A Game-Theoretic Approach for Cooperative Feature Extraction in Camera Networks

Alessandro E. C. Redondi[1,*], **Luca Baroffio**[1], **Matteo Cesana**[1], **Marco Tagliasacchi**[1]
[1]Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Via Ponzio 34/5, Milano

Abstract. Visual Sensor Networks consist of several camera nodes with wireless communication capabilities that can perform visual analysis tasks such as object identification, recognition and tracking. Often, VSNs deployments result in many camera nodes with overlapping fields of view. In the past, such redundancy has been exploited in two different ways: (i) to improve the accuracy/quality of the visual analysis task by exploiting multi-view information or (ii) to reduce the energy consumed for performing the visual task, by applying temporal scheduling techniques among the cameras. In this work, we propose a game theoretic framework based on the Nash Bargaining Solution to bridge the gap between the two aforementioned approaches. The key tenet of the proposed framework is for cameras to reduce the consumed energy in the analysis process by exploiting the redundancy in the reciprocal fields of view. Experimental results in both simulated and real-life scenarios confirm that the proposed scheme is able to increase the network lifetime, with a negligible loss in terms of visual analysis accuracy.

Keywords: Features Extraction, Object Recognition, Game Theory, Nash Bargaining Solution.

*Alessandro Enrico Redondi, alessandroenrico.redondi@polimi.it

1 Introduction

Wireless camera networks, also known as Visual Sensor Networks (VSN), are becoming more and more popular nowadays, supporting advanced pervasive tasks such as surveillance and environmental monitoring.[1,2] Composed of several battery-operated wireless camera nodes, VSNs are expected to play a key role in the context of Smart Cities where they can be used to implement traffic congestion detection, parking lot monitoring and several other tasks.[3] Moreover, being battery operated and capable of wireless communication, VSNs require no fixed power and communication infrastructures, thus they can be deployed rapidly and with limited installation costs. Camera networks are particularly challenging from the research point of view: acquisition, processing and transmission of visual data are resource-eager operations which are at odds with the resource-constrained environment typical of such networks, characterized by limited transmission

bandwidth, processing power and energy budget. This clash between application requirements and platform resources calls for novel solutions in the design process of VSNs.

The traditional system design for VSNs follows a *compress-then-analyze* (CTA) paradigm, where images (or videos) are acquired and compressed locally at the camera nodes, and then transmitted to one or multiple information sinks which perform the specific analysis tasks (video surveillance, face detection, object recognition, etc). Recently, a paradigm shift has emerged: according to the *analyze-then-compress* (ATC) paradigm, the visual content is processed locally at the camera nodes to extract a concise representation constituted by local visual features. In a nutshell, salient keypoints are detected in the acquired image, and a visual feature is computed for each keypoint by properly summarizing the photometric properties of the patch of pixels around the keypoint. Such features are then compressed and transmitted to the remote processing center for further analysis. Since the features-based representation is usually more compact than the pixel-based one, the ATC approach is particularly attractive for those scenarios in which bandwidth is scarce, like VSNs.[4]

Here we consider a reference scenario where a VSN is deployed to perform object recognition according to the ATC paradigm. In this scenario, each camera extracts visual features from the detected objects and transmits them to a central controller. There, the received features are matched with a database of labeled features from known objects to find the most similar ones. We focus on the case of cameras with overlapping fields of view (FoVs): such a case may be the result of a dense random deployment, as it often happens in surveillance applications,[5] or may be enforced to increase the robustness of the camera network (e.g., ensuring visual coverage even in case of camera failures). In such a scenario, the redundancy existing between cameras' fields of view may be exploited in two different ways. On the one hand, the availability of multiple cameras captur-

ing overlapping views of a scene may be induced and/or exploited to improve the performance of the specific visual application. As an example, for the case of object recognition, multiple views of the same scene can provide obvious walk-arounds to occlusions. However, such performance improvement requires the cameras to be active (acquiring and processing) concurrently with additional costs in terms of network infrastructure and overall energy consumption. On the other hand, such redundancy may be exploited with the purpose of extending the network lifetime rather than the visual task accuracy. Typically, this is achieved by organizing redundant cameras in clusters and by putting to sleep some nodes while other sense the environment. These two objectives, incrementing the accuracy and extending the lifetime, are clearly in contrast and therefore generally addressed separately.

This work addresses the two aspects jointly and analyses the accuracy/energy consumption trade-off involved in object recognition tasks performed by multiple cameras with partially overlapping FoVs. To this extent, we propose a game theoretic framework to model the cooperative visual feature extraction process, and we resort to the Nash Bargaining Solution (NBS) to steer the cooperative processing. The key tenet of the proposed framework is to reduce the energy consumption for visual feature extraction by exploiting the redundancy in the reciprocal FoVs. The proposed scheme is then applied to different multi-view image datasets and implemented in a real-life VSN testbed to assess its performance. Experimental results confirm that the proposed coordination scheme reduces the energy consumption with respect to the case in which multiple cameras process the whole input image, with a negligible loss in the achieved quality.

The paper is organized as follows: Section 2 overviews the related work; Section 3 introduces the reference VSN scenario, including empirical conditions under which the cooperative framework may be enabled. Section 4 formally describes the game theoretic-framework for cooperative

object recognition in VSNs and derives a closed-form Nash Bargaining Solution. Section 5 contains the performance evaluation of the proposed scheme in different network/dataset conditions, on both simulated and real-life data. Finally, Section 6 concludes the paper.

2 Background and Related Work

In the last few years, an increasing number of works have faced the problem of managing VSNs featuring cameras with overlapping FoVs. The main focus is on area coverage and task assignments problems, which are critical in monitoring applications. Ai and Abouzeid[6] propose solutions to maximize the visual coverage with the minimum number of sensor, assuming to have cameras with tunable orientations. Wang and Gao[5] propose a novel model called full-view coverage, observing that the viewing direction of a camera willing to recognize an object should be sufficiently close to the facing direction of that object. The concept is leveraged by the same authors[7] presenting a method to select camera sensors from a random deployment to form a virtual barrier made of cameras for monitoring tasks. As a result, many redundant cameras (i.e., cameras with overlapping FoVs) might be selected. Since visual sensor nodes are battery operated, it is imperative to optimize their operation, thus maximizing their lifetime. In most of the works that deal with coverage, lifetime is defined as the amount of time during which the network can satisfy its coverage objective. With this definition, the approach traditionally used is to leverage the redundancy resulting from random deployment and organize redundant cameras in clusters. Then, coordination can be applied among cluster members, by putting to sleep some nodes while others sense the environment.[8] A similar idea is applied when dealing with task assignment problems, where visual tasks have to be assigned to different cameras taking into consideration each task requirements and each camera available resources (e.g., task frame rate, camera residual energy).[9–11]

On the contrary, fusing information from multiple views of the same object, can improve the performance of visual analysis tasks. As an example, the works in[12] and [13] propose face recognition systems based on local features extracted from multiple views of the same face with clear improvements compared to a single camera system. The availability of multiple views can be leveraged also in application scenarios different from face recognition. As an example, the work by Naikal et al.[14] proposes a distributed object recognition system for VSNs where visual features extracted from multiple views of the same objects are used to improve the efficiency of object recognition. Summarizing, there exists a dichotomy between the need of extending lifetime (which calls for de-activating camera nodes) and the need of improving the accuracy of the specific visual task (which requires many cameras to be active at the same time). To our knowledge, the available literature tends to focus on one of these two contrasting objectives; differently, we aim at gauging a more thorough analysis of the quality/lifetime tradeoff in VSNs by relying on a game theoretic framework.

Game theoretic bargaining frameworks have recently been applied to VSNs; in,[15] Li *et al.* consider a visual sensor network composed of multiple camera for tracking moving people and introduce a bargaining framework to orchestrate the camera-to-person assignment. Similarly, in [16] a decentralized control method for tracking multiple targets using game theory is proposed. Pandremmenou *et al.* focus on video applications and resort to bargaining approaches to optimally allocate the source coding rates, channel coding rates, and power levels among several camera nodes running Direct Sequence Code Division Multiple Access (DSCDMA)[17][18] ; the key tenet is to derive bargaining solutions based on Nash Bargaining[19] and Kalai-Smorodinsky[20] theory, such that camera nodes reach an efficient cooperation point. The Nash Bargaining Solution is used also in[21] to solve a multi-camera radio resource allocation problem in VSN for video transmission. To

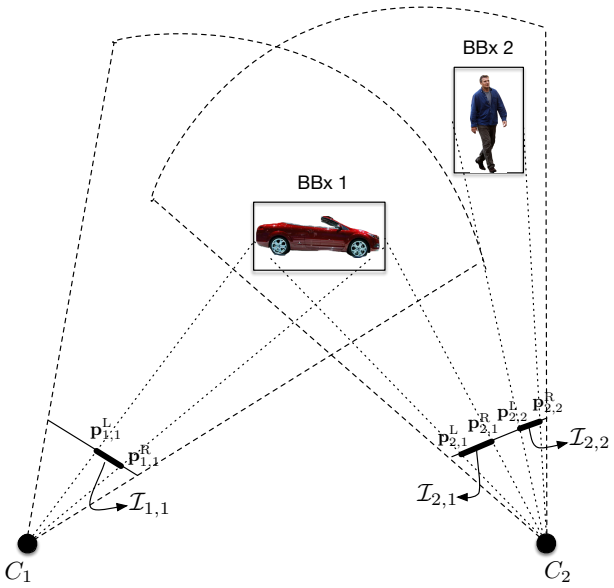


Fig 1 The two cameras C_1 and C_2 detect objects in their fields of view, and identify a BBx for each detected object. Note that the same object may not be present in both cameras FoVs.

the best of our knowledge, ours is the first attempt to apply game-theoretic bargaining approach to the problem of cooperative visual feature extraction.

3 Reference Scenario

We consider a scenario where N wireless camera nodes with fully or partially overlapping FoVs are able to communicate with each other, i.e., they are in direct transmission range. Figure 1 shows a descriptive example for $N = 2$: each camera acquires an image from the environment and has to perform local visual features extraction.

3.1 Non-cooperative visual sensor networks

In the case of non-coordinated networks, each camera independently performs the following steps:

- *Object detection:* the acquired image is pre-processed with foreground detection/background subtraction techniques to detect possible objects in the field of view.[22] Typically, a bounding box (BBx) is drawn around each detected object.[23] Note that multiple objects may be detected

in one image, as it is the case of camera C_2 in Figure 1. Therefore, we denote as $\mathcal{I}_{i,j}$ the portion of image contained in the j-th BBx of the i-th camera, $i \in 1 \dots N$. Each BBx $\mathcal{I}_{i,j}$ is defined in terms of its top-left and bottom right coordinates, namely $\mathbf{p}^{\mathrm{L}}_{i,j}$ and $\mathbf{p}^{\mathrm{R}}_{i,j}$, and the number of pixels contained in a BBx is denoted as $P_{i,j}$

- *Feature extraction:* the pixels corresponding to each BBx are processed by means of a local feature extraction algorithm. Such step encodes the photometric properties of each detected object in a representation which is: (i) generally more concise than the pixel-domain one and (ii) robust to several image transformations (scale, rotation, illumination changes) and thus ideal for being used for recognition tasks.

- *Features transmission:* the features extracted from each BBx are transmitted to a remote controller, where they are matched against a database of labelled features and object recognition is performed.

Clearly, all the aforementioned steps require non negligible energy to be performed. Since in VSNs camera nodes are battery-operated, it is imperative to optimize the process of feature extraction and transmission to limit the corresponding energy consumption. This is precisely the goal of this work.

3.2 Cooperative visual sensor networks

One may leverage the fact that cameras have overlapping FoVs to set up a cooperative framework that allows to save energy without sacrificing the visual analysis performance. In particular, we posit that features extracted from different views of the same object share a high degree of similarity. Therefore, different cameras may agree to share the feature extraction task by processing only

sub-portions of the detected object: since the energy needed to perform feature extraction depends primarily on the number of processed pixels,[24] such a cooperative approach is expected to provide notable energy savings with respect to the non-cooperative case, i.e., when each camera processes the entire bounding box.

Clearly, one main condition that must hold in order to enable the cooperative framework is that the cameras willing to cooperate are indeed looking at the same face of one object. Only in this case the features extracted from multiple views will share an high degree of similarity, thus allowing the different cameras to split the feature extraction task between them. Two options are possible to check this condition, depending on whether the cameras are *calibrated* or not.

3.2.1 Calibrated cameras

In the former case, the geometrical relationship between the two cameras (that is, relative translation and rotation) is algebraically represented by the so-called fundamental matrix $\mathbf{F}$, available to both cameras and allowing to check if a point $\mathbf{p}$ (in pixel coordinates) in the first view corresponds to a point $\mathbf{p}'$ in the second view (see Figure 1), through the well known fundamental matrix equation:[25]

$$\mathbf{p'^{T}}\mathbf{F}\mathbf{p} = 0. \tag{1}$$

However, such a case is practically possible if cameras are perfectly calibrated and if the coordinates of the BBx are estimated without errors, two conditions which are often very difficult to obtain. Therefore, in the following section, we propose an alternative, yet effective method to practically assess if calibration may be enabled.

3.2.2 *Non-calibrated cameras*

Even without precise calibration, cameras may rely on the exchange of pixel-based, scene-dependent information to infer if the collaborative framework may be set up. As an example, the cameras may exchange a downsampled version of their acquired image and compute a measure which is representative of the degree of similarity of the two views, hence on the expected redundancy between the extracted visual features. Here we rely on the Common Sensed Area (CSA) value, defined in,[26] as the ratio between the number of pixels belonging to the common area of two images i and j and the overall number of pixels of the image captured by camera i. In practice, such common area is approximated by applying a correlation-based image registration technique and counting the number of correctly registered pixels. For the computation of the CSA, the two cameras may exchange a thumbnail version of each bounding box contained in the acquired images and estimate their correlation based on them. The "optimal" size of the exchanged thumbnails comes from a trade-off choice between the need of limiting the signalling overhead (smaller thumbnail) and the need of being accurate in measuring the overlap between fields of view (larger thumbnail).

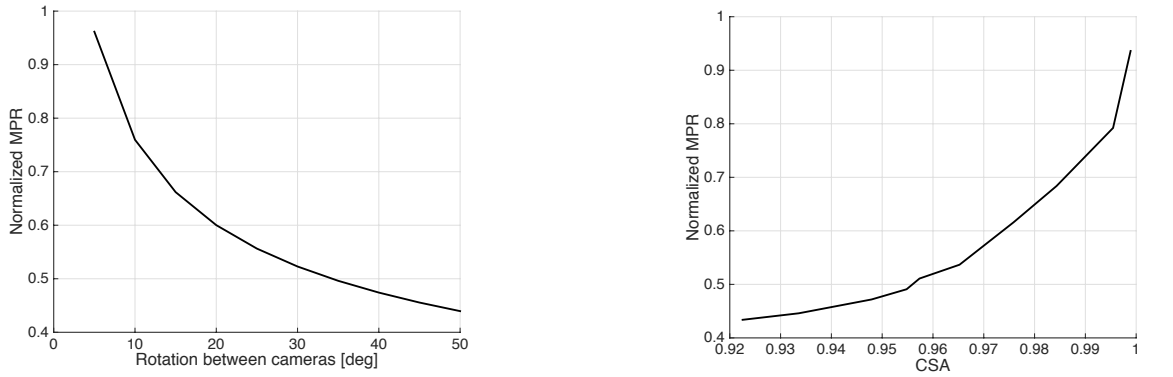
3.2.3 *Empirical conditions for cooperation*

As mentioned before, cooperation may be enabled when the features extracted by cameras with overlapping fields of view share a high degree of similarity. Such redundancy may be measured with the number of matches post RANSAC (MPR):[27] let $\mathcal{F}_1$ and $\mathcal{F}_2$ be the sets of features extracted from the images acquired by two cameras with overlapping FoVs. Features from $\mathcal{F}_1$ are matched against the features from $\mathcal{F}_2$ to find correspondences, and a geometric consistency check with RANSAC is performed to remove matches that are not consistent with a global geometric transformation. The MPR is defined as the number of inliers (e.g., those correspondences which

are consistent with the estimated geometric transformation). Therefore, the higher the MPR, the higher the redundancy between the two sets of features. We performed a set of experiments to assess how the MPR changes when varying the inter-camera geometry: to cope with both the calibrated and non-calibrated case, we observe the relationship of the MPR with the angle between the cameras viewing direction (available in the calibrated case) or the Common Sensed Area (CSA) value. To perform such analysis, we used the COIL-100 and ALOI datasets (see Section 5.3), which contain images of different objects, each captured at 72 different poses obtained by rotating the object by 5 degrees each time. For each object, we selected $\mathcal{F}_1$ as the set of BRISK[28] local features extracted from the image taken at 0 degrees. For what concerns $\mathcal{F}_2$, we computed it each time incrementing the rotation by five degrees. For each couple of images and sets of features, we computed the MPR value (normalized with respect to the minimum number of features extracted from the two images), and the CSA between the two images (downsampled to 22×18 pixels).

Figures 2(a) and 2(b) show the average normalized MPR value when varying the angle between the two views and at different values of the computed CSA, respectively. From the inspection of such figures, one may establish practical conditions under which the cooperative framework may be enabled by fixing an MPR threshold. As an example, in order to obtain an MPR greater than 0.5 (that is, 50% of the local features are common between the two views) a CSA value greater than 0.95 should be obtained. In an ideal situation where calibration can be performed without errors and the angle between the cameras may be obtained, its value should be less than 30 degrees[1] Note that such threshold values may be changed according to the application requirements. As an example, one may want to make the CSA constraint more tight and enable cooperation only when

[1]Note, however, that this particular value depends on the distance between the cameras and the object, thus on the datasets used in this evaluation. We therefore strongly suggest to rely only on the CSA computation in practical scenarios.



(a) (b)

Fig 2 (a) Normalized MPR value at different rotation angles; (b) Normalized MPR value at different computed CSA

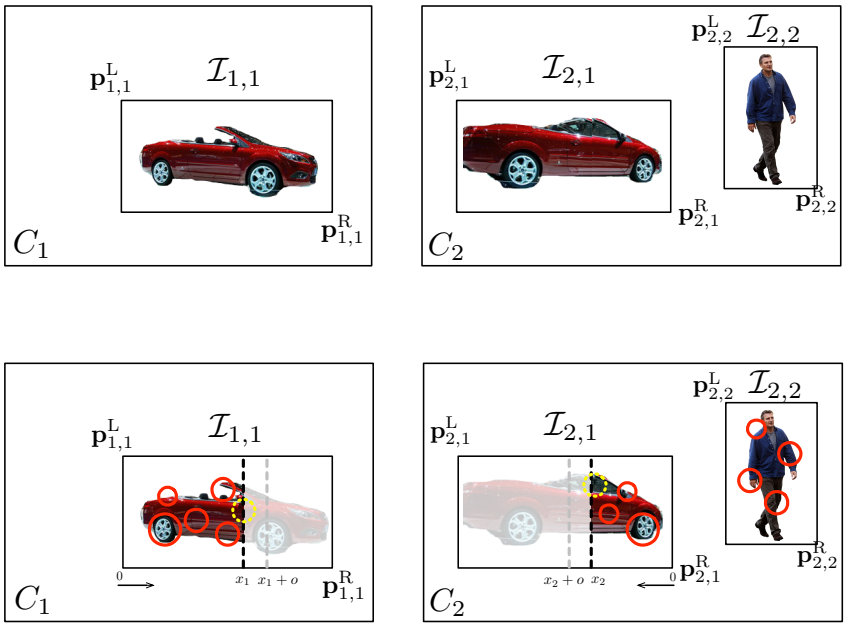


Fig 3 Top - The BBx detected on the images acquired by C_1 and C_2, referring to the scenario in Fig. 1. The two cameras exchange the top-left and bottom-right coordinates of the BBx (in case of calibrated cameras), or a sub-sampled version of each BBx (in case the camera are non-calibrated). This allows to establish a correspondence between the BBx. Bottom - Only for the corresponding BBx (the ones containing the car), an image sub-portion is selected from each camera for features extraction (red solid circles). The exact location of the sub-portions is computed relying on game theory as explained in Section 4. Note that features near the splitting border (yellow dashed circles) are correctly detected, but can't be computed unless an offset region is added to the BBx.

more than 50% of the features are recognized as common, that is, using a higher CSA value.

According to such results, we propose additional steps to set up the cooperative framework, which is briefly illustrated in figure 3:

- *Geometric consistency:* After object detection, cameras exchange a thumbnail of the acquired BBx and compute the CSA: cooperation is then established if the CSA is greater than 0.95. Note that this process is not limited to pairs of cameras, but may be easily extended to networks of multiple camera nodes with overlapping FoVs by relying on transitivity.

- *Bounding box splitting:* having identified a common object in their FoVs, each camera may select a sub-portion of its own BBx instead of processing it entirely. As illustrated in Figure 3, for the case of two cameras, we assume that the leftmost camera selects from the leftmost region of its BBx up to x_1, while the rightmost camera processes from the right end to x_2, where x_1 and x_2 can be expressed as proportions of the BBx area that are processed by the first and the second camera, respectively. That is, x_i is in the range from 0 (when the i-th camera does not perform any processing) to 1 (when the i-th camera processes its entire BBx). Note that, without loss of generality, splitting may be applied in the vertical direction as well. In the following section, we propose a game theoretic approach for determining the proportion of the BBx to be processed on each camera, i.e., the values of x_i, $i = 1 \dots N$. Once such values are computed, each camera may extract features from the reduced BBx and transmit them to the central controller. A reasonable constraint for the variables x_i is that they sum up to 1, i.e., visual features are extracted from the entire object, although in different views. However, it is important to note that image splitting may negatively affect the features extraction process. As illustrated in the bottom part of Figure 3, this is due to the fact that the extraction of one visual feature requires the processing of a patch of pixels around the corresponding keypoint. If the keypoint is detected close to the splitting line, there may not be enough pixels to perform the feature extraction. In the case of a very discriminative feature being close to the splitting line,

such approach may negatively affect the performance of object recognition. To overcome this issue, an offset is added to the variables x_i. In the following, we denote by o such required offset, normalized with respect to the total image size.

4 Game-Theoretic Models

The reference scenario described in Section 3 can be modeled as a game among N cameras which have to decide the portion of the common bounding box they need to process. Let $\mathbf{x} = (x_1, x_2, \ldots x_N) : \sum_{i=1}^{N} x_i \leq 1$ be an outcome of the game, being x_i the portion of the BBx which is assigned for processing to camera i, with $0 \leq x_i \leq 1$.

Let $\mathcal{X}$ be the set of all possible outcomes of the game. Let us further define an utility function $u_i(\mathbf{x})$ which represents the *preference* for camera i on the outcome $\mathbf{x}$. The set of possible payoff vectors is defined as $\mathcal{U} = \{u_1(\mathbf{x}), u_2(\mathbf{x}), \ldots, u_N(\mathbf{x})\}$.

In the reference scenario, it is reasonable to bind the utility function $u_i(\mathbf{x})$ to the energy consumed in the feature extraction process. The energy consumed for extracting features increases linearly with the number of processed pixels.[24] Thus, we model the energy consumption as:

$$E_i(x_i) = P_i(a_i x_i + b_i), \tag{2}$$

where the parameters a_i and b_i depend on the particular processor available on the i-th camera, and P_i is the size in pixels of the bounding box currently under processing.

We define the utility function for camera i as:

$$u_i(\mathbf{x}) = \begin{cases} E_i(\sum_{k=1,k\neq i}^{N} x_k) & \text{if } \sum_{i=1}^{N} x_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Intuitively, the utility function for the i-th camera is the amount of energy that camera i saves through cooperative processing, if the cameras agree to collaboratively process the entire image, i.e., if $\sum_{i=1}^{N} x_i = 1$. If the cameras fail to reach an agreement, i.e., if $\sum_{i=1}^{N} x_i < 1$, each camera has to process the image on its own and thus its utility function is equal to zero. In such settings, the following proposition holds true.

Proposition 4.1 Each task assignment vector $\mathbf{x} = (x_1, x_2, \dots x_N) : \sum_{i=1}^{N} x_i = 1$ is a Nash Equilibrium.

Proof. Consider a task assignment vector $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_N^*)$ such that $\sum_{i=1}^{N} x_i = 1$, resulting in a payoff of $\mathcal{U}^* = \{u_1(\mathbf{x}^*), u_2(\mathbf{x}^*), \dots, u_N(\mathbf{x}^*)\}$. In order for such generic task assignment vector to be a Nash Equilibrium, one must show that all the players have not any incentive in unilaterally modifying their strategy. Assume that player i decides to decrease her contribution to $x_i < x_i^*$: the condition $\sum_{i=1}^{N} x_i = 1$ is no more satisfied, i.e. the cameras are not able to reach an agreement to process the entire image and thus player i's utility function drops to zero. Hence, the player does not profit from an unilateral change of strategy. □

To select which of the equilibria has to be implemented, the scenario under consideration can be modeled as a bargaining problem with two main ingredients:

- **Feasibility set:** the convex set $\mathcal{F} \subseteq \mathbb{R}^N$ including all the possible payoff vectors $\mathbf{u} = (u_1, u_2, \dots, u_N)$ defined by Eq. (3);

- **Disagreement point:** the value of the utility function the players are expected to receive if the negotiation breaks down $\mathbf{d} = (d_1, \ldots, d_N)$; in our case, if negotiation breaks down, each camera has to process the entire BBx, thus its utility (e.g., spared energy) at the disagreement point is null, $d_i = 0$, $i = 1 \ldots N$.

The Feature Extraction Bargaining Problem (FEBP) can be defined as the tuple $(\mathcal{F}, \mathbf{d})$. A solution concept which can be applied to such game theoretic scenario is the generalized Nash Bargaining Solution (NBS) which provides an axiomatic solution to the bargaining, further providing an operative method to derive it. Formally, the NBS defines an agreement point (bargaining outcome) $\mathbf{x}_{\text{NBS}}$ which verifies the following four axioms:

1. **Rationality:** $u_i(\mathbf{x}_{\text{NBS}}) > d_i$, $i = 1 \ldots N$, i.e., no player would accept a payoff that is lower than the one guaranteed to him under disagreement;

2. **Pareto optimality:** under the optimality conditions, the payoff of each player cannot be further improved without reducing other players' ones;

3. **Symmetry:** if the players are undistinguishable, the agreement should not discriminate between them;

4. **Independence of irrelevant alternatives:** the solution of a bargaining problem does not change as the set of feasible outcomes is reduced, as long as the disagreement point remains the same, and the original solution feasible.

Since $\mathcal{X}$ is compact and convex and the utility functions $u_i(\mathcal{X})$ are concave and upper bounded, the generalised NBS for the bargaining problem is the unique solution of the following optimiza-

tion problem:[29]

$$\begin{aligned} \text{maximize} \prod_{i=1}^{N} (u_i(\mathbf{x}) - d_i)^{\alpha_i} \\ \sum_{i=1}^{N} x_i = 1 \\ x_i \geq 0 \quad \forall i \end{aligned} \tag{4}$$

The exponents α_i represent the *bargaining power* of each camera, and are chosen such that $\sum_{i=1}^{N} \alpha_i = 1$. A natural choice for the bargaining powers α_i is to relate them to the residual energy of each camera $\mathcal{E}_i$. In particular, a desirable condition is that a camera bargaining power increases as its residual energy decreases (i.e., cameras close to deplete their energy are eagerer to cooperate). Thus, we define the bargaining powers as:

$$\alpha_i = \frac{\mathcal{E}_i^{-1}}{\sum_{i=1}^{N} \mathcal{E}_i^{-1}} \tag{5}$$

4.1 Characterizing the Two-player Nash Bargaining Solution

Finding the Nash Bargaining Solution scales down to solving the following non-linear constrained optimization problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) = -u_1(\mathbf{x})^{\alpha_1} u_2(\mathbf{x})^{\alpha_2} \\ \text{subject to} \quad & x_1 \geq 0,\ x_2 \geq 0 \\ & x_1 + x_2 = 1. \end{aligned} \tag{6}$$

The Karush-Kuhn-Tucker conditions (KKT) are first order necessary conditions for a solution in nonlinear programming to be optimal. See Appendix A for a quick review on KKT conditions. In

our case, the following Lagrangian J and KKT conditions can be written:

$$\mathcal{L} = -u_1(\mathbf{x})^{\alpha_1} u_2(\mathbf{x})^{\alpha_2} + \lambda_1(1 - x_1 - x_2) - \mu_1 x_1 - \mu_2 x_2$$

1. **Stationarity**: $\nabla_{\mathbf{x}} \mathcal{L} = 0$
2. **Primal feasibility**: $\begin{cases} -x_1 \leq 0 \\ -x_2 \leq 0 \\ x_1 + x_2 - 1 = 0 \end{cases}$
3. **Dual feasibility**:

   $\mu_1 \geq 0$

   $\mu_2 \geq 0$

4. **Complementary slackness**:

   $\mu_1 x_1 = 0$

   $\mu_2 x_2 = 0$

Consider the non-trivial case in which both x_1 and x_2 are non-zero. From the complementary slackness conditions it is possible to observe that $\mu_1 = 0,\ \mu_2 = 0$. Substituting for μ_1 and μ_2, the stationarity condition can be simplified as follows:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}} \lambda_1(1 - x_1 - x_2) = 0 = \begin{cases} \frac{\partial f(\mathbf{x})}{\partial x_1} - \lambda_1 \\ \frac{\partial f(\mathbf{x})}{\partial x_2} - \lambda_1 \end{cases} \tag{7}$$

By enforcing primal feasibility, i.e. $x_1 = 1 - x_2$, the problem corresponds to a system of two equations and two unknowns, namely λ_1 and x_1. The only viable solution, enforcing both the other

KKT conditions and the NBS axioms is the following:

$$\begin{cases} x_1 = \frac{a_1 a_2 \mathcal{E}_1 - a_1 b_2 \mathcal{E}_2 + a_2 b_1 \mathcal{E}_1}{a_1 a_2 \mathcal{E}_1 + a_1 a_2 \mathcal{E}_2} \\ x_2 = \frac{a_1 a_2 \mathcal{E}_2 + a_1 b_2 \mathcal{E}_2 - a_2 b_1 \mathcal{E}_1}{a_1 a_2 \mathcal{E}_1 + a_1 a_2 \mathcal{E}_2} \end{cases} \tag{8}$$

Basically, the amount of pixels to be processed by a camera is proportional to the amount of residual energy of such camera compared to the other one, and depends on the characteristics of the cameras processors, i.e. a_i and b_i. The same approach can be employed to solve the bargaining problem for multiple cameras.

Consider the case in which the two cameras have the same processing characteristics and no processing bias, i.e. $a_1 = a_2 = a > 0$ and $b_1 = b_2 = 0$. In this setting, the Nash Bargaining Solution reduces to

$$\begin{cases} x_1 = \frac{\mathcal{E}_1}{\mathcal{E}_1 + \mathcal{E}_2} \\ x_2 = \frac{\mathcal{E}_2}{\mathcal{E}_1 + \mathcal{E}_2} \end{cases} \tag{9}$$

In this case the system resembles a voltage divider circuit in which the workload assigned to the cameras and their residual energy correspond to the voltage on the two components and their resistance, respectively.

We conclude this section with a numerical example of the NBS solution. Consider the scenario illustrated in Figure 1 and in the bottom part of Figure 3, and assume that cameras C_1 and C_2 have the same processing characteristics but different residual energies. Assume that the computed CSA between the two cameras for the BBx containing the red car is high enough to enable cooperation. The two cameras exchange their residual energies $\mathcal{E}_1$ and $\mathcal{E}_2$ to compute the fraction of common

BBx to process. Assume, e.g. $\mathcal{E}_1 = 2\mathcal{E}_2$: according to (9) $x_1 = 2/3$ and $x_2 = 1/3$. Therefore, C_1 will process from the leftmost end up to two thirds of its BBx, while C_2 will process from two thirds of its BBx to the rightmost end.

5 Performance Evaluation

We are interested in assessing the performance of the cooperative framework in terms of object recognition accuracy and energy efficiency. To this extent, we have implemented the full pipeline of a typical object recognition task based on BRISK[28] visual features: camera nodes acquire a *query* image, extract visual features from it, and transmit the features to a sink node where object recognition is performed. There, the received features are matched against features extracted from a database of images. Matching consists in pair-wise comparisons of features extracted from, respectively, the query and database image. The Hamming distance is adopted to measure the similarity between BRISK visual features extracted from the image and the ones contained in the database. Two features are labeled as matching if their distance is below a pre-defined threshold. Additionally, a geometric consistency check step based on RANSAC is applied to filter out outliers. Hence, the images in the database can be ranked according to the number of matches with the query image.

5.1 Accuracy Evaluation

Average Precision (AP) is commonly adopted to assess the performance of object recognition/image retrieval. Given a query q, AP is defined as:

$$AP_q = \frac{\sum_{k=1}^{n} P_q(k) r_q(k)}{R_q}, \tag{10}$$

where $P_q(k)$ is the precision (i.e., the fraction of relevant images retrieved) considering the top-k results in the ranked list; $r_q(k)$ is an indicator function which is equal to 1 if the item at rank k is relevant for the query, and zero otherwise; R_q is the total number of relevant documents for the query q and n is the total number of documents in the list. The Mean Average Precision (MAP) for a set of Q queries is the arithmetic mean of the APs across different queries:

$$MAP = \frac{\sum_{q=1}^{Q} AP_q}{Q} \tag{11}$$

The MAP value ranges from 0 (no queried object was correctly recognized) to 1 (perfect recognition).

5.2 Energy Evaluation

Energy efficiency is captured by estimating the lifetime L of the system, that is the number of consecutive queries (images) which can be processed until one of the camera nodes depletes its energy. That is:

$$L = \min \frac{E_i^{\text{budget}}}{\frac{1}{Q}\sum_{q=1}^{Q} E_{i,q}}, \tag{12}$$

where E_i^{budget} is the energy budget of the i-th camera and $E_{i,q}$ is the energy required for processing the q-th query on the i-th camera. To characterize the per-query energy consumption of a camera, we rely on the following energy model:

$$E_{i,q} = E^{\text{acq}} + P^{\text{cpu}}\left[t_{i,q}^{\text{bb}} + t_{i,q}^{\text{det}} + t_{i,q}^{\text{desc}}\right] + E^{\text{tx}} r M_i, \tag{13}$$

Table 1 Parameters used for the energy evaluation

Name	Symbol	Value
CPU power	P^{cpu}	1.75 W
Energy budget	E^{budget}	20 KJ
Acquisition cost	E^{acq}	10^{-3} J/frame
Transmission cost	E^{tx}	2.2×10^{-7} J/bit
Feature size	r	512 bit

being E^{acq} the energy required for acquiring one image, P^{cpu} the power consumption of the CPU of each camera and $t_{i,q}^{\text{bb}}, t_{i,q}^{\text{det}}$ and $t_{i,q}^{\text{desc}}$ the times taken by the i-th camera to identify the bounding boxes, detect keypoints and extract features for the q-th query, respectively. The energy cost of transmitting the extracted features is captured by the last term of (13), where E^{tx} is the energy cost of transmitting one bit, r is the dimension in bit of each visual feature, and M_i is the number of features detected by the i-th camera. The values used for the energy costs are based on a Visual Sensor Node platform based on a BeagleBone linux computer[30] and are reported in Table 1.

5.3 Experimental Methodology

The evaluation has been carried out on several VSN topologies, each one consisting of a pair of camera nodes characterized by a different geometrical relationship, and by relying on different image datasets. From each dataset, we selected one common set of images as the reference database for the object recognition task and several set of images as query datasets. The query datasets are selected so as to mimic different camera geometries:

- *COIL100*[2]: this image database contains 100 objects, each captured at 72 different poses. Each pose of an object is obtained by rotating the object by 5 degrees. For each object, the reference database contains three images corresponding to the views at 0° and $\pm 10^\circ$. Five different camera geometries are tested as query datasets, taking for each object the couple of images at $\pm 5^\circ$, $\pm 15^\circ$,

[2] http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php

$\pm 20^\circ$,$\pm 25^\circ$ and $\pm 30^\circ$. We refer to such experiments with the label COIL-X, where X is in the set {5,15,20,25,30}.

- *ALOI*[3]: an image collection of one-thousand small objects. Similarly to the COIL-100 dataset, each object is captured at 72 different poses obtained by rotating the object by 5 degrees each time. The reference database and the test sets are obtained in the same way as for the COIL-100 dataset (i.e., five different camera configurations, each one with an increasing rotation). Again, we refer to such experiments with the label ALOI-X, where X is in the set {5,15,20,25,30}.

- *ANT66*[4]: A novel image database containing 66 objects, each one captured by two camera pairs with overlapping FoVs. The reference database contains one image per object and two camera geometries are available, which we refer to as ANT-0 and ANT-15. In ANT-0, the inter-camera geometry is a pure translation, while in ANT-15 the two cameras are translated and rotated by $+15^\circ$ and -15° degrees with respect to the object's main axis.

For each one of the twelve different camera topologies, the NBS-based cooperative framework is evaluated as follows:

1. For each of the two cameras, load a query image q from the current test set.

2. Find the image splitting x_1 and x_2 by solving the features extraction bargaining problem through the generalized NBS according to equation (4).

3. Extract BRISK features from the sub-portions defined by $x_1 + o$ and $x_2 + o$. Compute the per-camera energy consumption E_1 and E_2 as in (13), and the average precision AP_q as in (10). To compute the AP, the feature sets from the two camera views are independently matched against

[3] http://aloi.science.uva.nl/
[4] http://www.greeneyesproject.eu/

the reference dataset, and geometrically verified through RANSAC. The number of matches for the q-th query couple is then computed by summing the matches from the two independent views.

4. Update residual energies for the two cameras, $\bar{E}_1$ and $\bar{E}_2$.

5. Repeat steps 1-4 until (i) one of the two cameras deplete its energy or (ii) all queries in the datasets have been processed. Compute the MAP as in (11) and the system lifetime as in (12).

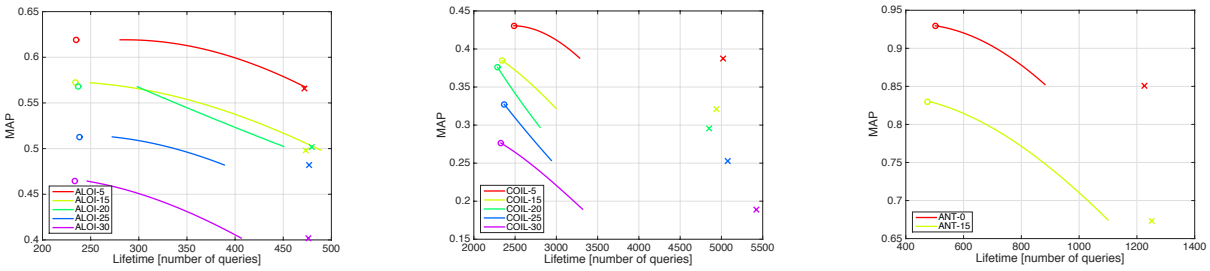
The entire process is repeated for increasing values of the offset o. For each camera topology, we also compared the NBS-based cooperative framework against the following two baseline scenarios:

- *Temporal Scheduling (TS):* at each query, only one camera among the ones with overlapping FoVs acquires the image and performs object recognition while the ether stays in an idle state, as done e.g., in the work by Alaei et al.[8] We select which camera should operate according to the maximum residual energy.

- *Multi View object recognition (MV):* following the approach by Naikal et al.,[14] at each query all cameras acquire an image, extract the corresponding features and transmit them to the sink node. Similarly to the NBS, features matching is performed by using the features extracted from the two camera views.

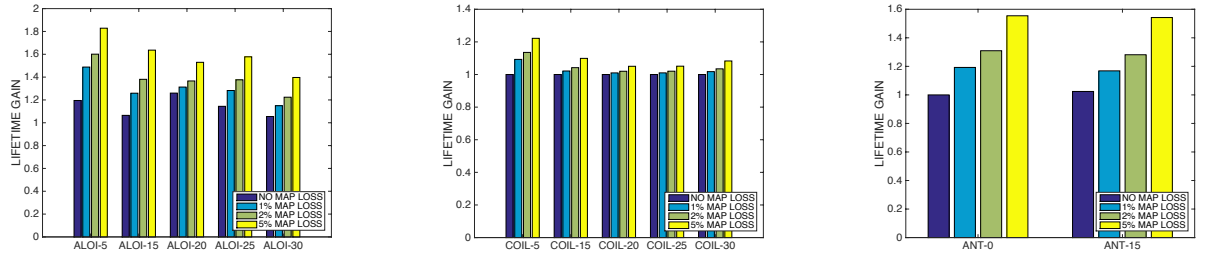
5.4 Experimental Results

Figure 4 illustrates the lifetime/accuracy trade-offs obtained by running the aforementioned experiments on the three reference datasets. Each figure reports:



(a) (b) (c)

Fig 4 Accuracy (MAP) vs energy (Lifetime) trade-off for the TS approach (cross), the MV approach (circle) and the NBS (solid line) for different datasets ((a) ALOI, (b) COIL-100 and (c) ANTLAB-66). Different camera geometries are illustrated with different colors.



(a) (b) (c)

Fig 5 NBS lifetime gain with respect to the MV approach for different values of tolerated MAP loss, for different datasets ((a) ALOI, (b) COIL-100 and (c) ANTLAB-66).

1. Colored circle - the performance of the multi-view approach in terms of achievable lifetime and accuracy.

2. Colored cross - the performance of the temporal scheduling approach.

3. Solid line - the performance of the NBS cooperative framework for different values of the overlap o.

Each color represents a different camera topology, obtained by changing the cameras geometry. As one can see, the best approach in terms of accuracy is the multi view approach (MV - colored circle). This is expected, as fusing features from multiple views increase the amount of information available for object recognition. However, this happens at the highest cost in terms of energy con-

sumption, as the different cameras need to be active and to process their acquired images entirely. At the other extremum, the temporal scheduling approach (TS - colored cross) exhibits the lowest energy consumption, as cameras are alternatively switched off to perform recognition. However, this comes with the drawback of decreased recognition accuracy due to the limited information available (only the features extracted from one camera are available at each query). Overall, the proposed NBS-based solution allows to efficiently trade-off recognition accuracy for energy: as one can see from Figure 4(a) by appropriately tuning the value of the overlap o, one can reduce the consumed energy by a factor of 2 approximately (with respect to the MV approach), with a limited accuracy loss (around 5%). Note that as the reference topologies become more challenging, that is, with increasing angle of displacement between the two cameras, the trade off between the MAP and system lifetime becomes less favorable, and the energy savings diminishes for a given MAP loss. Figure 4(b) and 4(c) report the same analysis for the COIL100 and the ANT66 datasets, respectively. Finally, Figure 5 summarizes the numerical analysis by reporting, for all the considered data sets, the gain in the system lifetime for a given tolerable MAP loss, always with respect to the MV scenario (two cameras always active). For a given dataset configuration, the lifetime gain increases as the tolerated level of MAP loss also increases. In general, for a given tolerated MAP loss the achievable lifetime gain tends to decrease as the camera geometry becomes more challenging.

5.5 Practical implementation

In addition to the experiments on simulated image data described in Section 5.4, the proposed framework has been implemented and tested on a real-life visual sensor network. The VSN is composed of two BeagleBone Black nodes equipped with WiFi interfaces and USB cameras, de-

ployed in such a way that the fields of view are overlapping and the empirical conditions for cooperation are fulfilled. A third node acting as central controller is also deployed. Each camera is operated with the open-source software proposed in,[31] which allows to easily program all steps of operation of the camera nodes, including image acquisition and processing, feature extraction and data transmission. The software run by each camera already includes the code needed for BRISK feature extraction and transmission. Additionally, it was modified with the following additional features:

- *Object detection and bounding box extraction*: upon deployment, the two camera nodes perform background estimation, so that it is possible to detect foreground objects (Figure 6). Since the used software framework for VSN relies on OpenCV for the image processing part, we used the in-built Gaussian Mixture-based background/foreground segmentation algorithm. After background subtraction, if an object is detected in the foreground a corresponding bounding box is computed. The estimated BBx (or part of it) is then given as input to the BRISK feature extraction algorithm available in the software framework.

- *Information exchange between camera nodes*: To implement the NBS approach, cameras periodically exchange a thumbnail of the detected BBx together with an estimate of their residual energy. According to,[26] each bounding box is resized to 22x18 pixels before transmission. To compute the residual energy, each camera starts with a predefined energy budget and decrements it according to the estimated consumed energy, computed with the same model used in (13) (that is, considering image acquisition, processing and features transmission). The BBx thumbnail (in uncompressed 8-bit resolution format) and residual energy are encapsulated in a specific signalling message and exchanged by the two cameras every

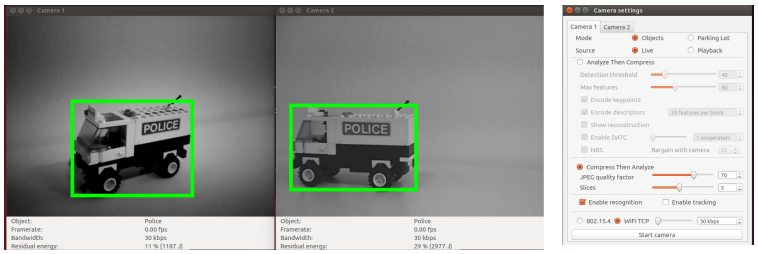


Fig 6 An object in the fields of view of both cameras, and the corresponding BBx. The computed CSA for such a scenario is equal to 0.9545. Therefore, cooperation can be enabled.

second. Assuming to store the residual energy in a 32 bits floating-point variable, the additional network overhead for such signalling communication is as low as 3.2 kbps. With such information, each camera can independently assess if the cooperation if possible through computation of the CSA value. According to Figure 2 the CSA threshold for cooperation is set equal to 0.95. If cooperation is possible, the two cameras also compute the solution of the NBS bargaining problem. Hence, cameras can independently select a portion of their bounding box to process with the feature extraction algorithm. Note that the availability of the residual energy value on both cameras allows to implement the temporal scheduling (TS) technique described in Section 5.3. In particular, upon reception of the signalling message, only the camera with the highest residual energy will process its bounding box.

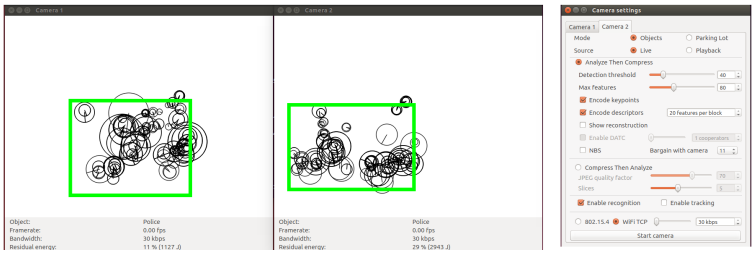


Fig 7 The Multi View (MV) approach, in which both cameras process the detected BBx

Such a setup allows to test the performance of the TS, NBS and MV approaches in a real scenario (see Figures 7 and 8). We select 20 objects from the ANTLAB-66 dataset, and place

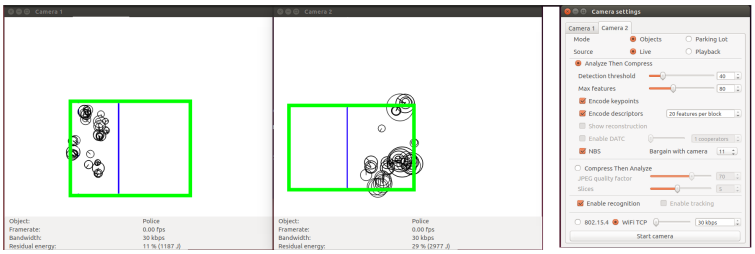


Fig 8 The proposed NBS approach, where cameras agrees to process only a portion of their common BBx.

Table 2 MAP and average consumed energy of the different approaches

	Avg. Energy per Query	MAP
Temporal Scheduling (TS)	0.765 J	0.73
Proposed Approach (NBS)	2.359 J	0.84
Multi-View (MV)	4.719 J	0.86

each object in the fields of view of the two cameras. For each approach (TS, NBS or MV), the features extracted by the cameras are transmitted to the central controller, where they are used to rank a database of labelled images and to compute the MAP value. In the NBS approach, we empirically verified that a CSA threshold equal to 0.95 corresponds roughly to an MPR of 50% also for the ANTLAB-66 dataset. Moreover, the overlap o was fixed to 10% of the BBx width. Concurrently, each camera keeps track of the energy consumed per frame. This allows to compute the average energy spent by the two cameras per query. Table 2 shows the results obtained in the three scenarios.

As one can see, the experiment on a real-life VSN confirms the results obtained through simulations. The TS approach is the one that exhibits the lowest energy consumption, at the cost of limited accuracy. At the other end, the MV approach in which both cameras process their BBx allows to reach the highest accuracy. This happens at the highest observed energy consumption. In between the two approaches, the proposed NBS-based scheme allows to obtain an accuracy which is just 2% less than the MV case at half the energy consumption.

6 Conclusions

Wireless camera networks and their applications are characterized by peculiar constraints and challenges that require to depart from traditional ways visual data is sensed, processed and transmitted. In this work, we have presented a cooperative approach for performing object recognition through features extraction in those cases where multiple cameras have overlapping FoVs. First, we derived empirical conditions under which cooperation may be enabled. Then, we relied on game-theory and modeled the scenario as a bargaining problem, which can be solved relying on the Nash Bargaining Solution. The resulting non-linear constrained optimization problem was solved in closed-form through the KKT conditions. We demonstrated with both simulations and a real-life experiment on a VSN testbed that the proposed approach is able to trade off system lifetime for task accuracy, improving the network lifetime with a negligible loss in the achieved visual analysis accuracy, compared to a traditional non-cooperative approach.

Appendix A: The Karush-Kuhn-Tucker conditions

Consider the following nonlinear optimization problem:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0,\ i = 1, \ldots, m. \\ & h_j(x) = 0,\ j = 1, \ldots, l. \end{aligned} \tag{14}$$

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point x^*. For x^* to be a local minimum, the following regularity conditions must hold:

$$\forall i = 1, \ldots, m \ \exists\ \mu_i, \quad \forall j = 1, \ldots, n \ \exists\ \lambda_j \quad s.t.$$

1. **Stationarity**: $\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{m} \lambda_j \nabla h_j(x^*) = 0$

2. **Primal feasibility**: $\begin{cases} g_i(x^*) \leq 0 \ \forall i = 1, \ldots, m \\ h_j(x^*) = 0 \ \forall j = 1, \ldots, l \end{cases}$

3. **Dual feasibility**: $\mu_i \geq 0 \ \forall i = 1, \ldots, m$

4. **Complementary slackness**: $\mu_i g_i(x^*) = 0 \ \forall i = 1, \ldots, m$

Acknowledgments

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 J. SanMiguel, C. Micheloni, K. Shoop, G. Foresti, and A. Cavallaro, "Self-reconfigurable smart camera networks," *Computer* **47**, 67–73 (2014).

2 K. Abas, C. Porto, and K. Obraczka, "Wireless smart camera networks for the surveillance of public spaces," *Computer* **47**, 37–44 (2014).

3 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).

4 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 *IEEE Multimedia Signal Processing (MMSP), 2013*, 278–282 (2013).

5 Y. Wang and G. Cao, "On full-view coverage in camera sensor networks," in *INFOCOM, 2011 Proceedings IEEE*, 1781 –1789.

6 J. Ai and A. A. Abouzeid, "Coverage by directional sensors in randomly deployed wireless sensor networks," *Journal of Combinatorial Optimization* **11**, 21–41 (2006).

7 Y. Wang and G. Cao, "Barrier coverage in camera sensor networks," in *Proceedings of the Twelfth ACM Intl. Symposium on Mobile Ad Hoc Networking and Computing*, 12:1–12:10 (2011).

8 M. Alaei and J. M. Barcelo-Ordinas, "Node clustering based on overlapping fovs for wireless multimedia sensor networks.," in *WCNC*, 1–6, IEEE (2010).

9 B. Rinner, B. Dieber, L. Esterle, P. R. Lewis, and X. Yao, "Resource aware configuration in smart camera networks," in *in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 58–65 (2012).

10 B. Dieber, L. Esterle, and B. Rinner, "Distributed resource-aware task assignment for complex monitoring scenarios in visual sensor networks," in *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*, 1–6 (2012).

11 C. Kyrkou, T. Theocharides, C. Panayiotou, and M. Polycarpou, "Distributed adaptive task allocation for energy conservation in camera sensor networks," in *Proceedings of the 9th International Conference on Distributed Smart Cameras*, *ICDSC '15*, 92–97, ACM, (New York, NY, USA) (2015).

12 L. An, B. Bhanu, and S. Yang, "Face recognition in multi-camera surveillance videos," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2885–2888 (2012).

13 J. Rambach, M. F. Huber, M. R. Balthasar, and A. M. Zoubir, "Collaborative multi-camera

face recognition and tracking," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, 1–6 (2015).

14 N. Naikal, A. Yang, and S. Sastry, "Towards an efficient distributed object recognition system in wireless smart camera networks," in *Information Fusion (FUSION), 13th Conf. on*, 1–8 (2010).

15 Y. Li and B. Bhanu, "Utility-based camera assignment in a video network: A game theoretic framework," *Sensors Journal, IEEE* **11**, 676–687 (2011).

16 B. Song, C. Soto, A. Roy-Chowdhury, and J. Farrell, "Decentralized camera network control using game theory," in *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, 1–8 (2008).

17 K. Pandremmenou, L. P. Kondi, K. E. Parsopoulos, and E. S. Bentley, "Game-theoretic solutions through intelligent optimization for efficient resource management in wireless visual sensor networks," *Signal Processing: Image Communication* **29**(4), 472 – 493 (2014).

18 K. Pandremmenou, L. P. Kondi, and K. E. Parsopoulos, "Fairness issues in resource allocation schemes for wireless visual sensor networks," *Proc. SPIE* **8666**, 866601–866601–12 (2013).

19 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* **23**, 1388–1401 (2013).

20 K. Pandremmenou, L. P. Kondi, and K. E. Parsopoulos, "Kalai-smorodinsky bargaining solution for optimal resource allocation over wireless ds-cdma visual sensor networks," *Proc. SPIE* **8305**, 83050T–83050T–11 (2012).

21 H. Park and M. van der Schaar, "Multi-user multimedia resource management using nash

bargaining solution," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, **2**, II–717–II–720 (2007).

22 T. Bouwmans, F. E. Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection a survey," in *Recent Patents on Computer Science*, 219–237 (2008).

23 T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review* **1112**, 31 – 66 (2014).

24 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 Intl. Conf. on*, 1–7 (2013).

25 R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second ed. (2004).

26 S. Colonnese, F. Cuomo, and T. Melodia, "Leveraging multiview video coding in clustered multimedia sensor networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, 475–480 (2012).

27 M. Makar, V. Chandrasekhar, S. Tsai, D. Chen, and B. Girod, "Interframe coding of feature descriptors for mobile augmented reality," *Image Processing, IEEE Transactions on* **23**, 3352–3367 (2014).

28 S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Proceedings of the 2011 Intl. Conf. on Computer Vision*, 2548–2555 (2011).

29 C. Touati, E. Altman, and J. Galtier, "Generalized nash bargaining solution for bandwidth allocation," *Computer Networks* **50**(17), 3242 – 3263 (2006).

30 A. Redondi, D. Buranapanichkit, M. Cesana, M. Tagliasacchi, and Y. Andreopoulos, "Energy consumption of visual sensor networks: Impact of spatio-temporal coverage," *Circuits and Systems for Video Technology, IEEE Transactions on* **24**, 2117–2131 (2014).

31 L. Bondi, L. Baroffio, M. Cesana, A. Redondi, and M. Tagliasacchi, "Ez-vsn: an open-source and flexible framework for visual sensor networks," *Internet of Things Journal, IEEE* **PP**(99), 1–1 (2015).

Alessandro E. C. 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). Currently he is an Assistant Professor at the "Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano" and his research activities are focused on algorithms and protocols for Visual Sensor Networks.

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, Informazione e Bioingegneria - 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 Associate 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 Associate Professor at the "Dipartimento di Elettronica, Informazione e Bioingegneria - 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 co-authored more than 120 papers in international journals and conferences, including award winning papers at MMSP 2013, MMSP2012, ICIP 2011, MMSP 2009 and QoMex 2009. 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.