A Novel Differential Coding Scheme for a Compact Image Descriptor with applications to Mobile Visual Search

Abhishek Nagar[2], Gaurav Srivastava[2], Felix C.A. Fernandes[2] and Aggelos K.Katsaggelos[1]

[1]Dept. of EECS, Northwestern University, Evanston, IL, USA
[2]Samsung Research America (SRA), Richardson, TX, USA

ABSTRACT

The ongoing MPEG standardization of Compact Descriptors for Visual Search (CDVS) focuses on image search for mobile applications and in that process, the extraction of local descriptors constitutes an important step. These local descriptors extracted from an image are further aggregated into global descriptors that are used for efficient retrieval of matching images from a database for a given query image. Current CDVS Test Model (TM) implements the global descriptor using the uncompressed Scale Invariant Feature Transform (SIFT) points. At the mobile devices, the global descriptor (GD) is computed as the quantized Fisher Vector of up to 300 SIFT points w.r.t a SIFT space Gaussian Mixture Model (GMM). It is noted that such an approach requires significant overhead in communication to transmit the global descriptor, especially at low bit rate. Hence, we propose an alternative and efficient way to re-construct the global descriptor from the local descriptors at the server side. The difference between the reconstructed GD and the original GD, are then selectively coded to strike a balance between bit rate cost and performance. The experiments on CDVS datasets shows around 0.5% increase in true positive rate and 1% decrease in false positive rate.

Index Terms— compact visual descriptor, global descriptor, mobile visual search

1. INTRODUCTION

In visual search applications using images, and consequently in the ongoing MPEG Compact Descriptors for Visual Search (CDVS) standardization which currently focuses on image search, extraction of local descriptors constitutes an important step [1,3,6]. These local descriptors extracted from an image are further converted to global descriptors that are used for efficient retrieval of matching images from a database. Visual search requires two steps in the retrieval part: (i) Using the global descriptors [6] for the query image to shortlist the database images, (ii) By using the local descriptors within a geometric verification step, calculating the matching scores between the query image and all the database images in the retrieved shortlist [1]. Therefore both the local and the global descriptors are important for visual search and Fig.1 shows the pipeline for extracting these descriptors from an image. In the figure, DoG refers to Difference of Gaussian that is used for keypoint detection.

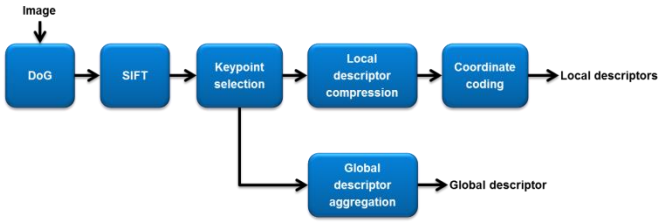


Figure 1. Current CDVS Pipeline for Descriptor Extraction.

Current CDVS Test Model (TM) implements the global descriptor using the uncompressed Scale Invariant Feature Transform (SIFT) points. At the mobile devices, the global descriptor (GD) is computed as the quantized Fisher Vector of up to 300 SIFT points w.r.t a SIFT space Gaussian Mixture Model (GMM). The GD extraction process is illustrated in Fig. 2. It is noted that such an approach requires significant overhead in communication to transmit the global descriptor, especially at low bit rate. Hence, we propose an alternative and efficient way to re-construct the global descriptor from the local descriptors at the server side. The difference between the reconstructed GD and the original GD, are then selectively coded to strike a balance between bit rate cost and performance.

The main problem with the current GD solution is that it is sending redundant information vis-à-vis the SIFT points already coded in the local descriptor. How to utilize this information, and construct an alternative GD from this noisy set of SIFTs and have a differential coding scheme that can selectively correct GD errors, while meeting bit budget constraint, is the focus of this paper.

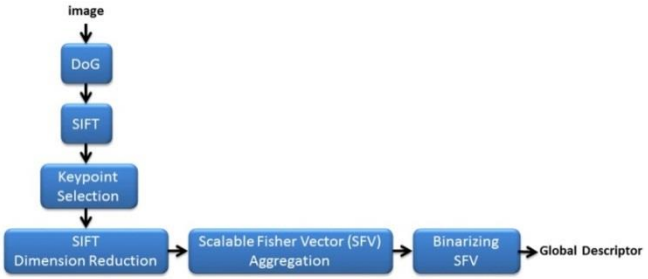


Figure 2. Pipeline of generating the global descriptor

2. Benefits of the Proposed Method

- More compact bitstream is obtained because instead of sending the full GD[9, 10, 11,12,13,14], we send the selective residual difference between the original GD (i.e. the one which is aggregated from uncompressed local descriptors) and the GD reconstructed from compressed local descriptors. The difference is selectively coded so that it occupies less bits than the original GD.

3. Proposed method for differential coding of GD (Global Descriptor)

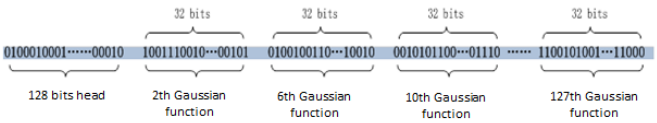


Figure 3: An example of current CDVS global descriptor bit stream.

As shown in Figure 3, the GD stream is coded using fixed length coding and is byte aligned. First 128-bit header is used to indicate which clusters are turned on i.e. which clusters are encoded as part of the GD bit stream. Let the number of the activated cluster be n. Then n 32-bit vectors are used to represent the GMM functions. As known, GD is generated using the uncompressed SIFT features.

In this paper, we first propose to derive the GD from the compressed LD and code the "residual" between using uncompressed SIFTs and reconstructed SIFTs from compressed LD bit stream. Therefore, we don't need to explicitly signal n*32 bits for GD. A block diagram of the coding scheme is shown in Figure 4.

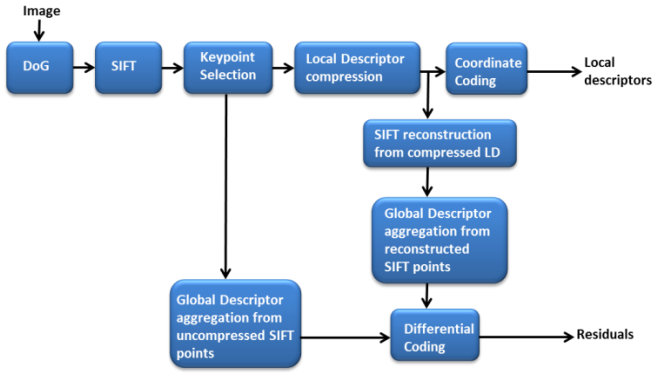
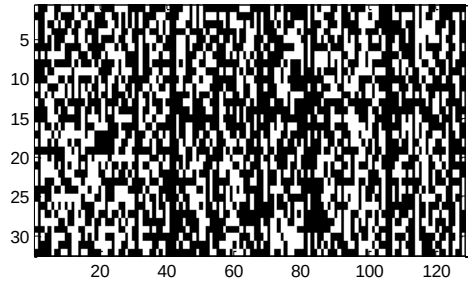


Figure 4: Differential coding of global descriptor

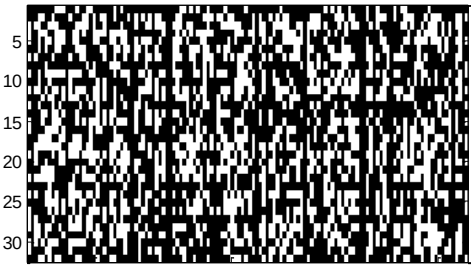
The scheme shown in Figure 4 is different from the SKIP mode scheme for handling the global descriptors (contribution #) in that in the current scheme, in addition to sending the coded local descriptors, the mobile device also sends selective residual information obtained from the differential coding between the global descriptors aggregated from the uncompressed SIFT points and those aggregated from the reconstructed SIFT points (from the compressed local descriptors). Obviously due to the compression of the local descriptors, the reconstructed GDs do not perform optimally with respect to the matching between two images that is performed to generate a short list of images for a query image during the retrieval. The residual information needed to reconstruct the original global descriptor is additionally supplied by the differential coding scheme.

4. Simulation Results

As aforementioned in Figure 3, for each image, the global descriptor can be represented as 32*128 binary matrix as illustrated in Figure 5(a). Meanwhile, reconstructed global descriptor from compressed local descriptor is shown in Figure 5(b)



(a) Original GD using uncompressed SIFTs



(b) Proposed GD reconstructed from compressed LD

Figure 5: Illustration of global descriptor for current TM implementation using uncompressed SIFTs and proposed solution using reconstructed SIFTs from compressed local descriptor. Each column represents the 32 bits corresponding to one of the 128 clusters.

As shown in Figure 5, we see that proposed solution looks similar as original scheme, however, there still exists difference for certain Gaussian functions i (0< i <127) . We further picture the GD difference between original and proposed solution, shown in Figure 6. As we can see, if we can efficiently compress such binary GD difference, we only need to signal the GD difference at client side and shift GD reconstruction and retrieval indexing to the server. To this end, we have proposed the following GD difference

encoding scheme (Section 4.1). Please note that in order to obtain a more compact bitstream [15,16,17,18 19, 20,21,22], we should make sure that the R(GD_difference) < n*32+128 where R(GD_difference) represent the number of the bits for encoding the binary GD difference, and n*32+128 are the bits for current TM GD implementation with n being the number of active GMM clusters.

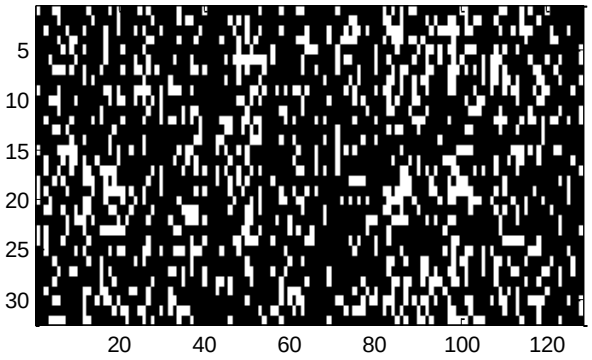


Figure 6: Global descriptor difference for the two GDs in Figure 5.

4.1 Bit-flip Significance based Correction (n_bit_flipped)

As shown in Figure 6, each "white pixel" represents the "1" which indicates the bit-flip at this position. We can calculate the number of "1"s in each cluster which shows the total number of bits flipped in this cluster. Figure 7 plots the number of the flipped bits in each cluster for a random image picked from the CDVS test databases. Other images share the similar pattern.

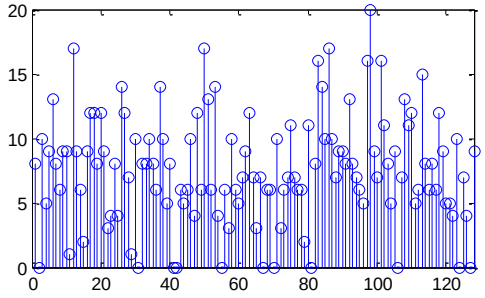


Figure 7: Number of flipped bits for each GMM cluster

Assuming that more the number of bit flips, the bigger the distortion, a simple scheme is to correct the k clusters which have the most number of bit flips noted as the bit_flip_significance. To implement this, we can sort the distortion in Figure 7, and only correct the k clusters as shown in Figure 8. So we need 32 correction bits for each of k clusters and 128 bits to signal which k of the 128 clusters would be corrected. Therefore R(GD_difference) = k*32+128. From the requirement that R(GD_difference) < n*32+128, we have k<n. We can adjust k to obtain the optimal trade-off between the retrieval performance and the transmission cost.

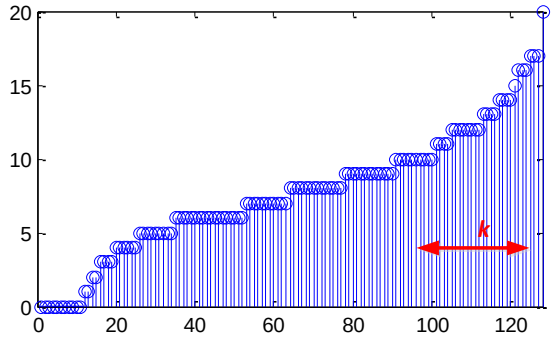


Figure 8: Bit-flip significance based correction

In summary, the whole differential coding framework can be generalized as two major steps: the first step is to select effective GMM clusters to be corrected (as demonstrated above), and the second step is to code these GMM clusters as well as their indices. Please note that we are using the fixed length code to illustrate the selective differential coding for global descriptor difference. However, the same method can be applied to the variable length codes as well (such as Huffman codes, universal codes as well as the more context-adaptive advanced binary arithmetic codes). Probabilistic model can be used to reduce the entropy for better coding efficiency.

For the effective GMM cluster indices coding, it is suggested to use the loss-less coding strategy through fixed-length codes, or VLC or arithmetic codes. For each selected GMM cluster, either loss-less or lossy coding schemes can be applied.

We test our algorithm on 7 CDVS datasets using the CDVS pipeline. We perform our experiments on CDVS dataset, which is provided by multiple universities and institutions. The datasets covers different categories including mixed text, Video Frames, Paintings, Frames captured from video clips, buildings, common object and graphics of CDs, DVDs, Books, Documents. Totally, there are 30256 images and 12,566,632 SIFT features with our SIFT extraction parameters setup. Some of the sample images are listed in Fig. 8. The images have different aspect ratios.

Figure 11. Sample images of the CDVS datasets

The results are shown in Table 1. As shown in the table, we can see that the overall precision is increased with fewer bits needed. The average true positive rate (TPR) is increased by 0.4796 while the false positive rate is reduced by 1%.

Experiment Number	Descriptor Lengths	Baseline		Selective Diff. Coding	
		TPR	FPR	TPR	FPR
1a	512	85.97	1.98	87.33	1.25
	1k	92.33	2.15	93.3	1.7
	2k	95.33	1.93	95.63	1.18
	1k,4k	90.67	1.1	91.57	1.31
	2k,4k	95.47	1.8	95.9	0.95
	4k	97.23	1.73	97.23	0.85
	8k	97.97	1.41	98	0.44
	16k	98.53	1.53	98.5	0.54
1b	512	86.07	2.04	87.4	1.35
	1k	92	2.15	93.43	1.76
	2k	95.73	1.93	96	1.18
	1k,4k	90.77	1.19	91.63	0.36
	2k,4k	95.37	1.87	95.57	1.17
	4k	97.1	1.7	97.17	0.78
	8k	97.9	1.34	97.87	0.36
	16k	98.33	1.51	98.37	0.51
1c	512	83.4	1.96	85.6	1.24
	1k	90.83	2.03	92.03	1.61
	2k	94.67	1.98	95	1.13
	1k,4k	88.67	1.14	89.67	1.27
	2k,4k	94.23	1.64	94.63	0.76
	4k	96.27	1.52	96.3	0.63
	8k	97.37	1.25	97.4	0.28
	16k	98	1.31	97.93	0.36
2	512	92.03	0	92.86	0.25
	1k	95.6	0.14	96.43	0.44
	2k	96.7	0.22	97.25	0.25
	1k,4k	94.51	0.14	95.33	0.27
	2k,4k	96.7	0.22	97.25	0.3
	4k	96.7	0.25	96.7	0.25
	8k	96.98	0.08	96.7	0.08
	16k	96.7	0.08	96.7	0.08
3	512	97	0.63	97	0.68
	1k	97.75	0.48	98	0.68
	2k	98.75	0.6	98.75	0.65
	1k,4k	97	0.6	97.5	0.88
	2k,4k	99	0.7	99	0.75
	4k	98.75	0.73	98.75	0.8
	8k	98.75	0.55	98.75	0.55
	16k	98.75	0.55	98.75	0.55
4	512	70.96	0.71	71.61	0.89
	1k	74.63	0.55	75.43	0.75
	2k	79.68	0.67	80.2	0.7
	1k,4k	73.61	0.51	74.81	0.61
	2k,4k	80.62	0.65	81.17	0.68
	4k	81.9	0.62	82.02	0.65
	8k	84.44	0.39	84.52	0.39
	16k	85.69	0.45	85.72	0.47
5	512	81.06	0.25	82.31	0.4
	1k	85.29	0.21	86.55	0.37
	2k	89.14	0.16	90.04	0.2
	1k,4k	80.9	0.27	82.04	0.37
	2k,4k	88.39	0.22	88.78	0.22
	4k	90.59	0.17	91.06	0.18
	8k	92.82	0.09	92.9	0.1
	16k	94.12	0.09	94.24	0.1

Table 1. Results on 7 datasets

5. Memory and Computational Complexity Considerations

Since the global descriptor is not a part of the bit stream therefore there is a saving of 128+32n bits where n is the number of active GMM clusters. But instead we need to send differential coding information which has a size of 128+32k bits where k is the number of clusters which have the largest number of bit flips between the original and reconstructed Fisher vector components. k can be selected to be smaller than n. Therefore there is a saving of (n-k)*32 bits.

6. Conclusions

In this paper, we proposed a scheme to save bits by reconstructing global descriptors from local descriptors. This scheme is proven to be effective in preserving the performance while saving bits.

Reference

[1] CDVS, *Description of Core Experiments on Compact descriptors for Visual Search, N12551.* San Jose, California, USA: ISO/IEC JTC1/SC29/WG11, Feb 2012

[2] S. Lepsoy, G. Francini, G. Cordava and P. P. Gusmao, *Statistical modeling of outliers for fast visual search*, in Proc. IEEE Workshop on Visual Content Identification and Search, July 2011.

[3] X. Xin, Li, Z., and A.K.Katsaggelos. "Robust feature selection with self-matching score". In Proc. of IEEE ICIP (2013).

[4] ISO/IEC JTC1/SC29/WG11/ M22672, *Telecom Italia's response to the MPEG CfP for Compact Descriptors for Visual Search*, Geneva, CH, Nov. 2011.

[5] Xin, X., Nagar, A., Srivastava, G., Li, Z., Fernandes, F., and Katsaggelos., A. K. Large visual repository search with hash collision design optimization. MultiMedia, IEEE 20, 2 (2013), 62-71.

[6] CDVS. *Evaluation Framework for Compact Descriptors for Visual Search, N12202.* Turin, Italy: ISO/IEC JTC1/SC29/WG11, 2011.

[7] X. Xin, Z. Li, and A. K. Katsaggelos. "Query compression for mobile visual search". In Proc. of IEEE ICIP (2012)

[8] CDVS. *Examples of feature selection to boost retrieval performance*, M23938. , San Jose, California, USA: ISO/IEC JTC1/SC29/WG11, Feb 2012.

[9] Xin, X., Li, Z., Ma, Z., and Katsaggelos, A. K. Spectral approximation to point set similarity metric. In Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on (2013), pp. 1-4.

[10] CDVS *Improvements to the Test Model Under Consideration with a Global Descriptor*, M23938. , San Jose, California, USA: ISO/IEC JTC1/SC29/WG11, Feb 2012.

[11] Xin, X., Z. Li, and A. K. Katsaggelos, "Laplacian Embedding and Key Points Topology Verification for Large Scale Mobile Visual Identification", Signal Processing: Image Communication, In Press.

[12] Liu, T. - J., H. J. Han, X. Xin, Z. Li, and A. K. Katsaggelos, "A Robust and Lightweight Feature System for Video Fingerprinting",European Signal Processing Conference (EUSIPCO), Bucharest, Romania, Aug. 27-31, 2012.

[13] Xin, X., and A. K. Katsaggelos, "A Novel Image Retrieval Framework Exploring Inter Cluster Distance", Int. Conf. Image Processing, 26/09/2010.

[14] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction", *IEEE Trans. Pattern Anal. Mach*. Intell. 2007.

[15] [M. Iliadis, S. Yoo, X. Xin and A. K. Katsaggelos. "Virtual touring: A content based image retrieval application". In Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on (2013), pp. 1-4.

[16] Z. Li, X. Xin, and A. K. Katsaggelos. Image topological coding for visual search. In US Patent (2012).

[17] J. Springer, X. Xin Z. Li, J. Watt and A. K. Katsaggelos. "Forest hashing: Expediting large scale image retrieval". In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on (2013), pp. 1681-1684.

[18] TJ. Liu, H. Han, X. Xin, Z. Li, and A. K. Katsaggelos. "A robust and lightweight feature system for video fingerprinting". In Proc. of EUSIPCO (2010).

[19] X. Xin, Li, Z., and A.K.Katsaggelos. Laplacian embedding and key points topology verification for large scale mobile visual identification. Signal Processing: Image Communication 28, 4 (2013), 323-333.