This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing

Compress-then-Analyze vs Analyze-then-Compress: what is best in Visual Sensor Networks?

Alessandro Redondi, Luca Baroffio, Lucio Bianchi, Matteo Cesana, Marco Tagliasacchi

Abstract—Visual Sensor Networks (VSNs) have attracted the interest of researchers worldwide in the last few years, and are expected to play a major role in the evolution of the Internet-of-Things (IoT). When used to perform visual analysis tasks, VSNs may be operated according to two different paradigms. In the traditional *compress-then-analyze* paradigm, images are acquired, compressed and transmitted for further analysis. Conversely, in the *analyze-then-compress* paradigm, image features are extracted by visual sensor nodes, encoded and then delivered to a remote destination where analysis is performed. The question this paper aims to answer is *What is the best visual analysis paradigm in VSNs?* To do this, first we empirically characterize the rate-energy-accuracy performance of the two aforementioned paradigms. Then, we leverage such models to formulate a resource allocation problem for VSNs. The problem optimally allocates the specific paradigm used by each camera node in the network and the related transmission source rate, with the objective of optimizing the accuracy of the visual analysis task and the VSN coverage. Experimental results over several VSNs instances demonstrate that there is no “winning” paradigm, but the best performance are obtained by allowing the coexistence of the two and by properly optimizing their utilization.

Index Terms—Visual Sensor Networks, Local Visual Features, Resource Allocation, SIFT, BRISK.

1 Introduction

Reading a book or recognizing a familiar face are actions that characterize people’s everyday life and require processing of visual stimuli. The early visual system, comprising the pathway from the eye to the visual cortex, is responsible for processing such stimuli in a rich, yet energy-efficient manner, so that they can be interpreted, memorized, communicated and finally analyzed and converted into high level semantic concepts. The physiology of the early visual system is the result of a sophisticated balancing between information coding, i.e., the transfer of the data captured by the photoreceptor field to the visual cortex, and energy efficiency. Indeed, sight is characterized by extremely low metabolic energy expenditure.

Digital cameras have been developed mimicking a simplified model of the human visual system in a two step process following a *compress-then-analyze* (CTA) paradigm: images are acquired in digital format by sampling and quantizing the light-field on a discrete lattice of pixels. Images, or image sequences, are then compressed in order to be stored and/or transmitted for further analysis. A large body of research has focused on the analysis of visual data to accomplish high level tasks, e.g., recognizing letters, faces, objects, detecting events, etc. In this paradigm, image analysis is often based on a compressed and hence lossy representation of the original image, which might significantly impair its efficiency [3], [4], [5], [6], [7]. Nonetheless, such a compress-then-analyze paradigm is being employed in many application scenarios where energy constraints are not overwhelming (e.g., video surveillance).

The integration of low-power wireless networking technologies such as IEEE 802.15.4-enabled transceivers [8] with inexpensive camera hardware [9], [10] has enabled the development of the so-called wireless multimedia sensor networks (WMSNs), also known as visual sensor networks (VSNs). VSNs can be thought of as networks of wireless devices capable of sensing visual content [11], such as still images and video, depth maps, etc. Due to their flexibility and low-cost, VSNs have attracted the interest of researchers worldwide in the last few years, and are expected to play a major role in the evolution of the Internet-of-Things (IoT) paradigm [12], [13].

The compress-then-analyze paradigm can be adapted to VSNs by properly accounting for the additional energy constraints posed by the resource-constrained sensor platform and the limited nominal bandwidth of current standards for low-power communication among sensor nodes. Several research efforts have been put in place to design wireless sensor networks for supporting still image and video delivery [14], [15]. However, when only the result of the visual analysis matters, transmitting image or video data retaining a pixel-level representation is inefficient in terms of the computational and network resources used, especially when the analysis is based on data sensed by more than one camera.

Alternatively, it is possible to consider a scenario where the bitstream flowing in the visual sensor network is reduced by some sort of local processing which extracts and

- *The authors are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecncio di Milano, P.zza Leonardo da Vinci, 32 - 20133, Milano - Italy, e-mail: firstname.lastname@polimi.it*

- *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.*
- *The material in this paper has been partially presented in [1] and [2].*

Manuscript received April 14, 2015; revised November 24, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing

encodes visual features, rather than compressing and transmitting a representation of the sensed images in the pixel domain. The key tenet is that most visual analysis tasks can be carried out based on a succinct representation of the image, which entails both global and local features, while it disregards the underlying pixel-level representation, thus leading to a joint *analyze-then-compress* (ATC) paradigm; image features are collected by visual sensor nodes, processed (compressed), and then delivered to the final destination(s) in order to enable higher level visual analysis tasks. Extracting features from visual data is, however, a computationally intensive task. The process entails detecting image keypoints and computing the corresponding descriptors; the related computational complexity grows linearly with the image size and with the required number of descriptors [16].

The question this paper aims to answer is *What is the best visual analysis paradigm in VSNs?* We consider a scenario in which a visual sensor network is deployed to support applications based on image retrieval and object recognition. Our specific contributions are:

1) We empirically characterize the rate-accuracy and the rate-energy behavior of the two visual analysis paradigms. For the rate-accuracy models, we build up the full visual analysis pipeline for image retrieval under the two paradigms, which is then used to derive experiment-based models for the achieved accuracy at different target transmission rates. For the rate-energy models, we implement a working visual sensor platform based on a BeagleBone Linux computer and a IEEE802.15.4 compliant radio transceiver, and we derive experimentally the consumed energy to accomplish the image retrieval task according to the two paradigms.
2) The rate-accuracy and rate-energy models are then leveraged to design a resource and paradigm allocation problem for visual sensor networks. The problem returns as a solution the specific paradigm used by each camera node in the network (ATC or CTA) and the related transmission rate. The target is to maximize a multi-objective function which comprises the accuracy of the visual analysis task and the VSN coverage, i.e., the number of cameras which can be concurrently activated, under a predefined target lifetime constraint.
3) The resource and paradigm allocation problem is solved for several VSN instances, characterized by different parameters in terms of network topology, lifetime constraints and application frame rate requirements. The obtained solutions show that there is no clear winner between CTA and ATC, and depending on the particular parameters, each camera in the network may select a different paradigm. Thus, the coexistence of both paradigms allows to operate the VSN at its best.

In Section 2, we review the related work on the design of wireless networks for video transmission and analysis. In Section 3 we introduce the two visual paradigms subject of our analysis and the accuracy measure used for their comparison. Sections 4 and 5 present the rate-accuracy and rate-energy models for both the CTA and ATC paradigms, and compare their behavior. Section 6 introduces the proposed resource and paradigm allocation problem for VSNs whose solution for several network instances is commented in the experimental evaluation in Section 7. Finally, Section 8 concludes the paper and discusses future works.

2 Related Work

The design of energy efficient wireless sensor networks has been largely debated and addressed in the literature [17], [18] [19], [20], [21], [22]. The proposed solutions generally scale down to finding the resource allocation strategies which lead to minimal energy consumption, and thus maximal network lifetime. The resources under consideration may include the transmitted power at the wireless transceiver, Medium Access Control (MAC) parameters as well as the routes to deliver the information to the final destination(s).

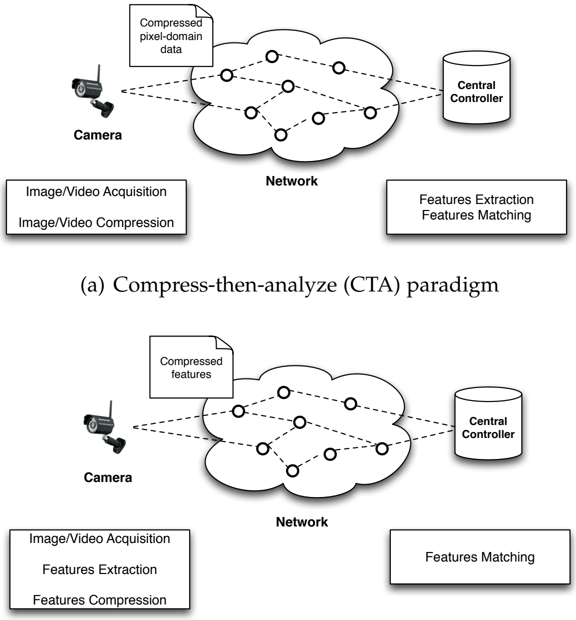
The very same problem of energy efficiency becomes even more relevant in visual sensor networks for two main reasons: first, wireless nodes are now required to perform additional energy-greedy multimedia processing tasks (acquisition, encoding, etc.); second, multimedia applications may have in general more stringent requirements in terms of expected quality of service (QoS), which turns into higher energy to be consumed to effectively support them.

A good deal of work has recently focused on designing effective and long-lasting wireless networks to support video transmission. The interested reader may refer to [23] for a survey on the topic. The problem of resource allocation for supporting video streams in wireless network is addressed in [24] and [25]. In [24], the focus is on the design of a dynamic video encoder which can be adapted to the current status of the network conditions, whereas, [25] proposes an optimization framework to maximize the peak signal to noise ratio (PSNR) in cooperative wireless networks; namely, centralized and distributed PSNR-optimal strategies are proposed to jointly control the video encoding rate, the selection of relaying nodes and the allocated power level to perform wireless transmissions.

In [26] and, successively, in [27] and [28] the authors introduce an optimization framework to jointly optimize the coding rate and the routes in wireless sensor networks where correlated visual sensors operate under distributed source coding. The proposed problem formulation uses an objective function which is the combination of the overall distortion and the lifetime of the wireless sensor network. A distributed algorithm is further proposed to heuristically solve the aforementioned problem.

A similar contribution and networking scenario is considered in [29] and [30] where power control is also included in the optimization problem formulation. In [31], distributed algorithms based on Lagrangian duality are proposed to maximize the network lifetime of wireless video sensor networks by properly setting the video source rates, the encoding powers and the routing in the network. The collected video quality is finally assessed against the achieved maximal lifetime.

A two-step heuristic is introduced in [32] to prolong the network lifetime in wireless video sensor networks. The proposed algorithm first selects the routes between the video sources and the sink nodes by resorting to an energy-aware routing metric, and then properly sets the encoding rate at the video sources.



(a) Compress-then-analyze (CTA) paradigm

(b) Analyze-then-compress (ATC) paradigm

Fig. 1: The two different approaches to implement image analysis in visual sensor networks

Differently from the aforementioned work, which is generally targeting video delivery optimization, we focus on image retrieval and object recognition applications, which are based on local features extraction algorithms. In particular, we propose a thorough comparison between two complementary paradigms to support image retrieval and object recognition applications. The importance of these two paradigms were already partially noticed in the field of mobile visual search. In [33], Girod *et al.* compare the two paradigms in terms of retrieval accuracy, system latency and energy consumption in the case of an image retrieval application for mobile phones. Here we make a further step, proposing rate-energy and rate-accuracy models for the two paradigms, which are then used to optimally allocate the resources in a more complex scenario, such as the one encountered in visual sensor network deployments.

3 Compress-then-analyze vs analyze-then-compress

The reference literature on visual sensor networks generally evaluates the "quality" of multimedia encoding and transmission based on rate-distortion models. These models capture the effect of varying the output bit-rate R on the visual content distortion D, which is usually evaluated through the computation of the signal-to-noise ratio (SNR) or peak signal-to-noise ratio (PSNR). However, such models cannot be used for capturing the effect of varying the transmission rate on the accuracy of a visual analysis tasks such as object recognition. Thus, we propose compact rate-accuracy models for the two paradigms usable for visual analysis (CTA or ATC). To this extent, we have implemented a full image retrieval pipeline, as illustrated in Figure 1.

In the compress-then-analyze case, query images are first compressed with JPEG at different rates by varying the *Quality Factor* QF from 1 to 100. The images are then transmitted to a central controller, which extracts local features from the compressed query images and matches them against the features extracted from a reference database of (uncompressed) labeled images. Since features extraction is performed at a remote, powerful central controller without particular computational limits, we assume that the SIFT algorithm [34] is used to extract features from the JPEG encoded images in the CTA paradigm. SIFT features are considered as the gold standard in visual analysis, as they typically achieve state-of-the-art performance in most applications. At the same time, extracting and matching SIFT features is costly.

In the analyze-then-compress case, features are extracted from uncompressed query images, encoded with a suitable algorithm and transmitted to a remote controller, where they are matched with the features extracted from the reference database images. Since feature extraction is now performed on a resource-limited camera node, we consider the state-of-the-art BRISK algorithm [35], which is optimized for fast computation, and thus suitable for low-power and low-complexity hardware, while guaranteeing accuracy performance close to the one of SIFT. To encode BRISK features, we use the method proposed in [36].

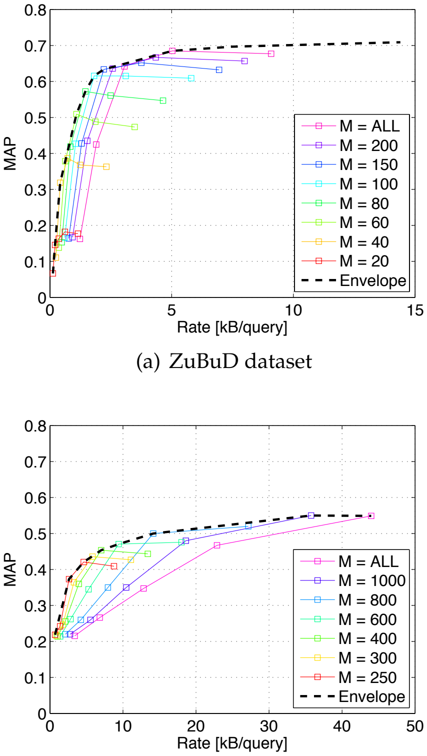
In the CTA paradigm, rate is controlled by operating on the JPEG quality factor of the image queries. Conversely, in the ATC paradigm, rate is controlled by tuning : (i) the dimension D (in bits) of each BRISK descriptor; and (ii) the number of features M to be transmitted to the central controller. Note that the number of detected keypoints is content-dependent and might exceed M. Therefore, a subset of M keypoints needs to be selected for the computation and transmission of the associated descriptors. According to [1], we sorted the features in descending order of their associated strength, and we selected the top-M features. This operation allows to obtain better rate-accuracy performance than selecting the subset of features randomly.

For both CTA and ATC approaches, features matching is performed computing either the Euclidean or Hamming distances (for SIFT and BRISK, respectively), and filtering matches using the ratio-test and a geometric consistency check with RANSAC [34]. The Mean of Average Precision (MAP) measure is used to assess the accuracy of the retrieval process.

Two data sets have been considered in the evaluation:

- *ZuBuD*: The Zurich Building Database[1] contains 1005 color images of 201 buildings of the city of Zurich. Each building has five VGA images (640x480), taken at random arbitrary view points under different seasons, weather conditions and by two different cameras. A separate archive containing 115 images (at a resolution of 320x240) of the same buildings (with different imaging conditions) is available as query dataset.
- *Oxford*: The Oxford Building Database[2] consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually

[1] http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html
[2] http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/



(a) ZuBuD dataset

(b) Oxford dataset

Fig. 2: Rate-accuracy curves for (a) the ZuBuD dataset and (b) the Oxford dataset when using BRISK features. Each of the solid colored lines represents the rate-accuracy curve for a different number of features M, when varying the dimension D of each BRISK feature in the range {32, 64, 128 256, 512} bits. The black dashed line is the envelope of the rate-accuracy curves family, and represents the best accuracy that can be obtained for a target rate. The curve corresponding to the label ALL is obtained by using all the detected features for matching.

annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries, for a total of 55 queries. The resolution of each query image is 768x1024, much greater than in the ZuBuD dataset case.

4 Rate-Accuracy Modeling

4.1 ATC rate-accuracy model

Figure 2(a) shows a family of rate-accuracy curves referring to the ATC paradigm for the ZuBuD dataset when using BRISK features; each curve is obtained with a different value of M, varying the dimension of BRISK descriptor D (in bits). Similar curves have been obtained for the Oxford dataset (see Fig. 2(b)). It is clear from the figures that different combinations of M and D induce different performance. To derive a synthetic rate-accuracy model for the ATC paradigm, we focus on the envelope of the rate-accuracy curves family, which represents the best operational trade-off that can be obtained, that is, the minimum rate to be used to achieve a given MAP.

In [1], we showed that it is possible to derive an analytic expression for the envelope of the family of curves, which can be written as:

$$A(\rho) = p_1\rho + p_2\sqrt{\rho^2 + \rho} + p_3, \qquad (1)$$

where $A(\rho)$ is the MAP for a given rate ρ (in kB/query), and p_1, p_2 and p_3 are application-specific parameters. Moreover, in [1] we also showed that the derivation of the analytic rate-accuracy model for the ATC paradigm allows to define the concept of *internal allocation*: namely, under a particular bitrate budget ρ, the internal allocation determines the optimal number $M(\rho)$ of local visual features to transmit and their corresponding size (in bits) $D(\rho)$ to obtain the maximum accuracy.

Table 1 reports the values of the rate-accuracy model parameters for all the tested datasets. To quantify the goodness of our model we compute the Pearson's correlation coefficient R^2 between the real and the estimated envelope, obtaining a value equal to or greater than 0.97.

4.2 CTA rate-accuracy model

Similar to the case of ATC, it is worthwhile to analyze the performance of CTA at different rates. In this paradigm, the images acquired by a camera node are compressed with JPEG and transmitted to a central controller. Thus, rate may be varied by properly modifying the JPEG quality factor. High quality factors minimize the distortion introduced by the encoding process, but the resulting image size may struggle with the limitation imposed by the available bandwidth. Conversely, low quality factors allow to efficiently encode the input image at the cost of increasing distortion in the pixel domain. Such artefacts may impact on the object recognition task performance. To model the CTA rate-accuracy performance, we have run the object recognition pipeline on the ZuBuD and Oxford datasets, each time varying the query images JPEG quality factor. Results are reported in Figure 3(a), and 3(b). Two cases are considered for CTA:

- SIFT features are extracted from the query image at the central controller (green solid lines). Such a case allows to obtain the best accuracy performance, at the cost of increasing complexity at the central controller.
- BRISK features are extracted at the central controller (red dashed line). This case is suitable when computational resources at the server are limited, but it results in poorer visual accuracy.

The experimental results show that the rate-accuracy curve for CTA paradigm is well represented by the following functional form:

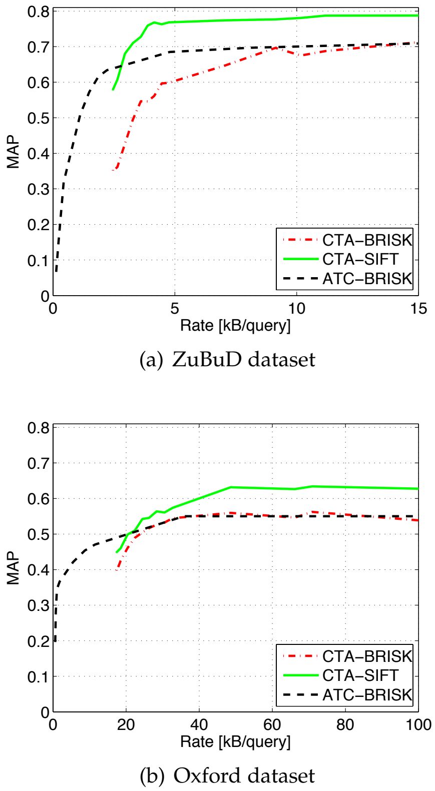
$$A(\rho) \quad = \quad \frac{(q_1\rho + q_2)}{\rho + q_3}, \qquad (2)$$

whose parameters are reported in Table 1 (columns 6-9) for the ZuBuD and Oxford data sets.

4.3 Comparison of CTA and ATC rate-accuracy

Figure 3(a), and 3(b) show the results for the ZuBuD, and Oxford datasets, respectively. In each figure, we include the rate-accuracy curves obtained for the two aforementioned

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing



(a) ZuBuD dataset

(b) Oxford dataset

Fig. 3: Comparison between the rate-accuracy curves of the *compress-then-analyze* and the *analyze-then-compress* paradigms

CTA configurations, and the curve for ATC (black dashed line), when BRISK features are extracted and compressed at the remote nodes. In this latter case, the curves correspond to the envelopes in Figures 2(a), and 2(b). For SIFT, we used the OpenCV v.2.4.3 implementation, while for BRISK we used the implementation provided by the authors[3].

The results in Figure 3 indicate that the choice of the paradigm is dictated by the bandwidth constraints imposed by the network. Indeed, at low bitrates, the ATC approach is not only the preferable solution, but also the only one that can be adopted as there exists a lower bound to the source rate when operating in CTA which forces the source rate to be above 17.3 kbyte/query for Oxford data set and 2.4 kbyte/query for ZuBud data set respectively. Conversely, when the network allows to send high-quality query images at high bitrates, extracting features at the central controller is the best choice. However, note that this is a condition which is seldom met in visual sensor networks. Moreover, if the central controller is not subject to computational constraints, the use of non-binary features like SIFT is to be preferred to BRISK.

5 Rate-energy Modeling

CTA and ATC may be compared not only with respect to their rate-accuracy behaviors, but also looking at what is the energy that each visual sensor node consumes to operate in one or the other paradigm. Generally speaking, the per-node energy consumption is the sum of two components: the energy E_{cpu} for acquiring and processing data, and the energy E_{tx} for transmitting the data to a remote location. Typically, in generic wireless sensor networks only the energy needed for transmitting data is taken into account. This is motivated by the fact that processing is generally limited to simple operations and the energy spent for transmitting the sensed information dominates on the total energy consumption. However, this assumption does not hold when considering visual sensor networks, where the energy required to process multimedia data can not be neglected.

	Analyze-then-compress				Compress-then-analyze			
Dataset	p_1	p_2	p_3	R^2	q_1	q_2	q_3	R^2
ZuBuD	-2.83	2.83	-0.64	0.98	0.83	-1.68	-1.95	0.99
Oxford	-2.25	2.25	-0.67	0.97	0.65	-7.42	-8.57	0.97

TABLE 1: Values of the model parameters for the two different paradigms and datasets used. The parameters are derived expressing the rate ρ in kB and the MAP accuracy value between 0 and 1. We also report the value of the Pearsons's correlation value between the observed behaviors and the ones obtained with the proposed rate-accuracy models (1) and (2).

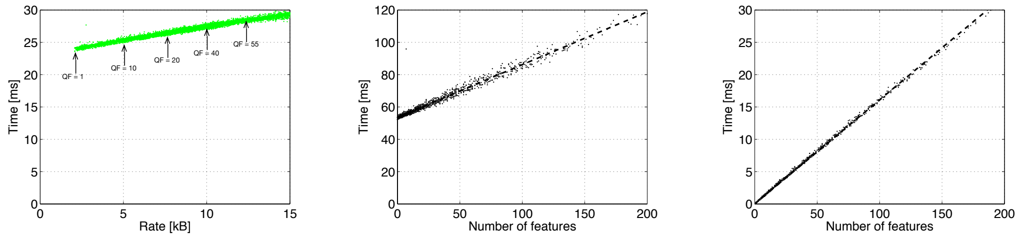
While comparing the required transmission energy in ATC and CTA is straightforward, as E_{tx} depends primarily only on the amount of data that is transmitted (i.e., it is function of ρ only), the energy E_{cpu} spent by a particular processing algorithm depends on several factors. First of all, the hardware architecture on which the algorithms are evaluated plays a role of primary importance. Second, different implementations of the tested algorithms may produce very different results. It follows that the obtained results might not be easily generalizable. A different approach could be to dissect the algorithms in a series of simple operational blocks (e.g., sums, multiplications, memory accesses, etc...), and then compare the overall complexity based on the number of used operational blocks. However, even in the case, the elementary blocks may be implemented very differently from architecture to architecture and they may have different energy characteristics, thus making this approach impractical.

We take here a practical approach to derive the energy-rate characteristics of the two paradigms under investigation. We implemented a visual sensor platform composed by a BeagleBone Linux computer[4], which is equipped with a low-power camera and a IEEE 802.15.4 compliant radio transceiver, as illustrated in Figure 5(a) and 5(b). Then, we measured the power P_{cpu} consumed by the processor of the visual sensor node using an Adafruit INA219 DC current sensor as done in [37] and observed that (i) the power consumption is constant over time, and (ii) does not depend on the particular activities performed by the processor (JPEG compression, BRISK detection and description). Therefore, the energy consumption may be estimated by keeping track of the time taken by the processor for performing different processing tasks and multiplying the result by the estimated

[3] http://www.asl.ethz.ch/people/lestefan/personal/BRISK

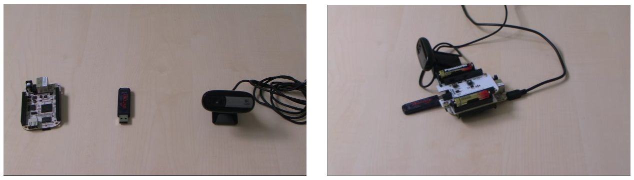
[4] http://beagleboard.org/static/beaglebone/latest/Docs/Hardware/BONE_SRM.pdf

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing



(a) CTA - JPEG encoding time (b) ATC - BRISK detection time (c) ATC - BRISK description time

Fig. 4: Indirect estimation of the processing energy through the measurement of the time needed to operate in the CTA or ATC paradigm. Fitting the lines relative to detection, description and JPEG encoding allows to retrieve the parameters in Table 2



(a) (b)

Fig. 5: (a) The main components of the reference visual sensor node platform: a BeagleBone computer (left), a Shimmer Span IEEE 802.15.4-compliant transceiver (center) and a low-power USB camera. The radio transceiver and the camera are attached to the BeagleBone to form a visual sensor node (b).

Parameter	Description	Value
P_{cpu}	CPU processing power [W]	2.1
E_{tx}	Energy for transmitting one bit [J/bit]	2.6×10^{-6}
E_{rx}	Energy for receiving one bit [J/bit]	2.9×10^{-6}
τ_{off}	Time for initializing the detector [ms/pixel]	1.6×10^{-4}
τ_{det}	Time for detecting one BRISK feature [ms]	0.31
τ_{desc}	Time for describing one BRISK feature [ms]	0.16

TABLE 2: Energy and time parameters measured from extensive experiments on a BeagleBone-based visual sensor node.

constant power consumption.

- For the CTA paradigm, we kept track of the time $t_{\text{cpu}}^{\text{CTA}}(\rho)$ spent by the visual sensor node to encode an input image with JPEG at different quality factors, as illustrated in Figure 4(a). The energy consumption due to processing can be consequently estimated as:

$$E_{\text{cpu}}^{\text{CTA}}(\rho) = P_{\text{cpu}} \cdot t_{\text{cpu}}^{\text{CTA}}(\rho), \tag{3}$$

  Therefore, the total energy for operating in the CTA paradigm at a particular target bitrate ρ can be estimated as:

$$E^{\text{CTA}}(\rho) = E_{\text{cpu}}^{\text{CTA}}(\rho) + E_{\text{tx}}(\rho). \tag{4}$$

  Note that we do not consider the energy needed for feature extraction, as this step is performed on the central controller.
- For the ATC paradigm, first we analyze the time needed by the BeagleBone to detect and describe BRISK features. As illustrated in Figure 4(b), the detection time can be modeled as a function of the number of features detected M:

$$t_{\text{det}}(M) = \tau_{\text{off}} + M\tau_{\text{det}}, \tag{5}$$

  where τ_{off} is an offset initialization time (needed to initialize the detector) and τ_{det} is the time needed to detect one feature. Note that the initialization time τ_{off} depends also on the resolution of the input image. Similarly, as illustrated in Figure 4(c), we can model the description time as:

$$t_{\text{desc}}(M) = M\tau_{\text{desc}}, \tag{6}$$

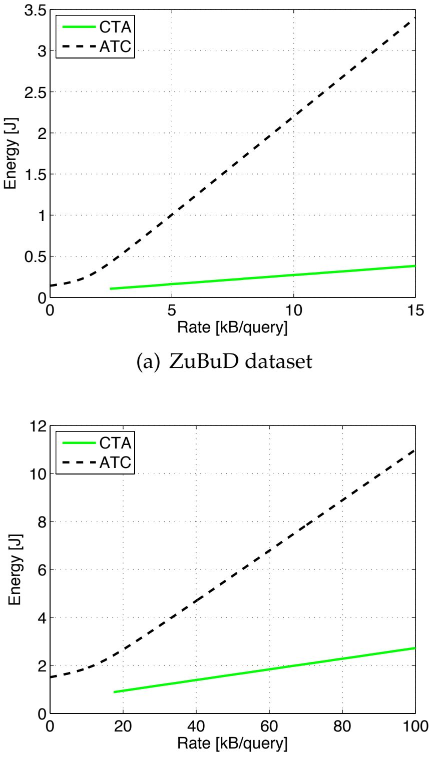
  where τ_{desc} is the time needed to describe one feature. As explained in Section 4, the ATC paradigm is operated through the use of a rate-accuracy model that computes the optimal *internal allocation* for a given bitrate target ρ, comprising the optimal number of features $M(\rho)$ to be used and the corresponding encoding parameters. It follows that, to compute the processing energy as a function of the rate ρ for the ATC paradigm we can use the following equation:

$$E_{\text{cpu}}^{\text{ATC}}(\rho) = P_{\text{cpu}} \cdot \left[\tau_{\text{off}} + M(\rho)(\tau_{\text{det}} + \tau_{\text{desc}})\right]. \tag{7}$$

  Finally, the total energy consumption in the ATC mode, is:

$$E^{\text{ATC}}(\rho) = E_{\text{cpu}}^{\text{ATC}}(\rho) + E_{\text{tx}}(\rho). \tag{8}$$

Table 2 reports the values of the energy parameters obtained in our tests, and Figures 6(a) and 6(b) shows the energy-wise comparison of the ATC and CTA paradigms for the ZuBuD and Oxford datasets. As one can see, the experiments on the BeagleBone platform reveal that the ATC paradigm generally requires a camera node to consume more energy for processing than CTA. This is due to the high cost of extracting features on the camera node, while such step is performed on the central controller in CTA. However, it is important to consider that CTA is based on JPEG encoding, whose implementation leverages more than 20 years of optimizations and refinements; on the other side, the algorithms that enable the ATC paradigm are very recent and far from being at their maximum performance energy-wise. Recently, some attention has been given to this problem, with research studies on Application Specific Integrated Circuits (ASIC) and Field Programmable Gate



(a) ZuBuD dataset

(b) Oxford Building Dataset

Fig. 6: Comparison between the rate-energy curves of the *compress-then-analyze* and the *analyze-then-compress* paradigms

Arrays (FPGA) capable of detecting and extracting features in a very efficient way [38]. As a quantitative example, the work in [39] present a feature extraction algorithm based on FAST [40] that can run at 62.5 frame per second on a Xilinx Vertex 5 FPGA chip. With the same architecture and image resolution, the work in [41] reports a peak frame rate of 45 fps for JPEG compression. This demonstrates that ATC may constitute a viable option not only when bandwidth availability is scarce, but also when the available energy is limited. The shape of the curves in Figures 6(a) and 6(b) also shows that there exists an upper bound to the camera node lifetime when operating in CTA and ATC. In fact, the ATC paradigm shows an offset energy consumption for small values of the rate ρ, which is due to the fixed energy consumption for initializing the features detector (see [16]), and which is observed to be approximately 0.15 J and 1.6 J for the ZuBuD and Oxford datasets, respectively (the energy requirements for operating with the Oxford dataset are greater due to the higher resolution of the query images with respect to ZuBuD). Similarly, the minimum consumed energy under CTA (corresponding to the minimum feasible rate for CTA) is observed to be approximately 0.1 J for ZuBud and 0.9 J for Oxford. This leads to a maximum lifetime for a camera given by $\frac{\bar{E}}{E_{\min}}$, being $\bar{E}$ the energy budget available at the camera node and $E_{\min}$ the energy offset for a particular configuration of paradigm and dataset. As an example, assuming an energy budget of $\bar{E} = 32.4$ kJ, the maximum feasible lifetimes for different configurations of paradigm and visual content are reported in Table 3.

Dataset	CTA maximum lifetime	ATC maximum lifetime
ZuBuD	311×10^3	228.5×10^3
Oxford	30.6×10^3	20.5×10^3

TABLE 3: Upper bounds to the lifetime (measured in maximum number of processed images) for camera nodes operating at different visual analysis paradigms with different visual content with an initial power budget of $\bar{E}$=32.4 kJ.

6 Resource and Paradigm Allocation in VSNs

This section leverages the rate-accuracy and rate-energy models developed in Sections 4 and 5 to cast a paradigm and resource allocation problem in visual sensor networks. The general target is to find the specific paradigm each active camera node should adopt and the related transmission rate to maximize the accuracy of the analysis task for a specific target lifetime of the visual network. The problem formulation explicitly considers bandwidth, energy, and routing constraints dictated by the individual nodes and network topology, as well as the costs of operating each camera node in the two reference paradigms.

Let $G = (V, E)$ be a directed graph that models a visual sensor network, in which V denotes the set of nodes and E denotes the set of wireless links. Two nodes i and j with $i, j \in V$ are in communication range if the directed link starting from node i to node j, $(i, j) \in E$. Without loss of generality, we consider the case of symmetric links only, that is, if $(i, j) \in E$ then $(j, i) \in E$. Here we consider heterogeneous networks, i.e., composed by both camera nodes and generic nodes. Camera nodes acquire images and depending on the operational paradigm, either encode them with JPEG or perform visual feature extraction locally. The data resulting from such visual processing is then transmitted to a remote central controller or sink node, possibly with the aid of multiple generic nodes that act as relays. Hence, let $V = C \cup N \cup S$, being C the set of camera nodes, N the set of generic/relay nodes and S the set of sink nodes. To simplify the discussion, here we assume the presence of only one sink node.

Let y_i denote a binary variable defined as

$$y_i = \begin{cases} 1 & \text{if camera } i \text{ is active} \\ 0 & \text{if camera } i \text{ is inactive} \end{cases} \tag{9}$$

and x_i the binary variables which specify the visual paradigm used by camera i and defined as:

$$x_i = \begin{cases} 1 & \text{if camera } i \text{ operates in CTA mode} \\ 0 & \text{if camera } i \text{ operates in ATC mode} \end{cases} \tag{10}$$

Finally, let $f_{(i,j)}$ denotes the flow over directed link $(i, j) \in E$, and ρ_i denotes the source rate generated by camera i with $i \in C$.

In the following, we will gradually introduce and explain the objective function and the constraints used for the problem formulation.

6.1 Objective function

We are interested in maximizing the accuracy of the analysis task performed by the visual sensor network. However, we should point out that the VSN can be composed of several camera nodes, each one possibly being active or inactive, and, if active, acquiring different visual contents. In such setting, a high level objective is to maximize the number of active cameras while, at the same time, getting the "best" accuracy out of the active cameras. Such high-level objective is composed of two contrasting sub-objectives targeting respectively "coverage" and "accuracy" of the visual task. Generally speaking, the accuracy of active cameras decreases as their number increases; this is due to the fact that a larger number of active cameras leads to higher traffic in the network. To account for both aspects, we leverage a multi-component objective function to be maximized. Let $A_i(\rho_i)$ be the rate-accuracy function of the i-th camera, the aforementioned objective can be implemented through the following combination of objective function and constraints:

$$\max \left[\alpha A^* + (1-\alpha)\left(\frac{1}{|C|}\sum_{i\in C} y_i\right)\right] \tag{11}$$

s.t.

$$A_i(\rho_i) \leq y_i[x_i A_i^{\text{CTA}}(\rho_i) + (1-x_i)A_i^{\text{ATC}}(\rho_i)]\ \forall i\in C \tag{12}$$

$$y_i[A_i(\rho_i) - A^*] \geq 0 \qquad \forall i \in C \tag{13}$$

$$A^* \leq A_i(\rho_i) \qquad \forall i \in C, \tag{14}$$

where the constraints (12-14) force the accuracy of active cameras to be higher than a reference threshold variable A^* and the objective function aims at maximizing a convex combination of A^* (minimum accuracy) and the fraction of cameras which are activated in the solution (second term, coverage). Note that other "measures" of the visual task accuracy can be used in the objective function depending on the specific application scenario: as an example, the average accuracy of active cameras can be used in place of the minimum accuracy, or, if all the cameras have overlapping field-of-views, thus they are "seeing" the same visual content, one may want that at least on one camera the visual analysis task have the maximum accuracy. In that case, one could decide to maximize the maximum accuracy, thus allocating all the available bandwidth to only one camera. Note that, in case the application executed on the VSN has particular accuracy requirements, these can be enforced by adding proper constraints on the minimum accuracy (e.g, $A^* > \bar{A}$, with $\bar{A}$ the required accuracy threshold). The same can be done on a per-camera basis, by properly inserting constraints to ensure that each camera gets a minimum accuracy (e.g., $A_i(\rho_i) \geq \bar{A}_i$). However, we do not explore this possibility in the rest of this paper.

6.2 Flow conservation constraints

The formulation of the resource allocation problem is based on a "fluidic" model, with flows of data streaming from the sources of the network (camera nodes), to a remote destination (sink nodes), through one or multiple relay nodes. Clearly, one should ensure that all the data produced by the cameras is correctly received by the sink node. This fact can be conveniently expressed using the following constraints:

$$\sum_{\substack{(i,j)\in E\\ j\in N\cup S}} f_{i,j} = y_i\rho_i \qquad \forall i \in C \tag{15}$$

$$\sum_{\substack{(k,j)\in E\\ j\in N\cup S}} f_{k,j} - \sum_{\substack{(j,k)\in E\\ j\in N\cup C}} f_{j,k} = 0 \qquad \forall k \in N \tag{16}$$

$$\sum_{i\in C} y_i\rho_i = \sum_{\substack{(j,i)\in E\\ j\in N\cup C,\, i\in S}} f_{j,i} \tag{17}$$

$$\sum_{\substack{(j,i)\in E\\ j\in V}} f_{j,i} = 0 \qquad \forall i \in C. \tag{18}$$

Constraint sets (15), (16) and (17) impose that the flow is conserved across camera nodes, relay nodes and sink nodes respectively. In this formulation, we assume that camera nodes cannot act as relays of information, thus incoming flow into camera nodes has to be set to 0 by constraints (18).

6.3 Interference a constraints

The available bandwidth in the network is limited and must be shared among sensor nodes. To ensure that transmissions of multiple nodes do not interfere with each other, one should carefully allocate the camera source rates. Such allocation should then permit to schedule the transmission of multiple nodes in such a way that neither interferences nor delays reduce the overall quality of delivery. Here, we translate this requirements by identifying subsets of interfering links in the network. The main idea is to constraint the total amount of data streamed over those links, so that scheduling is possible and interference or collisions are avoided. We assume that nodes use a mechanism similar to RTS/CTS prior to packets transmission so that two links (i,j) and (h,k) interfere with each other if and only if i) $(i,j) = (h,k)$; ii) (i,j) is adjacent to (h,k); or iii) (i,j) is adjacent to another link which is adjacent to (h,k). We can then introduce the set $I_{(i,j)}$ which includes all the links interfering with link (i,j). If the generic link (i,j) has capacity $C_{(i,j)}$, the interference constraint can be expressed as:

$$f_{i,j} + \sum_{(h,k)\in I_{(i,j)}} f_{(h,k)} \leq C_{(i,j)} \qquad \forall (i,j) \in E. \tag{19}$$

6.4 Lifetime constraints

As explained before, our formulation attains the objective of maximizing a convex combination of accuracy and coverage subject to a predefined network lifetime constraint. Here, we will use a classical definition of network lifetime, that is the period of time from the beginning of the operation of the system to the instant when the first sensor node fails due to energy depletion. To correctly express the lifetime constraints we assume that (i) the i-th node in the network starts its operation with a pre-defined energy budget $\bar{E}_i$, and (ii) the lifetime L of the VSN is expressed in terms of the number of consecutive images that can be acquired and transmitted by each camera. Let T be the period of acquisition (i.e., the inverse of the frame rate of the system), in seconds, which may be tuned according to the specific

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing

applications. With these assumptions, the lifetime of the network can be expressed as either L image acquisitions or $L \times T$ seconds. As explained in Section 5, the energy spent by each node in the sensor network is determined by two components, namely the transmission/reception and processing energy. Depending on the role of each node in the network, we identify two possibilities:

1) *Relay nodes:* relay nodes in the sensor network consume E_{tx} Joules per bit transmitted and E_{rx} joules per bit received. The lifetime constraint for relay nodes can be thus expressed as:

$$E_{\text{tx}} \sum_{(k,j)\in E} f_{k,j} + E_{\text{rx}} \sum_{(j,k)\in E} f_{j,k} \le \frac{\bar{E}_k}{L} \qquad \forall k \in R. \tag{20}$$

2) *Camera nodes:* camera nodes consume energy to acquire and process images, and to transmit the multimedia content (whether it refers to a compressed-image or to compressed features). Depending on the visual analysis paradigm used by each camera (i.e., compress-then-analyze or analyze-then-compress), the processing energy will follow the behavior illustrated in Section 5. Without loss of generality, let $E^{\text{CTA}}_{\text{cpu},i}$ and $E^{\text{ATC}}_{\text{cpu},i}$ be the processing energy consumed by the i-th camera node in either the CTA or ATC paradigm (expressed in Joules per bit); the lifetime constraint for camera nodes can be expressed as:

$$E_{\text{tx},i}\rho_i + E^{\text{CTA}}_{\text{cpu},i} \le x_i \frac{\bar{E}_i}{L} + (2 - x_i - y_i)K \qquad \forall i \in C \tag{21}$$

$$E_{\text{tx},i}\rho_i + E^{\text{ATC}}_{\text{cpu},i} \le (1 - x_i) \frac{\bar{E}_i}{L} + (1 + x_i - y_i)K \quad \forall i \in C \tag{22}$$

where K is set to a sufficiently big value in order to (i) satisfy the constraint which is not active in the particular solution (i.e., constraint (22) if the camera is in CTA mode or constraint (21) if the camera is in ATC mode), or (ii) satisfy both constraints (21) and (22) when a camera is inactive, regardless to the value of x_i.

The complete formulation for the resource and paradigm allocation problem in VSNs is reported hereafter:

$$\max \left[\alpha A^* + (1-\alpha)\left(\frac{1}{|C|}\sum_{i\in C} y_i\right)\right], \tag{23}$$

$$\begin{aligned}
&\text{s.t.}(12) - (22) && \\
&\rho_i \le y_i K && \forall i \in C \quad (24)\\
&y_i x_i(\rho_i - \rho^{\text{CTA}}_{\text{min}}) \ge 0 && \forall i \in C \quad (25)\\
&y_i(1-x_i)(\rho_i - \rho^{ATC}_{\text{max},i}) \le 0 && \forall i \in C \quad (26)\\
&A^* \in [0,1] && (27)\\
&A_i(\rho_i) \in [0,1] && \forall i \in C \quad (28)\\
&y_i \in \{0,1\} && \forall i \in C \quad (29)\\
&x_i \in \{0,1\} && \forall i \in C \quad (30)\\
&f_{i,j} \in \mathbb{R}^+ && \forall (i,j) \in E \quad (31)
\end{aligned}$$

where constraint (24) forces the source rate of inactive camera to be 0. Constraints (25) and (26) can be explained by referring to Figure 3, which reports the rate-accuracy curves when operating in CTA or ATC mode; it is clear from the figure that there is a minimum rate below which CTA cannot be used ($\rho^{\text{CTA}}_{\text{min}}$); moreover, it is also clear that if the achievable rate of a camera exceeds $\rho^{\text{ATC}}_{\text{max}}$, that camera is better off operating in CTA as this provides higher accuracy. In other words, if the rate achievable by a camera is below $\rho^{\text{CTA}}_{\text{min}}$ or above $\rho^{\text{ATC}}_{\text{max}}$, the camera is forced to run ATC or CTA paradigms, respectively. We can thus exploit such properties to introduce in the formulation the two additional constraints (25), which impose minimum source rate for active cameras operating in CTA, and constraint (26), which impose that the source rate of an active camera operating in ATC cannot exceed to maximum rate $\rho^{\text{ATC}}_{\text{max}}$. Finally, constraints (27)-(31) define the decision variables of the formulation. Operatively, the resource allocation problem may be solved in a centralized fashion on the base station once the topology and the application requirements are known. The solution (source rate and operative paradigm) can be then transmitted to each camera in the network. Moreover, the base station may compute and transmit a new optimal solution whenever changes in topology or application requirements occur.

6.5 Generalization to other visual analysis tasks and hardware platforms

The proposed resource and paradigm allocation problem is built around network specific constraints and the rate-accuracy and rate-energy models derived in Section 4 and 5, respectively[5]. Although such rate-energy-accuracy models were specifically derived for the case of image retrieval and object recognition applications, the resource allocation problem may be easily generalized to other application scenarios, comprising different system assumptions or even totally different visual analysis tasks. As an example, the model can be easily adapted to the case where visual queries are obtained starting from video streams rather then from still images. In that case, a camera operating according to the CTA paradigm would acquire and transmit a compressed video stream (e.g., by using a state-of-the-art encoder such as MPEG or H.264/AVC). In the case of ATC, recent schemes for encoding visual features extracted from video sequences may be adopted [42]. Clearly, adapting the resource allocation problem to this scenario requires the experimental derivation of the rate-accuracy and rate-energy models for both CTA and ATC starting from video sequences. Following the same idea, it would be possible to use the proposed framework also when another hardware platform, different from the BeagleBone, is available. In this case, in order to re-use the proposed resource allocation and paradigm problem, it is required to derive experimentally a proper rate-energy model for the platform under consideration. As an example, we performed the experimental energy evaluation proposed in Section 5 using a Raspberry PI model B platform in place of the BeagleBone. Our analysis revealed that the same energy model can be used, that is the

[5] Note also that the proposed rate-accuracy and rate-energy models can be also fused to produce a parametric energy-accuracy model, using the rate ρ as parameter. Such representation may be easier to use for comparing the two paradigms in specific scenarios, e.g., when rate cannot be directly controlled/measured [37].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2016.2519340, IEEE Transactions on Mobile Computing

processing power P_{cpu} remains constant (and equal to 2.31 W in this case) and therefore energy can be estimated solely based on the processing time.

7 Experimental results

This section comments on the solutions of the resource allocation problem for simulated visual sensor networks instances. The simulation instances are created with Matlab and characterized by tunable input parameters including the number of deployable camera nodes, $c = |C|$, the number of relay nodes, $n = |N|$, and the number h of routing hops along the paths between each camera node and the information sink. Moreover, on each camera node, we randomly select one visual content between the ZuBuD dataset and the Oxford dataset characteristics. Depending on the chosen dataset, the energy parameters and the rate-accuracy behavior of camera nodes are set according to the characteristics of the BeagleBone-based visual sensor nodes presented in Sections 4 and 5. Assuming that each BeagleBone (either camera or relay) is powered by 4 AA batteries, each node in the network starts its operations with an energy budget $\bar{E}_i = \bar{E} = 32.4$ kJ. The capacity of each link in the network is set to 31.25 kilobytes per second (i.e., 250 kbps), and the application frame rate is set to $f = 1$ query/second or $f = 0.1$ query/second We formalized the optimization problem defined in (23)-(31) through AMPL [43] and for each network instance characterized by a specific parameters tuple <*c*,*n*,*h*,*f*>, we produced 10 realizations of the relative network topology, and the corresponding AMPL data files. Then, we solved each problem instance using the Bonmin solver [44], and averaged the results. All tests were carried out on a 2.3 GHz Intel Core 2 Duo PC under Windows.

The quality of the solutions is evaluated with respect to the following three performance metrics:

1) *Coverage (COV)*: the fraction of cameras which are active in the solution, that is, the second term of the objective function in (11):
$$COV = \frac{\sum_{i \in C} y_i}{c};$$
2) *Minimum Guaranteed Accuracy (MGA)*: the minimum value of accuracy over all the cameras which are active in the solution, that is, A^* in the objective function (11);
3) *Lifetime Accuracy Ratio (LAR)*: the two aforementioned performance measures, namely COV and MGA, are able to capture the performance of a specific solution of the resource allocation problem, that is, for a particular lifetime constraint L. Let $\mathcal{J}$ be the optimal value of the objective function defined in (11). Clearly, $\mathcal{J}$ is a function of the independent variable L. The function $\mathcal{J}(L)$ corresponds to the solid line in the upper row of Figure 7. To generalize such measures to any possible lifetime constraint, we compute the (normalized) area under the curve defined by $\mathcal{J}(L)$, that is:
$$LAR = \frac{\int_0^{L_{\text{MAX}}} \mathcal{J}(L)dL}{L_{\text{MAX}}},$$
where the upper integral limit L_{MAX} can be selected arbitrarily. By definition, the LAR ranges between 0 and 1, and allows to compare different solutions of the resource allocation problem in terms of their energy efficiency and accuracy, considering all possible values for the lifetime constraint.

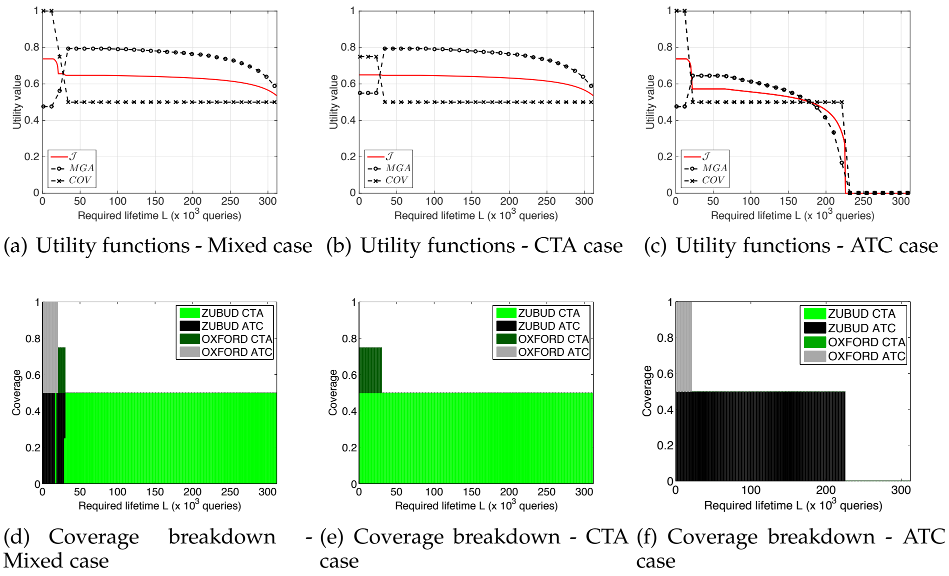
7.1 Walk-through Example

To clarify the dynamics behind the solutions to the resource allocation problem, we start off by considering a sample scenario where we simulated a network instance composed of four camera nodes directly connected to a single sink node.

We assume that two cameras work according to the ZuBuD dataset parameters and the other two according to the Oxford dataset parameter. Moreover, we set the parameter α in (11) to 0.5.

Fig. 7(a) and 7(d) summarize the results obtained by solving the optimization problem in the aforementioned sample topology. Figure 7(a) shows the behavior of the coverage, COV, the minimum guaranteed accuracy, MGA and the utility function $\mathcal{J}$ as a function of the required lifetime L, whereas Fig. 7(d) gives the corresponding breakdown of the active camera types distinguished in ZuBud and Oxford cameras (zoomed for the sake of readability in Figure 8). Three operation regions can be appreciated in the curves of Fig. 7(a):

- $0 \leq L \leq 20 \times 10^3$: all four cameras are active ($COV = 1$) and they are all using ATC. The choice of ATC is driven by the fact that the available bandwidth per camera is below the lower bound of the bandwidth for operating in CTA (see Section 4.3); namely, the maximum available bandwidth per camera is $31.25/4 = 7.81$ kB/query, which is below the minimum required rate for operating Oxford cameras in CTA, that is, 17.3 kB/query. As illustrated in Fig. 8, for $L = 17 \times 10^3$ and $L = 18 \times 10^3$, one of the ZuBuD cameras switch to CTA mode. This doesn't violate any problem constraints, and does not modify the value of the objective function, which is lower bounded by the value of accuracy given by the Oxford cameras.
- $21 \times 10^3 \leq L \leq 30 \times 10^3$, having all cameras active and operating in ATC mode is no longer feasible as the target lifetime L is now higher than the upper bound of the lifetime for Oxford cameras operating in ATC (see Table 3). Conversely, operating the Oxford cameras in CTA is not feasible as the required rate would exceed the available bandwidth. The solution thus turns off one of the Oxford cameras and changes the operation paradigm of the other Oxford camera to CTA as more bandwidth is now available (the bandwidth has to be split among three cameras only). Also in this case, as showed in Fig. 8, for $L = 29 \times 10^3$ and $L = 30 \times 10^3$ one of the two ZuBuD cameras switches to CTA mode. Again, this does not violate any constraints and does not modify the value of the objective function.
- $L \geq 31 \times 10^3$, the required rate to operate the Oxford camera in CTA at the target lifetime is no longer feasible as the target lifetime is higher than the lifetime upper bound for Oxford cameras operating in CTA (see Table 3); in the solution, the Oxford camera is turned off and the two Zubud cameras switch to the CTA paradigm as more bandwidth in now available (the bandwidth has to be split among only two cameras now).



(a) Utility functions - Mixed case (b) Utility functions - CTA case (c) Utility functions - ATC case

(d) Coverage breakdown - Mixed case (e) Coverage breakdown - CTA case (f) Coverage breakdown - ATC case

Fig. 7: Utility functions and coverage breakdown for (a,d) the mixed case, (b,e) the CTA case and (c,f) the ATC case.

We also compare the cases where all the cameras in the network use only the CTA (Figures 7(b) and 7(e)) or the ATC (Figures 7(c) and 7(f)) paradigm, against the solution of the proposed formulation, which allows solutions where the two paradigms coexist. Operatively, this is achieved by forcing the value of the binary variables x_i in (30) to 1 or 0 respectively. By looking at the solid line (i.e., the curve relative to $\mathcal{J}$) in Figure 7(a) and 7(b), it is clear that the mixed case outperforms CTA for low values of the required lifetime. This is due to the smaller coverage achievable by CTA due to bandwidth constraints. This is generally true, as CTA cannot cope with very low-bandwidth regimes (differently from ATC) and sacrifices the functioning of the most bandwidth-eager cameras to free bandwidth for the other cameras. Hence, in the CTA-only case, coverage is penalized. Conversely, by looking at the solid line in Figure 7(a) and 7(c), it is clear that the mixed approach outperforms ATC for high values of the require lifetime: this is again generally true, as ATC is more energy-eager than CTA. Generally speaking, ATC should be used when bandwidth is the primary constraint, while CTA when energy is the primary constraint. In a situation where the two dimensions (energy/bandwidth) are equally important and changing with time, the mixed paradigm allows to obtain the best results by adapting to the particular network conditions.

7.2 General Topologies

To generalize the results obtained for the sample topology, we tested the solution of the proposed optimization model on several synthetic network topology instances. Specifically, we varied the number of cameras c, the number of network hops h and the application frame rate f. The number of network hops set as input parameter impacts on the number of relay nodes that each camera uses to deliver the visual data to the sink node. Each topology is created by randomly deploying camera nodes and relay nodes in a variable-sized area, whose dimensions are adjusted based on the number of network hops h. The communication range of each node in the topology is set to 15 m and some of the radio links were deleted randomly to simulate an indoor environment.

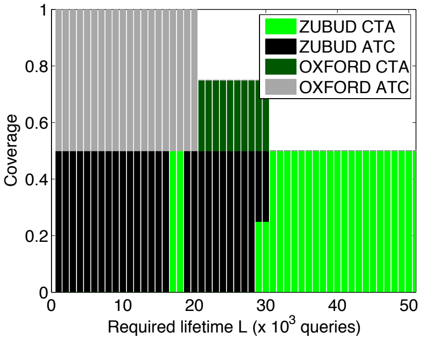


Fig. 8: Detailed (zoomed) version of Fig. 7(d)

Then, for each topology, we create and solve several instances of the resource allocation problem, by varying the lifetime constraint L, from 1 to L_{MAX}. For each solved instance, we analyze the lifetime-accuracy tradeoff that results from the solution of the problem. Again, we compare the solution obtained with the proposed model with the cases in which only the CTA and ATC paradigms are used (i.e., a traditional scenario). We report LAR values for two representative cases: one in which L_{MAX} is set to a small value, representing a loose lifetime constraint, and one in which L_{MAX} is high. For each configuration of the topology parameters, the reported LAR value has been obtained by averaging over 10 different topologies realizations.

Table 4 reports the LAR values for different numbers of camera nodes c, different values of the network diameter (network hops h) and different image retrieval query rate f, when L_{MAX} is set to 18×10^3 (e.g. 5 days of network utilization at 1 query per second). Similarly, Table 5 reports

f	h	c	Mixed	ATC Only	CTA Only
1 query/second	1	2	**0.74**	0.70	0.73
		4	**0.70**	**0.70**	0.61
		6	**0.69**	**0.69**	0.60
	2	2	**0.70**	**0.70**	0.61
		4	**0.66**	**0.66**	0.61
		6	**0.67**	**0.67**	0.60
0.1 query/second	1	2	**0.77**	0.70	**0.77**
		4	**0.77**	0.70	**0.77**
		6	**0.77**	0.70	**0.77**
	2	2	**0.77**	0.70	**0.77**
		4	**0.77**	0.70	**0.77**
		6	**0.76**	0.70	0.75

TABLE 4: Lifetime Accuracy Ratio (LAR) values for different network topologies when $L_{\text{MAX}} = 18 \times 10^3$ queries

f	h	c	Mixed	ATC Only	CTA Only
1 query/second	1	2	**0.66**	0.56	0.65
		4	**0.65**	0.56	0.64
		6	**0.65**	0.56	0.64
	2	2	**0.65**	0.56	0.64
		4	**0.64**	0.55	0.63
		6	**0.64**	0.55	0.62
0.1 query/second	1	2	**0.67**	0.56	**0.67**
		4	**0.67**	0.56	**0.67**
		6	**0.67**	0.56	**0.67**
	2	2	**0.66**	0.56	**0.66**
		4	**0.66**	0.55	**0.66**
		6	**0.66**	0.55	0.64

TABLE 5: Lifetime Accuracy Ratio (LAR) values for different network topologies when $L_{\text{MAX}} = 180 \times 10^3$ queries

the LAR values when L_{MAX} is increased by ten times (i.e., 180×10^3).

From our results, the following comments can be made:

1) When a loose lifetime constraint is set (e.g., Table 4) and for high application frame rates (1 query per second, rows from 1 to 6), *bandwidth* is the primary limitation and dominates over the energy constraint. In most of these cases, ATC outperforms CTA as it allows for low-bitrate utilization. Conversely, for low application frame rates (0.1 query per second, rows from 7 to 12) *energy* is the primary limitation and dominates over bandwidth. Thus, in these cases, CTA outperforms ATC.
2) Conversely, when a tight lifetime constraint is set (e.g., Table 5) and regardless to the input application frame rate, CTA outperforms ATC. Again, this is due to the fact that energy is the dominating constraint and CTA shows better rate-energy performance than ATC.
3) In general, and not surprisingly, increasing the number of cameras c, the network hops h or the application frame rate f decreases the achievable LAR value.
4) In all cases, the proposed mixed approach which allows the coexistence of the two paradigms shows the best performance. Clearly, this is due to the flexibility of the proposed solution, which can adapt to the particular network topology conditions and setup of problem parameters.

Cameras	Mixed	ATC-only	CTA-only
2	**0.88**	0.84	**0.88**
4	**0.82**	**0.82**	0.675
6	**0.79**	**0.79**	0.55

TABLE 6: Values of the objective function $\mathcal{J}$ computed on experiments with a real-life VSN testbed based on BeagleBone camera nodes

7.3 Real-life testbed

We implemented both CTA and ATC approaches on several BeagleBone-based visual sensor nodes coupled with IEEE 802.15.4 compliant transceivers. Visual sensor nodes acquire and transmit visual data (compressed images or features) to a sink node [45]. There, a graphical user interface allows to control the operative paradigm to use on each camera, including the possibility to tune operational parameters such as the JPEG quality factor or the number of features to transmit. Additionally, an object recognition engine is implemented on the sink, thus allowing to compute performance measures such as the MAP. Such information, together with the rate-energy model validated with real-life measurements (see Section 5), allow to solve the resource allocation problem for any value of desired lifetime. To demonstrate the feasibility of the proposed method in a real-life experiment, we deployed three different VSN topologies composed by 2, 4 and 6 BeagleBone camera nodes The camera nodes were in direct communication range with the sink, and the bandwidth achievable with the IEEE 802.15.4 transceivers was estimated to be equal to 4 kBytes per second. Assuming a frame rate of f = 1fps, it is clear that such bandwidth does not allow to use the Oxford dataset in CTA mode (for which the minimum bandwidth is 17.3 kBytes per second, see Fig. 3(b)). Therefore, we pre-loaded on each camera node query images from the ZuBuD dataset only. For each of the three topologies, we solved the resource allocation problem (with $\alpha = 0.5$, $L = 18 \times 10^3$ and $f = 1$ fps) and operate the camera nodes according to the optimal solution returned by the problem. That is, each camera is operated according either to CTA or ATC. If the former is selected, the JPEG quality factor is tuned so that the output rate matches the rate returned by the optimal solution. The same is done for ATC, properly selecting the number of features to transmit. Note that some cameras may be also turned-off. In such conditions, we computed the MAP over the ZuBuD query dataset for each camera and the number of active cameras, and evaluated the optimal value of the objective function $\mathcal{J}$, defined in (11) and capturing both the object recognition accuracy and the VSN coverage. Results were compared with the ATC-only or CTA-only approaches. Table 6 reports the computed value of $\mathcal{J}$ for the three topologies. As one can see, the proposed mixed approach allows to obtain the best results in terms of minimum guaranteed accuracy and coverage. Also, it can be observed that the performance of the CTA-only approach rapidly decrease as the number of camera increases: due to the limited bandwidth available, one has to either switch off some camera nodes or use a very low quality factor, thus resulting in poor application performance.

8 Conclusions

In this paper, we considered VSNs deployed to support applications based on object recognition. We focused on two complementary paradigms for performing remote visual analysis: the traditional *compress-then-analyze* (CTA) paradigm, where images are compressed and transmitted to a central controller for analysis, and a novel, alternative paradigm called *analyze-then-compress* (ATC). In this latter paradigm, camera nodes extract visual features from the acquired images. Such features are then transmitted to a central controller for further analysis. Through experiments on a real visual sensor node testbed, we characterized the two paradigms in terms of their rate-energy and rate-accuracy performance, showing that ATC allows for low-bandwidth visual analysis at the cost of higher energy consumption. Then, we formulated a resource allocation problem leveraging the proposed rate-energy-accuracy models and showed through simulations that the best results are obtained when the two paradigms are allowed to coexist in the network. Future works will address the problem of filling the gap between the energy performance of the two paradigms: this may be achieved by properly optimizing the execution of state-of-the-art features extraction algorithms, either with software or hardware design methodologies. Also, we plan to extend the presented rate-energy-accuracy modeling to other tasks that may be supported by VSNs, such as pedestrian detection and tracking.

References

[1] A. Redondi, M. Cesana, and M. Tagliasacchi, “Rate-accuracy optimization in visual wireless sensor networks,” in *Proc. 2012 19th IEEE Int. Conf. on Image Processing (ICIP)*. IEEE, 2012, pp. 1105–1108.

[2] A. Redondi, L. Baroffio, M. Cesana, and M. Tagliasacchi, “Compress-then-analyze vs. analyze-then-compress: Two paradigms for image analysis in visual sensor networks,” in *Proc. 2013 IEEE 15th Int. Workshop on Multimedia Signal Processing (MMSP)*, Sept 2013, pp. 278–282.

[3] A. Zabala and X. Pons, “Effects of lossy compression on remote sensing image classification of forest areas,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 1, pp. 43 – 51, 2011.

[4] ——, “Impact of lossy compression on mapping crop areas from remote sensing,” *International Journal of Remote Sensing*, vol. 34, no. 8, pp. 2796–2813, 2013.

[5] A. Tsifouti, M. M. Nasralla, M. Razaak, J. Cope, J. M. Orwell, M. G. Martini, and K. Sage, “A methodology to evaluate the effect of video compression on the performance of analytics systems,” *Proc. SPIE*, pp. 85 460S–85 460S–15, 2012.

[6] A. D. Bagdanov, M. Bertini, A. Del Bimbo, and L. Seidenari, “Adaptive video compression for video surveillance application,” in *Proc. of IEEE International Symposium on Multimedia (ISM)*, Dana Point, CA, USA, 2011.

[7] G. Gualdi, A. Prati, and R. Cucchiara, “Video streaming for mobile video surveillance,” *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1142–1154, 2008.

[8] G. Lu, B. Krishnamachari, and C. Raghavendra, “Performance evaluation of the IEEE 802.15.4 mac for low-rate low-power wireless networks,” in *IEEE Internat. Conf. on Perform., Comp., and Comm.*. IEEE, 2004, pp. 701–706.

[9] Y. Kwon and D. Shin, “The security monitoring system using IEEE 802.15.4 Protocol and CMOS Image Sensor,” in *Proc. IEEE Internat. Conf. on New Trends in Inf. and Serv. Sci., NISS ’09*. IEEE, 2009, pp. 1197–1202.

[10] P. Chen, P. Ahammad, C. Boyer, S.-I. Huang, L. L., E. Lobaton, M. Meingast, S. Oh, S. Wang, P. Yan, A. Yang, C. Yeo, L.-C. Chang, J. D. Tygar, and S. Sastry, “CITRIC: A low-bandwidth wireless camera network platform,” in *ACM/IEEE Int. Conf. on Distrib. Smart Cam., ICDSC*, Sept., pp. 1–10.

[11] Y. Charfi and B. Canada, “Challenging issues in visual sensor networks,” *IEEE Wireless Comm.*, vol. 16, no. 2, pp. 44–49, Apr. 2009.

[12] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, “From today’s INTRAnet of things to a future INTERnet of things: a wireless- and mobility-related view,” *IEEE Wireless Comm.*, vol. 17, no. 6, pp. 44–51, Dec. 2010.

[13] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions,” *Future Gen. Comp. Syst. J.*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013.

[14] S. Paniga, L. Borsani, A. Redondi, M. Tagliasacchi, and M. Cesana, “Experimental evaluation of a video streaming system for wireless multimedia sensor networks,” in *IEEE/IFIP Annual Ad Hoc Netw. Worksh.* IEEE, 2011, pp. 165–170.

[15] W. Yu, Z. Sahinoglu, and A. Vetro, “Energy efficient JPEG 2000 image transmission over wireless sensor networks,” in *Proc. IEEE Global Telecom. Conf., GLOBECOM*, vol. 5. IEEE, 2004, pp. 2738–2743.

[16] A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi, J. Ascenso, and R. Cilla, “Evaluation of low-complexity visual feature detectors and descriptors,” in *Digital Signal Processing (DSP), 2013 18th International Conference on*. IEEE, 2013, pp. 1–7.

[17] R. Madan, S. Cui, S. Lall, and A. Goldsmith, “Cross-layer design for lifetime maximization in interference-limited wireless sensor networks,” *Wireless Communications, IEEE Transactions on*, vol. 5, no. 11, pp. 3142–3152, 2006.

[18] I. Dietrich and F. Dressler, “On the Lifetime of Wireless Sensor Networks,” *ACM Transactions on Sensor Networks (TOSN)*, vol. 5, no. 1, pp. 1–39, February 2009.

[19] S. Cicalo and V. Tralli, “Distortion-fair cross-layer resource allocation for scalable video transmission in ofdma wireless networks,” *Multimedia, IEEE Transactions on*, vol. 16, no. 3, pp. 848–863, April 2014.

[20] K. Pandremmenou, L. Kondi, and K. Parsopoulos, “Geometric bargaining approach for optimizing resource allocation in wireless visual sensor networks,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 8, pp. 1388–1401, Aug 2013.

[21] K. Lin, W.-L. Shen, C.-C. Hsu, and C.-F. Chou, “Quality-differentiated video multicast in multirate wireless networks,” *Mobile Computing, IEEE Transactions on*, vol. 12, no. 1, pp. 21–34, Jan 2013.

[22] H. Park and M. van der Schaar, “Bargaining strategies for networked multimedia resource management,” *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3496–3511, July 2007.

[23] Y. He and L. Guan, “Optimal resource allocation for video communication over distributed systems,” in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 1414–1423.

[24] Z. He, Y. Liang, L. Chen, I. Ahmad, and D. Wu, “Power-rate-distortion analysis for wireless video communication under energy constraints,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 5, pp. 645–658, 2005.

[25] Z. Guan, T. Melodia, and D. Yuan, “Jointly Optimal Rate Control and Relay Selection for Cooperative Wireless Video Streaming,” *IEEE/ACM Transactions on Networking*, vol. 21, no. 4, pp. 1173–1186, August 2013.

[26] C. Li, J. Zou, H. Xiong, and Y. Zhang, “Joint coding/routing optimization for correlated sources in wireless visual sensor networks,” in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, 2009, pp. 1–8.

[27] C. Li, J. Zou, H. Xiong, and C. W. Chen, “Joint coding/routing optimization for distributed video sources in wireless visual sensor networks,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 2, pp. 141–155, 2011.

[28] J. Zou, H. Xiong, C. Li, R. Zhang, and Z. He, “Lifetime and distortion optimization with joint source/channel rate adaptation and network coding-based error control in wireless video sensor networks,” *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 3, pp. 1182–1194, 2011.

[29] B. Peng, J. Zou, C. Tan, and M. Wang, “Network lifetime optimization in wireless video sensor networks,” in *Wireless Mobile and Computing (CCWMC 2009), IET International Communication Conference on*, 2009, pp. 172–175.

[30] Y. Chen, X. Hu, H. Yang, and L. Ge, “Power control routing algorithm for maximizing lifetime in wireless sensor networks,” in *Advances in Mechanical and Electronic Engineering*, ser. Lecture Notes in Electrical Engineering, D. Jin and S. Lin, Eds., 2013, vol. 178, pp. 129–136.

[31] Y. He, I. Lee, and L. Guan, “Distributed algorithms for network lifetime maximization in wireless visual sensor networks,” *Circuits*

and Systems for Video Technology, IEEE Transactions on, vol. 19, no. 5, pp. 704–718, 2009.

[32] J. Jang, G. Kim, and C.-M. Kyung, "Lifetime elongation of event-driven wireless video sensor networks," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, 2013, pp. 437–440.

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

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

[35] S. Leutenegger, M. Chli, and R. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 2548–2555.

[36] A. Redondi, L. Baroffio, J. Ascenso, M. Cesana, and M. Tagliasacchi, "Rate-accuracy optimization of binary descriptors," in *Proc. 2013 20th IEEE Int. Conf. on Image Processing (ICIP)*, Sept 2013.

[37] L. Bondi, L. Baroffio, M. Cesana, A. E. Redondi, and M. Tagliasacchi, "A visual sensor network for parking lot occupancy detection in smart cities," in *IEEE 2nd World Forum on Internet of Things*. IEEE, 2015.

[38] J.-S. Park, H.-E. Kim, and L.-S. Kim, "A 182 mw 94.3 f/s in full hd pattern-matching based image recognition accelerator for an embedded vision system in 0.13-mu cmos technology," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 5, pp. 832–845, 2013.

[39] K. Dohi, Y. Yorita, Y. Shibata, and K. Oguri, "Pattern compression of fast corner detection for efficient hardware implementation." in *FPL*, 2011, pp. 478–481.

[40] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ser. ECCV'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 430–443. [Online]. Available: http://dx.doi.org/10.1007/11744023_34

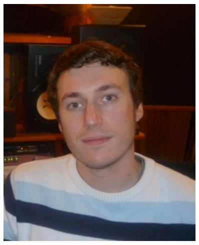
[41] D. Modrzyk and M. Staworko, "A high-performance architecture of jpeg2000 encoder," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 569–573.

[42] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding visual features extracted from video sequences," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2262–2276, May 2014.

[43] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL : a modeling language for mathematical programming*. South San Francisco: Scientific Press, 1993, computer disk label: AMPL student edition.

[44] P. Bonami, L. Biegler, A. Conn, G. Cornuejols, I. Grossmann, G. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Waechter, "An algorithmic framework for convex mixed integer nonlinear programs." in *IBM Research Report RC23771*, oct. 2005.

[45] L. Bondi, L. Baroffio, M. Cesana, A. E. Redondi, and M. Tagliasacchi, "Open-source and flexible framework for visual sensor networks," in *Proceedings of the 9th International Conference on Distributed Smart Camera*, ser. ICDSC '15. New York, NY, USA: ACM, 2015, pp. 197–198. [Online]. Available: http://doi.acm.org/10.1145/2789116.2802650



Alessandro Redondi received the MS in Computer Engineering in July 2009 and the Ph.D. in Information Engineering in 2014, both from Politecnico di Milano. From September 2012 to April 2013 was a visiting student at the EEE Department of the University College of London (UCL). His research activities are focused on algorithms and protocols for Visual Sensor Networks.

Lucio Bianchi Lucio Bianchi received the Bachelor degree in Electronic Engineering and the Master of Science in Computer Engineering from Politecnico di Milano, Milano, IT, in 2010 and 2012 respectively. He is currently a Ph.D student at Image and Sound Processing Group in Dipartimento di Elettronica, Informazione e Bioingegneria at Politecnico di Milano. His research focuses on sampling, processing and reconstruction of acoustic fields.

Luca Baroffio received the M.Sc. degree (2012, cum laude) in Computer Engineering from Politecnico di Milano, Milan, Italy. He is currently pursuing the Ph.D. degree in Information Technology at the "Dipartimento di Elettronica e Informazione - Politecnico di Milano", Italy. In 2013 he was visiting scholar at "Instituto de Telecomunicações", Lisbon, Portugal.

His research interests are in the areas of multimedia signal processing and visual sensor networks.

Matteo Cesana is currently an Assistant Professor with the Dipartimento di Elettronica, Informazione e Bioingegneria of the Politecnico di Milano, Italy. He received his MS degree in Telecommunications Engineering and his Ph.D. degree in Information Engineering from Politecnico di Milano in July 2000 and in September 2004, respectively. From September 2002 to March 2003 he was a visiting researcher at the Computer Science Department of the University of California in Los Angeles (UCLA). His research activities are in the field of design, optimization and performance evaluation of wireless networks with a specific focus on wireless sensor networks and cognitive radio networks. Dr. Cesana is an Associate Editor of the Ad Hoc Networks Journal (Elsevier).

Marco Tagliasacchi is currently Assistant Professor at the "Dipartimento di Elettronica e Informazione - Politecnico di Milano", Italy. He received the "Laurea" degree (2002, cum Laude) in Computer Engineering and the Ph.D. in Electrical Engineering and Computer Science (2006), both from Politecnico di Milano. He was visiting academic at the Imperial College London (2012) and visiting scholar at the University of California, Berkeley (2004).

His research interests include multimedia forensics, multimedia communications (visual sensor networks, coding, quality assessment) and information retrieval. Dr. Tagliasacchi coauthored more than 120 papers in international journals and conferences, including award winning papers at MMSP 2013, MMSP2012, ICIP 2011, MMSP 2009 and QoMex 2009. He has been actively involved in several EU-funded research projects. He is currently co-coordinating two ICT-FP7 FET-Open projects (GreenEyes - www.greeneyesproject.eu, REWIND - www.rewindproject.eu).

Dr. Tagliasacchi is an elected member of the IEEE Information Forensics and Security Technical committee for the term 2014-2016, and served as member of the IEEE MMSP Technical Committee for the term 2009-2012. He is currently Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technologies (2011 best AE award) and APSIPA Transactions on Signal and Information Processing. Dr. Tagliasacchi was General co-Chair of IEEE MMSP 2013 (Pula, Italy) and he was Technical Program Coordinator of IEEE ICME 2015 (Turin, Italy).