Multi-View Coding and Routing of Local Features in Visual Sensor Networks

Alessandro Enrico Redondi, Luca Baroffio, Matteo Cesana, Marco Tagliasacchi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Italy
Email: {name.surname}@polimi.it

***Abstract*—Visual Sensor Networks (VSNs) have been recently used for implementing automatic visual analysis tasks where local image features, instead of images, are compressed and transmitted to a central controller. Such features may also be compressed in a multi-view fashion, exploiting the redundancy between overlapping views. In this paper we analyze the problem of multi-view coding and routing of features in VSNs. We empirically analyze the relationship between the bitrate reduction obtained with a practical multi-view local features encoder and several geometry-based, image-based and feature-based predictors. The purpose of this analysis is to identify the most accurate, yet compact predictor of the achievable compression efficiency when jointly encoding correlated streams of local features. Then, we propose a robust optimization framework that exploits the aforementioned predictors. The proposed mathematical problem maximizes the amount of data extracted from the VSN by properly routing the streams of features, subject to capacity, interference and energy constraints, explicitly considering the uncertainty in the compression efficiency estimation. Extensive experiments on simulated VSNs show that multi-view coding maximizes the amount of data extracted from camera nodes, while the robust optimization approach provides significant improvement in uncertain scenarios compared to the optimal solution of a deterministic approach.**

I. Introduction

Visual Sensor Networks (VSNs) are composed of many low-cost, battery-operated wireless camera sensors with the ability of acquiring, processing and transmitting visual data. By extending the capabilities of traditional Wireless Sensor Networks (WSNs), VSNs will play a major role in the evolution of the Internet-of-Things (IoT) paradigm by enabling visual data gathering, processing and analysis. Especially in the scenario of smart cities, VSNs may be used to implement several visual analysis applications such as traffic and infrastructure monitoring, vacant parking lot detection, surveillance and many others. VSNs are particularly stimulating from the research point of view as they pose additional challenges compared to traditional WSN. Such challenges come from the struggle between applications requirements and technology constraints: on the one hand, applications based on visual data generally require intense processing and high bandwidth availability. On the other hand, VSNs are characterized by tight energy, processing and bandwidth constraints, thus calling for advanced solutions in the areas of data compression, processing and networking.

Very recently, a new trend has emerged as a possible solution to enable visual analysis in resource constrained VSNs. The key tenet is that many analysis tasks are carried out through the extraction of distinctive local features from the image data [1]. Each local feature is composed by a *keypoint*, i.e. a salient region of the image, and a *descriptor*, which summarizes the photometric properties of the image area around the keypoint. Being robust to several transformations (scale, rotation, illumination, viewpoint, etc...) such local features are particularly suited to performing analysis tasks such as object detection, recognition, tracking and classification.

Recent studies have shown that this kind of features have several interesting properties: first, they can be extracted very efficiently even on low-power architectures; second, they are able to summarize the salient parts of the visual content with a much more compact representation than traditional compressed image data; third, they can be further encoded in a global representation known as Bag of Visual Words (BoVW) which is extremely efficient for storage, transmission and retrieval purposes.

These characteristics have paved the way for a novel paradigm for visual analysis in wireless networked scenario: instead of acquiring, compressing and transmitting *images* to a central server for further analysis, camera nodes may extract, compress and transmit *features* to the server, thus greatly reducing the utilized bandwidth, yet without sacrificing the performance of the analysis that follows. Such a paradigm has been applied successfully to several networked scenarios, including VSNs [2] and mobile visual search [3], and constitutes the basis of the recently released MPEG-7 CDVS (Compact Descriptors for Visual Search) standard [4].

In VSNs, it often happens that multiple camera sensors are deployed in the same area and it is likely that their fields of view (FoVs) overlap. This redundancy is typically enforced to increase the robustness of monitoring (e.g, ensuring visual coverage even in case of camera failures) or its accuracy (e.g., avoiding occlusion). As a consequence, the visual data acquired by cameras with overlapping FoVs may exhibit a high degree of correlation, regardless of the paradigm chosen for visual data transmission (i.e., image-based or feature-based). Since bandwidth is generally constrained in VSNs, it is imperative to find an efficient way to remove the redundancy before data transmission. Several works in the past have addressed the problem of compressing correlated image data in networked scenarios. Exploiting recent advances in the fields of multi-view coding (MVC) [5] and distributed video coding (DVC) [6], such works generally aim at maximizing the overall compression efficiency following a two steps approach: first, the correlation existing between different cameras is predicted using either geometric [7] or content-dependent information [8]; then, based on this prediction, routing in the network is optimized so as to maximize different performance metrics (e.g., lifetime [9], [10] or quality-of-service [11]). Still, to the best of our knowledge, very limited work has been done for what concerns the compression and transmission of correlated streams of visual features, as previous works focused only on image/video data. In this paper we propose a robust optimization framework to jointly encode and transmit correlated streams of features in resource-constrained VSNs. In particular, the novel contribution of this paper is twofold:

1) We analyze empirically the relationship between the bitrate reduction obtained with a practical multi-view local feature encoder and several predictors. We consider different types of predictors, including topology-based, image-based, feature-based and mixed ones. The purpose of the analysis is to identify the most accurate, yet compact, predictor of the achievable compression efficiency when jointly encoding correlated streams of local features.
2) We propose a joint multi-view coding and routing robust optimization framework that exploits the aforementioned predictor. We introduce a mathematical formulation that seeks the optimal routing such that the amount of data extracted from the VSN is maximized. The formulation explicitly takes into account the uncertainty in the estimation of the compression efficiency, as well as network-related constraints (e.g., capacity and interference) and the energy costs of compressing and transmitting visual features, which are obtained through measurements on a real VSN testbed.

The rest of this paper is organized as follows: Section II discusses the related works in the area of compression and routing of correlated visual data in VSNs. In Section III we present an empirical analysis aimed at selecting the best predictor of the achievable multi-view feature compression efficiency. The resulting predictor is leveraged in Section IV to formulate a joint multi-view coding and routing robust optimization problem. Section V provides an extensive experimental evaluation of the proposed framework, while Section VI concludes the paper.

II. Related Work

Several solutions for joint compression and routing of correlated images in visual sensor networks are available in the literature. As mentioned earlier, such works follow a two steps approach: in the first step, an estimation process is performed to predict the possible gain resulting from the joint compression of images acquired by two or more cameras with overlapping FoVs. In the second step, the prediction is leveraged to perform network-related optimizations. In [7] a spatial correlation model is proposed to describe the redundancy existing between multiple homogeneous cameras (i.e., with the same focal length). The proposed model uses geometrical information of the cameras (e.g., their locations and sensing directions) to estimate a correlation coefficient between them. The correlation coefficient is used in [12] to predict the compression efficiency of H.264 with MVC extension, and to partition a VSN into a set of coding clusters such that the global coding gain is maximized. In [9], the correlation coefficient between cameras is leveraged to set up three different network optimization problems targeting (i) the placement of multimedia processing hubs to collect and encode correlated images in a VSN, (ii) the maximization of the global compression gain and (iii) the maximization of the VSN lifetime. In [13], a correlation-aware quality-of-service routing algorithm is proposed in order to minimize energy consumption in the network subject to delay and reliability constraint. The work in [10] proposes a joint coding/routing optimization problem which maximizes the lifetime of a VSN subject to image distortion and rate constraints. Again, the key parameter in evaluating the rate-distortion function of each camera is an inter-view spatial correlation coefficient, which is assumed inversely proportional to the distance between two cameras. Clearly, providing an accurate modeling of the relation between camera correlation and multi-view compression efficiency is of key importance in such works. Therefore, several efforts have been made to improve such a modeling, either taking into account camera heterogeneity [14], or departing from a geometric/spatial approach and taking a different approach which explicitly takes into account the visual content of the different views. In [8], the *common sensed area* (CSA) between different camera views is defined and used as a predictor for the compression efficiency of multi-view coding. Differently from previous works, the CSA is not computed based on geometric information, but is estimated starting from downsampled images which are exchanged between cameras. The main benefits in taking this approach is that it is robust to several scene-related factors (presence of moving objects, occlusions, illumination changes, etc...) that a geometric model may not capture accurately. The CSA is also leveraged in [15] to evaluate the possible benefits of a joint coding/routing scheme in multi-hop VSNs.

All the aforementioned works deal with coding and transmission of correlated *images*. To the best of our knowledge, there are no works targeting joint coding/routing of *features* data in networked scenarios, although some preliminary works have studied the problem of compressing local [16] or global features [17] extracted from multiple, correlated views. This multi-view features coding (MVFC) approaches form the basis for joint coding/routing of features data in VSNs.

III. MVFC Compression Efficiency Prediction

This section describes the approach taken to estimate the compression efficiency of a practical multi-view features encoder based on different predictors. First, we give a brief background on local and global features extraction and multi-view features coding. Then, we describe different predictors for the MVCF compression efficiency, including existing and novel approaches. We compare their performance in terms of accuracy of prediction and overhead transmission cost.

A. Background on local and global features

There exist several different algorithms for extracting local visual features from an image, all following a two-steps approach. First, a *detector* algorithm identifies salient keypoints $\mathbf{k}$ in the image. Each keypoint is generally characterized by its location, dimension (scale) and principal orientation of the surrounding patch of pixels. Then, for each keypoint, a *descriptor* vector $\mathbf{d}$ is computed, which summarizes the photometric properties of the image area around the keypoint. A visual feature f is then composed of a keypoint and the corresponding descriptor, i.e. $f = \{\mathbf{k}, \mathbf{d}\}$, and we denote as $\mathcal{F}$ the complete set of features extracted from an image. Without loss of generality, in this work we consider SIFT features, which are widely recognized as the gold standard in terms of performance for a broad range of visual analysis tasks.

The set $\mathcal{F}$ can also be transformed in a global representation known as Bag of Visual Words according to the following process. First, a vocabulary of W descriptors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_W$, (known as visual words) is learned from a large number of representative descriptors. Second, the set $\mathcal{F}$ is clustered using the W visual words as centers. Finally, a BoVW histogram is produced: the histogram has W bins $b_i, i = 1 \ldots W$ where b_i counts the number of descriptors from $\mathcal{F}$ mapped to the word $\mathbf{w}_i$. Usually, a descriptor is mapped to its nearest centroid in the descriptor space. Due to its compactness and ability to summarize efficienctly an image content, the BoVW representation is generally used in the field of content based image retrieval from very large databases [18].

B. Multi-view local features coding (MVFC)

Several compression algorithms have been proposed to efficiently encode a set of features $\mathcal{F}$. Mimicking the best practices in the field of video and image coding, such algorithms exploit either the redundancy of the set of features, or encode $\mathcal{F}$ using another set of features $\mathcal{F}_r$ as reference. As an example, the work in [19] proposes a coding scheme to encode set of features extracted from video sequences (i.e., $\mathcal{F}_r$ and $\mathcal{F}$ are extracted from temporally adjacent frames). In a similar fashion, the work in [16] propose a multi-view local features coding scheme, where $\mathcal{F}$ and $\mathcal{F}_r$ are extracted from two correlated views.

C. MVFC compression efficiency prediction

We have implemented a multi-view local features encoder following the design proposed in [16]. Given two sets of features $\mathcal{F}_i$ and $\mathcal{F}_j$ extracted from two cameras with overlapping fields of view, we define the MVFC coding efficiency as:

$$\eta_{i,j} = \frac{R_j - R_{i,j}^{\text{MVFC}}}{R_j}, \tag{1}$$

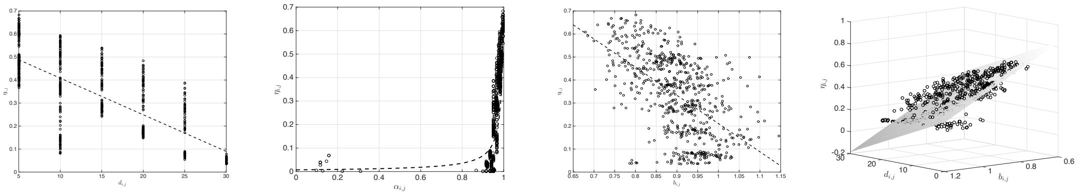
where R_j is the rate needed to encode $\mathcal{F}_j$ alone and $R_{i,j}^{\text{MVFC}}$ is the rate to encode $\mathcal{F}_j$ using MVFC with $\mathcal{F}_i$ as reference. Put in other way, $\eta_{i,j}$ is the achievable bitrate reduction (in percentage) on $\mathcal{F}_j$, when using $\mathcal{F}_i$ as reference. Note that $\eta_{i,j}$ can be computed exactly only after camera i transmits $\mathcal{F}_i$ to camera j for multi-view encoding. Our goal is to estimate $\eta_{i,j}$ without explicit transmission of $\mathcal{F}_i$ to camera j, so that network optimization can be put in place. Clearly, as it happens for multi-view coding of pixel-domain content such as images and video, we expect $\eta_{i,j}$ to be directly related to the the amount of correlation existing between the two cameras. Thus, we study the relationship of $\eta_{i,j}$ with the following predictors of inter-camera correlation:

- *Geometry-based:* similarly to [7], we consider geometric predictors only, such as the distance $d_{i,j}$ between the two camera centers or the angle $\theta_{i,j}$ between their sensing directions.
- *Image-based:* as proposed in [8], the two cameras may exchange a thumbnail version of the acquired images and estimate their correlation based on them. The CSA $\alpha_{i,j}$ between two cameras is computed on a per-frame basis as the number of pixels in the overlapping region between the view of camera i and a suitably displaced version of the view of camera j. The displacement is chosen so at to maximize the inter-view normalized bidimensional crosscorrelation function. The suggested size for the thumbnail to be exchanged between the two cameras is 22×18 pixels.
- *Feature-based:* The BoVW representation can be naturally used to understand the similarity of two images. After feature extraction, camera i may produce a BoVW histogram and transmit it to camera j, which also computes its own histogram. In practice, the BoVW histrograms are first normalized to unit length and then quantized before transmission. Finally, a distance measure between the two histograms can be computed and used as predictor of inter-view correlation. Clearly, several degrees of freedom are available, such as the size of the vocabulary and the distance measure to be used. Here, we use BoVW histograms with incresing vocabulary size W in the range {128, 256, 512, 1024, 2048}. We use the Euclidean distance as a measure of similarity between two histograms.
- *Mixed approaches:* we also evaluate multi-predictor approaches, where the MVFC compression efficiency is estimated based on the knowledge of both the geometry between the cameras and a content-based predictor.

To compare and evaluate the predictors, several tests have been performed on publicly available multi-view video sequences and image datasets. In particular, we relied on two different types of multi-view datasets:

a) Linearly spaced cameras with parallel sensing directions ($d_{i,j} > 0, \theta_{i,j} = 0$): The three datasets Akko&Kayo, Kendo and Balloons[1] all contain multi-view video sequences recorded with a linear array of cameras with 5-cm spacings. From each dataset, 6 camera pairs are chosen, corresponding to a linear spacing of 5,10,15,20,25 and 30 cm respectively.

[1] http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/



(a) $\hat{\eta} = p_1 \cdot d + p_2$ (b) $\hat{\eta} = \eta_{\max} \cdot \epsilon/(1 - \alpha + \epsilon)$ (c) $\hat{\eta} = p_1 \cdot b + p_2$ (d) $\hat{\eta} = p_1 \cdot d + p_2 \cdot b + p_3$

Fig. 1. Relationship between the compression efficiency η and different predictors for the Akko&Kayo dataset: (a) physical distance between cameras in cm, (b) CSA, (c) Euclidean distance between normalized BoVW histograms with 1024 bins and (d) mixed geometry- and feature-based approach. Each graph shows also the result of model fitting.

For each pair, 50 frames are chosen. This gives a total number of image pairs (samples) equal to $3 \times 6 \times 50 = 900$.

b) Image datasets where cameras have non-parallel sensing directions ($d_{i,j} > 0, \theta_{i,j} > 0$): The Columbia Object Image Library (COIL-100)[2] and Amsterdam Library of Image Objects (ALOI)[3] contain images of objects captured at 72 different poses obtained by rotating the object by 5 degrees each time. From each dataset, 8 camera pairs are selected, corresponding to the following angles between the camera sensing directions: $\{\pm 5°, \pm 10°, \pm 15°, \pm 20°\}$. For each camera pair, 50 images are selected, for a total of $2 \times 8 \times 50 = 800$ samples.

The tests have been performed according to the following steps: for each sample (i.e., each couple of frame (i, j) from the aforementioned datasets), we extract SIFT local features. Frame i is used as reference view, and the corresponding set of features $\mathcal{F}_i$ is used to encode the features extracted from the j-th view using the MVFC encoder. The resulting coding efficiency $\eta_{i,j}$ is stored. Simultaneously, for the same couple of frames (i, j) we stored the physical distance $d_{i,j}$ between the cameras, or the angle between the camera sensing directions $\theta_{i,j}$ (depending on the type of dataset under study), the CSA value $\alpha_{i,j}$ and the Euclidean distance between the BoVW representations of the two frames, $b_{i,j}$.

Figure 1 shows the relationship between the MVFC compression efficiency and the different predictors for the dataset characterized by linearly spaced cameras (similar results are obtained for the datasets with non-parallel sensing directions). Our goal is to find a model based on such predictors such that (i) the estimation accuracy is maximized and (ii) the cost of transmitting the predictor is minimized. We rely on linear regression for the geometric-based (Fig. 1(a)) and feature-based (Fig. 1(c)) predictors. For the image-based (CSA) predictor, we rely on the model defined in [15]:

$$\hat{\eta}_i = \eta_{\max} \cdot \frac{\epsilon}{1 - \alpha_i + \epsilon}, \tag{2}$$

where $\eta_{\max}$ is the maximum observed compression and ϵ is a parameter to be estimated. Finally, for the mixed case (Fig. 1(d)), we used multi-linear regression. To evaluate the accuracy of each model, we estimate the parameters with least-squares and we compute the root mean squared error (RMSE) between the observed and predicted compression efficiency, $\hat{\eta}_i$ and η_i, respectively:

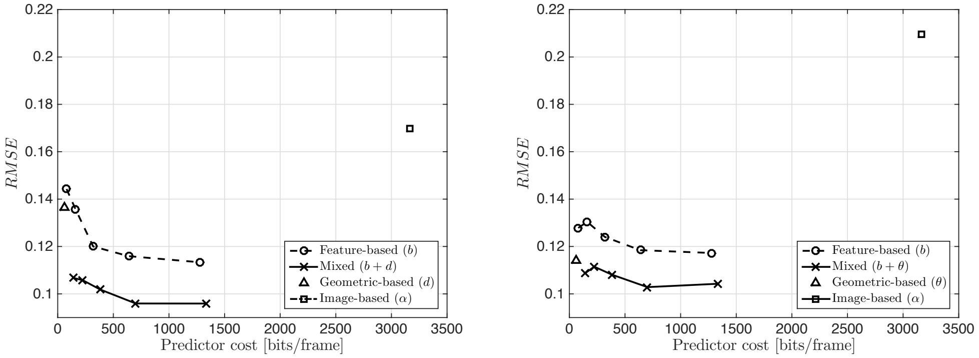
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}_i - \eta_i)^2}, \tag{3}$$

The RMSE computation was performed relying on k-fold cross-validation with $k = 5$. For each model, we also compute the cost of transmitting the predictor between the two cameras. For the geometry-based model, one may assume that the physical topology is known to all cameras and therefore there is no need of transmitting the predictors. For the image-based (CSA) model, we rely on what suggested in [15]: assuming the exchange of a thumbnail image of 22×18 pixels between the cameras, and a pixel depth of 8 bits, the cost of such a predictor is 3168 bits. For the feature-based models, the cost of the predictor depends on the BoVW size W. We assumed that each bin in the histogram is quantized uniformly using 8 bits, and the output symbols of the quantizer are lossless coded with an arithmetic encoder (whose symbols probabilities are learned from a great set of quantized BoVW histograms).

Figures 2(a) and 2(b) illustrate the accuracy-cost tradeoff obtained for the different predictors, on datasets characterized by cameras with parallel sensing directions (Akko&Kayo, Ballons and Kendo) and cameras with non-parallel sensing directions (Coil-100, ALOI), respectively. Basing on the inspection of such results, several consideration can be made:

- *Image-based methods seem to perform poorly with respect to the other predictors*, exhibiting the worst RMSE at the highest cost for transmission. The performance in terms of RMSE are even worse in the case of datasets where cameras have non-parallel sensing direction. This is expected, considering that the CSA is estimated by finding a suitable linear displacement between two views and maximizing their crosscorrelation. When the inter-geometry between cameras is not a pure translation, such a method fails.
- *methods based solely on geometric information perform particularly well on the tested dataset.* On the one hand, they are the cheapest solution in terms of data transmitted between the two cameras (if the geometry is known a-priori, there is not even the need to transmit such information). However, we posit that such a particularly good performance is due to the datasets used for the

[2] http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php
[3] http://aloi.science.uva.nl/



(a)

(b)

Fig. 2. Performance of the different predictors in terms of their estimation accuracy and overhead transmission cost for (a) camera with parallel sensing directions and (b) cameras with non-parallel sensing direction.

analysis, which are generated in controlled scenarios without occlusions and difference in lightings conditions.

- *The performance of feature-based methods generally increase with the BoVW size:* this is clearly expected, as a larger vocabulary can represent more accurately similarity between two images. Also, the performance seems to saturate for vocabularies larger than 1024 words.
- *Overall, the best performance is obtained fusing the information from geometric-based and feature-based methods*.

From such considerations, we select the mixed approach with BoVW size $W = 1024$ as predictor of compression efficiency. The transmission of the selected predictor requires about 700 bits and the RMSE is 0.096 for the case of linearly spaced cameras and 0.103 for cameras with non parallel sensing directions.

IV. Network Optimization

The predictor identified in Section III can be leveraged to set up a mathematical framework in order to optimize the operation of the VSN. In particular, given a VSN topology, we are interested in finding (i) which camera nodes should cooperate (that is, they jointly encode their set of visual features in a multi-view fashion) and (ii) what are the optimal routing paths from camera nodes to the sink node so that the amount of information extracted from the VSN is maximized.

A. Network Model

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph that models a visual sensor network, in which $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ denotes the set of wireless links. Let $\mathcal{V} = \mathcal{C} \cup \mathcal{S}$, being $\mathcal{C}$ the set of camera nodes and $\mathcal{S}$ the sink node. A directed link $(i, j) \in \mathcal{E}$ exists if nodes i and j (with $i, j \in \mathcal{V}$) are in communication range. Without loss of generality, we consider the case of symmetric links only, that is, if $(i, j) \in \mathcal{E}$ then $(j, i) \in \mathcal{E}$, as well. Each camera node acquires an image and extracts a set of visual features from it. Such features have to be transmitted to the sink node for analysis purposes. Let ρ_i be a variable that denotes the amount of visual feature data transmitted by the i-th camera. Finally, let $f_{i,j}$ be a variable denoting the flow over the directed link (i, j).

The optimization problem should decide whether two cameras i and j have to jointly encode their set of features or not. In case they do, we say that the two cameras *cooperate*. Let $x_{i,j}$ be a binary variable defined as:

$$x_{i,j} = \begin{cases} 1 & \text{if cameras } i \text{ and } j \text{ cooperate} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In case the two cameras cooperate, that is, if $x_{i,j}$ is equal to 1, camera i will transmit its set of features $\mathcal{F}_i$ to camera j, which will encode its own set $\mathcal{F}_j$ with the multi-view feature encoder, using $\mathcal{F}_i$ as reference view.

B. Objective function

The objective of the optimization problem is to maximize the amount of data extracted from the VSN. Such an objective function is a natural choice for VSNs, as the amount of data extracted from camera nodes is proportional to the visual quality of the content itself. When camera nodes transmit images or videos whose final purpose is to be perceived by human beings, the visual quality is captured by rate-distortion models. Conversely, when camera nodes transmit local features that have to be analyzed by a computer algorithm (e.g., object recognition, people identification), the quality of the visual features is capture by proper rate-accuracy models [20]. In any case, the higher the amount of data extracted by camera nodes, the better the visual quality/accuracy.

Although different functions can be used as objective (i.e., maximizing the sum or the mean of such rates over all cameras), here we rely on a fair max-min approach by maximizing the minimum source rate r, that is:

$$\max_{\rho, f, x} r \tag{5}$$

$$\text{s.t.}$$

$$\rho_i \geq r \tag{6}$$

This ensures that camera nodes in the VSN will receive a fair assignment for their output source rates.

C. Energy constraints

In VSNs, camera nodes are generally battery-operated and their energy resources are limited. Therefore, one may want to limit the energy consumption of each camera so that at least R rounds can be completed:

$$E_{\text{rx}} \sum_{(n,c)\in E} f_{n,c} + E_{\text{tx}} \sum_{(c,n)\in E} f_{c,n} + E_{\text{extr}} + E_{\text{c}}^{\text{inter}} \sum_{b\in\mathcal{C}} x_{b,c} \tag{7}$$

$$+ E_{\text{c}}^{\text{intra}}(1 - \sum_{b\in\mathcal{C}} x_{b,c}) \leq \frac{\bar{E}}{R},\ \forall c \in \mathcal{C},$$

where $\bar{E}$ is the initial energy budget of each camera. In particular, we consider that camera nodes consume energy for receiving and transmitting data (E_{rx} and E_{tx}, respectively), as well as for extracting features from the acquired images (E_{extr}) and compressing them, either relying on the set of features coming from another view or not ($E_{\text{c}}^{\text{inter}}$ or $E_{\text{c}}^{\text{intra}}$, respectively).

D. Cooperation-related constraints

As a first step, we impose some practical limits on cooperation. In particular, we define the following constraints:

$$\sum_{j\in\mathcal{C}} x_{i,j} \leq 1,\ \forall i \in \mathcal{C}, \tag{8}$$

$$\sum_{i\in\mathcal{C}} x_{i,j} \leq 1,\ \forall j \in \mathcal{C}, \tag{9}$$

$$x_{i,j} + x_{j,i} \leq 1\ \forall i, j \in \mathcal{C}. \tag{10}$$

Constraint (8) impose that a camera can be used as reference view only by at most one other camera. Similarly, constraint (9) impose that one camera can use at most on reference view. Note that, in the most general scenario, constraints (8) and (9) may also be avoided. For removing the latter, one should model the fact that the same set of features is used as reference view for different cameras. In a networked scenario, this require to transmit the same set of features to all the cameras willing to use it as a reference, thus making the network modeling more complex. We leave this as future work. Similarly, we do not deal with the fact that a single view may be encoded with reference to multiple views, as avoided by constraint (9). While this is possible in principle, we expect limited improvements in the coding efficiency when using multiple references.

Finally, constraint (10) states that cooperation may work only in one direction at a time, that is if camera i is used as reference for camera j, then the opposite should be avoided.

Note that, in case of multi-view encoding, there must exist a flow of (at least) ρ_i on the link from i to j, that is:

$$f_{i,j} \geq \rho_i \cdot x_{i,j}\ \forall (i,j) \in \mathcal{E}, i \in \mathcal{C}, j \in \mathcal{C} \tag{11}$$

Constraint (11) is not a strict equality: this allows to model the case of camera nodes which also act as relays. In this case, even if cameras i and j do not jointly encode their set of features (i.e., $x_{i,j} = 0$), still camera i may route its (or other) traffic through camera j.

The sum of outgoing flows from the j-th camera can be written as:

$$\sum_{k|(j,k)\in\mathcal{E}} f_{j,k} = \sum_{k|(k,j)\in\mathcal{E}} f_{k,j} + \rho_j \cdot (1 - \sum_{i\in\mathcal{C}} x_{i,j}\hat{\eta}_{i,j}),\ \forall j \in \mathcal{C} \tag{12}$$

Constraint (12) states that in case camera i and j cooperate, the outgoing flow of camera j is computed by incrementing its incoming flow with the quantity $\rho_j \cdot (1 - \sum_i x_{i,j}\hat{\eta}_{i,j})$, which represents the multi-view compressed bitstream of camera i given the reference flow of camera j. The compression efficiency $\hat{\eta}_{i,j}$ is estimated on a per-frame basis using the mixed-model presented in Section III, after the two cameras have exchanged their BoVW histograms.

E. Flow conservation constraints

The formulation of the optimization 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 node). 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:

$$\sum_{j\in\mathcal{C}} \rho_j \cdot (1 - \sum_{i\in\mathcal{C}} x_{i,j}\hat{\eta}_{i,j}) = \sum_{i|(i,s)\in\mathcal{E}} f(i,s) \tag{13}$$

Note that constraint (13) takes into account the possible cooperation between camera nodes to determine the total amount of flow injected in the network.

F. Interference and capacity 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 an 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 constrain 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 $\mathcal{I}_{i,j}$ 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:

$$f_{i,j} + \sum_{(h,k)\in\mathcal{I}_{i,j}} f_{(h,k)} \leq c_{(i,j)} \qquad \forall (i,j) \in \mathcal{E}. \tag{14}$$

G. Robust counterpart problem

The complete problem formulation is a mixed integer nonlinear problem (MINLP) in the optimization variables ρ, f and x and it is solved on a per-frame basis after camera nodes exchange BoVW histogram and communicate the estimated compression efficiency $\hat{\eta}_{i,j}$ to the solver. However, as shown in Figure 2, such estimates are not perfect. The values of $\hat{\eta}_{i,j}$ used during the optimization may in fact be different from the actual achieved compression efficiency $\eta_{i,j}$, causing the computed solution to be inefficient or even infeasible.

In order to cope with this uncertainty, we leverage a methodology known as robust optimization [21] and find a solution that is not optimal for the nominal value of the parameters $\eta_{i,j}$, but is robust with respect to its uncertain variation. Formally, let $\min f(x,u)$ s.t. $g(x,u) \leq 0$ be an optimization problem defined over the variables x and the uncertainty parameter $u \in \mathcal{U}$, where $\mathcal{U}$ is the uncertainty set. A *robust feasible* solution to such an optimization problem is a vector x such that all realizations of the constraints from the uncertainty set $\mathcal{U}$ are satisfied, i.e. $g(x,u) \leq 0, \ \forall u \in \mathcal{U}$. The robust counterpart solution can be then formalized as the robust feasible solution which optimizes a worst-case objective function:

$$\begin{aligned} &\min_x \max_u f(x,u) \\ &\text{s.t. } g(x,u) \leq 0 \ \forall u \in \mathcal{U}. \end{aligned} \tag{15}$$

In other words. the robust optimal solution is simply the best uncertainty-immunized solution we can associate with the uncertain problem. Moreover, in many scenarios finding the robust solution is no harder than solving the deterministic problem. Clearly, the definition of the uncertainty set $\mathcal{U}$ is key to be able to formulate a robust counterpart that can be solved efficiently. For the problem defined in Sections IV-A to IV-F, the uncertainty comes from the parameters $\eta_{i,j}$, whose lower and upper bounds are known from Figure 1(d). This uncertainty affects the flow conservation constraints (12),(13) and, indirectly, in the interference constraint in (14). If the estimated value of the compression efficiency used during optimization is higher than the actual value, such constraints may be violated (e.g., the resulting flow routed on one link may exceed the link capacity). Therefore, a robust solution must satisfy the worst realization of the contraints, that is when the parameters $\eta_{i,j}$ assume their minimal values in the uncertainty set. We use such a lower bound to solve the robust counterpart problem.

V. Experimental Evaluation

To evaluate the performance of the proposed framework, we have carried out extensive experimental simulations. In particular, we are interested in assessing how beneficial multi-view coding is in realistic scenarios, compared to the case in which camera nodes transmit their data to the sink independently. For the task at hand, we simulated several VSN instances characterized by different numbers of camera nodes deployed uniformly at random in a squared area of 100m×100m, with one sink node deployed at the center of the area. A link is established between two cameras if their distance is less than the communication range R_{comm}, which is set equal to 30 meters. Some of the links are removed with probability $p_r = 10\%$ to simulate asymmetries and noise in the radio environment. For each couple of cameras in one network instance, we select a couple of images in the datasets with parallel sensing directions and we generate a couple of nominal and actual compression efficiency parameters $\hat{\eta}_{i,j}$ and $\eta_{i,j}$, either by prediction or by running the MVFC encoder.

TABLE I
Measurements from a VSN testbed

Name	Symbol	Value
Link capacity	c	30 kbps
Transmission power	E^{tx}	$5.35{\times}10^{-5}$ J/bit
Reception power	E^{rx}	$5.35{\times}10^{-5}$ J/bit
Energy cost for features extraction	E^{extr}	$6.42\ {\times}10^{-2}$ J
Energy cost for independent coding	E^{intra}	$8.56{\times}10^{-2}$ J/bit
Energy cost for multi-view coding	E^{inter}	$2.14{\times}10^{-1}$ J/bit
Initial energy budget	$\bar{E}$	32.4×10^{-3} J

To obtain the energy costs in (7) we implemented the multi-view feature encoder proposed in [16] on a Linux-operated BeagleBone Black platform. Such a platform may be coupled with a IEEE 802.15.4-compliant dongle such as the Memsic TelosB for low-power wireless communication and with ad-hoc camera boards to provide vision capabilities and has already been used in the past to implement real-life VSN applications [22]. The resulting energy consumption was measured indirectly by keeping track of the time spent by the platform in each operative mode and multiplying this time by the platform power consumption. Similarly, we also measured the maximum data rate achievable by the TelosB dongle to model the capacity of each link. Table I summarizes the values resulting from such measurements campaign and which have been used in the optimization problem.

We formalized both the optimization problem introduced in IV and its robust counterpart with AMPL and generated different data instances varying each time the number of camera nodes and the maximum consumed energy $\bar{E}$. The instances were solved with BONMIN [23]. All results are presented in terms on the percentage gain G obtained by the proposed framework compared to a non-cooperative case, i.e., when camera nodes transmit all their data to the sink independently without relying to multi-view coding. That is,

$$G = \frac{r_{\text{c}} - r_{\text{nc}}}{r_{\text{nc}}}, \tag{16}$$

where r_{c} is the solution in the cooperative case) and r_{nc} is the solution obtained when fixing all the variables $x_{i,j}$ to zero (i.e., maximum flow achievable in the non-cooperative case).

As a first experiment, we computed the achievable gain for different energy constraints. We give to R in (7) two representative values in order to activate or avoid the energy constraint ($R = 5000$, $R = 5$). For both cases, we vary the number of cameras from 4 to 10. We repeat the experiments 10 times averaging the results, which are shown in Table II. As one can see, multi-view coding always provides better

TABLE II
GAIN OBTAINED FOR DIFFERENT ENERGY CONSTRAINT

Number of cameras	5	7	9	10
$R = 5000$	0.0718	0.0745	0.0978	0.1175
$R = 5$	0.0522	0.0691	0.0792	0.0914

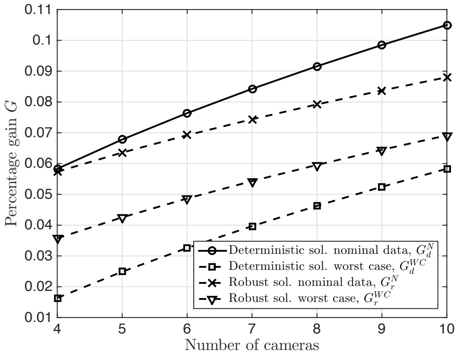


Fig. 3. Percentage gain achieved in different scenarios

performance with respect to the non cooperative case (the gain G is always positive). Moreover, multi-view coding is more attractive when the energy constraint is tight.

To evaluate the performance of the robust optimization approach, we rely on the approach presented in [24] and we compute the gain G in the following cases:

- *Deterministic solution in nominal scenario:* $G_{\mathrm{d}}^{\mathrm{N}}$, obtained by solving the deterministic solution using the parameters $\hat{\eta}_{i,j}$ as estimated by the predictor.
- *Deterministic solution in worst-case scenario:* $G_{\mathrm{d}}^{\mathrm{WC}}$, obtained using the routing solution returned in the deterministic scenario and computing the objective function value in the worst case scenario (i.e., when the $\eta_{i,j}$ have their minimum value). Such a measure captures the effect of overestimation of compression efficiency. In case of unfeasibility, we scale the source rates outgoing from each camera so that the solution is feasible.
- *Robust solution in worst-case scenario:* $G_{\mathrm{r}}^{\mathrm{WC}}$, obtained by solving the robust solution using the worst-case parameters.
- *Robust solution in nominal scenario:* $G_{\mathrm{r}}^{\mathrm{N}}$, obtained using the routing solution returned in the worst-case robust scenario and computing the objective function value in the nominal scenario. This value captures the effect of underestimation of compression efficiency.

We also compute the following ratios to compare the robust and deterministic solutions on the nominal and worst case data:

$$R_{\mathrm{ac}} = \frac{G_{\mathrm{d}}^{\mathrm{N}} - G_{\mathrm{r}}^{\mathrm{N}}}{G_{\mathrm{d}}^{\mathrm{N}}},\ R_{\mathrm{wc}} = \frac{G_{\mathrm{r}}^{\mathrm{WC}} - G_{\mathrm{d}}^{\mathrm{WC}}}{G_{\mathrm{d}}^{\mathrm{WC}}}$$

The latter ratio, R_{wc}, measures the relative benefit of the robust solution in the worst-case, i.e. the maximum protection that a robust solution can provide. The first ratio, R_{ac}, captures the percentage loss of optimality of the robust solution in the nominal case, i.e. the cost of protection. We compute the value of the gain in the different scenarios and the corresponding ratios on network instances characterized by an increasing number of camera nodes from 4 to 10. For a particular number of camera nodes used, we generate 20 different network instances and we average the obtained results. For the energy constraints, R was fixed to 5.

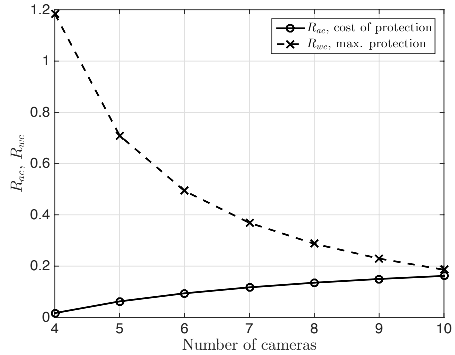


Fig. 4. Maximum protection and protection cost of the robust approach

Figure 3 illustrates the gain obtained in different scenarios. As expected, in the nominal scenario (i.e., when the compression efficiency parameters used in the optimization match with the actual data), the deterministic solution is the one that performs the best. Conversely, in the worst-case scenario, the deterministic solution performs poorly. The robust solution, instead, achieve a higher gain than the deterministic solution in the worst case scenario at the price of reduced performance in the nominal scenario. To better understand the performance of the two solutions, it is worth analyzing the ratios R_{ac} and R_{wc} shown in Figure 4. We observe that R_{wc} is always greater than R_{ac}, but this difference decreases as the number of cameras in the network increaeses. This means that the robust solution is particularly attractive for small-sized VSNs, when it is able to compensate for uncertainty while suffering very small performance losses. When the number of camera nodes increases, the performance increase of the robust solution compared to the deterministic solution in the worst case scenario decreases, and its cost increases.

We also present numerical results to illustrate the performance of the deterministic and robust approaches on a fixed network. In particular, we fix the network topology and obtain the deterministic and robust solution. Then, we generate a random set of compression efficiency parameters by perturbing the nominal estimate $\eta_{i,j}$ with increasing standard deviation σ so that they fall between the observed lower and upper bound. We compute the objective function of the deterministic and robust solution for such parameters, scaling the solutions to guarantee feasibility. We repeat the test 100 times and we show the average of the obtained values in Figure 5. As one can see, the robust solution always allow to obtain an higher gain compared to the deterministic one, for all uncertainty levels.

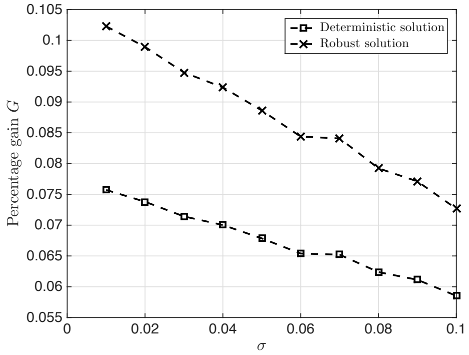


Fig. 5. Average gain of the deterministic and robust solution for different levels of uncertainty

VI. Conclusions

We addressed the problem of multi-view coding and routing of local features in VSNs. We studied the relation between the compression efficiency of a practical encoder and several predictors. We identified a novel predictor, comprising both content- and geometric- based information, which maximizes the prediction accuracy at the overhead transmission cost of about 700 bits. Then, we set up a robust optimization framework for maximizing the amount of information extracted from the VSN. The robust solution explicitly takes into account the uncertainty in the estimation of the compression efficiency and performs better than a deterministic solution in the worst case, at the cost of a small loss in optimality. Future works will address the study of distributed solution for the robust optimization problem and the extension of the proposed framework to more complex scenarios in which more than two cameras can cooperate at once.

Acknowledgment

This work has been supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open project GreenEyes (grant number: 296676), and by the Italian Ministry for Education, University and Research (MIUR) through the national cluster project SHELL, Smart Living technologies (grant number: CTN01_00128_111357).

References

[1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[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 *Multimedia Signal Processing (MMSP), 2013 IEEE 15th Intl. Workshop on*, Sept 2013, pp. 278–282.

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

[4] L.-Y. Duan, F. Gao, J. Chen, J. Lin, and T. Huang, “Compact descriptors for mobile visual search and mpeg cdvs standardization,” in *Circuits and Systems (ISCAS), 2013 IEEE Intl. Symposium on*, May 2013, pp. 885–888.

[5] A. Vetro, T. Wiegand, and G. J. Sullivan, “Overview of the stereo and multiview video coding extensions of the h.264/mpeg-4 avc standard.” *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/pieee/pieee99.html#VetroWS11

[6] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, “Distributed video coding,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, Jan 2005.

[7] R. Dai and I. Akyildiz, “A spatial correlation model for visual information in wireless multimedia sensor networks,” *Multimedia, IEEE Trans. on*, vol. 11, no. 6, pp. 1148–1159, Oct 2009.

[8] S. Colonnese, F. Cuomo, and T. Melodia, “Leveraging multiview video coding in clustered multimedia sensor networks,” in *Global Communications Conf. (GLOBECOM), 2012 IEEE*, Dec 2012, pp. 475–480.

[9] P. Wang, R. Dai, and I. Akyildiz, “Visual correlation-based image gathering for wireless multimedia sensor networks,” in *INFOCOM, 2011 Proceedings IEEE*, April 2011, pp. 2489–2497.

[10] 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 Trans. on*, vol. 21, no. 2, pp. 141–155, Feb 2011.

[11] R. Dai, P. Wang, and I. Akyildiz, “Correlation-aware qos routing for wireless video sensor networks,” in *Global Telecommunications Conf. (GLOBECOM 2010), 2010 IEEE*, Dec 2010, pp. 1–5.

[12] P. Wang, R. Dai, and I. Akyildiz, “Collaborative data compression using clustered source coding for wireless multimedia sensor networks,” in *INFOCOM, 2010 Proceedings IEEE*, March 2010, pp. 1–9.

[13] R. Dai, P. Wang, and I. Akyildiz, “Correlation-aware qos routing with differential coding for wireless video sensor networks,” *Multimedia, IEEE Trans. on*, vol. 14, no. 5, pp. 1469–1479, Oct 2012.

[14] M. Y. Mowafi, F. H. Awad, and W. A. Aljoby, “A novel approach for extracting spatial correlation of visual information in heterogeneous wireless multimedia sensor networks,” *Computer Networks*, vol. 71, pp. 31 – 47, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614002412

[15] S. Colonnese, F. Cuomo, and T. Melodia, “An empirical model of multiview video coding efficiency for wireless multimedia sensor networks,” *Multimedia, IEEE Trans. on*, vol. 15, no. 8, pp. 1800–1814, Dec 2013.

[16] L. Bondi, L. Baroffio, M. Cesana, A. Redondi, and Tagliasacchi, “Multiview coding of local features in visual sensor networks,” in *IEEE Intl. Conf. on Multimedia and Expo (ICME) - Workshop on Distributed and Cooperative Visual Recognition and Analysis (DCVRA)*, July 2015.

[17] N. Naikal, A. Y. Yang, and S. S. Sastry, “Towards an efficient distributed object recognition system in wireless smart camera networks,” in *13th Conf. on Information Fusion, FUSION 2010, Edinburgh, UK, July 26-29, 2010*. IEEE, 2010, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5711893

[18] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *Proceedings of the Intl. Conf. on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477. [Online]. Available: http://www.robots.ox.ac.uk/~vgg

[19] L. Baroffio, A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, “Coding local and global binary visual features extracted from video sequences,” *Image Processing, IEEE Trans. on*, vol. PP, no. 99, pp. 1–1, 2015.

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

[21] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*, ser. Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[22] H. Huang, C.-C. Ni, X. Ban, J. Gao, A. Schneider, and S. Lin, “Connected wireless camera network deployment with visibility coverage,” in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 1204–1212.

[23] 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.

[24] W. Ye and F. Ordonez, “Robust optimization models for energy-limited wireless sensor networks under distance uncertainty,” *Wireless Communications, IEEE Trans. on*, vol. 7, no. 6, pp. 2161–2169, June 2008.