Interframe Coding of Global Image Signatures for Mobile Augmented Reality

David M. Chen, Mina Makar, Andre F. Araujo, Bernd Girod
Department of Electrical Engineering, Stanford University, Stanford, CA 94305

Abstract

For mobile augmented reality, an image captured by a mobile device's camera is often compared against a database hosted on a remote server to recognize objects in the image. It is critically important that the amount of data transmitted over the network is as small as possible to reduce the system latency. A low bitrate global signature for still images has been previously shown to achieve high-accuracy image retrieval. In this paper, we develop new methods for interframe coding of a continuous stream of global signatures that can reduce the bitrate by nearly two orders of magnitude compared to independent coding of these global signatures, while achieving the same or better image retrieval accuracy. The global signatures are constructed in an embedded data structure that offers rate scalability. The usage of these new coding methods and the embedded data structure allows the streaming of high-quality global signatures at a bitrate that is less than 2 kbps. Furthermore, a statistical analysis of the retrieval and coding performance is performed to understand the tradeoff between bitrate and image retrieval accuracy and explain why interframe coding of global signatures substantially outperforms independent coding.

1 Introduction

Mobile augmented reality (MAR) systems process a stream of viewfinder frames captured by a mobile device's camera to recognize, track, and augment objects that appear in these frames [1–4]. Many of these systems compare the acquired frames to a database of labeled images to perform object recognition. For robust image description, local image features such as SIFT [5], SURF [6], CHoG [7], or RIFF [4] are extracted and can be reliably matched between images of the same objects, even when there are severe photometric and geometric variations.

To provide fast comparisons against a large database, a global signature can be constructed to summarize the most important statistics of the local features. Many popular global signatures can be categorized into two groups: feature histograms and feature residuals. Histogram-based methods [8, 9] generally use a large codebook of visual words to form a histogram that records how often each visual word is visited by an image's local features. In contrast, residual-based methods [10, 11] use a relatively small codebook of visual words to form a hash from the quantization errors and attain comparable retrieval performance as histogram-based methods.

If the database is hosted on a remote server, then the query signatures extracted from the frames on the mobile device need to be transmitted over a network. It is critically important that the amount of data sent over the network is as small as possible to minimize system latency. One emerging MPEG standard, Compact Descriptors for Visual Search (CDVS) [12], aims to design a low bitrate signature for general

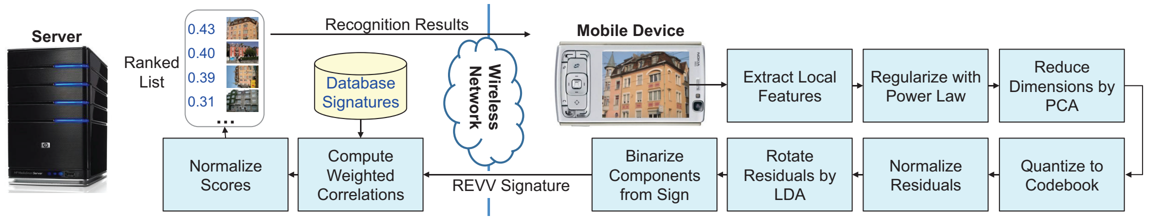

Figure 1: System for extracting a Residual Enhanced Visual Vector (REVV) signatures from a still image on a mobile device and matching against a database on a server.

image matching applications. The CDVS participants have shown that transmitting a combination of residual-based global signatures and local features attains the best image matching performance for a range of different image categories [13, 14] .

The problem of compressing local features for a motion video has only recently received attention. A temporally coherent keypoint detector and interframe feature coding techniques are presented in [15,16], which enables a continuous stream of local features to be sent from the mobile device to the server at a low bitrate. Since the features are sent for each frame, the recognition system can instantly adapt to sudden changes in the scene for ultra low-latency MAR applications.

The methods in [15,16] focused only on interframe coding of local features. In this paper, we develop efficient methods for interframe coding of residual-based global signatures. We design an embedded data structure that offers rate scalability and direct comparison of signatures encoded at different bitrates. Compared to independent coding of global signatures extracted from video frames, we can effectively reduce the bitrate of sending a continuous stream of global signatures to less than 2 kbps, with up to 88× bitrate reduction, while achieving the same or better image retrieval accuracy. A statistical analysis of the global signatures is performed to help understand how retrieval accuracy varies with bitrate and to help optimize the system design.

In Sec. 2, we review briefly the major processing stages and system structure for generation of a compact global signature. Then, in Sec. 3, we develop a new framework and new methods for interframe coding of a continuous stream of global signatures. In Sec. 4, we build a statistical model for residual-based global signatures and analyze the performance of independent and interframe coding methods. Experimental results on a large dataset of videos in Sec. 5 show our interframe coding methods yield substantial bitrate savings compared to independent coding of global signatures.

2 Compact Global Signatures for Image Retrieval

The image retrieval system that uses compact Residual Enhanced Visual Vector (REVV) signatures to achieve high recognition accuracy [11] is shown in Fig. 1. First, on the mobile device, scale- and rotation-invariant local features are extracted from a captured image. A power law is then applied to the feature descriptors to reduce the influence of peaky components. The dimensionality of the descriptors is subsequently reduced using Principal Component Analysis (PCA). Then, the descriptors are quantized using a small codebook, typically containing 64 - 256 codewords. The

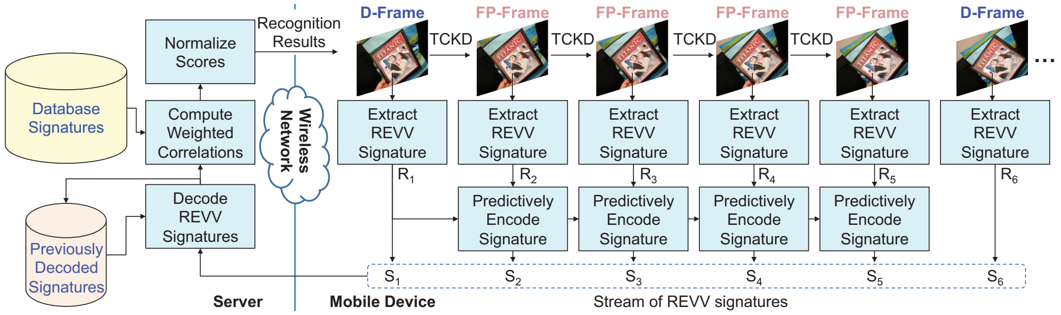

Figure 2: System for extracting interframe coded REVV signatures from video frames on a mobile device and matching against a database on a server. The block "Extract REVV Signature" outputs a binary REVV signature using the methods depicted in Fig. 1.

quantization residual vectors are aggregated for each codeword, and each aggregated residual vector is normalized to have unit magnitude. After a rotation by Linear Discriminant Analysis (LDA) to optimally separate matching and non-matching residual vectors, each residual component is binarized depending on its sign. The binarized signature is transmitted from the mobile device to a remote server, where a database of REVV signatures is stored. Weighted correlations between the query signature and the database signatures are efficiently computed using bitwise XOR and POPCNT instructions and lookup tables. Finally, the scores are normalized to take into account that some database images visit more codewords than other database images. Recognition results derived from the ranked list of database images are sent back to the mobile device for augmentation.

The system in Fig. 1 is designed for compressing global signatures extracted from still images. If global signatures are extracted from consecutive frames of a motion video, the temporal correlation between the signatures should be carefully exploited to achieve much higher bitrate savings. In the next section, we develop several efficient methods for interframe coding of a continuous stream of REVV signatures.

3 Interframe Compression of Global Signatures

3.1 Temporally Coherent Keypoint Detection

To exploit the correlation between neighboring REVV signatures, we use a temporally coherent keypoint detector (TCKD) [15,16] which divides frames into two categories: Detection Frames (D-Frames) and Forward Propagation Frames (FP-Frames). For each D-Frame, TCKD detects SIFT keypoints [5]. Then, each SIFT keypoint is propagated into the subsequent FP-Frame by searching across a set of similarity transforms to minimize the sum of absolute differences (SAD) in the canonical patch for the keypoint. This propagation continues until the next D-Frame appears. In the interframe coding framework depicted in Fig. 2, a D-Frame appears once every 5 frames as an example, but in practice a D-Frame is usually inserted once every 30 frames, or once every second assuming a frame rate of 30 frames/second.

3.2 Predictive Coding of Global Signatures

To construct temporally coherent global signatures, SIFT descriptors [5] are extracted from the TCKD keypoints, and then REVV signatures are generated from these SIFT descriptors. Each D-Frame's REVV signature is independently coded. In contrast, each FP-Frame's REVV signature is predictively coded using the previously transmitted REVV signature as reference. We design and implement three different predictive coding methods to adapt to various types of scene content and achieve the best coding efficiency for the current video stream.

3.2.1 Selective Codeword Propagation (SCP)

For a codebook of k visual words, let the original REVV signature of the i^{th} frame be denoted as $\mathbf{R}_i = \{(U_{i,1}, R_{i,1}), \cdots, (U_{i,k}, R_{i,k})\}$. Here, $U_{i,j} \in \{0,1\}$ is a binary variable indicating if the j^{th} codeword is visited by the i^{th} frame and $R_{i,j}$ 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 $\mathbf{S}_i = \{(V_{i,1}, S_{i,1}), \cdots, (V_{i,k}, S_{i,k})\}$. The SCP method assigns $V_{i,j} = U_{i,j}$ AND $V_{i-1,j}$ for $1 \leq j \leq k$. If $V_{i,j} = 1$, then the SCP method further assigns $S_{i,j} = S_{i-1,j}$, which propagates the previously sent binary residual vector for the j^{th} codeword. We do not encode the difference between the residual vectors $R_{i,j}$ and $S_{i-1,j}$, because these small differences are due to small temporal fluctuations in the feature descriptors and do not noticeably affect image retrieval results. Note that only $V_{i,j}$ needs to be sent, because $S_{i,j} = S_{i-1,j}$ has been previously received at the server. Additionally, $V_{i,j}$ needs to be sent only if $V_{i-1,j} = 1$, because $V_{i,j} = 0$ when $V_{i-1,j} = 0$.

3.2.2 Selective Frame Propagation (SFP)

When the scene content changes gradually, two consecutive frames visit mostly the same codewords and have similar residual vectors at these codewords. Taking the idea behind SCP one step further, SFP propagates all of the residual vectors between two frames if these two frames' REVV signatures have a high degree of similarity. As before, let $\mathbf{R}_i = \{(U_{i,1}, R_{i,1}), \cdots, (U_{i,k}, R_{i,k})\}$ denote the original REVV signature and $\mathbf{S}_i = \{(V_{i,1}, S_{i,1}), \cdots, (V_{i,k}, S_{i,k})\}$ denote the predictively coded REVV signature for the i^{th} frame. We define the interframe codeword similarity rate between the i^{th} and $(i-1)^{\text{st}}$ frames as $r_k(i, i-1) = \left(\sum_{j=1}^{k} \text{AND}\left(U_{i,j}, V_{i-1,j}\right)\right) / \left(\sum_{j=1}^{k} U_{i,j}\right)$. If $r_k(i, i-1)$ exceeds a high threshold t_{r_k}, then SFP assigns $V_{i,j} = V_{i-1,j}$ and $S_{i,j} = S_{i-1,j}$ for $1 \leq j \leq k$, and only a single bit is sent to indicate that the previous frame's residual vectors should be entirely propagated. Otherwise, SFP uses SCP, and only a single bit is sent to the server to indicate a temporary switch to the SCP mode, followed by the bits generated by SCP.

3.2.3 Selective Frame Propagation + Local Search (SFP + LS)

Since REVV signatures are compact, they can be easily stored in a database on a mobile device with a small memory capacity. The SFP + LS scheme exploits

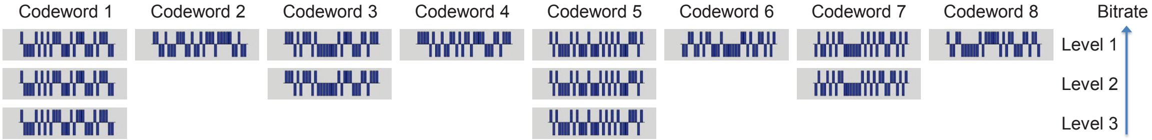

Figure 3: Three separate embedding levels for the REVV signature, where residual vectors are progressively discarded as the bitrate is reduced.

this valuable property of REVV to incrementally update a local database. When the server replies with the recognition results for a query, the server also sends the REVV signature and a small set of compressed features for the top-ranked database image, if this data has not been previously sent back to the mobile device. The mobile device updates its local database with the received REVV signature. On a subsequent query, the local REVV database is first searched and top candidates within a shortlist are then verified for geometric consistency with the Random Sample Consensus (RANSAC) algorithm. If a local database candidate attains a high number of RANSAC inliers, then the search terminates locally on the mobile device and the best local database candidate is retrieved; otherwise, a REVV signature is sent by SFP encoding to expand the query on the server. By using a compact local database, SFP + LS effectively reduces the amount of query data sent through the uplink by occasionally sending feedback data through the downlink, which is well suited to most wireless networks where uplink speeds are much lower than downlink speeds.

3.3 Embedded Global Signatures for Rate Scalability

To enable rate scalability, we develop an embedded data structure for the REVV signature. Fig. 3 shows three separate embedding levels of the same REVV signature, corresponding to three different target bitrates. The embedding at Level 1 corresponds to the highest-quality signature and requires the largest bitrate. At Level 2, the residual vector for every other codeword is discarded to reduce the bitrate by approximately 2×. Similarly, at Level 3, only the residual vector for every 4th codeword is retained to reduce the bitrate by approximately 4×. The advantages of this rate adjustment scheme are that (1) signatures generated at 2 different levels can be directly compared to each other, (2) a lower bitrate embedding can be derived from a higher bitrate embedding, and (3) the server only needs to restore the embedding at Level 1. This rate scalability provided by the embedded data structure allows for a convenient dynamic selection of the proper bitrate in response to various changing system conditions, such as network transmission quality and scene contents.

4 Analysis of Retrieval and Coding Performance

4.1 Modeling Image Retrieval Accuracy

First, we analyze the correlations between binary residual vectors. The correlation C_{nm} for a non-matching image pair and correlation C_m for a matching image pair are computed as $C_q = \sum_{j=1}^{N_{\text{visit},q}} C_{q,j}$ for $q \in \{nm, m\} = \{\text{non-matching}, \text{matching}\}$

where $N_{\text{visit},q}$ is the number of codewords visited in common by the pair of images and $C_{q,j}$ is the correlation at a particular codeword. Equivalently, if we know the Hamming distance $H_{q,j}$ between two binary residual vectors at a codeword, we can compute the codeword-level correlation as $C_{q,j} = d_{\text{PCA}} - 2H_{q,j}$, where d_{PCA} is the descriptor dimensionality after PCA. The image-level correlation can be rewritten as $C_q = N_{\text{visit},q}\, d_{\text{PCA}} - 2H_q$, where $H_q = \sum_{j=1}^{N_{\text{visit},q}} H_{q,j}$.

The number of codewords available is $N_{k,l} = \lfloor k/2^{l-1} \rfloor$, where $l = 1, 2, 3$ indicates the embedding level described in Sec. 3.3. We model $N_{\text{visit},q} \sim \text{Binomial}\,(N_{k,l},\, p_q)$ for $q \in \{nm, m\}$. The parameters p_{nm} and p_m are the probabilities that a codeword is visited by both images for the non-matching and matching cases, respectively.

If $N_{\text{visit},nm} = n$, then $n\, d_{\text{PCA}}$ bits are compared between two non-matching images. Let H_{nm} denote the sum of the $n\, d_{\text{PCA}}$ bits. For a non-matching image pair, the bits are independent, so H_{nm} can be modeled conditionally as a binomial random variable: $H_{nm} | \{N_{\text{visit},nm} = n\} \sim \text{Binomial}\,(n\, d_{\text{PCA}}, \theta_{nm})$. Here, $\theta_{nm} \approx 0.5$ is the probability that two non-matching images differ in any single bit of the binary residual vector.

Similarly, if $N_{\text{visit},m} = n$, then $n\, d_{\text{PCA}}$ bits are compared between two matching images. Let H_m denote the sum of the $n\, d_{\text{PCA}}$ bits. For matching images, the bits are strongly dependent, so we employ a Generalized Binomial Distribution (GDB) [17] to capture this dependence. For a GBD, the probabilities of success S_N and failure F_N on the N^{th} Bernoulli trial depend on the number of successes $h - 1$ in the previous $N - 1$ trials:

$$\begin{aligned} p\,(S_N | h-1, N-1) &= (1-\alpha_m)\theta_m + \alpha_m\,(h-1)\,/\,(N-1) \qquad (1) \\ p\,(F_N | h, N-1) &= (1-\alpha_m)(1-\theta_m) + \alpha_m\,(1 - h/\,(N-1)) \qquad (2) \end{aligned}$$

where θ_m is the probability of success on the first trial in the sequence and $\alpha_m \in [0, 1]$ is a parameter controlling the amount of dependence between the separate trials. The probability of h successes in N trials is then defined recursively for $h = 0, \cdots, N$:

$$\begin{aligned} p_{\text{GBD}}(h|N) &= p(S_N|h-1, N-1)\, p_{\text{GBD}}(h-1|N-1) \;+ \\ &\quad\; p(F_n|h, N-1)\, p_{\text{GBD}}(h|N-1). \qquad (3) \end{aligned}$$

The conditional distribution for H_m is then $p_{H_m|N_{\text{visit},m}}(h|n) = p_{\text{GBD}}(h\,|\,n\, d_{\text{PCA}})$.

Since $C_q = n\, d_{\text{PCA}} - 2H_q$ when $N_{\text{visit},q} = n$, the conditional distribution for C_q is $p_{C_q|N_{\text{visit},\,q}}(c|n) = p_{H_q|N_{\text{visit},q}}\,(0.5\,(n\, d_{\text{PCA}} - c)\,|n)$. Then, the distribution for C_q is $p_{C_q}(c) = \sum_{n=0}^{N_{k,l}} p_{N_{\text{visit},\,q}}(n)\; p_{C_q|N_{\text{visit},\,q}}(c|n)$.

In Fig. 4, we plot the distributions for C_{nm} and C_m for the three different embedding levels. The empirical distributions are extracted from video frames in the Stanford Streaming MAR Dataset [16], which will be described in greater detail in Sec. 5. Each non-matching distribution is centered around a zero correlation score and is precisely described by a mixture of binomials. In contrast, each matching distribution has a long tail skewed toward high correlation scores, caused by the dependence between the different bits, and requires a mixture of GBDs for accurate modeling. From Fig. 4, we also see that as we move from Level 3 to Level 2 to Level 1 and thus increase the bitrate, the area of overlap between the non-matching and matching distributions decreases and image retrieval accuracy increases.

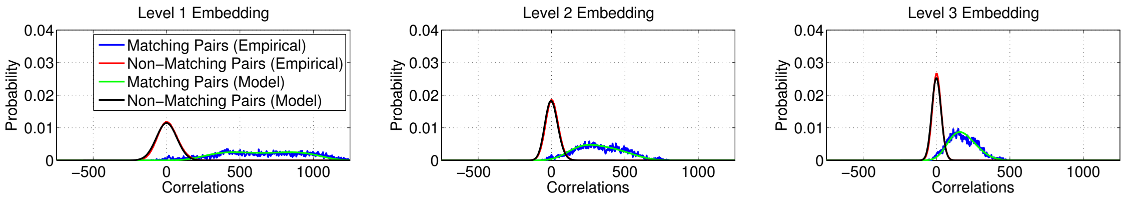

Figure 4: Distributions of correlations between binary residual vectors for matching and non-matching image pairs.

In querying a large database of images with REVV signatures, a ranked list of correlation scores is generated. In this list, $N_{db,m}$ correlation scores $C_m^{(1)}, \cdots, C_m^{(N_{db,m})}$ and $N_{db,nm}$ correlation scores $C_{nm}^{(1)}, \cdots, C_{nm}^{(N_{db,nm})}$ belong to the matching and non-matching database images, respectively. Generally, $N_{db,m} \ll N_{db,nm}$. We assume the scores $C_m^{(1)}, \cdots, C_m^{(N_{db,m})}$ and $C_{nm}^{(1)}, \cdots, C_{nm}^{(N_{db,nm})}$ are distributed i.i.d. according to the matching and non-matching correlation models derived previously.

Mean precision at rank 1 (PA1) and mean average precision (MAP) are commonly used measures of retrieval accuracy. As we show in Sec. 5, the PA1 and MAP values are close to each other, so we focus on predicting the PA1 value. Let $C_m^{\max} = \max\left\{C_m^{(1)}, \cdots, C_m^{(N_{db,m})}\right\}$ and $C_{nm}^{\max} = \max\left\{C_{nm}^{(1)}, \cdots, C_{nm}^{(N_{db,nm})}\right\}$. Then, $\text{PA1} = P\left(C_m^{\max} \geq C_{nm}^{\max}\right)$. The cumulative distribution function (CDF) for $C_m^{\max}$ is

$$F_{C_m^{\max}}(c) = P\left(C_m^{\max} \leq c\right) = P\left(C_m \leq c\right)^{N_{db,m}} = F_{C_m}(c)^{N_{db,m}} \tag{4}$$

using the assumption that $C_m^{(1)}, \cdots, C_m^{(N_{db,m})}$ are i.i.d. random variables. Similarly, the CDF for $C_{nm}^{\max}$ is $F_{C_{nm}^{\max}}(c) = F_{C_{nm}}(c)^{N_{db,nm}}$. We can compute the probability mass functions (PMFs) $p_{C_m^{\max}}(c)$ and $p_{C_{nm}^{\max}}(c)$ by taking discrete differences of $F_{C_m^{\max}}(c)$ and $F_{C_{nm}^{\max}}(c)$, respectively. Then, we have

$$\text{PA1}_{\text{rem}} = P\left(C_m^{\max} \geq C_{nm}^{\max}\right) = \sum_c p_{C_m^{\max}}(c) P\left(c \geq C_{nm}^{\max}\right) = \sum_c p_{C_m^{\max}}(c) F_{C_{nm}^{\max}}(c). \tag{5}$$

Eq. 5 predicts the PA1 values for the independent, SCP, and SFP coding methods which all search a database on a remote server.

To predict the PA1 value for the SFP + LS method, we define P_{LS} as the probability that the local database search succeeds for a query. If the local database search is insufficient, a REVV signature is transmitted to the server and the remote query will succeed with probability given in Eq. 5. Hence, the overall probability of success for the SFP + LS method is given by $\text{PA1}_{\text{loc}} = P_{LS} + \left(1 - P_{LS}\right) \text{PA1}_{\text{rem}}$. In Sec. 5, we will observe that PA1 values predicted by our model match well with PA1 values obtained from actual retrieval experiments.

4.2 Modeling Uplink Coding Bitrate

Now, we model the uplink bitrate for the different coding methods. The uplink bitrate for independent coding of REVV signatures is predicted to be $R_{\text{Indep}} =$

$N_{\text{Frames}}\, N_{k,l}\, (1 + \rho_{\text{visit}}\, d_{\text{PCA}})$ bits/second, where as before $N_{k,l}$ is the number of codewords available for the l^{th} embedding level and d_{PCA} is the descriptor dimensionality after PCA. Additionally, $\rho_{\text{visit}} \in [0, 1]$ is the average fraction of codewords visited by an image's feature descriptors and N_{Frames} is the number of frames per second.

The uplink bitrate for predictive coding with SCP is predicted to be

$$R_{\text{SCP}} = \underbrace{N_{\text{D-Frames}}\, N_{k,l}\, (1 + \rho_{\text{visit}}\, d_{\text{PCA}})}_{\text{bitrate for D-Frames}} + \underbrace{N_{\text{FP-Frames}}\, N_{k,l}\, \rho_{\text{visit}}}_{\text{bitrate for FP-Frames}}. \tag{6}$$

Here, $N_{\text{D-Frames}}$ and $N_{\text{FP-Frames}}$ are the number of D-Frames and FP-Frames, respectively, per second. Note that $N_{\text{Frames}} = N_{\text{D-Frames}} + N_{\text{FP-Frames}}$. Similarly, the uplink bitrate for SFP is predicted to be

$$R_{\text{SFP}} = \underbrace{N_{\text{D-Frames}}\, N_{k,l}\, (1 + \rho_{\text{visit}}\, d_{\text{PCA}})}_{\text{bitrate for D-Frames}} + \underbrace{N_{\text{FP-Frames}}\, (1 + P\left(r_k < t_{r_k}\right)\, N_{k,l}\, \rho_{\text{visit}})}_{\text{bitrate for FP-Frames}}, \tag{7}$$

where r_k is the interframe codeword similarity rate and t_{r_k} is the threshold deciding when to switch between the SFP and SCP methods, defined in Sec. 3.2.2. Finally, the uplink bitrate for SFP + LS is predicted to be $R_{\text{SFP+LS}} = (1 - P_{LS})\, R_{\text{SFP}}$, where as in the last section P_{LS} is the probability that the local database search succeeds. As we show in the next section, R_{SCP}, R_{SFP}, and $R_{\text{SFP+LS}}$ are much smaller than R_{Indep} when the number of D-Frames per second is small, e.g., $N_{\text{D-Frames}} = 1$.

5 Experimental Results

We evaluate the performance of the independent coding and predictive coding methods on the Stanford Streaming MAR Dataset [16]. This dataset contains 32 VGA-resolution query videos recorded with a camera-phone showing books, DVDs, CDs, and other product packages. Sometimes, several different objects appear in the same video sequence. Each query frame is matched against a database of 1M images to evaluate retrieval accuracy. For every query D-Frame, we extract 250 SIFT features using a feature selector optimized for matching accuracy [18]. For REVV, we use a codebook of $k = 190$ codewords trained on an independent dataset and $d_{\text{PCA}} = 32$ PCA eigenvectors. For interframe coding, we use 1 D-Frame for every 30 frames and an interframe codeword similarity threshold of $t_{r_k} = 0.9$.

Fig. 5 plots the mean precision at rank 1 (PA1) and the mean average precision (MAP) versus the uplink bitrate for independent coding, SCP, SFP, and SFP + LS. The uplink bitrate is varied by changing the embedding level. Both SCP and SFP attain similar PA1 and MAP values as independent coding, but they substantially reduce the uplink bitrate by 14× and 24×, respectively. The best performing method is SFP + LS, which reduces the uplink bitrate by 88× to less than 2 kbps and attains higher PA1 and MAP values because the geometric verification for the local database search improves retrieval accuracy. SFP + LS uses a small downlink bitrate of 14 kbps to update the local on-device database.

In SFP + LS, the best interframe coding method is automatically chosen depending on the current video contents. When local search suffices, nothing is transmitted

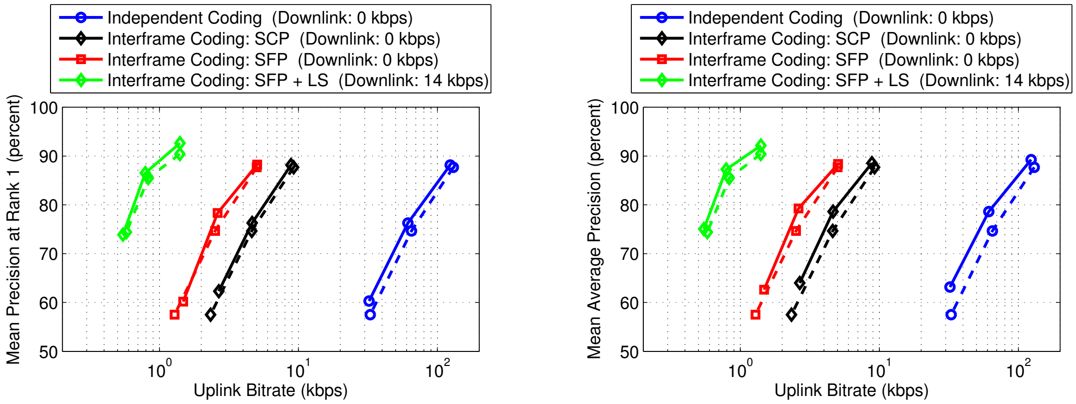

Figure 5: Image retrieval accuracy measured on the Stanford Streaming MAR Dataset. Solid and dashed lines correspond to empirical and model data, respectively.

to the server. Otherwise, if $r_k > t_{r_k}$, the SFP method is used to propagate residuals. Finally, if local search is insufficient and $r_k \leq t_{r_k}$, the SCP method activates to accommodate more rapid changes in the scene contents.

The solid lines in Fig. 5 represent empirical data, while the dashed lines represent model data. Our analysis from Sec. 4 accurately predicts how retrieval precision varies with the uplink bitrate for each method. As we move to a higher-quality embedding level, the matching and non-matching score distributions have smaller area of overlap, leading to higher retrieval precision but also a larger bitrate. The analysis captures the relative advantages between independent coding, SCP, SFP, and SFP + LS.

6 Conclusions

In this paper, we have developed three new methods for interframe coding of global signatures extracted from a continuous sequence of frames acquired on a mobile device. The usage of these new methods enables sending a low bitrate stream of global signatures from a mobile device to a server, which is important for accurate large-scale image retrieval in mobile augmented reality applications. By exploiting the correlation between global signatures of neighboring frames through selective propagation of codeword residuals, we can substantially reduce the uplink bitrate by as much as 88× compared to independent coding of global signatures while achieving the same or better image retrieval accuracy. Less than 2 kbps is required to continuously stream high-quality global signatures from the mobile device to a server, which is practical for even slow wireless links. The global signatures are encoded in an embedded data structure that offers rate scalability. We have also performed a detailed statistical analysis of the global signature's retrieval and coding performance. The analysis reveals how adjusting the rate in the embedded data structure changes the area of overlap between matching and non-matching score distributions and consequently affects the image retrieval accuracy. Intuitive expressions for the bitrates of independent and interframe coding methods are derived to explain why interframe coding obtains substantial bitrate savings compared to independent coding.

References

[1] Kooaba, "Kooaba Augmented Reality," http://blog.kooaba.com/2010/04/mobile-augmented-reality-the-next-level-of-sophistication.

[2] Amazon, "Amazon Flow," https://itunes.apple.com/us/app/flow-powered-by-amazon/id474664425.

[3] D. Chen, S. Tsai, R. Vedantham, R. Grzeszczuk, and B. Girod, "Streaming mobile augmented reality on mobile phones," in *International Symposium on Mixed and Augmented Reality*, October 2009.

[4] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod, "Rotation-invariant fast features for large-scale recognition and real-time tracking," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 334–344, April 2013.

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

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

[7] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Compressed histogram of gradients: a low bitrate descriptor," *International Journal of Computer Vision*, vol. 96, no. 3, pp. 384–399, May 2012.

[8] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, October 2003.

[9] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2006.

[10] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[11] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Processing*, vol. 93, no. 8, pp. 2316–2327, August 2013.

[12] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. Reznik, "Mobile visual search: architectures, technologies, and the emerging MPEG standard," *IEEE Multimedia*, vol. 18, no. 3, pp. 86 –94, March 2011.

[13] D. Chen, V. Chandrasekhar, G. Takacs, S. Tsai, M. Makar, R. Vedantham, R. Grzeszczuk, and B. Girod, "Compact Descriptors for Visual Search: Improvements to the test model under consideration with a low-memory global descriptor," in *ISO/IEC JTCI/SC29/WG11 M24757*, April 2012.

[14] J. Lin, L.-Y. Duan, S. Yang, J. Chen, T. Huang, A. C. Kot, and W. Gao, "Compact Descriptors for Visual Search: Performance improvements of the Scalable Compressed Fisher Vector," in *ISO/IEC JTCI/SC29/WG11 MPEG2013/M28061*, January 2013.

[15] M. Makar, S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for mobile augmented reality," in *IEEE International Symposium on Multimedia*, December 2012.

[16] ——, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *International Journal of Semantic Computing*, vol. 7, no. 1, pp. 5–24, March 2013.

[17] Z. Drezner and N. Farnum, "A generalized binomial distribution," *Communications in Statistics*, vol. 22, no. 11, pp. 3051–3063, 1993.

[18] G. Francini, S. Lepsoy, and M. Balestri, "Selection of local features for visual search," *Signal Processing: Image Communications*, vol. 28, no. 4, pp. 311–322, April 2013.