CODING VIDEO SEQUENCES OF VISUAL FEATURES

Luca Baroffio, Matteo Cesana, Alessandro Redondi, Stefano Tubaro, Marco Tagliasacchi

Dipartimento di Elettronica e Informazione, Politecnico di Milano

ABSTRACT

Visual features provide a convenient representation of the image content, which is exploited in several applications, e.g., visual search, object tracking, etc. In several cases, visual features need to be transmitted over a bandwidth-limited network, thus calling for coding techniques to reduce the required rate, while attaining a target efficiency for the task at hand. Although the literature has recently addressed the problem of coding local features extracted from still images, in this paper we propose, for the first time, a coding architecture designed for local features extracted from video content. We exploit both spatial and temporal redundancy by means of intra-frame and inter-frame coding modes. In addition, we propose a coding mode decision based on rate-distortion optimization. Experimental results demonstrate that, in the case of SIFT descriptors, exploiting temporal redundancy leads to substantial gains in terms of coding efficiency.

Index Terms— Visual features, video coding.

1. INTRODUCTION

Visual features are effectively used in many tasks, ranging from image/video retrieval, object recognition, object tracking, image registration, structure-from-motion, etc. In the last few years, local features have attracted the interest of the scientific community, due to their ability to provide a concise and robust representation of the underlying image content. This is achieved by identifying a set of salient keypoints by means of a detector and, for each keypoint, a descriptor is computed representing the content of the image patch around the keypoint.

In some applications, visual features need to be transmitted over a bandwidth-limited network. This is the case, for example, in scenarios that include mobile visual search [1] or wireless multimedia sensor networks [2]. In both cases, a device with an embedded camera computes a set of visual features of a still image and performs compression so as to minimize the number of bits to represent them. This has stimulated quite a few works in this area, which aim at efficiently encoding existing state-of-the-art descriptors (e.g., SIFT or SURF [3, 4]), or at revisiting the design of existing descriptors, thus leading to a representation that is more suitable for compression (e.g., CHoG [5]). In this context, an MPEG ad-hoc group on Compact Descriptors for Visual Search (CDVS) is currently working towards the definition of a standard covering this scenario [6]. In parallel, the design of binary descriptors (e.g., BRISK [7], FREAK [8], DBRIEF [9]) is being investigated, which are meant to be both compact and fast to compute. However, they are not able to achieve the same performance as SIFT [10] in terms of matching precision.

The project GreenEyes acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number:296676.

In the case of video content, local features can be extracted on a frame-by-frame basis. However, matching sets of visual features is expensive in terms of computational resources, and it does not scale with the size of the collection. For this reason, in the context of content-based video retrieval, more compact representations are built, which map each frame to a fixed-dimensional vector of visual words [11], so that multi-dimensional indexing schemes can be adopted. The use of bag-of-visual words decreases the computational complexity at the cost of reduced precision. Therefore, a re-ranking step restricted to the top-k results is appended at the end of the retrieval pipeline, taking advantage of the full set of local features [12]. In other applications, such as object tracking or structure-from-motion, the bag-of-visual words representation is inappropriate, since it disregards the underlying spatial configuration of the local features. Conversely, matching needs to be performed between sets of keypoints and their corresponding descriptors, calling for the processing of the full set of local features.

These exemplary application scenarios reveal that it might be necessary to store or transmit local features extracted from video sequences. Due to the large number of features per frame, large volumes of data are necessarily generated. Therefore, the investigation of suitable coding schemes is even more important than in the case of still images. As mentioned before, in the literature, coding has been explored for the case of features extracted from a single image, thus leveraging redundancy within the same descriptor [3], or among descriptors of the same image [4]. Instead, in this paper we propose a coding architecture designed for local features extracted from video sequences. To the best of the authors' knowledge, this has never been done before. Specifically, we exploit both spatial and temporal redundancy by means of intra-frame and inter-frame coding modes, respectively. In addition, inspired to the best practices in traditional video coding, we propose an algorithm for coding mode decision which is based on rate-distortion optimization. That is, it explicitly takes into account not only the distortion due to quantization, but also the rate needed to encode a descriptor according to the different coding modes.

The rest of this paper is organized as follows. Section 2 introduces the problem, defining the properties of the source to be coded. Section 3 illustrates the proposed coding architecture for both intra- and inter-frame coding. Experimental results on real video sequences are reported in Section 4 and conclusions in Section 5.

2. PROBLEM STATEMENT

Let $\mathcal{I}_n$ denote the n-th frame of a video sequence of size $N_x \times N_y$, which is processed to extract a set of local features $\mathcal{D}_n$. First, a scale invariant detector is applied, to identify stable keypoints in the scale-space domain. The number of detected keypoints $M_n = |\mathcal{D}_n|$ depends on both the image content and on the type and parameters of the adopted detector. Then, the (oriented) patches around the detected keypoints are further processed to compute the correspond-

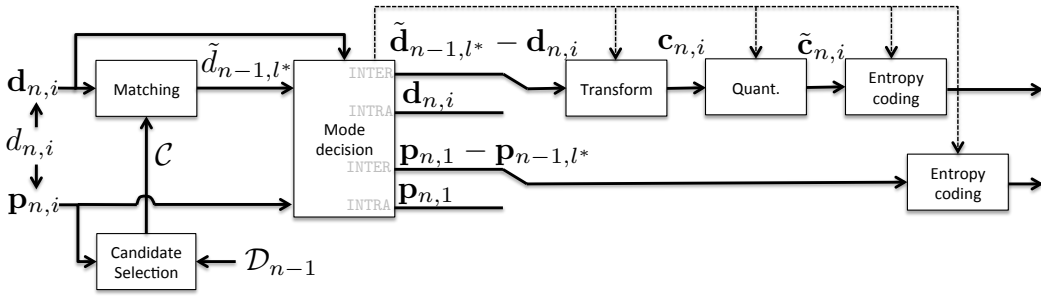

Fig. 1. Block diagram of the proposed coding architecture.

ing descriptors. Hence, each element of $d_{n,i} \in \mathcal{D}_n$ is a visual feature, which consists of two components: i) a 4-dimensional vector $\mathbf{p}_{n,i} = [x, y, \sigma, \theta]^T$, indicating the position (x, y), the scale σ of the detected keypoint, and the orientation angle θ of the image patch; ii) a P-dimensional vector $\mathbf{d}_{n,i}$, which represents the descriptor associated to the keypoint $\mathbf{p}_{n,i}$.

In this paper, we propose a coding architecture which aims at efficiently coding the sequence $\{\mathcal{D}_n\}_{n=1}^N$ of descriptors. Specifically, we consider lossy coding techniques that enable to reconstruct, at the decoder, an approximation $\tilde{\mathcal{D}}_n$ of the local features extracted from $\mathcal{I}_n$. Each decoded descriptor can be written as $\tilde{d}_{i,n} = \{\tilde{\mathbf{p}}_{n,i}, \tilde{\mathbf{d}}_{n,i}\}$. The number of bits necessary to encode the visual features of frame $\mathcal{I}_n$ is

$$R_n = \sum_{i=1}^{M_n} R^c_{n,i} + R^d_{n,i}. \tag{1}$$

That is, we consider the rate used to represent both the location of the keypoint, $R^c_{n,i}$, and the descriptor itself, $R^d_{n,i}$. The distortion is measured in terms of the mean square error between the original and decoded descriptor, averaged over the descriptors extracted from $\mathcal{I}_n$

$$D_n = \frac{1}{M_n P} \sum_{i=1}^{M_n} \|\tilde{\mathbf{d}}_{i,n} - \mathbf{d}_{i,n}\|_2^2, \tag{2}$$

where $\|\cdot\|_2$ denotes the l-2 norm. As for the component $\tilde{\mathbf{p}}_{n,i}$, we decided to encode the coordinates of the keypoint and its scale, i.e., $\tilde{\mathbf{p}}_{n,i} = [\tilde{x}, \tilde{y}, \tilde{\sigma}]^T$. At the decoder, this information is necessary when the matching score between image pairs is computed based on the number of matches that pass the spatial verification step using RANSAC [12]. Although most of the detectors produce as output coordinates represented in floating point precision, we decided to round the coordinates to quarter-pixel precision, which is typically sufficient for spatial verification. The scale parameter is also quantized with a step equal to 0.25. The vector $\tilde{\mathbf{p}}_{n,i}$ does not contain the orientation of the keypoint, as it is not employed to compute the matching score. Note that it might be necessary when using alternative spatial verification schemes, e.g., when weak geometry checking [13] is enforced.

The main contribution of this paper is the investigation of an inter-frame coding scheme, which aims at exploiting the temporal redundancy in sets of local features extracted from consecutive video frames. The same coding architecture can be adapted in a straightforward manner to encode sets of descriptors acquired from multiple cameras observing the same scene. In Section 3 we provide the details of the proposed coding architecture, which leverages some of the coding tools that are successfully employed in traditional video coding.

3. CODING OF LOCAL FEATURES

3.1. Intra-frame coding

The simplest coding scheme consists in performing a frame-by-frame processing, in which the local features extracted from frame $\mathcal{I}_n$ are encoded independently from those extracted from other frames. In our baseline approach, each dexel (descriptor element) of $\mathbf{d}_{n,i}$ is encoded by applying scalar quantization with step size Δ_j. That is,

$$\tilde{d}_{n,i,j} = \Delta_j \cdot \text{round}(d_{n,i,j}/\Delta_j) \tag{3}$$

Here, we fix the same quantization step size for all dexels, i.e., $\Delta_j = \Delta$, $j = 1, \ldots, P$.

In our previous work [4], we studied and compared different intra-frame coding schemes for SURF features. We observed that a significant coding gain was achieved by exploiting the inherent redundancy among the dexels of the same descriptor, confirming the findings obtained by Chandrasekhar et al. [3]. For SIFT, a coding gain was observed only at low bitrates [3]. Hence, we consider an intra-descriptor coding scheme that applies the Karhunen-Loève Transform matrix $\mathbf{T} \in \mathbb{R}^{P \times P}$ to each descriptor $\mathbf{d}_{n,i}$. The matrix $\mathbf{T}$ is determined based on the descriptors collected from a large set of training images. Let $\mathbf{c}_{n,i} = \mathbf{T}\mathbf{d}_{n,i}$ denote the descriptor in the transform domain, and $\tilde{\mathbf{c}}_{n,i}$ the result of scalar quantization. Similarly to the case above, the output symbols of the quantizer are entropy coded. In this paper, we do not consider inter-description coding scheme within the same frame, since we showed in [4] that it does not bring significant coding gains.

The output symbols of the quantizer are entropy coded, e.g., using arithmetic coding, using $R^d_{n,i}$ bits. The probabilities of the symbols used by the entropy coder are learned from descriptors extracted from a training set images for each dexel j and different values of Δ. In addition, the coordinates of each keypoint are encoded using $R^c_n \simeq M_n \cdot (\log_2 4N_x + \log_2 4N_y + S)$ bits, where S is the number of bits use to encode the scale parameter.

3.2. Inter-frame coding

In the case of inter-frame coding, we consider the set of local features extracted from a reference frame when encoding the set $\mathcal{D}_n$. In this work, we consider the features extracted from the previous frame, i.e., $\mathcal{D}_{n-1}$, thus mimicking P-frames in traditional video coding. For each descriptor $d_{n,i}$, $i = 1, \ldots, M_n$, encoding proceeds as follows:

- *Motion estimation*: Compute the best matching descriptor in the reference frame, i.e.,

$$\mathbf{d}_{n-1,l^*} = \arg\min_{l \in \mathcal{C}} J^{\text{INTER}}(\mathbf{d}_{n,i}, \mathbf{d}_{n-1,l}), \tag{4}$$

where

$$J^{\text{INTER}}(\mathbf{d}_{n,i}, \mathbf{d}_{n-1,l}) = \frac{1}{P}\|\mathbf{d}_{n,i} - \mathbf{d}_{n-1,l}\|_2 + \lambda R_{n,i}^{c,\text{INTER}}(l) \quad (5)$$

The set $\mathcal{C}$ contains the indexes of the local features in the reference frame to be used as candidate predictors. In this work, $\mathcal{C}$ is populated with the local features whose coordinates are in a neighborhood of $d_{n,i}$, in a search range of $(\pm\Delta x, \pm\Delta y)$ and whose scale is in a range of $\pm\Delta\sigma$. In the objective function in (5), the first term represents the square root of the mean square error of the prediction residuals, whereas the second term is a penalty term $R_{n,i}^{\text{INTER}}(l)$ that takes into account the rate needed to encode the position of the keypoint associated to $d_{n,i}$ relative to $d_{n-1,l}$. This includes a fixed number of bits equal to $\lceil \log_2 |\mathcal{D}_{n-1}| \rceil$ necessary to index $l \in \mathcal{D}_{n-1}$, as well as the bits used to entropy code the differences $\tilde{\mathbf{p}}_{n,i} - \tilde{\mathbf{p}}_{n-1,l}$, i.e., the equivalent of motion vectors in traditional video coding.

- *Coding mode decision*: Compare the cost of inter-frame coding with that of intra-frame coding, which can be expressed as

$$J^{\text{INTRA}}(\mathbf{d}_{n,i}) = \frac{1}{P}\|\mathbf{d}_{n,i}\|_2 + \lambda(\log_2 4N_x + \log_2 4N_y + S) \quad (6)$$

If $J^{\text{INTER}}(\mathbf{d}_{n,i}, \mathbf{d}_{n-1,l^*}) < J^{\text{INTRA}}(\mathbf{d}_{n,i})$, then proceed with inter-frame coding. Otherwise, encode the descriptor in intra-frame mode.

- *Intra-descriptor transform*: Depending on the selected coding mode, compute either $\mathbf{c}_{n,i} = \mathbf{T}^{\text{INTRA}}\mathbf{d}_{n,i}$ or $\mathbf{c}_{n,i} = \mathbf{T}^{\text{INTER}}(\mathbf{d}_{n,i} - \tilde{\mathbf{d}}_{n-1,l^*})$. When no transform is used, $\mathbf{T}^{\text{INTRA}} = \mathbf{T}^{\text{INTER}} = \mathbf{I}$. Otherwise, two different KLT transform matrices are estimated from training data.

- *Quantization*: Scalar quantization with fixed step size Δ is applied to the P elements of $\mathbf{c}_{n,i}$.

- *Entropy coding*: In the case of intra-frame coding, entropy coding proceeds as described in Section 3.1. Otherwise, for inter-frame coding, it is necessary to encode: i) the identifier of the matching keypoint in the reference frame and the position and scale of the keypoint with respect to the matching keypoint, which require $R_{n,i}^{\text{INTER}}(l^*)$ bits; ii) the quantized elements of $\tilde{\mathbf{c}}_{n,i}$. For the latter, the probabilities of the symbols used for entropy coding are learned from a training set of images. To this end, we considered only descriptors for which a good match was found, i.e., $\|\mathbf{d}_{n,i} - \mathbf{d}_{n-1,l}\|_2 < \|\mathbf{d}_{n,i}\|_2$. For each possible value of the quantization step size Δ, we computed the quantized prediction residuals $\tilde{\mathbf{c}}_{n,i}$, possibly after an inter-descriptor transform $\mathbf{T}^{\text{INTER}}$. For each of the P dexels, we estimated the probability of the symbols observing the number of occurrences of each of the possible reconstruction levels of the quantizer.

4. EXPERIMENTS

In order to evaluate the proposed coding architecture, we extracted SIFT visual features from a set of six video sequences at CIF resolution (352×288) and 30 fps, namely *Foreman*, *Mobile*, *Hall*, *Paris*, *News* and *Mother*, each with 300 frames, which are characterized by diverse visual content and motion characteristics [14]. We adopted the software implementation provided by VLFEAT [15] for both the detector and the descriptor, which provides results very similar to the original implementation [10].

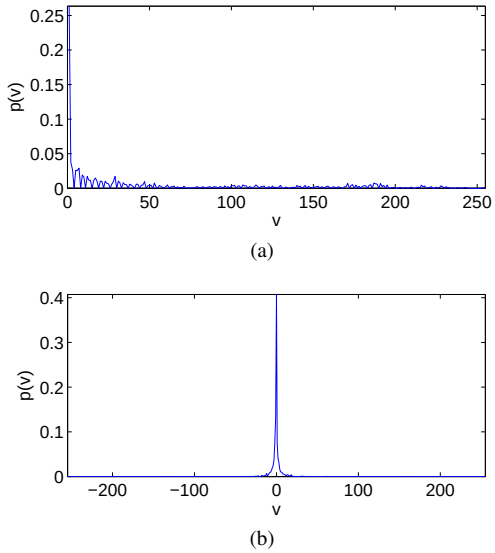

Fig. 2. Empirical distribution of: a) 77-th element of the SIFT descriptor; b) Corresponding prediction residuals.

Table 1. Average number of keypoints and minimum bitrate to achieve $SNR \geq 15$dB

	Foreman	*Hall*	*Mobile*
Num. of keypoints / frame	177.0	117.0	185.2
Uncompressed (kbps)	5590	3700	5850
`INTRA` (kbps)	1180	790	1130
`INTER` (kbps)	378	177	632
`INTRA/INTER` (kbps)	269	124	435
`INTRA` + $\mathbf{T}$ (kbps)	1710	1090	1680
`INTER`+ $\mathbf{T}$ (kbps)	533	235	845
`INTRA/INTER` + $\mathbf{T}$ (kbps)	345	144	489

Table 1 reports the average number of keypoints extracted from the video sequences, together with the corresponding bitrate in the case each descriptor element is represented using 8 bits. We split the data set into training (*Paris*, *News* and *Mother*) and testing (*Foreman*, *Mobile*, *Hall*) sequences. We used the descriptors extracted from the training set to learn the statistics of the symbols given as input to the entropy coder (which are stored in a lookup table both at the encoder and at the decoder), as well as to learn the transform. In the case of intra-frame coding, we considered a set of target quantization step sizes $\Delta \in \{170, 150, 120, \ldots, 2\}$. For each value of Δ, we estimated the probability of occurrence of the symbols obtained quantizing each of the $P = 128$ elements of the descriptor. The transform $\mathbf{T}^{\text{INTRA}}$ was estimated using the KLT, considering all descriptors extracted from the training set after subtracting the average of each element. As before, the statistics of the output of the quantizer are determined for each of the $P = 128$ transform coefficients, given a value of Δ.

In the case of inter-frame coding, we estimated the statistics of the differences between the keypoint locations as well as of the prediction residuals. To this end, we performed motion estimation as described in equations (4) and (5) on the uncompressed descriptors

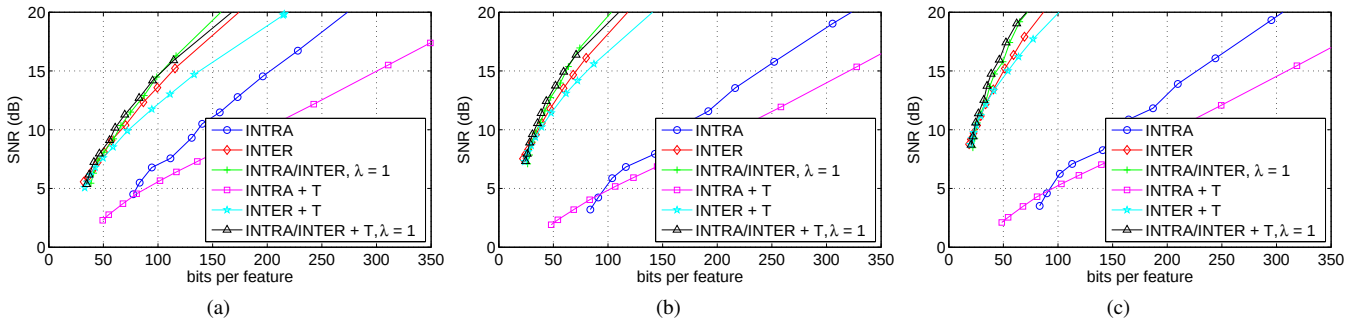

Fig. 3. Rate-distortion curves obtained with the proposed coding architecture. a) *Foreman*, b) *Mobile*, c) *Hall*.

extracted from the training sequences, setting $\lambda = 0$ and retaining only those descriptors for which $\|\mathbf{d}_{n,i} - \mathbf{d}_{n-1,l^*}\|_2^2 < \|\mathbf{d}_{n,i}\|_2^2$. Figure 2 shows, as an example, the statistics of the 77-th element of the SIFT descriptor and of the corresponding prediction residuals. We observe two facts: i) the variance of the prediction residuals is significantly smaller; ii) the distribution of the original descriptor element is rather complex and multi-modal. Conversely, the distribution of the prediction residuals can be approximately modeled as Laplacian. Similarly to the case of intra-frame coding, for each of the possible values of the quantization step size Δ, we estimated the probability of occurrence of the symbols obtained quantizing each of the $P = 128$ elements of the descriptor. The transform $\mathbf{T}^{\text{INTER}}$ and the corresponding statistics of the transformed prediction residuals are obtained following the same method as for intra-frame coding.

Figure 3(a), Figure 3(b) and Figure 3(c) show the rate-distortion curves obtained by encoding the visual features extracted from the test sequences *Foreman*, *Mobile* and *Hall*. The distortion is measured in terms of the signal-to-noise ratio (SNR), which is defined as

$$SNR = 10 \log_{10} \frac{\sum_{n=1}^{N} \sum_{i=1}^{M_n} \|\mathbf{d}_{n,i}\|_2^2}{\sum_{n=1}^{N} \sum_{i=1}^{M_n} \|\mathbf{d}_{n,i} - \tilde{\mathbf{d}}_{n,i}\|_2^2} \quad (7)$$

The rate includes the number of bits needed to encode both the locations of the keypoints and the descriptors and it is expressed in bits/feature. Regardless of the encoded content, inter-frame coding leads to significant coding gains with respect to intra-frame coding at all bitrates. It is particularly interesting to consider the bitrate necessary to achieve a SNR higher than 15dB. Indeed, in our previous work [4], we observed that in the case of object recognition performance saturates beyond this level. The number of bits per feature is reduced by 60%, 77% and 81% in *Foreman*, *Mobile* and *Hall*, respectively. The adoption of the coding mode decision, which determines if a descriptor is to be encoded with an inter-frame or intra-frame coding mode leads to a further rate reduction of 37%, 33% and 20% with respect to the case in which all descriptors are inter-frame coded. In our experiments, we fixed the value of $\lambda = 1$. We expect that additional gains can be achieved when adjusting λ depending on the target bitrate. This is left to future investigations. We also tested the coding efficiency when a transform is used to decorrelate the elements of the descriptor. In the case of intra-frame coding, this led to coding gains only at low bitrates. Conversely, in the case of inter-frame coding, the adoption of a transform did not seem to be beneficial. A similar observation was reported in [3], motivated by the fact that statistical distribution of the elements of SIFT descriptors does not follow a joint Gaussian distribution (see also Figure 2). As such, the KLT transform cannot be guaranteed to be optimal. In Table 1 we report the bitrate expressed in bits/second, which is obtained by considering the number of descriptors extracted from each sequence and the frame rate.

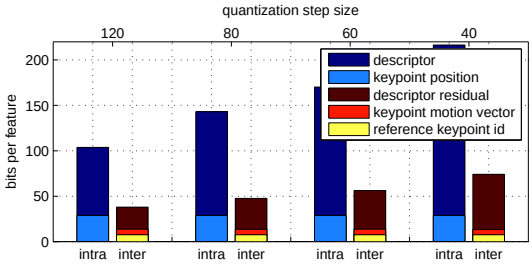

Fig. 4. Bit budget allocation between keypoint coordinates and descriptor elements.

In addition, we investigated the allocation of the bit budget between keypoint coordinates and descriptor elements. Figure 4 shows the allocation for different values of Δ, which correspond to a subset of the points below 15dB in the operational rate-distortion curve in Figure 3. In the case of inter-frame coding, the cost of encoding the keypoint coordinates includes the bits to represent the reference keypoint id, as well as the motion vectors. We observed that inter-frame coding reduced the number of bits necessary to represent both the keypoint coordinates and the descriptor elements

5. CONCLUSIONS

In this paper we presented a coding architecture that is designed for local visual features extracted from the frames of a video sequence. We considered both intra-frame and inter-frame coding, in order to exploit the redundancy also along the temporal dimension. We proposed a coding mode decision scheme, which is able to significantly improve the coding efficiency, by switching between intra- and inter-frame coding on a decriptor-by-descriptor basis. The proposed coding architecture is general and can be applied to any kind of local feature. In our experiments, we used SIFT features extracted from video sequences. In our future investigation we plan to use other types of local features, e.g., SURF, including the recent binary descriptors. We will also investigate more sophisticated mode decision strategies, e.g., by automatically determining the optimal value of λ depending on the target bitrate, leveraging the best practices adopted in video coding [16].

6. REFERENCES

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

[2] A. Redondi, M. Cesana, and M. Tagliasacchi, "Rate-accuracy optimization in visual wireless sensor networks," in *International Conference on Image Processing*, oct. 2012.

[3] Vijay Chandrasekhar, Gabriel Takacs, David Chen, Sam S. Tsai, Jatinder Singh, and Bernd Girod, "Transform coding of image feature descriptors," 2009, vol. 7257.

[4] A. Redondi, M. Cesana, and M. Tagliasacchi, "Low bitrate coding schemes for local image descriptors," in *International Workshop on Multimedia Signal Processing*, sept. 2012, pp. 124 –129.

[5] Vijay Chandrasekhar, Gabriel Takacs, David M. Chen, Sam S. Tsai, Radek Grzeszczuk, and Bernd Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," in *CVPR*, 2009, pp. 2504–2511.

[6] MPEG, "Compact descriptors for visual search," `http://mpeg.chiariglione.org/standards/mpeg-7/compact-descriptors-visual-search`.

[7] Stefan Leutenegger, Margarita Chli, and Roland Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *ICCV*, Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, Eds. 2011, pp. 2548–2555, IEEE.

[8] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst, "Freak: Fast retina keypoint," in *CVPR*. 2012, pp. 510–517, IEEE.

[9] Tomasz Trzcinski and Vincent Lepetit, "Efficient discriminative projections for compact binary descriptors," in *ECCV (1)*, Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, Eds. 2012, vol. 7572 of *Lecture Notes in Computer Science*, pp. 228–242, Springer.

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

[11] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. 2003, pp. 1470–1477, IEEE Computer Society.

[12] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*. 2007, IEEE Computer Society.

[13] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV (1)*, David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, Eds. 2008, vol. 5302 of *Lecture Notes in Computer Science*, pp. 304–317, Springer.

[14] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, march 2010, pp. 2430 –2433.

[15] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," `http://www.vlfeat.org/`, 2008.

[16] G.J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74 –90, nov 1998.