rc
3: repeat
4: $\quad Q_t = \min_Q\{J = D(Q_{t-1}, T_{t-1}) + \lambda R(Q_{t-1}, T_{t-1})\}$
5: $\quad T_t = \min_T\{J = D(Q_t, T_{t-1}) + \lambda R(Q_t, T_{t-1})\}$
6: until J converged
7: **Output:** Q_t, T_t

If $T_n^i = 1$, the corresponding coefficient C_n^i is included in the subsequent entropy coding, otherwise, C_n^i is discarded. We denote $T = \{T_n\}_{n=1}^N$ the coefficient thresholdings for all N blocks of an image. It should be noticed that coefficient thresholding T_n varies across image blocks.

To achieve this goal, we formulate the image compression as a rate-distortion optimized problem of finding out quantization table Q and thresholding T for each quantized DCT coefficient:

$$\min\{D(Q, T)\} \quad s.t. \quad R(Q, T) \leq Rc \quad (3)$$

$R(Q, T)$ and $D(Q, T)$ denote the coding bitrate and distortion, respectively.

The distortion $D(Q, T)$ is defined as the matching score between uncompressed image and its JPEG compressed image:

$$D(Q, T) = 1 - \frac{2 \times \#correctMatches}{\#originalFeats + \#compressedFeats} \quad (4)$$

$\#correctMatches$ measures the number of SIFT matching pairs between the uncompressed image and its JPEG compressed image, $\#originalFeats$ and $\#compressedFeats$ denote the number of detected SIFT features from the uncompressed image and its JPEG compressed image, respectively. Obviously, the choice of Q and T determines the matching score. Unlike traditional human perception oriented distortion measure, this criterion implicitly evaluates the ability of preserving SIFT features. Considering the state-of-the-art visual search approaches work on the SIFT features, this criterion is dealt with as an indicator of search accuracy.

We use Lagrange multiplier method to converts this rate-constrained problem into:

$$\min\{J = D(Q, T) + \lambda R(Q, T)\} \quad (5)$$

The Lagrange multiplier λ is a fixed constant that controls the rate-distortion trade-off, and J is the Lagrange cost.

3. JOINT OPTIMIZATION OF Q AND T

Since joint optimization of the quantization table Q and coefficient thresholding T is intractable, we propose to iteratively minimize 5, i.e., solving Q (or T) given that T (or Q) is fixed, as shown in Algorithm 1. The Lagrange cost J is non-increasing with each step, convergence is guaranteed.

Optimizing Q. With the coefficient thresholding T fixed, the optimal quantization table Q still entails too much computational complexity. Similar to [18], genetic algorithm is chosen to learn the

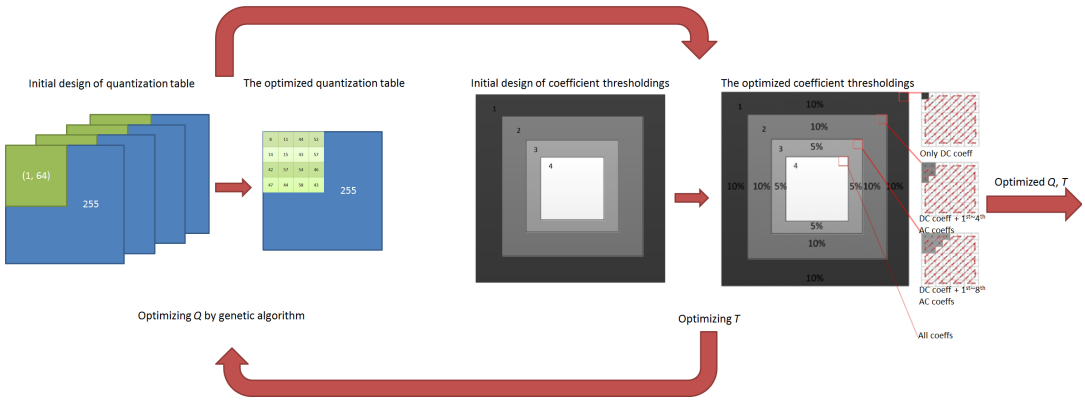


Fig. 3. The iterative optimization process of the quantization table Q and coefficient thresholding T.

suboptimal Q based on an initialized quantization table Q_0. Observing that informative visual features are mainly correlated to low frequency DCT coefficients of the uncompressed image, a set of quantization tables discarding high frequency information are employed as initial population of genetic algorithm, i.e., the low frequency components of each table are set to a small random integer between 1 and 64. With this initialization, the solution space of quantization table is largely reduced, meanwhile, the low frequency information are preserved with high probability.

The pipeline of genetic algorithm is inheritance with crossover and mutation, and evaluation. The crossover swaps the Q step from $\{Q_{xy}|x, y \in \{0, 2\}\}$ of two tables to form a new individual table, with a probability $P1$. Subsequently, the mutation randomly selects the Q step from $\{Q_{xy}|0 \leq x, y \leq 3\}$ and changes it to another possible value for that gene, with a probability $P2$. The evaluation uses the Lagrange cost J as fitness function to choose the better individuals for next iteration. We use the same parameters as [18] in the quantization table optimization process, i.e., $\lambda = 10$, $P1 = 0.5$ and $P2 = 0.1$. The initial quantization table Q_0, optimization process and the optimized table Q is illustrated in Fig.3.

Optimizing T. With the quantization table Q fixed, our goal is to find out the optimal coefficient thresholdings $T = \{T_n\}$ for all blocks of an image. Actually, the coefficient thresholdings play a key role to further eliminate the irrelevant low frequency information. Since optimizing the thresholding for all N blocks is difficult, we propose to optimize each block independently. Observing that the energy of quantized DCT coefficients for image reconstruction is decreasing in zigzag ordering and this rule is empirically applied to image compression for visual search, we choose to evaluate the 1 DC coefficient and 15 non-zero low frequency AC coefficients in the first 24 AC coefficients for preserving SIFT features. Particularly, denoting T_{n_k} as k consecutive thresholdings of n^{th} image block, we set

$$T_{n_k} = \{T^0_{n_k}, T^1_{n_k}, ..., T^{63}_{n_k}\}, 1 \leq k \leq 24 \tag{6}$$

where

$$T^i_{n_k} = \begin{cases} 0 & if \quad i \geq k \\ 1 & if \quad i < k \end{cases} \tag{7}$$

The DC coefficient is important for image reconstruction as well as the feature extraction, so we always set $T^0_n = 1$. Then the problem is simplified as how to choose k for each image block.

On the other hand, recent works [19][20] support the hypothesis that the probability of correct matched feature depends on its distance from the image center. The nearer the distance from image center, the quantized DCT coefficients from that image block are more important. In this paper, we take into account the localization information of image block for guiding the optimization of k in a heuristic manner. Specifically, we partition an image into m non-overlapping windows around the image center, e.g., $m = 4$ in Fig.3. The value k of image blocks located in outer windows should be smaller than those blocks in inner windows. Denoting $T_{k,m}$ the union of all image blocks T_{n_k} with windows m fixed, we enumerate a set of values for k and m respectively, and find out the combination of k and m that maximize the difference of Lagrange cost J_k between $T_{k,m}$ and $T_{k-1,m}$:

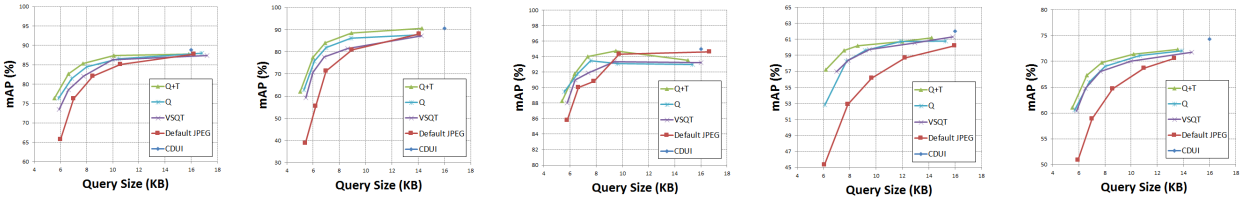
$$\nabla J_{k,m} = J(Q_{fixed}, T_{k,m}) - J(Q_{fixed}, T_{k-1,m}) \tag{8}$$

If $\nabla J_{k,m}$ reaches an extremum, the corresponding AC coefficients from 1^{st} to k^{th} are more valuable for preserving SIFT features than the rest AC coefficients with windows m fixed. The optimized coefficient thresholdings T is illustrated in Fig.3. Experimental results show that the optimized coefficient thresholdings T significantly improves the search performance. Furthermore, it brings about a bit extra computation cost in JPEG compression process, which is trivial compared to other feature detection based compression methods.

4. EXPERIMENTS

Datasets and Evaluation Protocols. We evaluate the retrieval perfomance of the proposed compression approach over the MPEG CDVS benchmark datasets [4][21][22][23][24][25]. The dataset consists of 8313 query images and 18440 reference images from five categories (mixed text+graphics , paintings, video frames, landmarks, common objects). A FLICKR1M dataset containing 1 million images is use as distracters, merging with the reference datasets to evaluate the scalability in dealing with large-scale image search. Our experiment follows the Test Model of MPEG CDVS evaluation framework [4]. SIFT descriptor is adopted as the local feature, mean Average Precision (mAP) is used to evaluate retrieval performance.

Baselines. We compare five baselines: (1) Default JPEG: Compress query images with JPEG baseline; (2) Visual Search-oriented Quantization Table (VSQT): Compress query images with the optimized quantization table mentioned in [18]; (3) Q: The proposed approach using only the optimized quantization table Q to compress query images. (4) Q+T: The proposed approach using both the optimized quantization table Q and coefficient thresholding T to compress query images. Besides baselines, we also extract MPEG CDVS standard compact descriptors at the operating point of 16KB



(a) Mixed text+graphics (b) Paintings (c) Video frames (d) Landmarks (e) Common objects

Fig. 4. Query size vs. retrieval performance in terms of mAP over the various types of datasets, combined with the distractor set FLICKR1M.

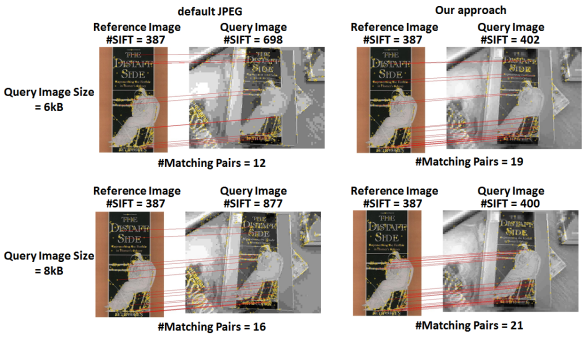
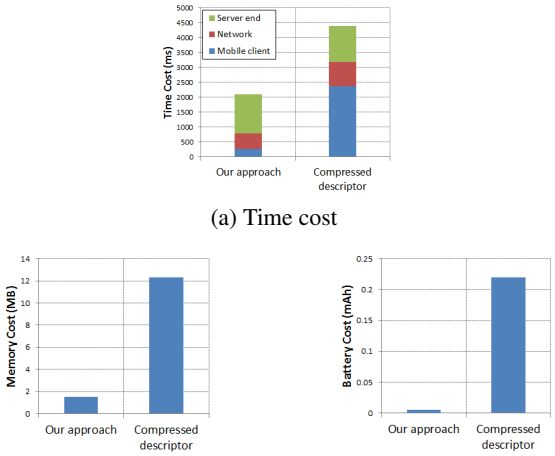


Fig. 5. Comparison of pairwise matching results using query image compressed with default JPEG and the proposed approach, respectively.



(a) Time cost

(b) Memory cost (c) Battery cost

Fig. 6. Comparison of time cost, memory cost and battery consumption between our approach and the CDUI method on smart phone HTC T328 by averaging from 1000 query images of 640x480. (a) Time cost, including the image processing on mobile client, network transmission and image search on the server end. (b) Memory cost on the mobile device. (c) Battery cost on the mobile device.

[4] from uncompressed query images and transmit them to the server (see Fig.1 (a)), named as Compact Descriptors of Uncompressed Image (CDUI), in order to provide an anchor result of the state-of-the-art benchmark retrieval pipeline [4][26].

Rate Distortion Analysis. We perform retrieval experiments over MPEG CDVS benchmark datasets plus 1 million FLICKR1M dataset. Fig.4 compares the mAP of baselines at varied query size over different benchmark datasets. All color images are converted to gray images and query size is controlled by adjusting the scale factor in JPEG codec. Our approach obtains better retrieval performance than VSQT and default JPEG. For example, the proposed method Q achieves +5% gain vs. default JPEG and +1% gain vs. VSQT. In addition, the proposed approach Q+T further improves the mAP with additional 1% for the same query size, verifying the effectiveness of the coefficient thresholding. Particularly, the results show that our approach has achieved superior performance at much lower bitrates from 5kB to 10kB, especially for the paintings dataset containing lots of low scale features. Experiments in [17] shows that low scale features are more vulnerable by JPEG compression. In a sense, our approach is robust for preserving vulnerable features. At the higher bitrates (e.g., 16kB), the mAP of our approach is comparable to the CDUI method adopted in MPEG CDVS evaluation framework. Fig.5 gives an example on the comparison of pairwise matching, using query image compressed with default JPEG and the proposed approach. This example shows that our approach obtains more matching pairs, which demonstrates the superior capability of preserving retrieval performance than the default JPEG compression.

Complexity Analysis. Fig.6 compares the time cost, memory cost and battery consumption between our approach and the CDUI method within MPEG CDVS standard on a smart phone HTC T328. Our approach compresses the query image to 16KB and CDUI extract compact descriptor in 16KB. The results show that directly sending a JPEG compressed query image provides prominent advantages in terms of time cost, memory cost and battery consumption, compared to local feature extraction on the mobile phone (CDUI).

5. CONCLUSION

In this paper, we propose to deeply compress query images by jointly optimizing of JPEG quantization table and coefficient thresholding for low bitrate mobile visual search. Instead of maintaining image quality from a perceptual perspective, the proposed approach is to preserve the informative visual features for improved search performance. In addition, the proposed method is compatible with the JPEG standard. The results show that our approach obtains superior performance than state-of-the-art at lower bitrates, meanwhile, it consumes low cost including processing time, memory and battery. Extension of the proposed method to other compression schemes (e.g. JPEG2000) will be addressed in future work.

6. REFERENCES

[1] Bernd Girod, Vijay Chandrasekhar, David M Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S Tsai, and Ramakrishna Vedantham, "Mobile visual search," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 61–76, 2011.

[2] L Duan, Jie Lin, Jie Chen, Tiejun Huang, and Wen Gao, "Compact descriptors for visual search," *IEEE MultiMedia*, In Print, 2014.

[3] ISO/IEC JTC1/SC29/WG11/N12201, "Call for proposals for compact descriptors for visual search," 2011.

[4] ISO/IEC JTC1/SC29/WG11/N12202, "Evaluation framework for compact descriptors for visual search," 2011.

[5] D. Lowe, "Distinctive image feature from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 10, no. 3, pp. 346–359, 2008.

[7] V. Chandrasekhar, G. Takacs, and D. Chen, "Transform coding of image feature descriptors.," in *Proceedings of Visual Communications and Image Processing*, 2009.

[8] H. Jégou, M. Douze, and C. Schmid, "Aggregating local descriptors into a compact image representation," in *Proceedings of Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[9] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3384–3391.

[10] J. Lin, L.-Y. Duan, T. Huang, and W. Gao, "Robust fisher codes for large scale image retrieval," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[11] Jie Lin, L-Y Duan, Yaping Huang, Siwei Luo, Tiejun Huang, and Wen Gao, "Rate-adaptive compact fisher codes for mobile visual search," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 195–198, 2014.

[12] David Chen, Sam Tsai, Vijay Chandrasekhar, Gabriel Takacs, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Processing*, vol. 93, no. 8, pp. 2316–2327, 2013.

[13] "Independent JPEG group," http://www.ijg.org.

[14] Matthew Crouse and Kannan Ramchandran, "Joint thresholding and quantizer selection for transform image coding: entropy-constrained analysis and applications to baseline JPEG," *Image Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 285–297, 1997.

[15] Viresh Ratnakar and Miron Livny, "An efficient algorithm for optimizing dct quantization," *Image Processing, IEEE Transactions on*, vol. 9, no. 2, pp. 267–270, 2000.

[16] Mina Makar, Chuo-Ling Chang, David Chen, Sam S Tsai, and Bernd Girod, "Compression of image patches for local feature extraction," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 821–824.

[17] Jianshu Chao and Eckehard Steinbach, "Preserving sift features in JPEG-encoded images," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 301–304.

[18] Ling-Yu Duan, Xiangkai Liu, Jie Chen, Tiejun Huang, and Wen Gao, "Optimizing JPEG quantization table for low bit rate mobile visual search," in *Visual Communications and Image Processing (VCIP), 2012 IEEE*. IEEE, 2012, pp. 1–6.

[19] G. Francini, S. Lepsy, and M. Balestri, "Selection of local features for visual search," *Signal Processing: Image Communication*, 2012.

[20] ISO/IEC JTC1/SC29/WG11/M22672, "Telecom italia's response to the MPEG CfP for compact descriptors for visual search," 2011.

[21] "Mvs," http://mars01.stanford.edu/mvs.

[22] "Ethz," http://www.vision.ee.ethz.ch/datasets.

[23] "Cturin180," http://pacific.tilab.com/download/CTurin180.zip.

[24] "Pkubench," http://url.cn/JF5vtm.

[25] "Ukbench," http://vis.uky.edu/ stewe/ukbench.

[26] ISO/IEC JTC1/SC29/WG11/W13145, "Test model 4: Compact descriptors for visual search," 2012.