A Hybrid Mobile Visual Search System With Compact Global Signatures

David M. Chen, *Member, IEEE*, and Bernd Girod, *Fellow, IEEE*

***Abstract*—Mobile visual search systems typically compare a query image against a database of annotated images for accurate object recognition. On-server database matching can search a large database hosted in the cloud, but the query latency could suffer with slow network transmissions or server congestion. On-device database matching can ensure fast recognition responses regardless of network or server conditions, but a small amount of memory on the mobile device can severely limit the number of images that can be stored in an on-device database. This paper presents a new hybrid system that combines the advantages of on-device and on-server database matching. At the core of this system, we first develop a compact and discriminative global signature to characterize each image. Our global signature uses an optimized local feature count that is derived from a statistical analysis of the retrieval performance. We additionally create two extensions that exploit color information within images and relationships between similar database images to improve retrieval accuracy. Then, we propose methods for efficient interframe coding of a sequence of global signatures which are extracted from the viewfinder frames on the mobile device. A low bitrate stream of global signatures can be sent to the server at an uplink bitrate of less than 2 kbps to broaden the search range of the current query and to update the on-device database to help future queries.**

***Index Terms*—Compact signatures, image retrieval, interframe compression, mobile visual search.**

I. Introduction

MOBILE VISUAL SEARCH (MVS) systems recognize objects in the user's vicinity using the camera and other sensors of the mobile device. Examples include landmark recognition [1], product recognition [2], artwork recognition [3], and video recognition [4]. Typically, a query image captured through the camera is compared against a database of labeled images to recognize objects that appear in the query image. For robust image matching, scale- and rotation-invariant local image features [5]–[10] are extracted from each image. The most salient statistics of the local features for each image can be further analyzed and summarized to form a global image signature [11]–[15]. By comparing the query global signature against a database of global signatures, we can quickly determine which database images are most similar to the query image.

In a practical MVS system where the database is hosted on a server, we want to minimize the amount of data exchanged between the mobile device and the server to achieve a low query latency. Therefore, significant attention has been devoted to creating compact image features. Previous work in compressing local features include random projections [16], transform coding [17], compressed histogram of gradients (CHoG) [18], location histogram coding [19], feature-aware JPEG encoding [20], bag of hash bits [21], and interframe feature coding [22]. Similarly, there have been numerous prior works in creating compact global signatures, including the GIST signature [23], tree histogram coding [24], mini bag of features [25], inverted index coding [26], location-aware vocabulary boosting [27], and compact feature residual vectors [13]–[15]. Inspired by these research developments, two emerging MPEG standards, Compact Descriptors for Visual Search (CDVS) [28] and Compact Descriptors for Video Analysis (CDVA) [29], aim at the technologies for generating low bitrate image or video signatures, respectively, for general MVS applications.

Our prior work has shown that by performing image retrieval directly on the mobile device, we can achieve low query latencies regardless of network or server conditions [15], [30], [31]. The on-device MVS system operates as shown in Fig. 1(a). At the center of this system is a database of global signatures stored directly in the random access memory (RAM) of the mobile device. Since a mobile device has limited RAM, it is crucial that these global signatures are compact. At the same time, the global signatures must enable fast comparisons across a large database and reliably differentiate between images of many different objects. We have shown that a residual enhanced visual vector (REVV) signature [15], [30], [31] is well-suited to building a memory-efficient, low-latency, and accurate on-device MVS system.

An on-device MVS system cannot recognize an object unless it is represented in the local database. In this paper, we greatly improve the capabilities of on-device searching by developing a new hybrid MVS system as depicted in Fig. 1(b). The previous on-device system is now a subsystem of this new hybrid system. When a local database search is sufficient, we finish the query locally and display the best local result. Otherwise, we additionally query a remote server where a much larger database of images is hosted. In the uplink, we transmit the same compact global signatures that are used for the local database search. Then, in the downlink, we return the labels, local features, and global signatures for the top-ranked database candidates. The

Manuscript received June 23, 2014; revised December 28, 2014; accepted April 17, 2015. Date of publication April 29, 2015; date of current version June 13, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Qi Tian.

D. Chen was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: dmchen@alumni.stanford.edu).

B. Girod is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: bgirod@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2427744

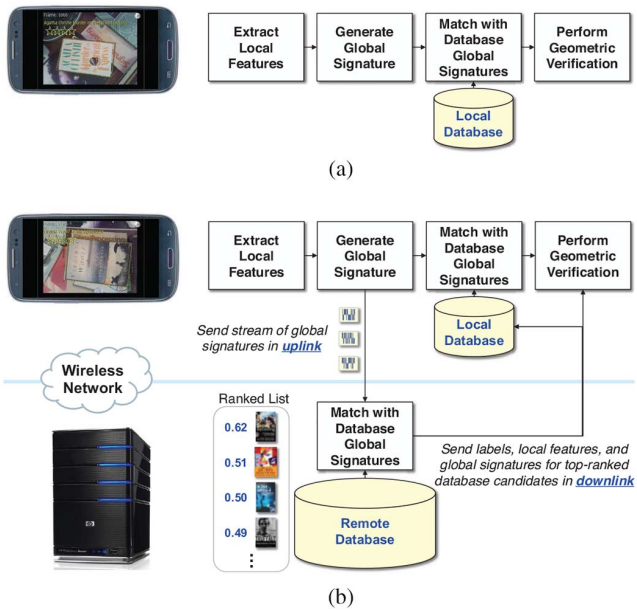


Fig. 1. Two different MVS system architectures. (a) On-device MVS system. (b) Hybrid MVS system.

new information from the remote server enables us to improve the response to the current query and to update the local database so that future local queries are more likely to succeed.

To achieve fast and robust recognition in the hybrid MVS system, our paper makes the following original contributions.

- *Robust recognition with global signatures*: We present a new version of the global REVV signature that has improved retrieval performance. Different from previous work [15], [30], [31], in this paper, we systematically analyze how to optimize the number of local features that are aggregated to generate a REVV signature, so that we maximize the retrieval accuracy for a target memory or bitrate budget. Our analysis develops a statistical model of the correlation scores between pairs of REVV signatures and the retrieval accuracy that can be obtained through REVV-based database comparisons. Based on this analysis, a new pipeline for extracting the optimized REVV signatures is constructed. We also develop two extensions that exploit the color information within images and relationships between similar database images to further improve retrieval accuracy.
- *Low-bitrate transmission of global signatures*: To instantly adapt to fast changes in the video contents, we perform image retrieval directly with the stream of viewfinder frames on the mobile device. If we ignore the temporal correlation between these viewfinder frames, then transmitting a continuous stream of REVV signatures to a remote server may require a large bitrate. Therefore, we propose several novel techniques for efficient interframe coding of REVV signatures. Local features are extracted from the viewfinder frames in a temporally coherent fashion. Then, REVV signatures are generated from these local features and predictively encoded. The resulting interframe-coded REVV stream is extremely compact and can be transmitted at very low bitrates to the server for the remote database query.

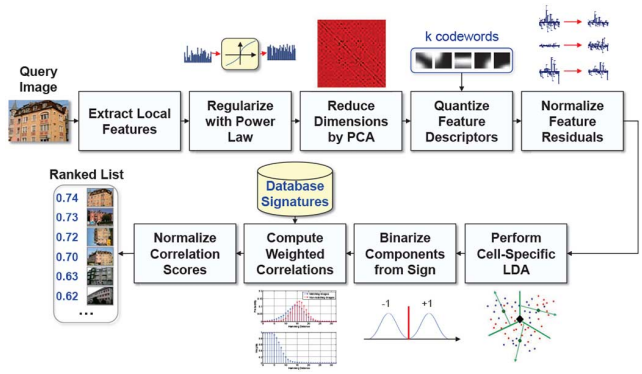


Fig. 2. Pipeline for generating and comparing global REVV signatures.

The remainder of this paper is organized as follows. Section II presents the design and optimization of the improved REVV signatures. Then, Section III develops two methods for interframe coding of REVV signatures. Finally, in Section IV, we utilize the newly developed techniques of Sections II-III to build a hybrid MVS system that combines the different advantages of on-device and on-server retrieval to achieve both fast and accurate large-scale image search.

II. Retrieval With Compact Global Signatures

The pipeline for generating and comparing global REVV signatures is shown in Fig. 2. Section II-A describes the steps in this pipeline. Then, Section II-B analyzes the retrieval performance of REVV and optimizes the local feature count to achieve the maximal retrieval accuracy under a memory or bitrate constraint. Comparisons between the improved REVV signature and several other high-performing global signatures are provided in Section II-C. Finally, extensions that exploit color information within images and relationships between similar database images are created in Section II-D.

A. Global Signature Generation and Comparison

In the first step of the pipeline in Fig. 2, N_{feat} local image features are extracted from an image. We will show later in Section II-B that N_{feat} should be carefully optimized to yield the best retrieval performance. Next, each extracted feature descriptor $\mathbf{v} = [v_1, \cdots, v_l]$ is regularized with a power law transformation [13], [32] that reduces the detrimental effects of peaky components, yielding $\mathbf{v}_\alpha = [\text{sign}(v_1)|v_1|^\alpha, \cdots, \text{sign}(v_l)|v_l|^\alpha]$. Typically, $\alpha \in [0.4, 0.7]$ yields good retrieval performance.

Then, the dimensionality of each feature descriptor is reduced by principal component analysis (PCA): $\mathbf{v}_{\text{pca}} = \mathbf{T}_{\text{pca}}(\mathbf{v}_\alpha - \mathbf{v}_{\alpha,\text{mean}})$, where $\mathbf{T}_{\text{pca}} \in \mathbb{R}^{l_{\text{pca}} \times l}$ is the PCA matrix and $\mathbf{v}_{\alpha,\text{mean}} \in \mathbb{R}^l$ is the mean of power law-transformed feature descriptors. Performing PCA enables us to more effectively learn a codebook for vector quantization in a lower-dimensional space using only a moderate number of training samples.

At this point, the PCA-transformed descriptors are quantized to a small codebook of k codewords, where typically $k \in [64, 512]$, using either hard binning with k-means or soft

binning with a Gaussian mixture model (GMM). After vector quantization, the quantization errors or residuals are added for each codeword. The l_{pca}-dimensional aggregated residual vector $\mathbf{f}_i$ for the ith codeword is normalized to have unit L_2 norm, so that the influences of visual burstiness and repetitive patterns in an image are effectively reduced [33].

To create an even more compact signature, a second dimensionality reduction operation maps each l_{pca}-dimensional residual vector to a shorter l_{lda}-dimensional coefficient vector. This time, however, we use the cell-specific statistics available after quantization and design a different linear discriminant analysis (LDA) transform for each Voronoi cell: $\mathbf{f}_{\mathrm{lda},i} = \mathbf{T}_{\mathrm{lda},i}\mathbf{f}_i$, where $\mathbf{T}_{\mathrm{lda},i} \in \mathbb{R}^{l_{\mathrm{lda}} \times l_{\mathrm{pca}}}$ is the LDA matrix for residual vectors extracted from matching and non-matching image pairs for the ith Voronoi cell. The benefits of using a cell-specific transform have been independently observed in [34].

Next, we binarize each l_{lda}-dimensional coefficient vector based on the sign of each coefficient. For the ith codeword, the binary vector is $\mathbf{b}_i = [\mathrm{sign}\{(\mathbf{f}_{\mathrm{lda},i})_1\}, \cdots, \mathrm{sign}\{(\mathbf{f}_{\mathrm{lda},i})_{l_{\mathrm{lda}}}\}]$. These binary vectors can be stored compactly in the system RAM or transmitted efficiently over the network. In addition to signed binarization, we also experimented with several recent hashing approaches–spectral hashing [35], max margin hashing [36], and iterative quantization hashing [37]–but did not observe any improvements over signed binarization. Another possible approach at this step is scalar quantization, which has shown good performance for local feature descriptors [38], but we prefer binarization over scalar quantization due to the smaller size of the binarized signature.

Binary vectors can be compared quickly in the compressed domain using XOR and POPCNT instructions. To compare the query REVV signature to a database REVV signature, we compute a weighted correlation

$$C = \frac{1}{(|I_q| \cdot |I_{db}|)^{\gamma}} \sum_{i \in I_q \cap I_{db}} w(h_i)\,(l_{\mathrm{lda}} - 2h_i) \tag{1}$$

where I_q and I_{db} are the sets of indices for the codewords visited by the query and database images, respectively; h_i is the Hamming distance between the query and database binary residual vectors at the ith codeword; $w(h_i)$ is a weighting function that rewards low Hamming distance values; and γ is an exponent that controls the power law regularization of the normalization factor. To achieve good recognition performance, we have found that the weighting function $w(h)$ can be derived from the relative likelihood of Hamming distances for matching and non-matching image pairs and that γ can be chosen from the range $0 < \gamma < 0.5$ [31]. By ranking the database images in terms of their correlation scores, we can quickly and reliably obtain a shortlist of the best matching database images.

The codebook, PCA and LDA transforms, and correlation weights used in the REVV pipeline are estimated from a set of training images that is composed of the INRIA Holidays Dataset, the Oxford Buildings Dataset, and the Pasadena Buildings Dataset. These training images are entirely separate from the datasets that we use later to evaluate retrieval performance.

B. Optimization of Local Feature Count

A very important control parameter in the REVV pipeline which has a significant impact on the retrieval accuracy, memory usage, and data transmission rate is N_{feat}, the number of local features extracted from each image. As we will show in this section, for a given codebook size k, there is an optimal number of local features $N^*_{\mathrm{feat}}(k)$ which achieves the maximal retrieval accuracy $P^*(k)$. To determine this optimal local feature count, we first develop a statistical model of the correlation scores for REVV signatures. From this model, we can then analyze the retrieval accuracy of REVV signatures and predict $N^*_{\mathrm{feat}}(k)$ and $P^*(k)$.

Analysis of Correlation Scores: Our model for the correlation scores captures the most important first-order effects in how the correlation scores change as the number of codewords k or the number of local features N_{feat} varies. Some second-order effects caused by score normalization and correlation weighting are ignored in this analysis to concentrate on the important trade-offs and to keep the number of parameters in the model to a minimum. When excluding the score normalization and correlation weighting steps, the correlation C_{nm} for a non-matching image pair and C_{m} for a matching image pair are given by

$$C_\theta = \sum_{i=1}^{N_{\mathrm{visit},\theta}} C_{\theta,i} \tag{2}$$

where $\theta \in \{\mathrm{nm}, \mathrm{m}\} = \{\mathrm{non-matching}, \mathrm{matching}\}$, $C_{\theta,i}$ is the correlation between binary residual vectors at a single codeword, and $N_{\mathrm{visit},\theta}$ is the number of codewords visited in common by the pair of images. Equivalently, if we know the Hamming distance $H_{\theta,i}$ between binary residual vectors at a single codeword, we can compute the codeword-level correlation as $C_{\theta,i} = l_{\mathrm{lda}} - 2H_{\theta,i}$. The image-level correlation can then be rewritten as

$$C_\theta = N_{\mathrm{visit},\theta} l_{\mathrm{lda}} - 2H_\theta \tag{3}$$

$$H_\theta = \sum_{i=1}^{N_{\mathrm{visit},\theta}} H_{\theta,i} \tag{4}$$

where H_θ is an image-level Hamming distance.

The number of codewords $N_{\mathrm{visit},\theta}$ visited in common by the pair of images can be modeled as a binomial random variable

$$p_{N_{\mathrm{visit},\theta}}(n) = \binom{k}{n} p^n_{\mathrm{visit},\theta}(1 - p_{\mathrm{visit},\theta})^{k-n} \tag{5}$$

where $p_{\mathrm{visit},\theta}$ is the probability that a codeword is visited in common by the pair of images. For non-matching images, whether or not one image visits a codeword has no effect on whether or not the other image visits the same codeword. Therefore, we have $p_{\mathrm{visit,nm}} = p^2_{\mathrm{visit}}$, where p_{visit} is the probability that a codeword is visited by a single image. For matching images, if one image visits a codeword, then the other matching image will visit the codeword with a probability $p_{\mathrm{visit,other}}$ that is higher than p_{visit}. Therefore, we have

$p_{\text{visit,m}} = p_{\text{visit}} p_{\text{visit,other}}$. We estimate p_{visit} and $p_{\text{visit,other}}$ from the aforementioned set of training images.

If $N_{\text{visit,nm}} = n$, then nl_{lda} dimensions are compared between the binary vectors of the two non-matching images. Let $h_{\text{nm,u}} = 1$ if the two binary vectors have different signs in the u^{th} dimension and $h_{\text{nm,u}} = 0$ otherwise. Suppose $p_{h_{\text{nm,u}}}(1) = p_{\text{sign,nm}}$. For a non-matching image pair, the variables $h_{\text{nm,u}}$ and $h_{\text{nm,v}}$ are statistically independent for $u \neq v$. Hence, the Hamming distance $H_{\text{nm}} = \sum_{u=1}^{nl_{\text{lda}}} h_{\text{nm,u}}$ between the two binary vectors can be modeled conditionally as a binomial random variable

$$p_{H_{\text{nm}}|N_{\text{visit,nm}}}(h|n) = \binom{l_{\text{lda}}}{h} p_{\text{sign,nm}}^{h} (1 - p_{\text{sign,nm}})^{nl_{\text{lda}}-h}. \tag{6}$$

Since $p_{\text{sign,nm}} \approx 0.5$, the two binary vectors will agree in sign for half of the dimensions on average.

Similarly, if $N_{\text{visit,m}} = n$, then nl_{lda} dimensions are compared between the binary vectors of the two matching images. As before, let $h_{\text{m,u}} = 1$ if the two binary vectors have different signs in the u^{th} dimension and $h_{\text{m,u}} = 0$ otherwise, and let $p_{h_{\text{m,u}}}(1) = p_{\text{sign,m}}$. For two matching images, however, the variables $h_{\text{m,u}}$ and $h_{\text{m,v}}$ are dependent for $u \neq v$. Matching feature descriptors between the two images cause strong correlations between the various dimensions. If the binary residual vectors match in one dimension, then it becomes more likely that they also match in other dimensions. Hence, we cannot use a binomial random variable to accurately model the conditional distribution of the sum $H_m = \sum_{u=1}^{nl_{\text{lda}}} h_{\text{m,u}}$.

To effectively model the dependence between the different dimensions of the residual vector, we employ a generalized binomial distribution (GBD) [39]. Appendix A describes the procedure for computing the probability $p_{\text{GBD}}(h|N)$ of achieving h successes in N trials in the GBD. With the GBD, we can model the conditional distribution for H_m

$$p_{H_m|N_{\text{visit,m}}}(h|n) = p_{\text{GBD}}(h|nl_{\text{lda}}). \tag{7}$$

Because the GBD captures the dependence between Bernoulli trials in the sequence, we can accurately model the dependence between the different dimensions of the residual vector.

Since $C_\theta = nl_{\text{lda}} - 2H_\theta$ when $N_{\text{visit},\theta} = n$, the conditional distribution for C_θ is

$$p_{C_\theta|N_{\text{visit},\theta}}(c|n) = p_{H_\theta|N_{\text{visit},\theta}}(0.5(nl_{\text{lda}} - c)|n). \tag{8}$$

Then, the overall distribution for C_θ is a mixture distribution

$$p_{C_\theta}(c) = \sum_{n=1}^{k} p_{C_\theta|N_{\text{visit},\theta}}(c|n)\, p_{N_{\text{visit},\theta}}(n). \tag{9}$$

For non-matching and matching image pairs, $p_{C_\theta}(c)$ is a mixture of binomial distributions and a mixture of GBDs, respectively.

To evaluate the accuracy of our statistical model, we use the MPEG CDVS Dataset [28]. This dataset consists of five categories of images: Graphics, Paintings, Video Frames, Landmarks, and Common Objects. There is a query set of 8 K images and a database 18 K images. The dataset also provides a set of 1 M distractor images that can be merged with the 18 K database images to assess large-scale retrieval performance.

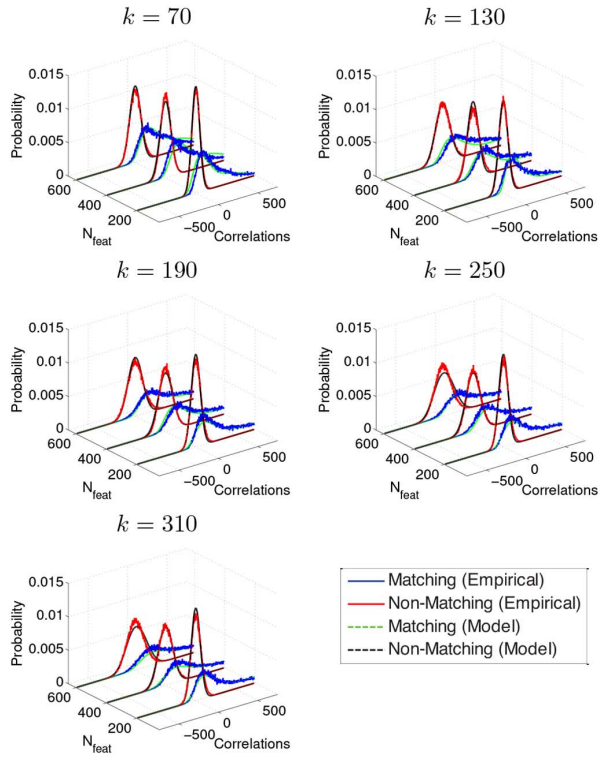


Fig. 3. Distributions of REVV correlation scores for codebook sizes $k = 70, 130, 190, 250, 310$ and feature counts $N_{\text{feat}} = 200, 400, 600$.

Fig. 3 plots the distributions for C_{nm} and C_{m} for codebook sizes $k = 70, 130, 190, 250, 310$ and local feature counts $N_{\text{feat}} = 200, 400, 600$. Similar results are obtained for other values of k and N_{feat}. The distributions computed according to the statistical model developed in this section are plotted together with the empirical distributions computed from the images in the MPEG CDVS Dataset. We can observe that the model distributions match well with the empirical distributions. The non-matching distributions are centered at a correlation value of 0 and are well modeled using a mixture of binomial distributions. Due to the dependence between the different dimensions of the residual vector, the matching distributions have a long tail skewed toward large positive correlation values, and these distributions are well modeled with a mixture of GBDs.

Analysis of Retrieval Accuracy: When querying a large database of images with REVV signatures, a ranked list of correlation scores is generated. In this list, $N_{\text{db,nm}}$ correlation scores $C_{\text{nm}}^1, \cdots, C_{\text{nm}}^{N_{\text{db,nm}}}$ belong to non-matching database images and $N_{\text{db,m}}$ correlation scores $C_{\text{m}}^1, \cdots, C_{\text{m}}^{N_{\text{db,m}}}$ belong to matching database images. Usually, $N_{\text{db,nm}} \gg N_{\text{db,m}}$, meaning there are substantially more non-matching images in the database than there are matching images. We assume the scores $C_{\text{nm}}^1, \cdots, C_{\text{nm}}^{N_{\text{db,nm}}}$ and $C_{\text{m}}^1, \cdots, C_{\text{m}}^{N_{\text{db,m}}}$ are distributed i.i.d. according to the non-matching and matching correlation models, respectively, developed above.

The mean precision at rank 1 (PA1) value measures how often the top-ranked database image is a correct match.

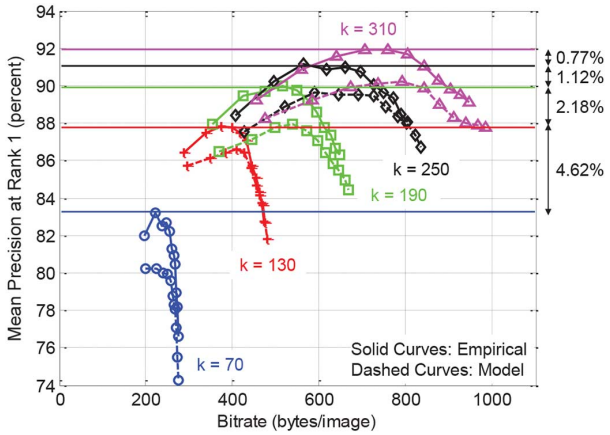


Fig. 4. PA1 versus the bitrate in bytes/image for the MPEG CDVS Dataset with a database of 18 K images. REVV signatures for five different codebook sizes $k = 70, 130, 190, 250, 310$ are evaluated. The horizontal colored lines mark the maximal precision values achieved for each codebook size.

Whenever the maximal value of the matching correlation scores $C_{\mathrm{m}}^{\max} = \max(C_{\mathrm{m}}^1, \cdots, C_{\mathrm{m}}^{N_{\mathrm{db,m}}})$ is larger than the maximal value of the non-matching correlation scores $C_{\mathrm{nm}}^{\max} = \max(C_{\mathrm{nm}}^1, \cdots, C_{\mathrm{nm}}^{N_{\mathrm{db,nm}}})$, then the top-ranked database image is correct. The cumulative distribution function (CDF) for $C_\theta^{\max}$ is found to be

$$F_{C_\theta^{\max}}(c) = P\left(C_\theta^{\max} \le c\right) \tag{10}$$

$$= P\left(C_\theta^1 \le c, \cdots, C_\theta^{N_{\mathrm{db},\theta}} \le c\right) \tag{11}$$

$$= P\left(C_\theta^1 \le c\right)^{N_{\mathrm{db},\theta}} \tag{12}$$

$$= \left(F_{C_\theta^1}(c)\right)^{N_{\mathrm{db},\theta}} \tag{13}$$

where the third line follows from the i.i.d. assumption for the correlation scores. Subsequently, the probability mass function (PMF) for $C_\theta^{\max}$ can be obtained by taking discrete differences of the CDF. Now, the PA1 value is predicted to be

$$\mathrm{PA1} = P\left(C_{\mathrm{m}}^{\max} \ge C_{\mathrm{nm}}^{\max}\right) \tag{14}$$

$$= \sum_c P\left(c \ge C_{\mathrm{nm}}^{\max}\right) p_{C_{\mathrm{m}}^{\max}}(c) \tag{15}$$

$$= \sum_c F_{C_{\mathrm{nm}}^{\max}}(c) p_{C_{\mathrm{m}}^{\max}}(c). \tag{16}$$

In Fig. 4, we plot the precision at rank one (PA1) value versus the bitrate per image in the MPEG CDVS Dataset, with a database of 18 K images, for five different codebook sizes $k = 70, 130, 190, 250, 310$. For each codebook size, the PA1-versus-bitrate curve is generated by increasing the number of local features N_{feat} from 100 to 600 in increments of 50. The mean bitrate associated with each PA1 value is $k(p_{\mathrm{visit}} l_{\mathrm{lda}} + 1)$ bits per image, because we need 1 bit per codeword to signal if that codeword has been visited and l_{lda} bits to describe each binary residual vector. Note that p_{visit} increases as N_{feat} increases. The solid curves in Fig. 4 represent the empirical measurements, while the dashed curves represent the values predicted by our statistical model.

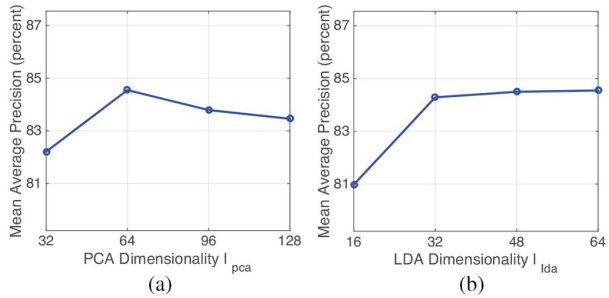


Fig. 5. MAP for REVV signatures on the MPEG CDVS Dataset with a database of 18 K images, as (a) PCA dimensionality l_{pca} is varied, and (b) LDA dimensionality l_{lda} is varied, when $l_{\mathrm{pca}} = 64$ is fixed. A codebook of $k = 190$ codewords is used in these experiments.

Our model accurately captures two important characteristics: (i) there is an optimal number of local features $N_{\mathrm{feat}}^*(k)$ to choose for each codebook size k that yields the maximal PA1 value $P^*(k)$, and (ii) as the codebook size k increases, the gaps between the maximal PA1 values decrease, so there are rapidly diminishing gains to using larger codebooks. Our model can be used to guide the selection of the optimal pair of k and $N_{\mathrm{feat}}^*(k)$ values for REVV-based image retrieval to meet a given memory or transmission bitrate constraints.

C. Comparisons to Other Global Signatures

Before we compare REVV to several other high-performing global signatures, we decide how to choose the PCA dimensionality l_{pca} and LDA dimensionality l_{lda}. First, we study the effect of varying l_{pca} on the retrieval performance of REVV signatures. Fig. 5(a) plots the mean average precision (MAP) on the MPEG CDVS Dataset as l_{pca} is varied. These experiments use a codebook of $k = 190$ codewords, although similar results are observed for other codebook sizes. For each value of l_{pca}, the LDA dimensionality l_{lda} is separately optimized to yield the highest retrieval accuracy. From Fig. 5(a), we can observe that when l_{pca} is too low, the retrieval accuracy drops because informative details of the feature descriptor are lost during dimensionality reduction. On the opposite end, when l_{pca} is too high, the retrieval accuracy drops because it becomes increasingly difficult to effectively learn a growing number of parameters for the vector quantizer using a fixed training set. Setting $l_{\mathrm{pca}} = 64$ yields the highest MAP value.

Next, we study the effect of varying the LDA dimensionality l_{lda} on retrieval performance, when the PCA dimensionality is fixed to $l_{\mathrm{pca}} = 64$. Fig. 5(b) plots the MAP values on the MPEG CDVS Dataset as l_{lda} is varied. As l_{lda} increases beyond 32, the MAP value increases only slightly. As l_{lda} decreases below 32, the MAP value drops noticeably because the binary residual vectors become too short to effectively discriminate between matching and non-matching image pairs. Thus, $l_{\mathrm{lda}} = 32$ provides a good tradeoff between signature length and retrieval accuracy.

Now, we can compare several high-performing compact global signatures including our improved version of REVV.

1) THC: This refers to tree histogram coding [24]. A vocabulary tree [12] with depth $d = 6$ and branch factor $k = 10$ is used, so the tree has 1 M leaf nodes. Term frequency inverse document frequency (TF-IDF) weighting, greedy-10

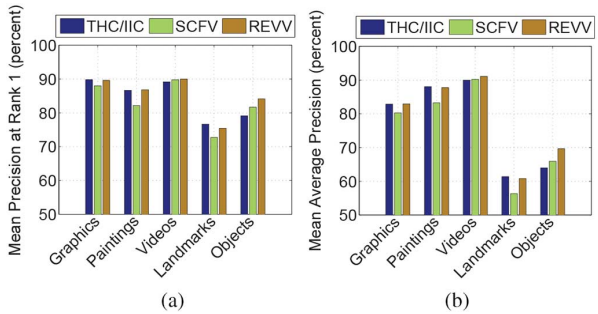


Fig. 6. Retrieval accuracy for several different global signatures on the MPEG CDVS Dataset with a database of 1 M images. The retrieval accuracy is measured in terms of (a) PA1 and (b) MAP.

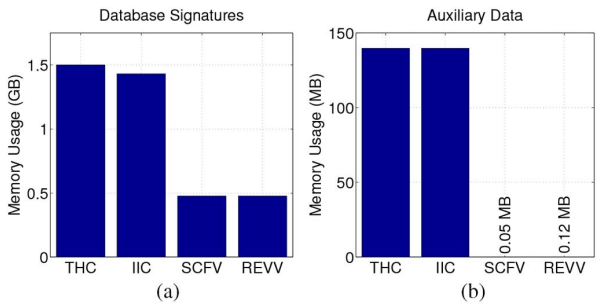


Fig. 7. Memory usage for several different global signatures on the MPEG CDVS Dataset with a database of 1 M images. The memory usage is split between (a) the database global signatures and (b) auxiliary data needed to generate the global signatures.

search [40], and soft binning [41] are used to obtain a high retrieval accuracy.

2) IIC: This refers to inverted index coding [26]. All parameters are the same as for THC, except that an inverted index is compressed rather than the tree histograms.
3) SCFV: This refers to the scalable compressed Fisher vector [42]–[45], which is a compact and enhanced version of the Fisher vector from [13]. We compare against a version of SCFV with $k = 128$ codewords that has similar memory usage as REVV.
4) REVV: We use the new REVV signature developed in this paper with a codebook of $k = 190$ codewords, $l_{\mathrm{pca}} = 64$ PCA eigenvectors, and $l_{\mathrm{lda}} = 32$ LDA eigenvectors.

Fig. 6 plots the PA1 and MAP values across the five categories of the MPEG CDVS Dataset for THC, IIC, SCFV, and REVV. A database of 1 M images is used to evaluate large-scale retrieval performance. The accuracies for REVV and THC/IIC are very similar across the different categories, although REVV outperforms THC/IIC by several percent in the objects category. Across all five categories, REVV consistently outperforms SCFV by a few percent.

Fig. 7 plots the memory usage for the 1 M database images in the MPEG CDVS Dataset, where the memory usage is separated into two types: (i) the memory used by the database signatures and (ii) the memory used by auxiliary data such as codebooks which are needed to generate the database signatures. First, Fig. 7(a) shows that the two feature residual methods, REVV and SCFV, significantly reduce the memory usage of the database signatures compared to the two tree histogram methods, THC and IIC. Second, Fig. 7(b) shows the large savings in the memory usage of the auxiliary data achieved by REVV or SCFV. The same high retrieval accuracy provided by THC and IIC can be achieved by REVV or SCFV using a codebook that is several orders of magnitude smaller. These substantial memory savings are important to deploy the retrieval system on a mobile device with a small memory capacity, as we will show in Section IV.

D. Extensions With Color and Query Expansion

In previous sections, we utilized only the grayscale information within images to extract REVV signatures, and we ignored the relationships between similar database images when comparing REVV signatures. This section develops two extensions that improve retrieval performance by utilizing each image's color information and exploiting similarities within groups of database images.

First, we create a new color version of REVV. The REVV pipeline depicted in Fig. 2 works generally for any multi-dimensional feature descriptor. In lieu of grayscale local features, we now compute RGB local features [46]. For each local keypoint, three separate d-dimensional feature descriptors are extracted from the RGB channels and concatenated into a longer $3d$-dimensional descriptor. PCA is subsequently performed in the new $3d$-dimensional feature space. In addition to correlations between different spatial and gradient bins in the feature descriptor, PCA can now additionally exploit correlations between the RGB channels. Other steps in the pipeline of Fig. 2 remain the same, except different codebooks and LDA eigenvectors are trained for color REVV compared to grayscale REVV.

Second, we extend REVV-based retrieval to incorporate a query expansion step that does not require geometric verification. Most prior query expansion methods required geometric verification to create secondary queries [47]–[49]. Query expansion without geometric verification has been explored for vocabulary trees with Hamming embeddding [50], but not before for feature residual vectors. For each query, we use the following procedure to score database images.

1) Compare the query REVV signature to the database REVV signatures to obtain an initial ranked list $(i_1, C[i_1])$, $(i_2, C[i_2])$, $\cdots$, $(i_{N_{\mathrm{db}}}, C[N_{\mathrm{db}}])$. Here, the similarity scores are sorted in decreasing order $C[i_1] > C[i_2] > \cdots > C[i_{N_{\mathrm{db}}}]$, and i_n is the index of the nth image in the ranked list.
2) Form a set of likely matching database images $I_{\mathrm{sim}} = \{i : C[i] > \alpha_{\mathrm{sim}} C_{\mathrm{mean}}\}$. Here, $C_{\mathrm{mean}} = \mathrm{mean}(C[i_1], \cdots, C[I_{N_{\mathrm{db}}}])$ is the mean of all similarity scores, and the factor $\alpha_{\mathrm{sim}} > 1$ is chosen during training to ensure a high likelihood of including only matching database images in I_{sim}.
3) For a shortlist size $N_s \ll N_{\mathrm{db}}$, update the similarity score for each image in the shortlist as

$$E(i_n) = \sum_{i_m \in I_{\mathrm{sim}}} C(i_m, i_n) \tag{17}$$

$$C[i_n] := \frac{C[i_n] + E(i_n)}{|I_{\mathrm{sim}}| + 1} \qquad n = 1, 2, \cdots, N_s \tag{18}$$

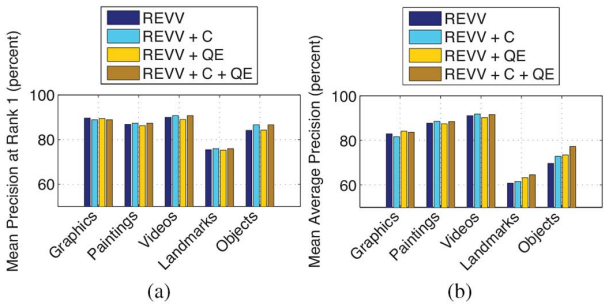


Fig. 8. Retrieval accuracy for several different versions of REVV on the MPEG CDVS Dataset with a database of 1 M images, measured in terms of (a) PA1 and (b) MAP.

where $C(i_m, i_n)$ is the similarity score between database images i_m and i_n.

4) Resort the similarity scores in the shortlist.

The query expansion occurs in step (3), where each database image in the set I_{sim} is used as a secondary query to compare against other database images in a shortlist. The updated similarity score for each image in the shortlist is an average of $|I_{\text{sim}}| + 1$ scores: the old similarity score to the actual query image and the new similarity scores to the database images in I_{sim}. Note that if I_{sim} is empty, i.e., no database images are judged to be highly similar to the query image in the initial ranking, then query expansion has no effect on the similarity scores.

Fig. 8 plots the retrieval accuracy on the MPEG CDVS Dataset for four different versions of REVV, listed as follows.

1) REVV: This is the same version as reported in Section II-C, with grayscale feature descriptors and no query expansion.
2) REVV+C: This version uses RGB feature descriptors but no query expansion.
3) REVV+QE: This version uses grayscale feature descriptors and query expansion.
4) REVV+C+QE: This version uses RGB feature descriptors and query expansion.

For all four versions, $k = 190$, $l_{\text{pca}} = 64$, $l_{\text{lda}} = 32$, and $N_{\text{feat}} = 250$.

First, REVV+C yields improvements over REVV in four of the five categories. REVV+C incurs a 1 percent drop in Graphics compared to REVV, because camera flashes in a large subset of images within Graphics cause severe color distortions. Second, REVV+QE yields improvements over REVV in Graphics, Landmarks, and Objects, categories in which there are multiple matching database images for each query image. REVV+QE produces no change in Paintings and Videos, because in those categories each query image has only one matching database image. Finally, REVV+C+QE gives the best performance and produces consistent improvements over REVV in all five categories. REVV+C+QE increases the MAP by 3 percent and the PA1 by 2 percent overall over REVV. The improvement is larger in MAP than in PA1, because query expansion affects images in later spots within the ranked list more than the image in the top spot.

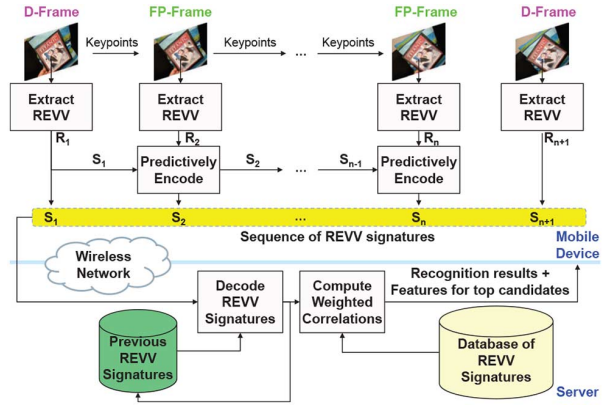


Fig. 9. Interframe coding of global REVV signatures. A stream of REVV signatures is extracted from the viewfinder frames on the mobile device, predictively encoded, transmitted to a server, and compared against a database of REVV signatures hosted on the server.

III. Interframe Coding of Global Signatures

In the last section, we developed an improved REVV global signature for large-scale image retrieval. Now, we will develop efficient interframe coding methods for these REVV signatures. This will enable a stream of REVV signatures, generated from a sequence of viewfinder frames on the mobile device, to be efficiently transmitted to a server. We build upon our preliminary work [51] and develop the interframe coding framework that is shown in Fig. 9. First, Section III-A describes a method for extracting temporally coherent local features. Then, Sections III-B through III-D present three interframe coding methods for REVV signatures.

A. Temporally Coherent Keypoint Detector

First, to exploit the correlation between neighboring REVV signatures, we use a temporally coherent keypoint detector (TCKD) [22] that divides the acquired video frames into two categories: detection frames (D-Frames) and forward propagation frames (FP-Frames). For each D-Frame, independent detection of local feature keypoints is performed. Then, each feature keypoint is propagated into the subsequent FP-Frame by searching across a small set of similarity transforms. The search attempts to minimize the sum of absolute differences (SAD) between a feature keypoint's canonical patch in the D-Frame and the transformed canonical patch in the FP-Frame. This propagation continues until the next D-Frame appears in the sequence, or until many of the feature keypoints are not successfully propagated. We extract a local feature descriptor from each keypoint determined by the TCKD and use these feature descriptors subsequently to create temporally coherent REVV signatures.

B. Selective Codeword Propagation

After generating REVV signatures from the temporally coherent local features, we want to predictively encode the stream of REVV signatures to substantially reduce the bitrate. For a codebook of k visual words, let the original REVV signature of frame t be denoted as $\mathbf{R}_t = \{(u^R_{t,1}, \mathbf{b}^R_{t,1}), \cdots, (u^R_{t,k}, \mathbf{b}^R_{t,k})\}$. Here, $u^R_{t,i} \in \{0, 1\}$ is a binary variable indicating whether or not the ith codeword is visited by frame t, and $\mathbf{b}^R_{t,i} \in \{-1, +1\}^{l_{\text{lda}}}$

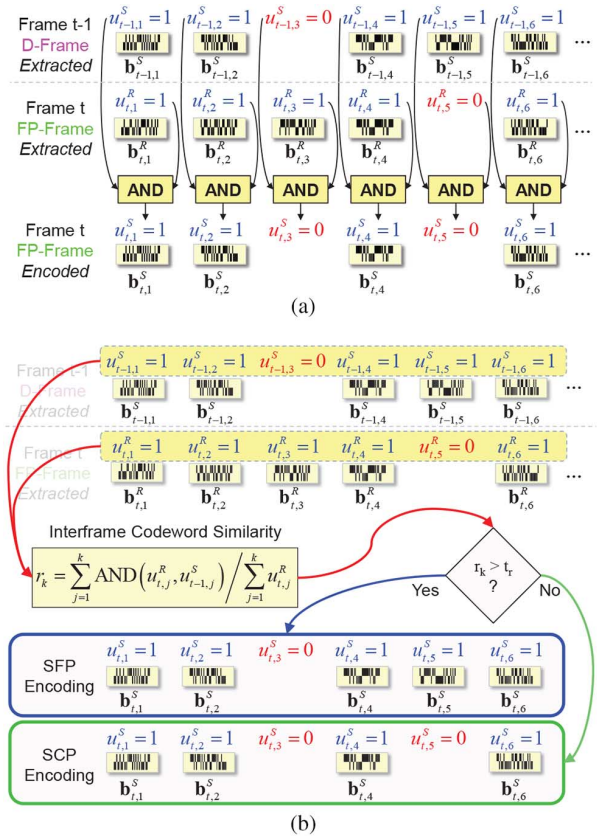


Fig. 10. Predictive coding of global REVV signatures with (a) SCP and (b) SFP.

is the corresponding binary residual vector if the codeword is visited. Similarly, let the predictively coded REVV signature which is sent to the server be denoted as $\mathbf{S}_t = \{(u^S_{t,1}, \mathbf{b}^S_{t,1}), \cdots, (u^S_{t,k}, \mathbf{b}^S_{t,k})\}$.

For a D-Frame, we set $\mathbf{S}_t = \mathbf{R}_t$, meaning the original REVV signature is transmitted without any predictive coding. For an FP-Frame, the selective codeword propagation (SCP) method assigns $u^S_{t,i} = \text{AND}(u^R_{t,i}, u^S_{t-1,i})$ for $1 \le i \le k$. If $u^S_{t,i} = 1$, then the SCP method further assigns $\mathbf{b}^S_{t,i} = \mathbf{b}^S_{t-1,i}$, which propagates the transmitted residual vector of frame $t-1$ at the ith codeword. An example of this predictive coding process with SCP is illustrated in Fig. 10(a). We do not encode the differences between the residual vectors $\mathbf{b}^R_{t,i}$ and $\mathbf{b}^S_{t-1,i}$, which are caused by small temporal fluctuations in the feature descriptors. These differences do not noticeably affect retrieval results and would require a considerable bitrate to communicate. Only $u^S_{t,i}$ needs to be sent for each FP-Frame, because $\mathbf{b}^S_{t,i} = \mathbf{b}^S_{t-1,i}$ has been previously received at the server. Additionally, $u^S_{t,i}$ needs to be sent only when $u^S_{t-1,i} = 1$, because $u^S_{t,i} = 0$ when $u^S_{t-1,i} = 0$.

For independent coding of REVV signatures where the temporal correlation is ignored, the uplink bitrate (in bits/second) is given by

$$R_{\text{Indep}} = N_{\text{Frames}} k \left(1 + p_{\text{visit}} l_{\text{lda}}\right) \tag{19}$$

where N_{Frames} is the number of viewfinder frames per second. As in Section II, p_{visit} is the probability that a codeword is visited by an image, and l_{lda} is the residual vector dimensionality at each codeword after cell-specific LDA. For SCP, the uplink bitrate (in bits/second) is given by

$$R_{\text{SCP}} = \underbrace{N_{\text{DF}} k \left(1 + p_{\text{visit}} l_{\text{lda}}\right)}_{\text{bitrate for D-Frames}} + \underbrace{N_{\text{FPF}} k p_{\text{visit}}}_{\text{bitrate for FP-Frames}} \tag{20}$$

where N_{DF} and N_{FPF} are the number of D-Frames and FP-Frames, respectively, per second. Note that $N_{\text{Frames}} = N_{\text{DF}} + N_{\text{FPF}}$. The bitrate savings between independent coding and SCP coding can be expressed as

$$\begin{aligned} \Delta R_{\text{SCP}} &= R_{\text{Indep}} - R_{\text{SCP}} && (21)\\ &= N_{\text{FPF}} k \left[1 + p_{\text{visit}} \left(l_{\text{lda}} - 1\right)\right]. && (22) \end{aligned}$$

Thus, the bitrate savings gained by SCP increase as the number of FP-Frames per second increases.

C. Selective Frame Propagation

When the scene content changes gradually, consecutive frames visit mostly the same codewords and have similar residual vectors at these codewords. Taking the idea behind SCP one step further, the selective frame propagation (SFP) method propagates all of the residual vectors between two frames if these frames' REVV signatures are very similar. An example of predictive coding with SFP is shown in Fig. 10(b). As in the previous section, let the original REVV signature of frame t be denoted as $\mathbf{R}_t = \{(u^R_{t,1}, \mathbf{b}^R_{t,1}), \cdots, (u^R_{t,k}, \mathbf{b}^R_{t,k})\}$ and the transmitted REVV signature be denoted as $\mathbf{S}_t = \{(u^S_{t,1}, \mathbf{b}^S_{t,1}), \cdots, (u^S_{t,k}, \mathbf{b}^S_{t,k})\}$. We define the interframe codeword similarity between frames t and $t-1$ as

$$r_k(t, t-1) = \frac{\sum_{i=1}^{k} \text{AND}\left(u^R_{t,i}, u^S_{t-1,i}\right)}{\sum_{i=1}^{k} u^R_{t,i}}. \tag{23}$$

If $r_k(t, t-1)$ exceeds a high threshold t_{r_k}, then SFP assigns $u^S_{t,i} = u^S_{t-1,i}$ and $\mathbf{b}^S_{t,i} = \mathbf{b}^S_{t-1,i}$ for $1 \le i \le k$. Only a single bit is sent to the server to indicate that the previous frame's REVV signature should be entirely propagated at every codeword. Otherwise, if $r_k(t, t-1)$ falls below t_{r_k}, then SFP switches back to SCP coding, and in this case, a single bit is sent to the server to indicate a temporary activation of the SCP mode, followed by the bits generated normally by SCP.

Similar to the SCP case, we can quantify the uplink bitrate (in bits/second) for SFP

$$\begin{aligned} R_{\text{SFP}} = &\underbrace{N_{\text{DF}} k \left(1 + p_{\text{visit}} l_{\text{lda}}\right)}_{\text{bitrate for D-Frames}} \\ &+ \underbrace{N_{\text{FPF}} \left(1 + k p_{\text{visit}} P\left(r_k < t_{r_k}\right)\right)}_{\text{bitrate for FP-Frames}} \end{aligned} \tag{24}$$

where $P(r_k < t_{r_k})$ is that probability that the interframe codeword similarity r_k falls below the threshold t_{r_k}. The bitrate savings between independent coding and SFP coding of REVV signatures can be written as

$$\begin{aligned} \Delta R_{\text{SFP}} &= R_{\text{Indep}} - R_{\text{SFP}} && (25)\\ &= \Delta R_{\text{SCP}} + N_{\text{FPF}} \left[k p_{\text{visit}} P\left(r_k \ge t_{r_k}\right) - 1\right] && (26) \end{aligned}$$

where we see that the bitrate savings offered by SFP are greater than the bitrate savings offered by SCP, as long as $P(r_k \geq t_{r_k}) > 1/(kp_{\text{visit}})$. When the interframe codeword similarity r_k increases, the probability $P(r_k \geq t_{r_k})$ increases and leads to larger bitrate reductions by SFP on top of the bitrate savings already provided by SCP.

D. Selective Frame Propagation + Local Search

The selective frame propagation + local search (SFP+LS) method fully combines the advantages of searching both a local database on the mobile device and a larger database on a remote server. If the local search results in a database match with a RANSAC inlier count N_{RANSAC} higher than a threshold $t_{\text{RANSAC}} = 25$ feature matches, then the search finishes locally on the mobile device. Otherwise, a REVV signature is transmitted to the server by SFP encoding. In this case, the first D-Frame in the coding chain occurs on the first frame during which the local search fails. When the server replies with the labels, REVV signatures, and local features of the top candidates, the local on-device database is opportunistically updated, so that (i) the current query can be improved by selecting from the best database candidates both locally and remotely, and (ii) future local searches are more likely to succeed with the updated local database.

The uplink bitrate (in bits/second) for SFP+LS is

$$R_{\text{SFP+LS}} = (1 - p_{\text{local}})\, R_{\text{SFP}} \tag{27}$$

where p_{local} is the probability that the local database search succeeds. Then, the savings between independent coding and SFP+LS can be written as

$$\Delta R_{\text{SFP+LS}} = R_{\text{Indep}} - R_{\text{SFP+LS}} \tag{28}$$

$$= \Delta R_{\text{SFP}} + p_{\text{local}} R_{\text{SFP}}. \tag{29}$$

Hence, the bitrate savings gained by SFP+LS are even larger than the bitrate savings gained by SFP. The additional bitrate savings increases as the probability of success in local search p_{local} increases.

IV. Hybrid Mobile Visual Search System

Having developed an improved REVV signature in Section II and several methods for efficient interframe coding of a REVV stream in Section III, we now put all the pieces together in a practical hybrid MVS system. First, Section IV-A presents an on-device image retrieval system with the improved REVV signature. Then, in Section IV-B, we utilize the interframe coding of REVV to query a remote server at a very low uplink bitrate, which greatly improves the search capabilities of the system when a local database search is insufficient.

A. On-Device Image Retrieval

To perform on-device image matching, we must first be able to store the database in the small amount of RAM available on a mobile device. Fig. 11 plots a histogram of the RAM capacities on 1,160 different mobile devices. In RAM, we store a local database of 100 K images, which is an order of magnitude larger than the local databases used in our previous work [30], [31]. These 100 K images could represent images of all products in a supermarket or bookstore, images of all landmarks in the local vicinity as estimated from the current GPS coordinates, or images of all the artwork in a museum, among many possible applications. When the database images are represented by the compact REVV signatures developed in Section II, the entire database consumes about 50 MB in RAM. Hence, this memory-efficient database can be stored on essentially 100 percent of these mobile devices which are available on the market today. For the majority of mobile devices, the 50 MB of RAM represents only a small fraction of the overall RAM capacity, so other applications running concurrently on the same device have most of the RAM still available. The small amount of memory usage also enables the database to be quickly transferable from the SD card or internal storage disk into RAM when the MVS application is initially launched on the mobile device or when the MVS application switches between different content databases.

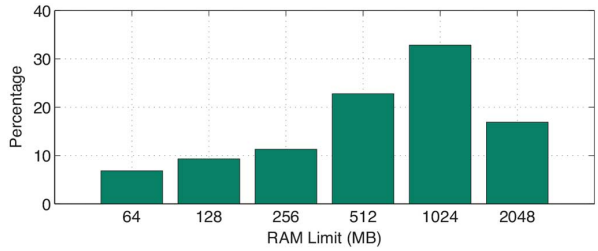


Fig. 11. Histogram of RAM capacities for 1,910 mobile devices available in December of 2014.

Using an efficient tracker based on RIFF features [8], we estimate the motion between viewfinder frames. The tracking enables us to determine periods of low motion that correspond to periods of user interest. At the beginning of every low-motion interval, $N_{\text{feat}} = 250$ SURF features [6] are extracted from each selected viewfinder frame. Then, a REVV signature is generated from the SURF feature set of each image using the pipeline described in Section II-A. For REVV, we use a codebook of $k = 190$ codewords, $l_{\text{pca}} = 64$ global PCA eigenvectors, and $l_{\text{lda}} = 32$ cell-specific LDA eigenvectors per codeword. Note that we optimize the local feature count to be $N_{\text{feat}} = 250$ to maximize retrieval accuracy for the $k = 190$ codewords, in accordance with our model in Section II-B. After the query REVV signature is compared against the database REVV signatures, a ranked list of the selected database candidates is created. The top 25 database images in this ranked list are further compared to each query frame using a distance ratio test [5] and RANSAC with an affine model [52] to determine the best matching database image(s). Finally, annotations are drawn on the viewfinder for the best matching object(s) and propagated using the RIFF-based tracker.

To create a highly interactive user experience, the image retrieval latency must be very low. We have implemented and tested the on-device MVS system illustrated in Fig. 1(a) on two Android devices: (i) a Samsung Galaxy S3 smartphone (1.4 GHz ARM Cortex A9 processor, 1 GB of RAM) and (ii) an Asus Google Nexus 7 tablet (1.5 GHz Qualcomm Snapdragon S4 processor, 2 GB of RAM). Fig. 12(a) plots a histogram of the image retrieval latency for 400 different queries executed by our system on each device. On average, each query takes

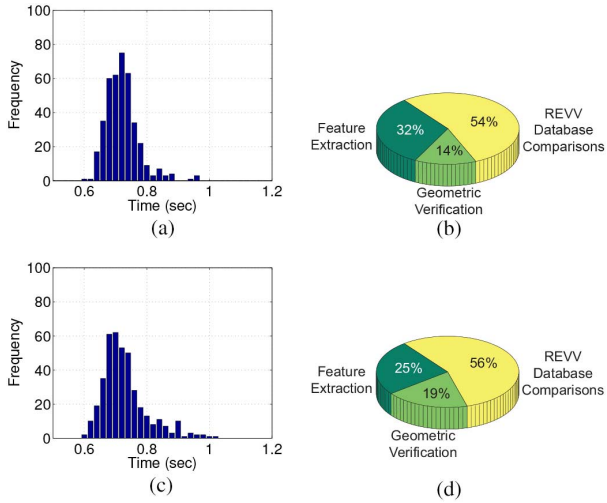


Fig. 12. Measurements of on-device image retrieval latency for 400 different queries against a database of 100 K images. (a), (c) Histogram of latencies. (b), (d) Percentage of time spent in feature extraction, database search, and geometric verification. The Samsung Galaxy S3 and Asus Google Nexus 7 have 1.4 GHz ARM Cortex A9 and 1.5 GHz Qualcomm Snapdragon S4 processors, respectively.

about 0.7 seconds on both devices, which enables near real-time augmentations to appear in the viewfinder. Fig. 12(b) plots the percentage of time spent in local feature extraction, database comparisons with REVV signatures, and geometric verification of the top database candidates. Because we are searching a large database of 100 K images on the mobile device, the REVV database comparisons take about half of the search time. Searching a smaller on-device database will reduce the REVV comparison latency accordingly. The feature extraction and geometric verification latencies are independent of the database size, assuming the shortlist length for geometric verification is fixed. Importantly, the quick response times shown in Fig. 12(a) can be achieved anywhere and anytime, independent of network or server conditions, because only fast on-device image retrieval is required.

B. Hybrid Image Retrieval

We now extend the on-device MVS system described in the previous section into the hybrid MVS system depicted in Fig. 1(b) using the interframe coding methods developed in Section III. We perform an evaluation on the Stanford Streaming Mobile Augmented Reality (SSMAR) Dataset [22]. In this dataset, there are 32 VGA-resolution query videos recorded at a frame rate of 30 frames/second with a mobile device and a database of 1 M still images. Whereas the database stored on the mobile device contained 100 K images, the database stored on the server contains 1 M images and hence provides much greater search coverage.

We test four different methods for the hybrid MVS system, listed as follows.

1) Independent: This refers to independent coding of a continuous stream of REVV signatures.
2) SCP: This refers to interframe coding with the selective codeword propagation method of Section III-B. We use 1 D-Frame and 29 FP-Frames for every 30 frames.

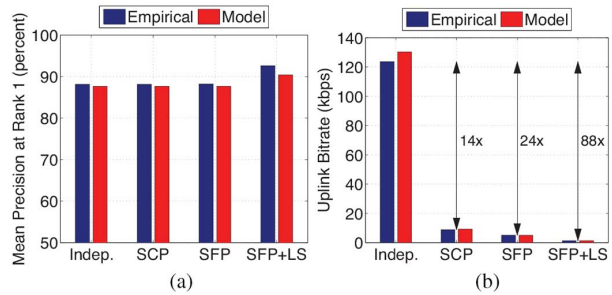


Fig. 13. (a) PA1 and (b) Uplink bitrate (in kbps) for four different coding methods on the SSMAR Dataset. SFP+LS uses a downlink bitrate of 14 kbps, while the other three methods require a downlink bitrate close to 0 kbps.

3) SFP: This refers to interframe coding with the selective frame propagation method of Section III-C. The same ratio of D-Frames to FP-Frames is used as for SCP. The interframe codeword similarity threshold is set to $t_{r_k} = 0.9$.
4) SFP+LS: This refers to the selective frame propagation + local search method of Section III-D. At the beginning of the experiment, none of the images in the SSMAR Dataset are contained in the local database, but over time, the local database is updated with the best matching database images for each new query.

First, Fig. 13(a) plots the mean precision at rank 1 (PA1) measurements averaged across the 32 different query videos contained in the SSMAR Dataset. It can be seen that SCP and SFP perform very similarly to independent coding, showing that SCP and SFP can effectively adapt to changes in the videos and preserve a high retrieval accuracy. The fourth method, SFP+LS, attains a slightly higher retrieval accuracy, because geometric verification is performed after comparing REVV signatures, which further removes some false database candidates. Fig. 13(a) plots the empirical PA1 values next to the predicted PA1 values derived from our statistical model in Section II-B. The model closely predicts the retrieval accuracy of all four methods being evaluated.

Then, Fig. 13(b) plots the uplink bitrate (in kbps) for the same four methods. Independent coding requires around 120 kbps. With interframe coding, we can substantially reduce the uplink bitrate required to communicate the continuous stream of REVV signatures: by a factor of $14\times$ for SCP, $24\times$ for SFP, and $88\times$ for SFP+LS. The bitrate reduction achieved by the best performing method, SFP+LS, is a combined result of two key factors: (i) the temporal redundancy between REVV signatures of viewfinder frames is carefully exploited, and (ii) when the local database search is sufficient, a query does not need to be sent to the remote server. With SFP+LS, we can query a remote server with a continuous stream of REVV signatures at an average uplink bitrate of less than 2 kbps.

The downlink bitrate for the independent coding, SCP, and SFP methods is close to 0 kbps, because the server replies with just the labels and metadata for the top ranked database candidates. The downlink bitrate for SFP+LS is 14 kbps, where the server replies with the labels, metadata, REVV signature, and a small set of compressed local features for the top ranked database candidates. The low ratio of uplink bitrate to downlink rate for SFP+LS is well suited to typical wireless networks that have much lower uplink speeds compared to downlink speeds.

V. Conclusion

On-device image-based retrieval can provide fast recognition responses regardless of conditions outside the mobile device. In contrast, on-server image-based retrieval can search a much larger database of images. This paper shows how a hybrid MVS system can combine the advantages of on-device and on-server retrieval. The hybrid system first performs a fast local query on the mobile. If the local search result is satisfactory, the query finishes locally. Otherwise, the hybrid system expands the query into the cloud by sending a compact stream of image signatures to a remote server.

First, to achieve robust recognition when searching a large database of images, we developed an improved version of the REVV global signature. The performance of the new REVV signature was optimized through a statistical analysis of the residual vector's most important retrieval characteristics. Our analysis showed that for each codebook size, there is an optimal local feature count which maximizes the REVV signature's retrieval accuracy. We then showed that the improved REVV signature has superior performance compared to several other global signatures. Furthermore, we developed two extensions that exploit color information within images and relationships between similar database images to improve retrieval accuracy.

In order to generate a compact stream of REVV signatures from a sequence of viewfinder frames that can be efficiently transmitted to a remote server, several effective interframe coding methods for REVV signatures are developed. Our best interframe coding method can reduce the uplink bitrate by almost two orders of magnitude compared to independent coding of REVV signatures. These large bitrate reductions are possible because our REVV signatures are carefully designed to be temporally coherent and the on-device REVV database is effectively leveraged to avoid transmissions whenever possible. The low bitrate of an interframe coded REVV stream permits querying the remote server even over networks with low transfer rates.

A robust, low-latency hybrid MVS system with a large cached database is practical today. Our system requires around 50 MB of RAM to store a database of 100 K images on the mobile device. When local search is insufficient, obtaining accurate query results from a remote server hosting a larger database of 1 M images requires less than 2 kbps on average to achieve highly accurate retrieval results. When the server replies, the local on-device database is opportunistically updated, which enables the local database to be populated with images related to the user's current and evolving interests.

Appendix A
Generalized Binomial Distribution

For a generalized binomial distribution (GDB) [39], let S_N and F_N denote success and failure, respectively, on the Nth Bernoulli trial in the sequence. Suppose $P(S_1) = p$ and $P(F_1) = 1 - p$. The probabilities for S_N and F_N depend on the number of successes h in the previous $N - 1$ Bernoulli trials as follows:

$$P(S_N|h) = (1 - \pi_m)\, p + \pi_m \frac{h}{N-1} \tag{30}$$

$$P(F_N|h) = (1 - \pi_m)\,(1 - p) + \pi_m \frac{N-1-h}{N-1}. \tag{31}$$

Hence, the probability for S_N is a mixture of the history-independent a priori probability of success p and the history-dependent empirical probability of success $h/(N-1)$, with a mixing weight π_m. As $\pi_m \to 0$, then the GDB becomes an ordinary binomial distribution. As $\pi_m \to 1$, the dependence between the different Bernoulli trials in the sequence becomes stronger. The probability of h successes in N trials is then defined recursively:

$$\begin{aligned} p_{\mathrm{GBD}}(h|N) = {} & P(S_N|h-1)\, p_{\mathrm{GBD}}(h-1|N-1) \\ & + P(F_N|h)\, p_{\mathrm{GBD}}(h|N-1). \end{aligned} \tag{32}$$

The initial conditions for the GBD are $p_{\mathrm{GBD}}(0|1) = 1 - p$ and $p_{\mathrm{GBD}}(1|1) = p$.

Acknowledgment

The authors thank M. Makar and A. Araujo for their software implementation of the temporally coherent keypoint detector. They also thank the reviewers for their insightful comments and suggestions, which helped them to improve the paper during revisions.

References

[1] G. Takacs *et al.*, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *Proc. ACM Multimedia Inf. Retrieval*, Oct. 2008, pp. 427–434.

[2] S. Tsai *et al.*, "Mobile product recognition," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1587–1590.

[3] P. Fockler, T. Zeidler, B. Brombach, E. Bruns, and O. Bimber, "PhoneGuide: Museum guidance supported by on-device object recognition on mobile phones," in *Proc. Int. Conf. Mobile Ubiquitous Multimedia*, Dec. 2005, pp. 3–10.

[4] W. Liu, T. Mei, Y. Zhang, J. Li, and S. Li, "Listen, look, and gotcha: Instant video search with mobile phones by layered audio-video indexing," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2013, pp. 887–896.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[7] V. Chandrasekhar *et al.*, "CHoG: Compressed histogram of gradients–a low bitrate feature descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 2504–2511.

[8] G. Takacs *et al.*, "Unified real-time tracking and recognition with rotation-invariant fast features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 934–941.

[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

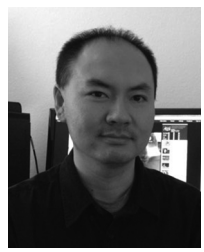
[10] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 510–517.

[11] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.

[12] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, pp. 2161–2168.

[13] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3384–3391.

[14] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3304–3311.
[15] D. Chen *et al.*, "Residual enhanced visual vectors for on-device image matching," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2011, pp. 850–854.
[16] C. Yeo, P. Ahammad, and K. Ramchandran, "Rate-efficient visual correspondences using random projections," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 217–220.
[17] V. Chandrasekhar *et al.*, "Transform coding of image feature descriptors," in *Proc. SPIE Conf. Vis. Commun. Image Process.*, Jan. 2009, pp. 1–10.
[18] V. Chandrasekhar *et al.*, "Compressed histogram of gradients: A low bitrate descriptor," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 384–399, 2012.
[19] S. Tsai *et al.*, "Location coding for mobile image retrieval," in *Proc. Int. Conf. Mobile Multimedia Commun.*, Sep. 2009, pp. 1–7.
[20] J. Chao and E. Steinbach, "Preserving SIFT features in JPEG-encoded images," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 301–304.
[21] J. He *et al.*, "Mobile product search with bag of hash bits and boundary reranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3005–3012.
[22] M. Makar, S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *Int. J. Semantic Comput.*, vol. 7, no. 1, pp. 5–24, Mar. 2013.
[23] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May 2001.
[24] D. Chen *et al.*, "Tree histogram coding for mobile image matching," in *Proc. IEEE Data Compression Conf.*, Mar. 2009, pp. 143–152.
[25] H. Jegou, M. Douze, and C. Schmid, "Packing bag-of-features," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2357–2364.
[26] D. Chen *et al.*, "Inverted index compression for scalable image matching," in *Proc. IEEE Data Compression Conf.*, Mar. 2010, p. 13.
[27] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, pp. 290–314, Feb. 2012.
[28] *Evaluation Framework for Compact Descriptors for Visual Search*, ISO/IEC JTCI/SC29/WG11 N12202, Jul. 2011.
[29] *Requirements for Compact Descriptor for Video Analysis*, ISO/IEC JTCI/SC29/WG11 N14095, Oct. 2013.
[30] D. Chen *et al.*, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Process.*, vol. 93, no. 8, pp. 2316–2327, Aug. 2013.
[31] D. Chen and B. Girod, "Memory-efficient image databases for mobile visual search," *IEEE Multimedia Mag.*, vol. 21, no. 1, pp. 14–23, Jan.–Mar. 2014.
[32] R. Arandjelovic, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2911–2918.
[33] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 1169–1176.
[34] J. Delhumeau, P.-H. Gosselin, H. Jegou, and P. Perez, "Revisiting the VLAD image representation," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2013, pp. 653–656.
[35] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Neural Inf. Process. Syst.*, Dec. 2008, pp. 1753–1760.
[36] A. Joly and O. Buisson, "Random maximum margin hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 873–880.
[37] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 817–824.
[38] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2012, pp. 169–178.
[39] Z. Drezner and N. Farnum, "A generalized binomial distribution," *Commun. Statist.*, vol. 22, no. 11, pp. 3051–3063, 1993.
[40] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–7.
[41] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[42] *Compact Descriptors for Visual Search: Performance Improvements of the Scalable Compressed Fisher Vector*, ISO/IEC JTCI/SC29/WG11 MPEG2013/M28061, Jan. 2013.
[43] J. Lin, L.-Y. Duan, T. Huang, and W. Gao, "Robust Fisher codes for large scale image retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1513–1517.
[44] J. Lin *et al.*, "Rate-adaptive compact Fisher codes for mobile visual search," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 195–198, Feb. 2014.
[45] L.-Y. Duan, J. Lin, Z. Wang, T. Huang, and W. Gao, "Weighted component hashing of binary aggregated descriptors for fast visual search," *IEEE Trans. Multimedia*, to be published.
[46] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
[47] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[48] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 889–896.
[49] R. Arandjelović and A. Zisserman, "Multiple queries for large scale specific object retrieval," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2012, pp. 1–8.
[50] "Visual query expansion with or without geometry: Refining local descriptors by feature aggregation," *Pattern Recog.*, vol. 47, no. 10, pp. 3466–3476, Oct. 2014.
[51] D. Chen, M. Makar, A. Araujo, and B. Girod, "Interframe coding of global image signatures for mobile augmented reality," in *Proc. IEEE Data Compression Conf.*, Mar. 2014, pp. 33–42.
[52] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.



David M. Chen (M'10) received the B.S. degree in 2006, the M.S. degree in 2008, and the Ph.D. degree in 2014, all in electrical engineering, from Stanford University, Stanford, CA, USA.

He was previously an Engineering Research Associate with the Department of Electrical Engineering, Stanford University. He also served as a Fellow with the Brown Institute for Media Innovation, Stanford University. His research interests include image and video retrieval for mobile visual search applications, computer vision, signal processing, and machine learning.

Dr. Chen received the Centennial TA Award in 2012 and the Outstanding TA Award in 2010. He received the Capocelli Prize for Best Student Paper from the Data Compression Conference in 2014.

Bernd Girod (S'80–M'80–SM'97–F'98) received the M.S. degree from the Georgia Institute of Technology, Atlanta, GA, USA, and the D.Eng. degree from the University of Hannover, Hannover, Germany.

He is the Robert L. and Audrey S. Hancock Professor of Electrical Engineering with Stanford University, Stanford, CA, USA. He is also Senior Associate Dean for Online Learning and Professional Development with the School of Engineering, Stanford University. Until 1999, he was a Professor with the Electrical Engineering Department, University of Erlangen-Nuremberg, Erlangen, Germany. He has authored or coauthored over 600 conference and journal papers and six books. His research interests include image, video, and multimedia systems.

Prof. Girod is a EURASIP Fellow, a member of the German National Academy of Sciences (Leopoldina), and a member of the National Academy of Engineering. He received 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, the EURASIP Technical Achievement Award in 2004, and the Technical Achievement Award of the IEEE Signal Processing Society in 2011.