MULTI-VIEW DISTRIBUTED SOURCE CODING OF BINARY FEATURES FOR VISUAL SENSOR NETWORKS

Nuno Monteiro, Catarina Brites, Fernando Pereira, João Ascenso

Instituto de Telecomunicações – Instituto Superior Técnico

ABSTRACT

Visual analysis algorithms have been mostly developed for a centralized scenario where all visual data is acquired and processed at a central location. However, in visual sensor networks (VSN), several constraints in computational power, energy and bandwidth require a radically different approach, notably a paradigm shift from centralized to distributed visual processing. In the new paradigm, visual data is acquired and features are extracted at the sensing nodes locations to be after transmitted to enable further analysis at some central location. In such scenario, one of the key challenges is to design suitable feature coding schemes that are able to exploit the correlation among the features corresponding to (partially) overlapped views of the same visual scene. To achieve efficient coding, it is proposed to employ the distributed source coding paradigm as it does not require any communication between the sensing nodes (rather expensive in VSN) and it is parsimonious in terms of computational resources. Experimental results show that significant accuracy and compression gains (up to 37.36%) can be achieved when coding features extracted from multiple views.

Index Terms **—** distributed source coding, feature coding, multi-view coding, visual sensor networks.

1. INTRODUCTION

In visual sensor networks (VSN), a visual scene is simultaneously acquired from multiple viewpoints by a network of distributed cameras [1]. Typically, VSNs have a large number of low-power sensing nodes, equipped with vision capabilities and enabling important image processing applications such as wireless visual surveillance, environmental monitoring and augmented reality. In a VSN, a large number of sensing nodes transmit data to sink nodes which have plenty of computational resources.

In this paper, the problem of object recognition in low-power and low-bandwidth distributed VSN is addressed, especially targeting applications such as visual surveillance where the tracking and identification of objects of interest is critical. By using a network of sensing nodes, it is expected that problems such as occlusions, illumination and pose variations can be solved and the object recognition accuracy improved by using information from several views of the visual scene at the sink node. However, the large bandwidth required by the nodes can far exceed the network resource constraints. This implies that new solutions for processing and communication of visual data are needed.

Nevertheless, the processing, coding and communication of visual data is shifting from centralized to distributed settings, due to the associated benefits in terms of scalability, reliability and task performance. Instead of transmitting compressed videos or images to a centralized location where the analysis is performed, sensing nodes (with cameras) perform a part of the processing by extracting and compressing local features. While the feature-based representation can be made more compact than the pixel-based representation, it is simultaneously possible to obtain significant energy and bandwidth savings that suits well the resource constrained VSN.

To address the VSN bandwidth limitations, efficient coding techniques are needed. The problem of efficiently compressing local features extracted from still images and video sequences has been already addressed, notably by exploiting the Intra-frame [2] and Inter-frame correlations [3]. This work proposes and evaluates a distributed coding architecture for the visual features extracted from multiple overlapping views, thus exploiting the Inter-view correlation. This feature coding scheme is inspired by the practices used in the field of multi-view video distributed coding and it can be applied to both real-valued features such as SIFT [4] and SURF [5] and binary features such as BRIEF [6], BRISK [7] and FREAK [8].

The proposed *Multi-view Distributed Feature Codec* (MDFC) is based on the popular DISCOVER codec proposed for (pixel-based) mono-view video coding [9]. In distributed feature coding, the correlation between sets of features extracted from different views (thus different sensing nodes) is exploited at the decoder side, which means that the cameras do not need to communicate among them. Since the Inter-view correlation is exploited only at the decoder side (sink node), a simplified network architecture with minimal routing overhead and lower bandwidth requirements can be achieved. In addition, the encoder architecture is rather simple, which better suits the VSN scenario where cameras are battery operated and cannot communicate among each other. In this paper, side information (SI) creation and correlation noise model (CNM) estimation techniques are proposed to exploit the Inter-view correlation. When the correlation between features extracted from different views is low, e.g. parts of the image from one view are occluded in other view, an Intra decoding mode is instead used.

The rest of this paper is organized as follows: in Section 2, related work is reviewed. After, Section 3 proposes the novel MDFC codec and Sections 4 and 5 the two novel tools, notably the SI creation mechanism and the CNM. Section 6 presents and discusses the experimental results while Section 7 concludes the paper.

2. RELATED WORK

Recently, several works in the literature have addressed the problem of efficiently compressing local features extracted from images and videos. The available coding schemes exploit the correlation between i) elements of each descriptor (Intra-descriptor) [2] or ii) descriptors (Inter-descriptor) [10]. Techniques to code the local features extracted from video sequences have also been proposed, exploiting the Intra-descriptor and Inter-frame correlation, i.e. descriptor predictions are created based on previously decoded descriptors [3]. Also, the MPEG group has recently finished a standard [11] for compact descriptors which provides to enable interoperability in the context of image retrieval. These descriptors are compact, discriminative and efficient to extract and mainly target mobile visual search applications. However, only a small number of works address the problem of coding local features extracted from multiple video views. This problem is rather important for VSNs where battery-operated sensing nodes capture the same visual scene from different perspectives. In [12], object recognition is improved by integrating information from multiple viewpoints considering a network of cameras with limited computational power, bandwidth and communication capabilities. To avoid sending redundant visual information, an unsupervised multi-

view feature selection algorithm based on a statistical model of the dependency between features is proposed. In this approach, local features are vector quantized into visual words and the frequency of each word is computed to form a histogram, a global representation of each image. In [13], an efficient object recognition system is proposed where the appearance of an object in multiple views is represented using feature histograms (global representation) and compressed using the theory of distributed compressed sensing. Thus, a sparsity-based distributed sampling scheme is employed, where SIFT features are extracted, quantized and then represented using a weighted histogram. In [14], a multi-view coding architecture suitable for non-binary and binary local features extracted from multiple views is proposed. Correlation between features extracted from multiple views is exploited using similar techniques to those used in the field of multi-view predictive video coding. In this work, the problem of coding local features is addressed but using a distributed coding approach which does not allow any communication between cameras and a more simplified encoder architecture with respect to [14].

3. PROPOSING A MULTI-VIEW DISTRIBUTED CODING SOLUTION FOR BINARY FEATURES

To obtain a low bitrate representation for local binary features, it is necessary to perform clustering/quantization based on an offline learning dictionary. This dictionary of words (the centroids) must effectively represent the entire space of all possible binary descriptors, typically a 512 dimensional space – the number of descriptor elements in each binary descriptor. In this case, each centroid represents a k binary cluster (set of similar descriptors), which is obtained using k-medians clustering with a k-means++ seeding [15]. The centroids of each cluster are available at both the sensing node (camera) and sink node, where object recognition is performed. Also, rate control is performed at the decoder via feedback channel depending on SI quality obtained for each descriptor. Figure 1 shows the MDFC architecture which is described next.

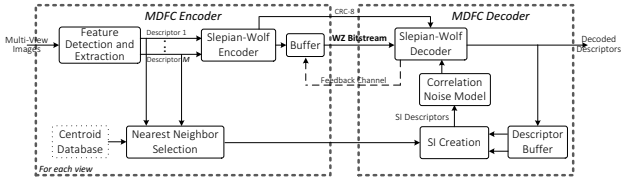

Figure 1 – MDFC codec architecture.

At the sensing nodes, the following operations are performed:

1. **Feature Detection and Extraction**: the most salient keypoints of the image are detected and the binary descriptors representing the patches centered at each keypoint location computed.
2. **Nearest Neighbor Selection**: The nearest centroid to the extracted descriptor is selected using as similarity metric the Hamming distance.
3. **Slepian-Wolf Encoder:** Each descriptor corresponds to a binary vector that is independently encoded using a channel code. In this case, two channel codes can be used: Turbo [16] and LDPC [17]. As usual, the systematic part is discarded and only the parity information is transmitted to the decoder. For the LDPC code, the parity-check matrix corresponds to a 3rd order regular matrix [17].
4. **Data Transmission:** In the first data packet, the centroid ID that was identified in Step 2 is transmitted along with some parity data. The centroid ID occupies 12 to 16 bits for 4096 to 65536 centroids. Whenever needed for decoding convergence, the decoder requests more parity data via the feedback channel.

The proposed MDFC decoder considers two modes that are adaptively selected: i) Intra and ii) Inter-view. In the Intra mode, only the correlation between each descriptor and the corresponding centroid is exploited while, in the Inter mode, the correlation between the descriptors in several views is exploited. With these two decoding modes, it is possible to efficiently model the binary descriptor statistics in a multi-view VSN scenario, where correspondences (similarities) between descriptors of different views are not always available, due to occlusions, field-of-view limitations and illumination variations. In such cases, the Intra mode is more efficient, while the Inter mode is more efficient when the correlation between descriptors from different views is high. The decoder attempts to decode the same descriptor twice, just as a predictive encoder performs Intra /Inter mode selection using rate-distortion optimization. The stop criterion of the channel decoder defines when the source is decoded (for the Intra or Inter modes) or when is necessary to ask for more parity bits, i.e. both Intra and Inter modes failed to decode the source. In summary, the following decoding steps are performed:

1. **SI Creation:** First, the centroid ID is used to retrieve the centroid descriptor for Intra decoding. Also, the centroid ID can be used to identify a set of already decoded descriptors from other neighboring views that can be used to decode the source using the Inter mode. For the Intra mode, the SI corresponds to the centroid value (descriptor). For the Inter mode, two novel SI creation techniques are presented in Section 4.
2. **Correlation Noise Model:** To make good use of the SI obtained in the previous step, the decoder needs to have a reliable CNM to characterize the statistical correlation between the original descriptors and the corresponding SI descriptors. The correlation noise corresponds to a virtual channel since the SI may be seen as a "corrupted" version of the original information. The soft information to be used by the Slepian-Wolf decoder is computed with the novel technique presented in Section 5.
3. **Slepian-Wolf Decoder:** An iterative decoding process is used where more parity bits are requested from the sensing node until the source is successfully Intra or Inter decoded. Notice that a feedback channel is usually available in many real VSN testbeds without introducing significant delay. When the turbo code is used, the stop criterion is based on the log-likelihood ratio (LLR) of each decoded bit. If the absolute value of a bit LLR is below a threshold of 4.6, the bit is considered as *uncertain*. When no more than 3 bits are defined as uncertain, the decoder claims that the source is fully decoded. In case a LDPC decoder is used, the stop criterion is the parity check of the sum-product algorithm [18]. In addition, an 8-bit CRC is used to guarantee a very small error decoding probability, i.e. near lossless decoding is achieved.

4. INTER-VIEW SIDE INFORMATION CREATION

As stated before, the Inter-view decoding mode tries to explore the spatial redundancy between views. To conditionally decode a new descriptor, it is necessary to create SI using already decoded descriptors for any of the other views (called reference views). The Inter-view SI creation module is therefore responsible to select which previously decoded descriptors (one or more) are well correlated with the descriptor being decoded. In the next two sections, two alternative Inter-view SI creation solutions proposed are described.

4.1 Centroid Based Strategy (CBS)

In a multi-view scenario, when the same feature is captured from two different views, the corresponding descriptors are quite similar. Thus, there is a high probability that descriptors representing the same point in the 3D space but acquired from different views belong to the same cluster, i.e. have the same centroid ID. Therefore, after receiving the centroid ID of a descriptor to be decoded, it is necessary to search the previously decoded descriptors from other reference views (transmitted from other sensing nodes) to select those that are represented by the same centroid ID. When no descriptor is found, the nearest populated centroid (using a Hamming distance metric) is looked for and its descriptors used as side information. In both cases, the selected descriptors are used by the CNM.

4.2 Geometry Based Strategy (GBS)

The second solution exploits the geometric position of the patch where the descriptors were extracted. When two sensing nodes capture the same objects, it is expected that similar pools of descriptors are transmitted by the sensing nodes but also that the geometric position of these descriptors can be characterized with a geometric model such as an affine or perspective model. The following procedure is applied:

1. **Centroid Matching:** The descriptors of the view to be decoded are represented by their centroids which are matched with each of the reference view descriptor centroids, i.e. the centroid ID is used to identify a set of similar descriptors for each new descriptor being decoded. This will provide coarse matching between descriptors when compared to the full descriptor representation that is now only available at the encoder.
2. **Affine Model Estimation:** Taking into account all the descriptor matches identified in the previous step, an affine model is estimated between the descriptors locations in the view being decoded and those in reference views. In this case, a set of homographies using the RANSAC algorithm with different maximum allowed re-projection errors is first estimated. The homography model corresponds to:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

   where $h_{11}, h_{12}, \cdots, h_{33}$ are the homography model parameters, x', y' the descriptor location in a previously decoded view and x, y the descriptor location in the view to be decoded. The re-projection error corresponds to the model error, i.e. the error between each previously decoded descriptor location after transformation and the descriptor location in the view to be decoded. After, the homography closer to an affine transformation is selected, i.e. when $h_{31} < \tau$ and $h_{32} < \tau$, i.e. last line of the transformation matrix is close to $[0,0,1]$. Using this approach, the transformation allowed is more restrictive (i.e. only translation, rotation, scale and shear) but also more robust. If no homography transformation fulfilling $h_{31} < \tau$ and $h_{32} < \tau$ is found, the current view is discarded and another decoded view is processed next with the algorithm returning to Step 1.
3. **Warping**: When a good homography is found, the transformation matrix is used to calculate a new set of locations on the reference view, i.e. the locations of the descriptors to be decoded are warped to the reference view. Afterwards, a window of 30×30 pixels is used, centered in each warped keypoint location and the previously decoded descriptors available in that window are retrieved. Then, the decoded descriptor, that is closer to the centroid value of the descriptor to decode, is used as side information.

After performing these same steps for all the reference views, the selected descriptors are sent to the CNM. For all the descriptors for which no good correlation with another view was found, the CBS decoding mode is used.

5. CORRELATION NOISE MODEL ESTIMATION

To make good use of the SI descriptors for decoding purposes, the decoder needs to have a reliable knowledge of the statistical model characterizing the correlation noise between the original descriptors *X* available at the encoder and the SI descriptors *Y* available at the decoder. In distributed video coding, the CNM between *X* and *Y* is typically modeled as a Laplacian distribution. However, this solution cannot be used here, since the source is a binary memoryless source where the symbols ('0' and '1') have the same probability of occurrence. Thus, a binary symmetric channel (BSC) is a more adequate approach to characterize the correlation between *X* and *Y*. The BSC channel has two input symbols (x_0 and x_1) and two output symbols (y_0 and y_1). The probability of observing y_1 at the decoder when x_0 is at the corresponding symbol at the encoder and the probability of observing y_0 at the decoder when x_1 is at the encoder are the same and equal to the error probability, p. In the proposed solution, the SI corresponds to a set of already decoded descriptors that are highly correlated with the source. However, instead of fusing the selected decoded descriptors to obtain a single descriptor (*hard decision*), it is proposed to use all selected descriptors to directly calculate the symbol probability, i.e. $p(y_0)$ and $p(y_1)$, thus performing a *soft decision*. To allow a rather fine granular estimation, these probabilities are calculated for each descriptor element in the following way:

$$\begin{aligned} P_- &= p(B_n = 0 \mid Y_n^0, \dots, Y_n^M) = \frac{N}{M} \\ P_+ &= p(B_n = 1 \mid Y_n^0, \dots, Y_n^M) = 1 - p(B_n = 0 \mid Y_n^0, \dots, Y_n^M) \end{aligned} \quad (1)$$

where B_n represents the n[th] bit (or descriptor element) of the descriptor to decode, *N* is the number of times that a descriptor element has the value y_0 and *M* is the total number of descriptors selected as SI. When the descriptor element of all SI descriptors has the same value, i.e. an error probability of 0%, the correlation model sets 0.99 and 0.01 as the limits for the probabilities P_+ and P_-. Finally, the soft information (probability) $L_{apriori}$ is computed as:

$$L_{apriori}(B_n) = ln(P_+/P_-) \quad (2)$$

The iterative soft decoding is performed with a LDPC or Turbo decoder using the soft information computed as in (2), and the encoder transmitted parity or syndrome information chunks. After, the (*a posteriori*) soft output information is obtained and thresholded (with 0) to obtain an estimate of the decoded descriptor. Finally, a CRC error detection technique is applied to assess if this estimate is reliable.

6. PERFORMANCE EVALUATION

The MDFC performance was studied in the context of an object recognition task, although the proposed solution may have other uses such as object tracking. In this Section, both the test conditions used and the experimental results obtained for bitrate compression and object recognition accuracy are presented.

6.1 Test Conditions

To evaluate the proposed MDFC solution, the Berkley Multiview Wireless (BMW) dataset [13] already used in the past to assess object recognition accuracy under networks with severe bandwidth constraints has been selected. The BMW dataset contains 20 different landmarks, each acquired from 16 different perspectives. For each perspective, 5 views are simultaneously acquired, taken with 5 different cameras where one of the cameras is in a central position and the others are placed around the central camera equally spaced, in a cross spatial configuration. Keypoint detection has been performed using the fast Hessian technique of the SURF detector [5]. Then, the BRISK [7] feature extractor has been used, which creates bit-strings with 512 bits. The SURF keypoint detector software was OpenCV 2.4.10 and the BRISK extraction software is available in [19]. An offline stage was used to cluster a set of descriptors and obtain meaningful and representative centroids. A total of 12456 images from the Paris [20], Oxford [21] and Stanford landmarks [22] images datasets, with a maximum of 300 features extracted from each image, were used to define 4096 centroids. Note that these datasets were not used in the evaluation of the proposed solution.

To perform object recognition, pair-wise matching between the decoded descriptors from the query images (all views) and the database descriptors is first performed. Then, wrong matches between the query and the database descriptors are filtered using the well-known ratio test: all matches in which the distance ratio is greater than 0.5 are rejected, which allows to obtain the best accuracy. The number of

matched descriptors (inliers) is taken as the relevancy score of the database image with respect to the query image. Also, when the query has multiple views from the same object, the number of inliers obtained for each database image (after matching and filtering) is added before ranking.

Since communication between the sensing nodes of the VSN cannot occur, the proposed MDFC can only be evaluated using as benchmark feature codecs which exploit Inter-view correlation at the decoder side or Intra feature codecs. Thus, the MDFC performance is compared to a predictive feature codec (PFC) that exploits the statistical correlation between each of the extracted descriptors and the corresponding centroid of the cluster to which it belongs, i.e. only Intra coding is performed. The encoder computes a descriptor residue which is the difference (XOR) between the descriptor to be coded and the corresponding centroid descriptor; this residue is after arithmetic coded to obtain better compression performance.

The codecs performance has been assessed not only in terms of compression factor but also in terms of average precision (AP), a widely used metric to assess retrieval accuracy [23]. MAP is calculated by taking the mean of the average precision (AP), where AP is calculated for each query by averaging the precision at each point a correct image is retrieved. To asses the accuracy, the BMW dataset was divided into two sets – the query and the database which are completely independent: the query images correspond to perspectives 0, 3, 6, 9 and 12 and the database images are the remaining ones. In the results shown, the number of reference views used to decode the current view varies from 0 (Intra), 1, 4 and 79. When 1 and 4 reference views are used, they have been acquired at the same time instant and correspond to views from the same perspective. Note that the first view is always coded as Intra. When more than 4 views are used, images acquired from different perspectives are exploited to achieve higher compression and object recognition accuracy, e.g. when 79 reference views are used, they correspond to all views from the 16 perspectives (1 landmark).

6.2 Experimental Results

The average bitrate savings results for all query images of the BMW dataset with respect to the uncompressed data rate are presented in Table I when the Centroid Based Strategy (CBS) and the Geometry Based Strategy (GBS) SI creation methods described in Section 4 are used; results are also shown when the LDPC and Turbo codes are used for Slepian-Wolf decoding. Note that the decoded descriptors are the same as the original descriptors and thus the retrieval performance is the same for a fixed number of reference views (columns). An insight of the operation modes used by the MDFC to decode the descriptors is shown in Figure 2, i.e. the number of times that the Intra and Inter modes are selected when the number of reference views increases. In Figure 3, the rate-accuracy results are presented with the accuracy of the object recognition measured by means of the MAP and AP@L with L=[1,5,10,20] metrics; for each query, a rank of only L images in the database is obtained to compute the AP. To obtain these results, the LDPC code and the GBS SI creation technique were employed and the number of coded views is increased to show the accuracy gains. From the results, the following conclusions can be made:

- **PFC versus MDFC:** As shown in Table 1, when all cameras are independently encoded and decoded (Intra), PFC slightly outperforms the MDFC – LDPC performance just using the Intra mode and obtains 7.56% higher bitrate savings comparing to MDFC – Turbo. This can be justified by the gap that distributed coding schemes typically have with respect to predictive coding (although the MDFC – LDPC gap is rather small). Note that, when using more reference views, the PFC codec cannot be used since it is assumed that cameras cannot communicate with each other.
- **MDFC – Turbo versus MDFC – LDPC:** As expected, for all evaluated cases, the LDPC codec is more efficient (see Table 1) since the same behavior was observed for distributed video coding schemes [9]. Moreover the MDFC – LDPC codec is able to outperform the PFC codec just when two cameras are used, which shows that exploiting the Inter-view redundancy is rather beneficial even only at the decoder side.
- **MDFC Intra-mode versus MDFC Inter-mode:** As shown in Figure 2, when more reference views are used to generate the SI, the number of descriptors decoded using the proposed Inter-view approach increases. This is expected considering the bitrate savings results presented in Table 1.
- **MDFC – LDPC rate-accuracy:** As shown, the MAP is below 30% for 118 kbit/query (one view) and continuously increases when more views are used, up to 70% for 2.64 Mbit/query (25 views); the same behavior is observed for the AP metrics – the number of views is shown at the top of the graph. Thus, it may be concluded that having multiple descriptors from different viewpoints has a big impact in the object recognition accuracy. When the number of retrieved objects is L, the average precision increases when compared with MAP, which means that the top ranking objects are quite often the same object as the query.

Table I – PFC and MDFC average Bitrate Reduction [%].

	Intra	CBS			GBS
Ref. Views	0	1	4	79	79
PFC	23.97				
MDFC - Turbo	16.41	22.47	27.57	28.44	33.23
MDFC - LDPC	23.04	28.05	32.45	33.23	37.36

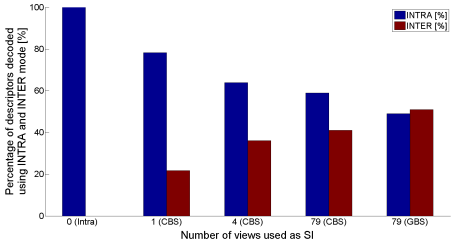

Figure 2 - Descriptors decoded as Inter and Intra [%].

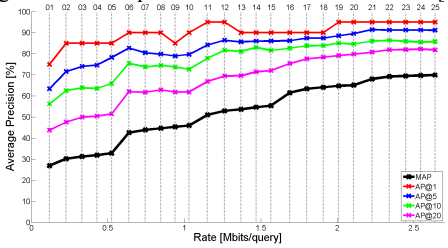

Figure 3 – MDFC – LDPC rate-accuracy performance.

7. CONCLUSIONS

In this paper, a distributed source coding solution able to exploit the Inter-view redundancy in binary features is proposed. This solution is suitable for a multi-view image acquisition system typical of a visual sensor network. Significant bitrate compression savings were obtained by exploiting the Inter-view redundancy at the decoder side (up to 37.36%). Also the accuracy of the object recognition was improved from 30% to 70% when more cameras are used. Future work will consider the design of a distributed descriptor selection coding scheme able to avoid the transmission of features which do not contribute to increase the accuracy of the visual analysis task.

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] S. Soro and W. Heinzelman, "A Survey of Visual Sensor Networks", Advances in Multimedia, vol. 2009, Article ID 640386, 2009.

[2] A. Redondi, L. Baroffio, J. Ascenso, M. Cesana and M. Tagliasacchi "Rate-accuracy Optimization of Binary Descriptors", IEEE International Conference on Image Processing, Melbourne, Australia, September 2013.

[3] L. Baroffio, J. Ascenso, M. Cesana, A. Redondi and M. Tagliasacchi, "Coding Binary Local Features Extracted From Video Sequences", IEEE International Conference on Image Processing, Paris, France, October 2014.

[4] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, November 2004.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding, vol. 110, no. 3, June 2008.

[6] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," European Conference on Computer Vision, Crete, Greece, September 2010.

[7] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary Robust Invariant Scalable Keypoints", IEEE International Conference on Computer Vision, Barcelona, Spain, November 2011.

[8] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast Retina Keypoint", IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA, June 2012

[9] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov and M. Ouaret, "The DISCOVER codec: Architecture, Techniques and Evaluation", Picture Coding Symposium 2007, Lisbon, Portugal.

[10] P. Monteiro and J. Ascenso, "Clustering based Binary Descriptor Coding for Efficient Transmission in Visual Sensor Networks", Picture Coding Symposium, December 2013, San Jose, CA, USA.

[11] L.-Y. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, W. Gao, "Overview of the MPEG-CDVS Standard," IEEE Transactions on Image Processing, vol. 25, no. 1, pp.179-194, Jan. 2016.

[12] C.M. Christoudias, R. Urtasun and T. Darrell, "Unsupervised Feature Selection via Distributed Coding for Multi-view Object Recognition", IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, June 2008.

[13] N. Naikal, A. Y. Yang and S. S. Sastry. "Towards an Efficient Distributed Object Recognition System in Wireless Smart Camera Networks", 13th Conference on Information Fusion (FUSION), Edinburgh, United Kingdom, July 2010.

[14] L. Bondi, L. Baroffio, M. Cesana, A. Redondi and M. Tagliasacchi, "Multi-View Coding of Local Features in Visual Sensor Networks", IEEE International Conference on Multimedia & Expo Workshops, June 2015, Turin, Italy.

[15] D. Galvez-López and J.D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences", IEEE Transactions in Robotics, vol.28, no.5, pp.1188-1197, October 2012.

[16] J. Ascenso, C. Brites and F. Pereira, "Design and Performance of a Novel Low-Density Parity-Check Code for Distributed Video Coding", IEEE International Conference on Image Processing (ICIP), San Diego, California, USA, October 2008.

[17] C. Brites, J. Ascenso and F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, May 2006.

[18] F. R. Kschischang, B. J. Frey and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm", IEEE Trans. Inform. Theory, February 2001.

[19] S. Leutenegger, M. Chli and R. Siegwart, http://www.asl.ethz.ch/people/lestefan/personal/BRISK/, 20 of May 2015.

[20] J. Philbin and A. Zisserman, The Paris Dataset, http://www.robots.ox.ac.uk/~vgg/data/parisbuildings/, 21 of April 2015.

[21] J. Philbin, R. Arandjelovic and A. Zisserman, The Oxford Buildings Dataset, http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/, 21 of April 2015.

[22] V. Chandrasekhar, D. Chen, S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, "The Stanford Mobile Visual Search Dataset", ACM Multimedia Systems Conference, San Jose, USA, February 2011.

[23] A. Canclini, R. Cilla, A. Redondi, J. Ascenso, M. Cesana and M. Tagliasacchi, "Evaluation of visual feature detectors and descriptors for low-complexity devices", IEEE/EURASIP Digital Signal Processing Conference, Santorini, Greece, 2013.