Low Latency Image Retrieval with Embedded Compressed Histogram of Gradient Descriptors

Vijay Chandrasekhar[1], Sam S. Tsai[1], Gabriel Takacs[1], David M. Chen[1], Ngai-Man Cheung[1],
Yuriy Reznik[2], Ramakrishna Vedantham[3], Radek Grzeszczuk[3], Bernd Girod[1]

[1]Information Systems Laboratory, Stanford University, Stanford, CA 94305
[2]Qualcomm Inc., San Diego, CA 92121
[3]Nokia Research Center, Palo Alto, CA 94304

ABSTRACT

Network latency remains the bottleneck for mobile visual search applications. We show how network latency can be reduced using Compressed Histogram of Gradient (CHoG) descriptors. We study the trade-off in Classification Accuracy (CA) and bitrate for different parameters of CHoG descriptors. We show how CHoG bitstreams can be used in a rate-scalable manner. The embedded representation of CHoG bitstreams reduces transmission delay and enables early termination on the server side. We obtain a 2-4× decrease in system latency using CHoG descriptors compared to transmitting uncompressed SIFT descriptors or JPEG images in a 3G network.

Keywords

mobile visual search, CHoG, content-based image retrieval

1. INTRODUCTION

Mobile phones have evolved into powerful image and video processing devices, equipped with high-resolution camera, color displays, and hardware-accelerated graphics. They are also equipped with location sensors, GPS receivers, and connected to broadband wireless networks allowing fast transmission of information. This enables a class of applications which use the camera phone to initiate search queries about objects in visual proximity to the user. Such applications can be used for identifying products, comparison shopping, finding information about movies, CDs, real estate or products of the visual arts. Google Goggles [1], Nokia Point and Find [2] and Snaptell [3] are examples of recently developed commercial applications. For these applications, a query photo is taken by a mobile device and compared against previously stored database photos. A set of image feature descriptors is used to assess the similarity between the query photo and each database photo.

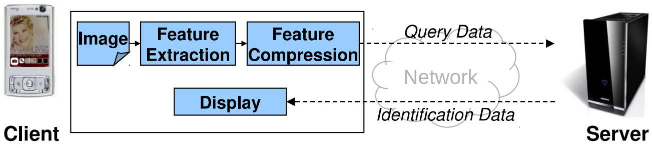


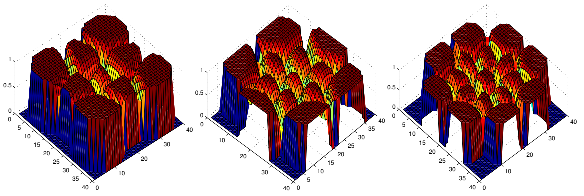
Figure 1: A mobile CD cover recognition system where the server is located at a remote location. Feature descriptors are extracted on the mobile-phone and query feature data is sent over the network. Once the CD cover is recognized on the server, identification data is sent back to the mobile-phone.

The system latency can be broken down into 3 components: (a) Processing time on mobile client (b) Network latency and (c) Processing time on server. In prior work [16, 15], we show that the processing time on the server and client is ∼1 second each, while the network transmission typically is the bottleneck in a 3G system. Hence, the size of the data sent over the network needs to be as small as possible to reduce latency and improve user interaction. To reduce network latency, we extract feature descriptors on the phone, compress the descriptors and transmit them over the network as illustrated in Figure 1. Such an approach has been demonstrated to reduce the amount of transmission data significantly compared to transmitting a JPEG compressed image [7, 6]. In this work, we focus on how transmission delay can be minimized using progressive transmission of compressed descriptors.

1.1 Prior Work

In [15], we present a state-of-the-art mobile product recognition system using a camera phone. The product is recognized through an image-based retrieval system located on a remote server. We provide experimental timings for different parts of the system, and show that transmission delay remains the bottleneck for 3G networks.

Low bitrate descriptors are critical for achieving low latency. In [7, 6], we propose a framework for computing low bitrate feature descriptors called Compressed Histogram of Gradients (CHoG). In [5], we perform a comprehensive survey of SIFT compression schemes and show that CHoG outperforms all schemes. We show that the CHoG descriptor at 60 bits matches the performance of the 128 dimensional



(*DAISY-9*) (*DAISY-13*) (*DAISY-17*)

Figure 2: The DAISY spatial binning configurations used for $n = 9, 13, 17$ spatial bins.

1024-bit SIFT descriptor [11].

Progressive transmission is common in the domain of image and video compression. E.g., JPEG2000 uses layering which allows for progressive transmission and rendering of images. This enables a client to display an image quickly by decoding only a portion of the image that it has received. As additional data is received by the client, the image can be progressively improved. Here, we show how progressive transmission can be applied to feature descriptors. In this work, we demonstrate progressive transmission of CHoG descriptors, but the ideas are also applicable to SIFT [11], SURF [4] and GLOH [12].

1.2 Contributions

In this work, we provide a thorough evaluation of CHoG descriptors in a large-scale retrieval system. We show how the CHoG bitstream can be used in a rate-scalable manner. The embedded representation of the CHoG descriptor reduces transmission delay and enables early termination on the server side. We report system latency in a 3G network for a CHoG based retrieval system. Using embedded CHoG descriptors, there's a 2-4× decrease in system latency compared to transmitting uncompressed SIFT descriptors or JPEG images.

1.3 Outline

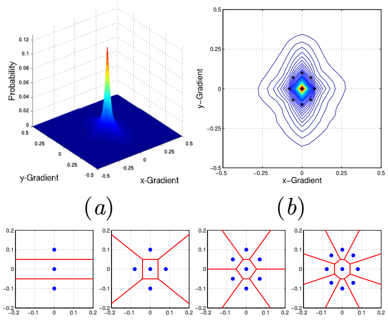
We organize the paper as follows. In Section 2, we review the CHoG descriptor. In Section 3, we provide an overview of the retrieval system. In Section 4, we provide experimental results for our retrieval system.

2. EMBEDDED CHOG DESCRIPTOR

CHoG [7] is a Histogram of Gradients descriptor that is designed to work well at low bitrates. We highlight some key aspects of the descriptor here and readers are referred to [7, 6] for more details.

2.1 Descriptor Review

First, we divide the patch into soft log polar spatial bins using DAISY configurations proposed in [18]. Next, the joint (d_x, d_y) gradient histogram in each spatial bin is captured directly into the descriptor. CHoG histogram binning exploits the skew in gradient statistics that are observed for patches extracted around keypoints. Finally, CHoG retains the information in each spatial bin as a distribution. This allows the use of more effective distance measures like KL divergence, and more importantly, enables efficient quantization and compression. Typically, 9 to 13 spatial bins and 3 to 9 gradient bins are chosen resulting in 27 to 117 dimensional descriptors. The spatial and gradient binning are illustrated in Figures 2 and 3.



(a) (b)

(*VQ-3*) (*VQ-5*) (*VQ-7*) (*VQ-9*)

Figure 3: The joint (d_x, d_y) gradient distribution (a) over a large number of cells, and (b), its contour plot. The greater variance in y-axis results from aligning the patches along the most dominant gradient after interest point The quantization bin constellations VQ-3, VQ-5, VQ-7 and VQ-9 and their associated Voronoi cells are shown at the bottom.

We quantize the gradient histogram in each cell individually and map it to an index. The indices are encoded with fixed length or entropy codes, and the bitstream is concatenated together to form the final descriptor. Fixed-length encoding provides the benefit of compressed domain matching at the cost of a small performance hit. In prior work [7, 6], we have explored several schemes for histogram compression: Huffman Trees, Type Coding and optimal Lloyd-Max VQ. In this work, we use Type Coding, which is linear in complexity to the number of histogram bins and performs close to optimal Lloyd-Max VQ [6].

2.2 Spatial Embedding

Since each spatial bin is encoded individually, the CHoG descriptor can be used in a rate-scalable manner. The DAISY spatial binning naturally lends itself to progressive transmission. For a set of descriptors, we first transmit the encoded data belonging to the inner spatial bins in the DAISY configuration, followed by data for the outer spatial bins. This is illustrated in Figure 4. Next, we show how embedded descriptors can be used in a retrieval system to achieve low latency.

3. RETRIEVAL SYSTEM

In this section, we provide an overview of the retrieval system for embedded descriptors. We show how progressive transmission on the client enables early termination on the server, and reduces transmission delay.

3.1 Client

We extract CHoG descriptors on the mobile device and transmit them over the network as illustrated in Figure 1. We extract 300 to 700 CHoG descriptors on the mobile device. We report results for DAISY-9 spatial binning, and VQ-5 and VQ-7 gradient binning in our experiments.

For DAISY-9 spatial binning, we refer to the inner 5 spatial bins as the *Base Layer* and the outer 4 spatial bins as the *Enhancement Layer*. The transmission order of bits is illustrated in Figure 4. The encoded data for *Base Layer* from all descriptors is transmitted followed by the encoded

Layer	Gradient binning	Bits / descriptor
Base	VQ-5	25
Base+Enhancement	VQ-5	48
Base	VQ-7	37
Base+Enhancement	VQ-7	70

Table 1: CHoG descriptor parameters for DAISY-9 spatial binning. *Base Layer* refers to the inner 5 spatial bins. *Enhancement Layer* refers to the outer 4 spatial bins.

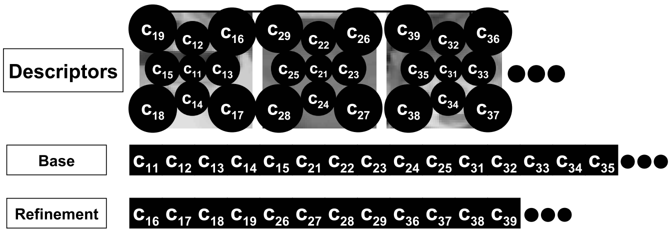

Descriptors
c19
c12
c16
c15
c11
c13
c14
c18
c17
c29
c22
c26
c25
c21
c23
c24
c28
c27
c39
c32
c36
c35
c31
c33
c34
c38
c37
●●●
Base
C11 C12 C13 C14 C15 C21 C22 C23 C24 C25 C31 C32 C33 C34 C35 ●●●
Refinement
C16 C17 C18 C19 C26 C27 C28 C29 C36 C37 C38 C39 ●●●

Figure 4: Transmission order of bits on the client. The DAISY-9 spatial bin configuration is overlaid on a set of scaled and oriented patches. c_{ij} refers to the encoded distribution in i^{th} feature and j^{th} spatial bin. The encoded distributions belonging to the inner spatial bins are transmitted first (*Base Layer*), followed by the outer spatial bins (*Enhancement Layer*).

data from *Enhancement Layer*. The location data for features is compressed using the scheme proposed in [17]. The location data is transmitted along with *Base Layer* descriptor data. The parameters for chosen CHoG descriptors are shown in Table 1.

3.2 Server

We briefly describe the retrieval pipeline for CHoG descriptors which builds on state-of-the-art proposed in [13, 14]. The server processing is illustrated in Figure 5. We train multiple Vocabulary Trees (VT) [13] with depth 6 and branch factor 10, resulting in trees with 10^6 leaf nodes. A VT is trained for *Base Layer* embedded descriptors and the full descriptors encompassing both *Base Layer* and *Enhancement Layer*. Once the data for *Base Layer* is received, the server can start the recognition process. If a match is found with *Base Layer* data, the server early terminates and send a response back to the client. In Section 4, we show how early termination can reduce application latency.

We use soft-assignment for quantization of descriptors to the 3 nearest centroids in each VT [14]. For each VT, we use the standard Term Frequency-Inverse Document Frequency (TF-IDF) scheme [13] that represents query and database images as sparse vectors of visual word occurrences, and compute a similarity between each query and database vector. We use the weighting scheme proposed in [13] which reduces the contribution of less descriminative descriptors. We use geometric constraints to re-rank the list of top 500 images [10]. Finally, we consider upto 50 images for pairwise matching with a RANSAC [9] affine consistency check.

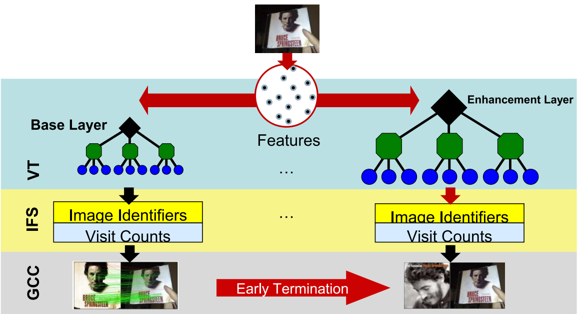


Figure 5: Server Architecture. We store multiple Vocabulary Trees on the server, one for each embedded layer. The server starts the recognition process once data for an embedded layer is received. If a match is found with *Base Layer* data, the server early terminates and returns a response to the mobile client.

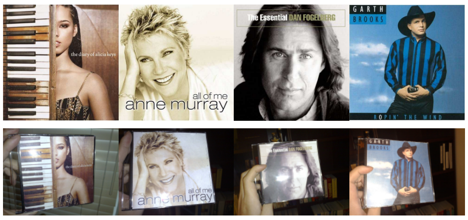


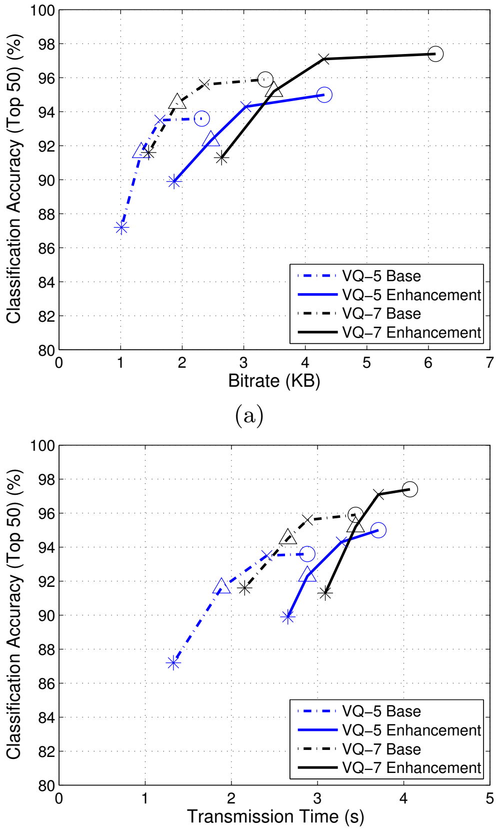
Figure 6: A clean database picture (*top*) is matched against a real-world picture (*bottom*) with various distortions.

4. RESULTS

For evaluation, we use a database of one million CD,DVD and book cover images, and a set of 1000 query images [8] exhibiting challenging photometric and geometric distortions, as shown in Figure 6. Each image has 500 × 500 pixels resolution. We define Classification Accuracy (CA) as the percentage of query images correctly retrieved.

First, in Figure 7 (a), we study the performance of embedded CHoG descriptors in a retrieval system. The maximum accuracy achieved by the system is $\sim$97.5%. For each configuration, the performance improves as the number of features increases and typically plateaus off. The higher bitrate descriptors plateau off at higher CA. For each configuration, there is typically a 2-4% increase in CA after the *Enhancement Layer* data are received compared to only the embedded *Base Layer* data. This implies that early termination can be achieved for a majority of query images once the embedded *Base Layer* data is received. Next, we show how early termination can reduce application latency.

In Figure 7 (b), we report the transmission delay for embedded CHoG descriptors over a typical 3G wireless network. The data transmission experiments are conducted over several days, with a total of more than 5000 transmissions at indoor locations where a image-based retrieval system would be typically used. The typical difference in transmission delay between *Base Layer* and the full data is 1-1.5 seconds. Figure 7 (b) can be interpreted as follows: At any operating point, the savings in latency is the difference



(a)

(b)

Figure 7: Retrieval results for embedded CHoG descriptors are shown in Figure (a). The corresponding transmission times in a 3G network are shown in Figure (b). The dotted and solid curves correspond to the performance of the system after *Base Layer* and *Enhancement Layer* data are received respectively. Points corresponding to the embedded and full data are represented by the same marker on the dotted and solid curves.

between the corresponding points on the *Base Layer* and *Enhancement Layer* curves, if a match is found with *Base Layer* data. E.g., for the lowest VQ-5 bitrate point, we can early terminate with 87% probability after the *Base Layer* data are received, and we reduce the end-to-end latency by 1.5 seconds. Once a match is found with *Base Layer* data, the server early terminates and sends a response to the mobile client. The different CHoG parameters allow trade-off in bitrate and CA. The operating point is chosen based on latency requirements of the system.

Next, we compare transmitting embedded CHoG descriptors to uncompressed SIFT descriptors or JPEG compressed images in Figure 8. For the JPEG scheme in Figure 8, the bitrate is varied by changing the quality of compression. For the SIFT scheme in Figure 8, each SIFT descriptor is transmitted uncompressed as 1024 bits (128 dimensions × 8 bits/dimension). For SIFT, we sweep the CA-bitrate curve by varying the number of descriptors transmitted. For the CHoG VQ-5 and VQ-7 schemes, we plot the average bitrate required for recognition with early termination. The average bitrate, in the presence of early termination, is computed as a weighted average of the *Base* and *Enhancement* layer bitrates in Figure 7 (a).

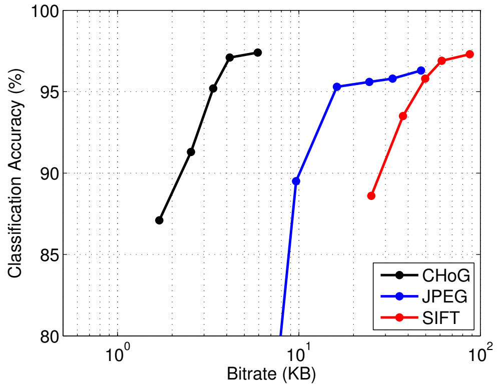


Figure 8: Bitrate comparisons of different schemes. Using embedded CHoG descriptors reduces the data by an order of magnitude compared to transmitting JPEG images or uncompressed SIFT descriptors.

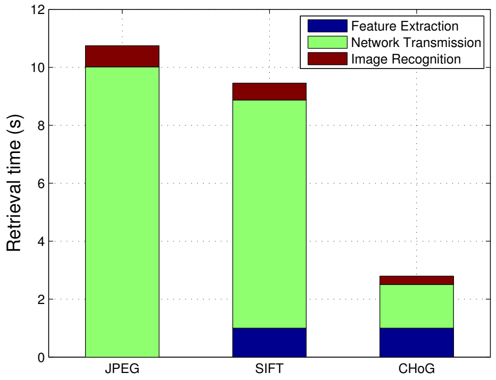


Figure 9: Processing time for different schemes. We achieve 2-4× reduction in end-to-end latency using embedded CHoG descriptors with early termination.

We observe that the performance of the JPEG scheme rapidly deteriorates at low bitrates. The performance suffers at low bitrates as the interest point detection fails due to blocking artifacts introduced by JPEG compression. We also note that transmitting uncompressed SIFT data is almost always more expensive than transmitting JPEG compressed images. We observe in Figure 8 that CHoG descriptors are an order of magnitude smaller than JPEG images or SIFT descriptors.

Finally, in Figure 9, we study the end-to-end latency at the highest accuracy point (97.5%) for the different schemes. For the JPEG scheme, there is no processing on the client. For the SIFT and CHoG schemes, ∼1 second is spent extracting features on the mobile client. For embedded CHoG descriptors, we compute the average transmission time as a weighted average of the *Base* and *Enhancement* layer timings in Figure 7 (b). We achieve a 2× reduction in system latency with embedded CHoG descriptors compared to JPEG images, and a 4× reduction compared to uncompressed SIFT descriptors.

5. CONCLUSION

We demonstrate how network latency can be reduced significantly using Compressed Histogram of Gradient (CHoG) descriptors. We show how CHoG bitstreams can be used in a rate-scalable manner. The embedded representation of CHoG bitstreams reduces transmission delay and enables early termination on the server side. We obtain a 2-4× decrease in system latency using embedded CHoG descriptors compared to transmitting uncompressed SIFT descriptors or JPEG images in a 3G network.

6. REFERENCES

[1] *Google Goggles.* http://www.google.com/mobile/goggles/.

[2] *Nokia Point and Find.* http://www.pointandfind.nokia.com.

[3] *SnapTell.* http://www.snaptell.com.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understing*, 110(3):346–359, 2008.

[5] V. Chandrasekhar, M. Makar, G. Takacs, D. Chen, S. S. Tsai, N. M. Cheung, R. Grzeszczuk, Y. Reznik, and B. Girod. Survey of SIFT Compression Schemes. In *Proceedings of International Mobile Multimedia Workshop (IMMW), IEEE International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010.

[6] V. Chandrasekhar, Y. Reznik, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod. Study of Quantization Schemes for Low Bitrate CHoG descriptors. In *Proceedings of IEEE International Workshop on Mobile Vision (IWMV)*, San Francisco, California, June 2010.

[7] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod. CHoG: Compressed Histogram of Gradients - A low bit rate feature descriptor. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 2009.

[8] D. M. Chen, S. S. Tsai, R. Vedantham, R. Grzeszczuk, and B. Girod. *CD Cover Database - Query Images*, April 2008.

[9] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of ACM*, 24(6):381–395, 1981.

[10] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 304–317, Berlin, Heidelberg, 2008.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] K. Mikolajczyk and C. Schmid. Performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[13] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, USA, June 2006.

[14] 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. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, June 2008.

[15] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod. Mobile Product Recognition. In *Submitted to ACM Multimedia (ACM MM)*, Florence, Italy, October 2010.

[16] S. S. Tsai, D. M. Chen, J. Singh, and B. Girod. Rate-efficient, real-time CD cover recognition on a camera-phone. In *Proc. of ACM Multimedia (ACM MM)*, Vancouver, British Columbia, Canada, October 2008.

[17] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, J. P. Singh, and B. Girod. Location coding for mobile image retreival systems. In *Proceedings of International Mobile Multimedia Communications Conference (MobiMedia)*, London, UK, September 2009.

[18] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, June 2009.