Low bitrate coding schemes for local image descriptors

A. Redondi, M. Cesana, M. Tagliasacchi

Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy
{redondi, cesana, tagliasa}@elet.polimi.it

***Abstract*—Efficient coding of local image descriptors is of paramount importance when they need to be transmitted to a remote destination on bandwidth constrained networks. This is a case that arises, e.g., in mobile visual search and visual wireless sensor networks. In this work we consider SURF, a popular descriptor suitable for low-complexity devices, and we provide a comparative study of lossy coding schemes operating at low bitrate (e.g., less than 128 bits / descriptor). Our investigation covers schemes that address both intra- and inter-descriptor redundancy, including methods that have not been tested before in this context, e.g., sparse coding, lifting-based coding on trees, and hybrid intra and inter-descriptor coding. The experimental evaluation is carried out on two publicly available datasets, in terms of both rate-distortion and rate-accuracy, for the specific task of object recognition. Our results show that a rate saving of 15-30% can be achieved by exploiting intra-descriptor redundancy. On the other side, addressing inter-descriptor redundancy does not lead to substantial gains when applied alone, whereas it leads to marginal gains (up to 3%) when used in hybrid schemes jointly with intra-descriptor coding.**

***Index Terms*—Local image descriptors coding, visual features compression, object recognition, mobile visual search**

I. Introduction

Local visual features have been the subject of several research studies during the past few years. Their ability to provide a concise and robust representation of the image content serves as a basis for a broad range of visual analysis tasks. Applications include, but are not limited to, people and object identification [1], traffic and parking monitoring [2] and scene classification [3]. Recently, visual features have been used in the field of mobile visual search [4], laying the foundations for a new class of multimedia retrieval applications. The general architecture of such applications consists in a client/server paradigm where a wireless device with an embedded camera (e.g., smart phones, smart cameras or visual sensor networks) extracts local features from the acquired content and sends the corresponding descriptors to a remote server in the form of a query. There, descriptors are matched against a large database to find similar images to be returned to the user application as search results. In a typical scenario the transmission of such descriptors is performed by means of a bandwidth constrained wireless network. Therefore, transmitting the query data to the server is demanding in terms of time and energy resources. Thus, it is imperative to design efficient coding schemes for image descriptors, in order to meet low-latency requirements and to minimize the cost of transmission and storage.

In this work, we are interested in studying and comparing the rate-distortion efficiency of different lossy coding schemes applied to bags of local descriptors extracted from still images. Specifically, we focus on SURF [5], as it iss designed with the purpose of reducing the computational complexity of state-of-the-art descriptors, e.g., SIFT [6], while retaining desirable invariance properties. To this end, we explore how to exploit correlation both within the elements of the same descriptor (i.e., intra-descriptor redundancy) and among descriptors extracted from different keypoints of the same image (i.e., inter-descriptor redundancy). In addition to methods already presented in the literature, we also include in our investigation coding schemes that have not been tested before in this context, e.g., sparse coding, lifting-based coding on trees, and hybrid intra- and inter-descriptor coding. Then, for the specific task of object recognition, we experimentally derive operational rate-accuracy curves on two widely used datasets. With this, we compare the task efficiency obtained by lossy coding SURF descriptors, with that of other state-of-the-art descriptors designed for low bitrate applications (i.e., CHoG [7]).

Our results show that a rate saving of 15-30% can be achieved by exploiting intra-descriptor redundancy. However, addressing inter-descriptor redundancy does not lead to substantial gains when applied alone, whereas it leads to marginal gains (up to 3%) when used in hybrid schemes jointly with intra-descriptor coding. This somewhat contradicts recent findings (on different datasets) suggesting significant potential benefits from predictive coding of sorted descriptors [8].

The rest of the paper is organized as follows: Section II summarizes the state-of-the-art on local descriptors coding. Section III illustrates the different coding schemes to be compared. Experimental results are reported in Section IV, while Section V concludes the paper and comments on future works.

II. Related Work

The problem of efficiently coding local visual features has been recently discussed by several works in the literature. It is possible to distinguish two different approaches: i) methods aimed at coding state-of-the-art descriptors; ii) methods aimed at finding alternative forms of descriptors, characterized by a more compact representation. A third class of approaches, which is not addressed in this paper, deals with the efficient

representation of bag-of-visual words descriptors obtained from local features [9].

In the first class of methods, Chandrasekhar et al. [10] address the problem of lossy coding SIFT and SURF descriptors, which are characterized by, respectively, 128 and 64 dexels (descriptor elements). They observe that the variances of the different dexels of the descriptor are different, mostly due to the non-uniform Gaussian weighting applied to the image patch. Also, dexels are correlated, suggesting the adoption of a transform coding scheme based on the Karhunen-Loeve Transform (KLT), which is learned from a large training set. A coding efficiency gain of 0.4 bits / dexel is observed for SURF, whereas the KLT seems to hurt coding efficiency for SIFT. Successful image matching (on the ZuBuDu dataset) is achieved with a rate allocation around 2 bits / dexel, i.e., 256 and 128 bits / descriptor for, respectively, SIFT and SURF. In [11], the authors exploit correlation among SIFT descriptors extracted from the same image by sorting them in such a way as to minimize the sum of pair-wise distances. Then, a predictive coding scheme based on DPCM is applied and inter-descriptor prediction residuals are quantized and transmitted. In a landmark search task, the methods attains the same efficiency (in terms of mean average precision) as SIFT encoded at full precision (i.e. $8 \times 128 = 1024$ bits / descriptor), when using at least 80 bits / descriptor.

As for the second class of methods, a compact descriptor, named PCA-SIFT, is described in [12] addressing the intra-descriptor redundancy of SIFT. There, the Principal Component Analysis is applied to the normalized gradient patch (before the computation of local gradient histograms) and only the first 20 principal components are retained. In this way, a more compact representation is obtained, which is also shown to be more robust to image deformations than SIFT. However, coding is not explicitly addressed, and each element is represented with 8 bits. Motivated by applications in the context of mobile visual search, the authors of [13] propose a CHoG (Compressed Histogram of Gradients) descriptor, which is specifically conceived to be compressed at low bitrates. The key tenet consists in gathering statistics around detected keypoints, so as to build a set of local histograms, which can be efficiently compressed by means of a carefully designed entropy coding scheme. CHoG descriptors are shown to be particularly promising, since they provide a compact, yet comprehensive, representation of the local image patches. The authors claim that CHoG can be compressed using as few as 53 bits / descriptor, while retaining the same matching accuracy as SIFT encoded at full precision. In [14], the same authors propose a low-latency image retrieval system with progressive transmission of CHoG descriptors. A feedback is used when a match is found on the server so that the querying client can stop the transmission process. In that case, descriptors are sorted based on the value of the local Hessian response. The underlying idea is that descriptors associated to a low value of the Hessian response are poorly localized and less discriminative. The BRIEF descriptor is introduced in [15], in order to speedup both computation and matching, while at the same time obtaining a rather compact representation. BRIEF can be computed using simple pixel intensity difference tests. The number of bits per descriptor is equal to the number of tests. In a simple recognition experiment, between 58 and 164 bits per descriptors are used to achieve the same results as U-SURF [5], a simpler although not rotational-invariant version of SURF.

III. Low-bitrate coding techniques

In this work, we focus on the SURF descriptor, which is widely used in applications where computational resources are scarce. Let I denote an image with M_I associated descriptors $\mathbf{d}_i \in \mathbb{R}^{64}, i = 1, \ldots, M_I$. When compression is not a concern, each of the 64 dexels of SURF is represented with 8 bits. In the following, we consider different lossy coding scheme aimed at reducing the average bitrate needed to represent $\mathbf{d}_i$. First, we consider intra-description coding schemes, in which each descriptor $\mathbf{d}_i$ is encoded independently from other descriptors. Then, we turn our attention to inter-descriptor coding schemes, which exploit correlation among the descriptors extracted from the same image. In all cases, let $\tilde{\mathbf{d}}_i$ denote the lossy coded version of $\mathbf{d}_i$ which can be reconstructed at the decoder side.

A. Intra-descriptor coding:

- *Scalar quantization (PCM).* The simplest approach consists in applying scalar quantization to each dexel. That is,
$$\tilde{d}_{i,j} = \Delta_j \cdot \text{round}(d_{i,j}/\Delta_j) \quad (1)$$
Here, we fix the same quantization step size for all dexels, i.e., $\Delta_j = \Delta$, $j = 1, \ldots, 64$. Let $\mathcal{D}_j = \{d_{1,j}, d_{2,j}, \ldots, d_{M,j}\}$ denote a set of dexels corresponding to the j-th dimension of M descriptors, which are collected from a large set of training images. Similarly to [10], we recognize the fact that dexels are characterized by different statistics. Therefore, optimal bit allocation calls for assigning an average number of bits $R_j \propto \log_2 \sigma_j^2$, where $\sigma_j^2 = \text{var}(\mathcal{D}_j)$. In practice, for a given step size Δ, we compute the output symbols of the quantizer, i.e., $\tilde{\mathcal{D}}_j^{\Delta} = \{\tilde{d}_{1,j}, \tilde{d}_{2,j}, \ldots, \tilde{d}_{M,j}\}$, and we consider the adoption of an optimal entropy coding scheme, i.e., such that the average code length is $R_j^{\Delta} = H(\tilde{\mathcal{D}}_j^{\Delta})$, where $H(\cdot)$ denotes the (empirical) entropy of a discrete memoryless source. Then, $R^{\Delta} = \sum_j^{64} R_j^{\Delta}$, where the superscript Δ emphasize the parameter used to adjust the rate.
- *Transform coding (KLT).* Chandrasekhar et al. [10] propose a transform coding scheme for local descriptors. The Karhunen-Loeve Transform matrix $\mathbf{T} \in \mathbb{R}^{64 \times 64}$ is determined based on descriptors collected from a large set of training images. Let $\mathbf{c}_i = \mathbf{T}\mathbf{d}_i$ denote the descriptor in the transform domain, and $\tilde{\mathbf{c}}_i$ the result of scalar quantization. Similarly to the case above, the rate allocated to each dimension is determined to be equal to $R_j^{\Delta} = H(\tilde{\mathcal{C}}_j^{\Delta})$, where $\mathcal{C}_j = \{c_{1,j}, c_{2,j}, \ldots, c_{M,j}\}$.
- *Sparse coding (k-SVD).* Descriptors can be compressed exploiting sparse coding techniques relying on redundant

representation of signals. We use the k-SVD algorithm, originally presented in [16], in order to learn a dictionary $\mathbf{D} \in \mathbb{R}^{64\times N}$. A sparse representation of each descriptor $\mathbf{d}_i$ is then obtained solving the following optimization problem:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}} ||\mathbf{D}\mathbf{a} - \mathbf{d}_i||_2^2 \text{ subject to } ||\mathbf{a}||_0 \leq l \quad (2)$$

That is, every descriptors $\mathbf{d}_i$ is approximated by means of a linear combination of l columns (a.k.a. atoms) of the redundant dictionary $\mathbf{D}$. In this work, we trained a dictionary composed by $N = 512$ atoms starting from a large set of image descriptors. When a solution to (2) is found, only the indexes of the selected atoms, together with their quantized magnitudes need to be transmitted. In order to control the allocated rate, we fix the quantization step size Δ and we let the number of atoms l vary. We found this approach more flexible than arbitrarily setting l and varying Δ. Then, the estimated rate for the k-SVD scheme is $R^l = l \cdot (\lceil log_2 N \rceil + H(\tilde{\mathcal{A}}))$, where $\tilde{\mathcal{A}}$ is the discrete memoryless source representing the quantized atoms weights.

B. Inter-descriptor coding:

- *Closed loop predictive coding (DPCM).* The basic scalar quantization (PCM) scheme is improved in [8] exploiting inter-descriptor correlation. Since the accuracy of the retrieval task is independent from the order of the received descriptors, they can be reordered such that the sum of inter-descriptor distances is minimized. Here, we adopt the same greedy algorithm as described in [8] to sort the descriptors, since optimal ordering is NP-hard. Note that the ordering does not need to be explicitly transmitted to the decoder. Then, for each dimension j, the sorted set $\mathcal{S}_j = \{d_{s(1),j}, d_{s(2),j}, \ldots, d_{s(M),j}\}$ is encoded with DPCM, in which the predictor is set to be equal to the previously encoded dexel, i.e., $\hat{d}_{s(i),j} = \tilde{d}_{s(i-1),j}$. The same quantization step size Δ is used for all dimensions, whereas the rate is set to be equal to $R_j^\Delta = H(\tilde{\mathcal{E}}_j^\Delta)$, where $\tilde{\mathcal{E}}_j^\Delta$ is the set of quantized prediction residuals.
- *Open loop predictive coding (TREE).* Inter-descriptor redundancy can be exploited by computing the pairwise similarities between descriptors, and leveraging a lifting-based transform coding scheme on trees [17]. Specifically, each descriptor is represented as nodes in a fully connected graph G, where the weight assigned to each edge is equal to the Euclidean distance between the descriptors associated to the linked nodes. Then, we compute the minimum spanning tree (MST) T of G, so that the coding scheme described in [17] can be applied. Lifting consists of K levels of decomposition. At each level k we divide T in two disjoint sets of neighboring nodes. Odd depth nodes ($n \in P_k$) are predicted using even depth nodes ($m \in U_k$) in the tree, according to the following equations:

$$\boldsymbol{\delta}_{n,k} = \boldsymbol{\sigma}_{n,k-1} - \sum_{m\in N(n,k)} p_k(n)\boldsymbol{\sigma}_{m,k}, \forall n \in P_k$$
$$\boldsymbol{\sigma}_{m,k} = \boldsymbol{\sigma}_{m,k-1} + \sum_{n\in N(m,k)} u_k(m)\boldsymbol{\delta}_{n,k}, \forall m \in U_k$$

where $\boldsymbol{\delta}_{n,k}$ and $\boldsymbol{\sigma}_{n,k}$ are the detail and approximation coefficients for node n at level k, $N(n,k)$ is the set of neighboring nodes of node n at level k, and p_k, u_k are the prediction and update operators, defined as:

$$p_k(n) = \frac{1}{|N(n,k)|}$$
$$u_k(m) = \frac{1}{2|N(m,k)|}$$

We initialize the transform assigning $\boldsymbol{\sigma}_{i,0} = \mathbf{d}_i, \forall i$. Then, at level $k+1$, a new graph G_{k+1} and MST T_{k+1} are created from the even depth nodes in T_k. The lifting scheme is critically sampled and perfect reconstruction is guaranteed by construction. The inverse transform is given by:

$$\boldsymbol{\sigma}_{m,k-1} = \boldsymbol{\sigma}_{m,k} - \sum_{n\in N(m,k)} u_k(m)\boldsymbol{\delta}_{n,k}, \forall m \in U_k$$
$$\boldsymbol{\sigma}_{n,k} = \boldsymbol{\delta}_{n,k} + \sum_{m\in N(n,k)} p_k(n)\boldsymbol{\sigma}_{m,k}, \forall n \in P_k,$$

with $k = K, \ldots, 1$. Note that, in order to invert the transform, the decoder needs to know: (i) the detail coefficients for each decomposition level, i.e., $\boldsymbol{\delta}_{n,k}$; (ii) the approximation coefficients for the last level of decomposition, i.e., $\boldsymbol{\sigma}_{m,K}$; and (iii) the sets of neighboring nodes for each level, i.e., all the spanning trees T_k with $k = 0, \ldots, K-1$. We quantize both approximation and detail coefficients at each level with the same step size Δ, leading to the quantized sets $\tilde{\boldsymbol{\sigma}}$ and $\tilde{\boldsymbol{\delta}}$. Then, we assume the adoption of a different entropy code for each level and each dimension. Let $\tilde{\mathcal{P}}_{j,k}^\Delta = \{\tilde{\delta}_{1,j,k}, \tilde{\delta}_{2,j,k}, \ldots, \tilde{\delta}_{M_P,j,k}\}$ and $\tilde{\mathcal{U}}_j^\Delta = \{\tilde{\sigma}_{1,j,K}, \tilde{\sigma}_{2,j,K}, \ldots, \tilde{\sigma}_{M_U,j,K}\}$ denote the sets of quantized lifting coefficients, where M_U and M_P are the number of nodes in the prediction and update sets at each level k. We compute the rate as $R_j^\Delta = H(\tilde{\mathcal{U}}_j^\Delta) + \sum_{k=1}^K H(\tilde{\mathcal{P}}_{j,k}^\Delta)$. In contrast with the DPCM scheme, in this case the encoder needs to explicitly send the structure of the transform, i.e., all the MSTs T_k, with $k = 0, \ldots, K-1$. For a given set of M_I descriptors, we compute the cost of transmitting the MSTs as:

$$C = M_I\lceil \log_2(M_I)\rceil + \sum_{k=1}^{K-1} M_{I,k}\lceil \log_2(M_I)\rceil, \quad (3)$$

where $M_{I,k}$ is the number of nodes at the k-th level of decomposition. In equation (3), we are assuming to transmit a tree of M_I nodes as a sequence $P(i), i = 1, \ldots, M_I$, where $P(i)$ is the index of the parent node of i.

C. Hybrid coding:

It is possible to fuse inter- and intra-descriptor coding to exploit the benefits of the two approaches. First, we transform the descriptors in the KLT domain. Then, we apply one of the aforementioned methods to address inter-descriptor redundancy. This method gives birth to two hybrid coding schemes, namely KLT-DPCM and KLT-TREE. Also, we propose a further modification, which takes into account the unequal energy distribution in the KLT domain and the different degree of inter-descriptor redundancy for the same KLT coefficient $c_{\cdot,j}$ across different descriptors. Therefore, we limit both inter-descriptor coding schemes to the first p KLT coefficients. Instead, the remaining $N-p$ coefficients of each descriptor are computed using scalar quantization (PCM). Hence, the rate for the hybrid scheme based on DPCM (called KLT-DPCM-p) is:

$$R_j^{\Delta} = \begin{cases} H(\tilde{\mathcal{E}}_j^{\Delta}), \text{ for } j = 1, \ldots, p \\ H(\tilde{\mathcal{C}}_j^{\Delta}), \text{ for } j = p+1, \ldots, N \end{cases} \quad (4)$$

where $\tilde{\mathcal{E}}_j$ represents the set of the quantized prediction residuals (now in the KLT domain) and $\tilde{\mathcal{C}}_j$ the quantized KLT coefficients. Similarly, for the hybrid scheme based on lifting (KLT-TREE-p), we compute the rate as:

$$R_j^{\Delta} = \begin{cases} H(\tilde{\mathcal{U}}_j^{\Delta}) + \sum_{k=1}^{K} H(\tilde{\mathcal{P}}_{j,k}^{\Delta}), \text{ for } j = 1 \ldots p \\ H(\tilde{\mathcal{C}}_j^{\Delta}), \text{ for } j = p+1 \ldots N \end{cases}, \quad (5)$$

where $\mathcal{U}_j^{\Delta}$ and $\mathcal{P}_{j,k}^{\Delta}$ now represent quantities expressed in the KLT domain. Obviously, when $p = 0$, KLT-DPCM-p and KLT-TREE-p are identical to the KLT scheme. As for the KLT-TREE-p scheme, the additional rate due to encoding the tree structures needs to be taken into account in the overall rate.

IV. Experimental results

In this section we show an experimental evaluation of the aforementioned compression schemes, considering results in terms of both rate-distortion and rate-accuracy, for the task of object recognition. All the schemes were tested on two publicly available datasets, characterized by different kind of contents, namely: (i) the ZuBuD dataset [18], and the (ii) 53 objects dataset [19] (Obj-53). The ZuBuD dataset contains 1005 color images of 201 buildings of the city of Zurich. Each building has five images, taken at random arbitrary view points under different seasons, weather conditions and by two different cameras. A separate archive containing 115 images of the same buildings (with different imaging conditions) is available as query dataset. The Obj-53 dataset contains 265 images of 53 different objects (toys, bottles, boxes, etc...). Again, for each objects there are five images taken at random view points and with different lighting condition. A dedicated query dataset is not available, so one image for object is used as query (removing it from the matching dataset). We chose such two different datasets in order to study the sensitivity of results to content.

For those schemes requiring parameter tuning, we carried out extensive simulations, and selected those parameters giving the best results in the average case. Specifically, in all open loop predictive coding schemes based on lifting (i.e., TREE, KLT-TREE and KLT-TREE-p), we set the maximum number of decomposition levels $K = 3$. Similarly, we selected $p = 9$ for KLT-DPCM-p and $p = 7$ for KLT-TREE-p. For those schemes adopting the KLT, the transform was trained using 10000 descriptors randomly selected from the tested datasets. The same set of descriptors is also used to build the redundant dictionary in the k-SVD approach.

A. Rate-distortion analysis

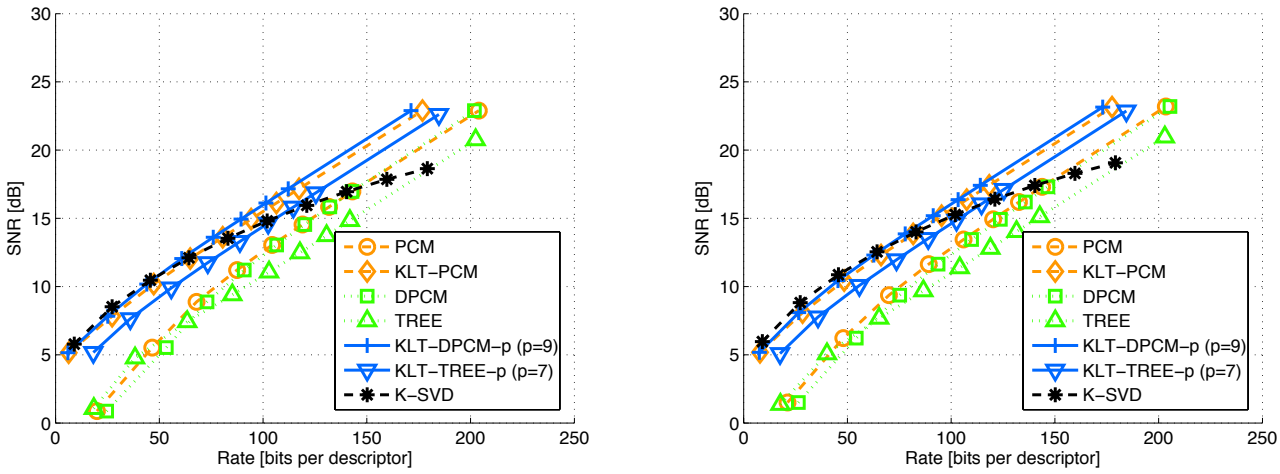
The rate-distortion analysis was carried out similarly for both datasets. We randomly selected 100 query images for the ZuBuD dataset and 50 for the Obj-53 dataset. For each image, we extracted SURF descriptors and compressed them with the different schemes. We limited to 250 the number of descriptors to be used, sorted by descending values of their associated Hessian response, so as to select the descriptors associated with the most stable keypoints. Distortion is measured using the quantization signal-to-noise ratio (SNR). Rate is indirectly controlled by varying the quantization step size Δ.

The operational rate-distortion curves are shown in Figure 1(a) and Figure 1(b) for the ZuBuD and the Obj-53 datasets, respectively. We are particularly interested in the range of bitrates for which SNR is smaller than 15 dB. Indeed, as shown below, this was found to be the distortion limit above which rate-accuracy curves saturate. In other words, no further performance increase is observed when distortion is decreased.

First, we point out that the rate-distortion curves share the same behavior across the two datasets, which means that the results of the analysis are content-independent. Then, we notice that methods addressing intra-descriptor correlation outperform those addressing only inter-descriptor correlation. This is clearly demonstrated by the gap between the three curves in the original domain (PCM, DPCM and TREE) and the corresponding ones in the transform domain (KLT-PCM, KLT-DPCM-p and KLT-TREE-p). As for inter-descriptor correlation, we point out that TREE and KLT-TREE-p are strongly penalized by the additional cost of transmitting the spanning trees at each level. This is the reason why their performance is generally surpassed by both PCM and DPCM based methods. Conversely, for DPCM and KLT-DPCM-p encoding is sequential. Hence descriptor order does not need to be transmitted to enable decoder-side reconstruction. We do not show the curves for KLT-DPCM and KLT-TREE (i.e., when $p = 64$) to avoid cluttering the figure, since they are outperformed respectively by KLT-TREE-p and KLT-DPCM-p. Finally, we point out that the k-SVD approach outperforms all other schemes at very low bitrates in terms of distortion.

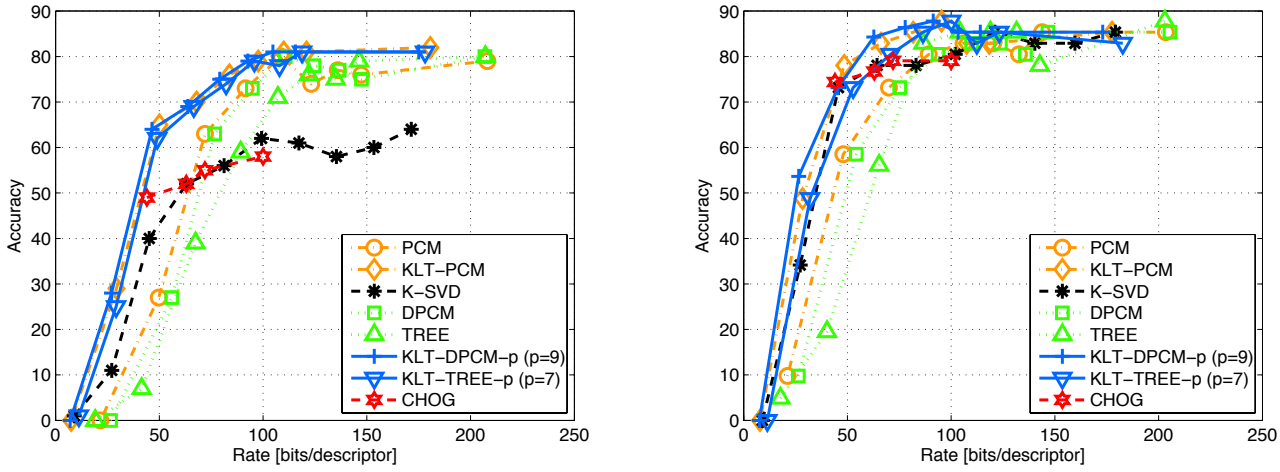
B. Rate-accuracy analysis

In order to evaluate the rate-accuracy performance of the aforementioned coding schemes, we implemented a state-of-the-art image retrieval system for the task of object recognition. For each query image, we repeated the same extraction and coding process as for the rate-distortion experiment. Then,



(a) (b)

Fig. 1. Rate-distortion curves for (a) ZuBuD and (b) Obj-53 dataset



(a) (b)

Fig. 2. Rate-accuracy curves for (a) ZuBuD and (b) Obj-53 dataset

for each coding method, we performed pairwise matching between the encoded descriptors of the query and those of each image in the database, assigning to each descriptor of the query its nearest neighbor (using the Euclidean distance as distance metric). Robust matching is achieved using two standard techniques, namely (i) the ratio test and (ii) a geometric consistency check with RANSAC. Both of them greatly reduce the number of mismatches, improving the retrieval performance of the system. Finally, we computed the value of accuracy as the number of correctly answered queries over the total number of queries. An answer is deemed correct whenever the image with the largest number of matching descriptors (i.e., the top-1 image) depicts the same object as the query. In addition, we compared all the techniques with the Compressed Histograms of Gradients (CHoG) approach. In this case we used the Kullback-Leibler divergence as distance metric between descriptors, as suggested by the authors [7].

Figures 2(a) and Figure 2(b) show the operational rate-accuracy curves for the ZuBuD and the Obj-53 datasets, respectively. First, as expected from the analysis of the rate-distortion curves, methods exploiting intra-descriptor redundancy (i.e., KLT-domain approaches) outperform methods leveraging only inter-descriptor coding. Second, the proposed hybrid method KLT-DPCM-p outperforms all other approaches at low bitrates. For example, at a target accuracy greater than 80% (i.e, at PCM-coding rate equal to 100 bits/descriptor for ZuBuD and 90 bits/descriptor for Obj-53), the KLT-DPCM-p approach exhibits a rate saving as large as 15 bits/descriptor for the ZuBuD dataset and 25 bits/descriptor for the Obj-53 dataset. That is, using the proposed approach can save

up to 15-30% with respect to PCM coding. The difference in performance between datasets is mainly due to the size of the database, which is greater for ZuBuD. The gain with respect to KLT-PCM is smaller (up to 3%), demonstrating the fact that most of the coding efficiency is obtained by means of intra-descriptor coding. Indeed, in the case of methods exploiting only inter-descriptor correlation (e.g., DPCM and TREE), the performance is worse than simple PCM coding, somehow contradicting the results previously appeared in the literature [8]. Despite using a different dataset, since the one adopted in [8] is not publicly available, we believe that this is due to the fact that entropy coding is explicitly addressed in our work.

A comparison of different coding schemes in terms of rate-distortion might not lead to the same conclusions when rate-accuracy is evaluated. Indeed, the k-SVD approach was shown to be the best performing scheme in Figure 1(a) and Figure 1(b) at low bitrates. Conversely, it is outperformed by other schemes in terms of accuracy. Finally, we found that for the specific task evaluated in this work, CHoG does not lead to the best accuracy.

V. Conclusion and future works

We have investigated several low bitrate coding techniques applied to SURF local descriptors. The results of our experiments show that intra-descriptor coding performs better than other approaches, both in terms of rate-distortion and rate-accuracy. Conversely, the exploitation of inter-descriptor redundancy alone does not seem to provide benefits, although little gains can be obtained in hybrid schemes. We have compared all the tested techniques with the CHoG approach, showing that, for the task considered in this work, optimizing the coding of existing descriptors (e.g, SURF) might perform better than using descriptors specifically designed for low bitrate applications.

References

[1] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999, pp. 1150–1157.

[2] J.-Y. Choi, K.-S. Sung, and Y.-K. Yang, "Multiple vehicles detection and tracking based on scale-invariant feature transform," in *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, 30 2007-oct. 3 2007, pp. 528 –533.

[3] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 4, pp. 712 –727, april 2008.

[4] B. Girod, V. Chandrasekhar, D. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham, "Mobile visual search," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 61 – 76, july 2011.

[5] H. Bay, T. Tuytelaars, and L. J. V. Gool, "Surf: Speeded up robust features," in *ECCV (1)*, 2006, pp. 404–417.

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

[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 *CVPR*, 2009, pp. 2504–2511.

[8] R. J. H. Y. J. Chen, L.-Y. Duan and W. Gao, "Sorting local descriptors for low bit rate mobile visual search," *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2011.

[9] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. P. Singh, and B. Girod, "Tree histogram coding for mobile image matching," in *DCC*, 2009, pp. 143–152.

[10] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod, "Transform coding of image feature descriptors," M. Rabbani and R. L. Stevenson, Eds., vol. 7257, no. 1, 2009. [Online]. Available: http://dx.doi.org/10.1117/12.805982

[11] R. J. H. Y. J. Chen, L.Y. Duan and W. Gao, "Sorting local descriptors for low bit rate mobile visual search," *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2011.

[12] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *CVPR (2)*, 2004, pp. 506–513.

[13] 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 *CVPR*, 2009, pp. 2504–2511.

[14] V. R. Chandrasekhar, S. S. Tsai, G. Takacs, D. M. Chen, N.-M. Cheung, Y. Reznik, R. Vedantham, R. Grzeszczuk, and B. Girod, "Low latency image retrieval with progressive transmission of chog descriptors," in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*, ser. MCMC '10, 2010, pp. 41–46. [Online]. Available: http://doi.acm.org/10.1145/1877953.1877966

[15] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *ECCV (4)*, 2010, pp. 778–792.

[16] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311 –4322, nov. 2006.

[17] G. Shen and A. Ortega, "Tree-based wavelets for image coding: orthogonalization and tree selection," in *Proceedings of the 27th conference on Picture Coding Symposium*, ser. PCS'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 265–268. [Online]. Available: http://dl.acm.org/citation.cfm?id=1690059.1690126

[18] H. Shao, T. Svoboda, and L. Van Gool, "ZuBuD — Zurich buildings database for image based recognition," Computer Vision Laboratory, Swiss Federal Institute of Technology, Tech. Rep. 260, Mar. 2003.

[19] (2003, Mar.). [Online]. Available: http://www.vision.ee.ethz.ch/showroom/zubud/