RATE-ACCURACY OPTIMIZATION IN VISUAL WIRELESS SENSOR NETWORKS

A. Redondi, M. Cesana, M. Tagliasacchi

Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy

ABSTRACT

We consider the problem of allocating the resources in a wireless sensor network, which is designed to perform visual analysis (e.g. object recognition). We depart from the traditional compress-then-analyze paradigm, in which nodes sense, compress and transmit visual data to a sink node. Instead, we study the case in which nodes extract and lossy code local features from pixel-domain representations of the sensed visual scene. The formulation of the allocation problem entails maximizing the lifetime of the visual sensor network subject to a target accuracy of the analysis task, together with energy, bandwidth and routing constraints. To this end, we contribute with the definition of a rate-accuracy model, which plays the role of the traditional rate-distortion model commonly adopted in visual communication. The proposed model captures the impact of: i) the number of selected local features; ii) the number of bits used for quantizing local features; iii) the criterion used to select the subset of local features to be transmitted. We verify the correctness of the models on two widely adopted visual dataset and we demonstrate the network lifetime gain that can be achieved by an optimal allocation of the resources.

Index Terms— Rate-accuracy optimization, Resources allocation, Visual Sensor Networks, Object recognition

1. INTRODUCTION

Wireless sensor networks (WSN) were originally designed to sense and transmit scalar data, e.g. temperature, humidity, etc. The goal has been to minimize the manufacturing cost of the hardware devices while, at the same time, enabling low power computation and communication primitives. This has led to the standardization of tailor-made physical and link layer protocols, e.g. IEEE 802.15.4, and to the availability of off-the-shelf battery-operated nodes. On top of that, several works have attempted to adapt WSNs to acquire and deliver visual signals captured through sensor cameras. However, recent results obtained experimenting with real world visual WSNs [1],[2] have demonstrated that it is currently infeasible to stream video of sufficient quality such as to enable further analysis tasks. Empowering energy-constrained WSNs with visual analysis requires departing from traditional solutions and pursuing a paradigm shift that affects the way visual data is sensed, processed and transmitted. Stimulated by the recent results in the field of mobile visual search [3], we posit that most visual analysis tasks can be carried out based on a succinct representation of the image based on local features [4], while it disregards the underlying pixel-level representation.

In this paper we consider the problem of allocating the resources in a visual WSN, with the objective of performing visual analysis. Specifically, we envisage a scenario in which sensing nodes extract, encode and deliver SURF (Speeded-Up Robust Features) local features [5] to a sink node that uses these features to perform object recognition. The choice of adopting SURF is motivated by the fact that the detector is significantly cheaper to compute than other state-of-the-art detectors (e.g. DoG, Hessian-Laplace, etc.), thus being tailored to the severe energy constraints that characterize WSNs.The relatively recent CHoG (Compressed Histograms of Gradients)[3] feature descriptors are also based on the fast-Hessian detector used by SURF and are promising for low-energy and low-bitrate applications. However, the CHoG feature vector dimension varies according to the quantization parameters used, so that matching between differently quantized descriptors is impractical.

The formulation of the allocation problem entails maximizing the accuracy of the object recognition task, subject to energy and bandwidth constraints. We claim that traditional rate-distortion models commonly adopted in visual communication are inadequate to capture the effect of lossy compression for the analysis task at hand. Therefore, the main contribution of this paper is the definition of a rate-accuracy model, which explicitly captures the impact of: i) the number of selected local features; ii) the number of bits used for quantizing local features; iii) the criterion used to select the subset of local features to be transmitted. To the best of our knowledge, this is the first time that rate-accuracy modeling is described in a systematic fashion. In order to demonstrate the general validity of the proposed model, we consider two publicly available datasets (COIL-100 [6] and ZuBuD [7]) widely adopted as benchmark for object recognition. As a second contribution, we formulate the resource allocation problem, which seeks the optimal distribution of rate to the individual nodes, with the goal of maximizing the lifetime of the visual network for a specific target accuracy. The problem formulation explicitly considers the cost of extracting, coding and transmitting the local features, as well as bandwidth, energy, and routing constraints dictated by the individual nodes and network topology.

We organize the main contributions in two parts. First we derive a rate-accuracy model for object recognition in Section 2. Then, we formulate and solve a resource allocation problem in Section 3 with the objective of maximizing the lifetime of the sensor network for a target level of accuracy. The problem formulation entails incorporating the rate-accuracy model with network-specific constraints. Experimental results are presented in Section 4, while Section 5 concludes the paper.

2. RATE-ACCURACY MODELING

Let $\mathcal{I}_i$, $i = 1, \ldots, N$, denote the image captured by the i-th sensor, where N indicates the number of sensors. Each image is processed by means of a scale-invariant detector, which identifies salient keypoints $\mathbf{k}_{i,j}$, $j = 1, \ldots, M_i$, where M_i is the (content-dependent) number of keypoints extracted by the i-th sensor. The vector $\mathbf{k}_{i,j} = [x_{i,j}, y_{i,j}, \sigma_{i,j}, H_{i,j}]^T$ indicates the location of the keypoint in the scale-space domain and, possibly, a measure related to the saliency of the keypoint (e.g. some function of the Hessian matrix com-

puted at $(x_{i,j}, y_{i,j}, \sigma_{i,j})$). Associated to each keypoint, a descriptor $\mathbf{d}_{i,j} \in \mathbb{R}^d$ is computed from a patch around $(x_{i,j}, y_{i,j})$, whose size is proportional to $\sigma_{i,j}$. In the case of SURF local features, $d = 64$. Each element of the vector $\mathbf{d}_{i,j}$ is quantized to $r^{\mathbf{d}}_{i,j}$ bits. In this work, we assume that all descriptors of the same image are quantized at the same rate, i.e. $r^{\mathbf{d}}_{i,j} = r^{\mathbf{d}}_i$. The overall rate needed to transmit the local features of image $\mathcal{I}_i$ is equal to

$$\rho_i = M_i(r^{\mathbf{k}}_i + d r^{\mathbf{d}}_i), \tag{1}$$

where $r^{\mathbf{k}}_i$ is the rate needed to (lossless) code the spatial coordinates of each keypoint.

We consider a state-of-the-art object recognition system. A camera node acquires a query image containing a particular object, extracts SURF local features and sends them to a sink node, where the recognition process takes place. We implemented a features matching scheme following the widely used approach suggested in [8], in which potential matches are determined based on the distances between local feature vectors of the query and the database images. To enforce geometric consistency among the matched local features, spatial verification is performed. In contrast to prior works, both the number M of features to be transmitted and the quantization bit rate $r^{\mathbf{d}}$ can be independently determined for each query image.

We notice that the number of detected keypoints is content-dependent and it might exceed M. Therefore, a subset of M keypoints needs to be selected for the computation and transmission of the associated descriptors. In the following, we either select a random subset of M keypoints, or we select the top-M keypoints sorted in decreasing order of the associated Hessian response H. We then compute the corresponding M feature vectors and transmit them to the sink node. Then, we evaluate the accuracy of the visual analysis task, which is defined as the number of correct objects recognized by the system over the total number of queries. We aim at obtaining an analytic model capturing the rate-accuracy curves resulting from such experiments. Interestingly, we observed that such curves share a common behavior across all the tested datasets. As an example, Figure 1 shows a family of curves for the ZuBuD dataset [7], where each curve is obtained with a different value of M, varying the number of bits $r^{\mathbf{d}}$ used to quantize each feature dimension.

As one can see, the final accuracy does not depend only on the number of features used for recognition, but also the quantization rate must be taken into account. For example, in Figure 1, using $M = 20$ features quantized with $r^{\mathbf{d}} = 8$ bits per element results in a lower accuracy than using 40 features with 4 bits per element, although the overall rate is the same in both cases. We are interested in finding the envelope of such family of curves, which represents the best operational rate-accuracy trade-off that can be obtained. Each curve in the family shown in Figure 1 can be divided in two parts. A monotone increasing part, where the accuracy rapidly rises, followed by a saturation part where the accuracy remains approximatively constant for increasing values of the rate. To derive the envelope of the family of curves, we approximate each curve $A_M(\rho)$ in the family with a piecewise linear function using two lines:

$$A_M(\rho) = \min(g_M(\rho), h_M(\rho)), \tag{2}$$
$$g_M(\rho) = m\rho + p_M, \quad h_M(\rho) = q_M. \tag{3}$$

To reduce the number of parameters to be estimated, we impose that the lines fitting the first part of the curve have all the same slope m, and the lines fitting the second part of the curve are all parallel to the x axis. Then, for each value of M we obtain p_M and q_M by means

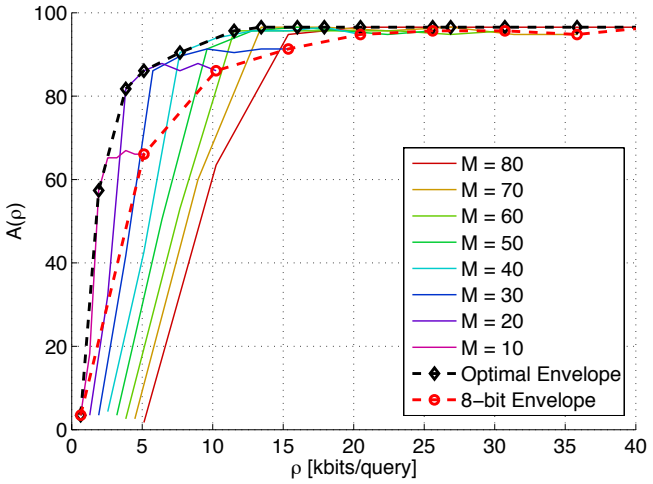


Fig. 1. Rate-accuracy curves for the ZuBuD datasets. Each of the solid colored lines represents the rate-accuracy curve for a different number of features M. The black dashed line is the envelope of the rate-accuracy curves family, and represents the best accuracy that can be obtained for a target rate. The red dashed line is the rate-accuracy curve obtained when the number of bits to quantize each feature element is fixed to $r^{\mathbf{d}} = 8$, and M varies.

of least squares and analyze their behavior when M varies, which can be modeled as follows:

$$p(M) = \alpha + \beta M, \quad q(M) = \delta + \frac{\gamma}{M} \tag{4}$$

where α, β, γ and δ are model parameters estimated from the observed family of curves. The envelope of such a family is the locus of points where $g(\rho) = h(\rho)$. That is,

$$m\rho + (\alpha + \beta M) = \delta + \frac{\gamma}{M}. \tag{5}$$

Solving for M, we get:

$$M(\rho) = \frac{-(\alpha - \gamma - m\rho) + \sqrt{(\alpha - \gamma - m\rho)^2 + 4\beta\delta}}{2\beta}. \tag{6}$$

Equation (6) gives the optimal number of features to be used at a given rate ρ, in order to maximize the accuracy of the analysis task. Substituting (6) in (1), we find the optimal number of bits to be allocated to each feature vector element. That is,

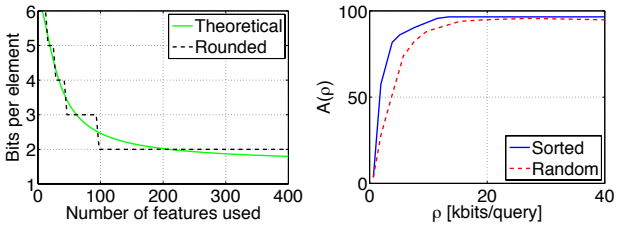
$$r^{\mathbf{d}}(\rho) = (\rho/M(\rho) - r^{\mathbf{k}})/d. \tag{7}$$

The pair $\langle M(\rho), r^{\mathbf{d}}(\rho)\rangle$ defines what we refer to as the internal allocation for a target rate ρ. Figure 2(a) shows the optimal internal allocation for the ZuBuD dataset. To further simplify the model, we approximate (6) with:

$$M(\rho) = a\rho + b\sqrt{\rho^2 + \rho} + c \tag{8}$$

Substituting back (8) in (2) we get a compact analytic form for the rate-accuracy model:

$$\begin{aligned} A(\rho) &= \alpha + \beta(a\rho + b\sqrt{\rho^2 + \rho} + c) + m\rho \\ &= (m + a\beta)\rho + b\beta\sqrt{\rho^2 + \rho} + \alpha + \beta c \\ &= p_1\rho + p_2\sqrt{\rho^2 + \rho} + p_3, \end{aligned}$$



(a) (b)

Fig. 2. (a) Optimal internal allocation for the ZuBuD dataset. The black dashed line represents the case in which we use an integer number of bits per feature element. (b) Comparison of features selection strategies for the ZuBuD datasets.

Interestingly, we find out that for all datasets, $p_1 \approx -p_2$, i.e. $(a + b) \approx -m/\beta$. To quantify the goodness of our model we compute the Pearson's correlation coefficient r between the real and the estimated envelope, obtaining a value greater than 0.99 for all datasets.

We also evaluate the impact of the feature selection strategy, i.e. transmitting M vectors at random, as opposed to transmitting the top-M vectors sorted based on the Hessian response. Figure 2(b) shows the resulting rate-accuracy envelopes for the two cases. As expected, a random selection strategy achieves poorer rate-accuracy efficiency.

3. RESOURCE ALLOCATION

3.1. Network model

Let $G = (V, E)$ be a directed graph that models a visual sensor network, in which V denotes the set of nodes and E the set of wireless links. We are particularly interested in heterogeneous networks, composed by both camera nodes and generic nodes with relay functions. Hence, let $V = C \cup N \cup S$, where C is the set of camera nodes, N the set of relay nodes and S is the set of sink nodes. To simplify the discussion we assume the presence of only one sink node. We define the node-link incidence matrix $\mathbf{A} \in \mathbb{R}^{|V|\times|E|}$, where each of its elements is:

$$a_{ij} = \begin{cases} 1 & \text{if link } j \text{ is outgoing from node } i \\ -1 & \text{if link } j \text{ is incoming into node } i \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Matrix $\mathbf{A}$ can be conveniently written in terms of two auxiliary matrices $\mathbf{A}^+$ and $\mathbf{A}^-$, whose elements are as follows:

$$a_{ij}^{\{+/-\}} = \begin{cases} 1 & \text{if link } j \text{ is } \{\text{outgoing from/incoming to}\} \text{ node } i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Hence $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$.

We decompose $\mathbf{A}$ in two sub-matrixes $\mathbf{A}_C \in \mathbb{R}^{|C|\times|E|}$ and $\mathbf{A}_N \in \mathbb{R}^{|N|\times|E|}$ by extracting, respectively, the rows corresponding to camera nodes and generic nodes. We point out that $\mathbf{A}_C = \mathbf{A}_C^+ - \mathbf{A}_C^-$ and $\mathbf{A}_N = \mathbf{A}_N^+ - \mathbf{A}_N^-$. Finally let $\mathbf{A}_{sink} \in \mathbb{R}^{|S|\times|E|}$ be the sink-links incidence matrix, which is a row vector in the simple case of a single sink node.

Finally, let $\mathbf{f} \in \mathbb{R}^{|E|}$ denote the link flow vector, which contains the flows associated to each link in the network, and $\boldsymbol{\rho} \in \mathbb{R}^{|C|}$ denote the source flow vector, which contains the source rates of each camera nodes, so that each element ρ_i in $\boldsymbol{\rho}$ is the source rate of camera $i \in C$.

3.2. Problem formulation

The resource allocation problem is formulated with the aim of maximizing the network lifetime. When the visual network is designed to perform an analysis task such as object recognition, it is convenient to express the network lifetime as the maximum number of visual queries Q_{max}, that can be successfully transmitted to the central controller, before one of the nodes in the network depletes its energy. We denote with R the channel rate for the visual network and we assume that all nodes in the network use the same transmission and reception power consumptions P_{tx} and P_{rx}. Moreover, we assume that each visual query correspond to a discrete time slot of length T. A typical approach in lifetime maximization problems [9]is to replace the variable Q_{max} with $q = 1/Q_{max}$, in order to simplify the solution. Accordingly, we formulate the resource allocation problem in the $|C| + |E| + 1$ variables $(\boldsymbol{\rho}, \mathbf{f}, q)$ as follows:

$$\text{min. } q \tag{11}$$

$$\text{s.t. } A(\boldsymbol{\rho}) \geq \bar{\mathbf{A}} \tag{12}$$

$$(P_{tx}T/R)\mathbf{A}_N^+\mathbf{f} + (P_{rx}T/R)\mathbf{A}_N^-\mathbf{f} \leq \mathbf{1} \cdot q\bar{E} \tag{13}$$

$$(P_{tx}T/R)\mathbf{A}_C^+\mathbf{f} + (P_{rx}T/R)\mathbf{A}_C^-\mathbf{f} \leq \mathbf{1} \cdot q\bar{E} - E_K(\boldsymbol{\rho}) \tag{14}$$

$$\mathbf{A}_N\mathbf{f} = \mathbf{0} \tag{15}$$

$$\mathbf{A}_C\mathbf{f} = \boldsymbol{\rho} \tag{16}$$

$$\mathbf{A}_{sink}\mathbf{f} = -\sum \boldsymbol{\rho} \tag{17}$$

$$\boldsymbol{\Gamma}\mathbf{f} \leq \mathbf{C} \tag{18}$$

where $\mathbf{1}$ and $\mathbf{0}$ are vectors of ones and zeros respectively. In this formulation, the variable q can be thought as the energy spent per visual query, normalized with respect to the total energy budget $\bar{E}$. Constraint (12) imposes that the accuracy level for each camera matches the target accuracy vector $\bar{\mathbf{A}}$. Here, we leverage the rate-accuracy model derived in Section 2. Constraint (13) limits the total energy per query spent by relay nodes, which takes into account the energy needed for receiving and transmitting visual features. Constraint (14) does the same for camera nodes, this time considering the cost of processing visual descriptors $E_K(\boldsymbol{\rho}) = E_{acq} + E_{det} + E_{desc}(\boldsymbol{\rho})$. This includes the cost of acquiring an image (E_{acq}), detecting key-points (E_{det}) and computing descriptors ($E_{desc}(\boldsymbol{\rho})$). We observe that E_{acq} and E_{det} depend only on the image resolution, while $E_{desc}(\boldsymbol{\rho})$ depends on the number of computed descriptors (i.e. $E_{desc}(\boldsymbol{\rho}) = E_{desc} \cdot M(\boldsymbol{\rho})$, being E_{desc} the cost of computing a single descriptor). Constraints (15), (16) and (17) enforce flow conservation for relay nodes, source nodes and the sink node, respectively. Constraint (18) limits the amount of flow on each link, in order to eliminate frame collisions. In this work we assume a mechanism similar to RTS/CTS, according to which two links i and j interferes with each others if: i) $i = j$; ii) j is adjacent to i; or iii) j is adjacent to another link k, which is adjacent to i. Hence, we introduce the matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{|E|\times|E|}$, where each of its entries γ_{ij} is equal to 1 if link i interferes with link j and 0 otherwise. We denote $\mathbf{C} \in \mathbb{R}^{|E|}$ as the link capacity vector, where we impose that each link i has the same capacity $C_i = R$.

We observe that the objective function and all but one of the constraints are linear in the optimization variables $(\boldsymbol{\rho}, \mathbf{f}, q)$. In (12), $A(\rho)$ is a concave function of ρ. Therefore, the optimization problem is convex and it can be conveniently solved using off-the-shelf solvers, e.g. CVX [10].

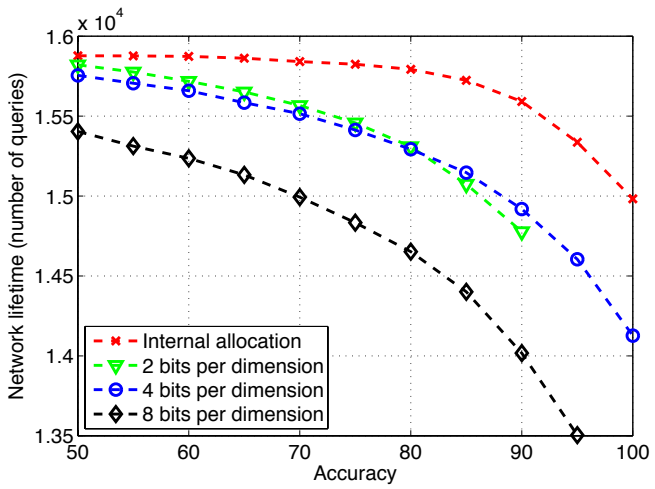


Fig. 3. Maximum network lifetime (number of visual queries transmitted) for different values of the target accuracy.

4. EXPERIMENTAL RESULTS

We implemented an object recognition system using SURF local features extracted with OpenSURF [11]. We estimated the rate-accuracy model parameters for both the ZuBuD [7] and COIL-100 [6] datasets, and plugged them into the resource allocation problem described in Section 3. We simulated a visual network composed by 35 relay nodes and 15 cameras, randomly deployed in an area of 2500 m^2. We set the parameters of the network nodes according to the characteristics of the IntelMote2 sensor node [12] , which can be equipped with a multimedia board to acquire images. The communication range of each node was set to $R_{comm} = 15m$, and we deleted links randomly with a probability $p_{del} = 0.1$, to simulate an indoor environment. We assumed that all nodes have the same energy budget $\bar{E}$ and that the required target accuracy is the same for each camera (i.e. $\bar{\mathbf{A}}$ in (12) is filled with identical values). In order to estimate the energy cost of the SURF algorithm we rely on the work in [13], where the functional blocks of the detector and descriptor algorithm are expressed in terms of atomic operations. The simulation parameters are reported in Table 1. To evaluate the benefits of the rate-accuracy optimization, we compared the proposed solution with a case in which we fixed the number of bits used to quantize each feature element, i.e. when a camera node does not perform internal allocation. Figure 3 shows the results obtained for the COIL-100 dataset. The proposed rate-accuracy model always leads to an increment in the maximum achievable number of queries. Moreover, when internal allocation is not used, the resource allocation problem may be unfeasible for a particular target accuracy level. The motivations for the unfeasibility of the problem are twofold: i) the rate-accuracy model used may never reach the target accuracy (e.g. the curve in Figure 3 corresponding to the case when feature descriptors are quantized with 2 bits per dimension); ii) the rate required to obtain the target accuracy violates one of the constraints of the problem (e.g the black curve in Figure 3, where 8 bits per dimension are used).

5. CONCLUSIONS

In this work we presented a rate-accuracy optimization framework that can be used to prolong the lifetime of a wireless visual sensor network designed to perform object recognition. We formalized a model describing the relationship between the optimal number of features to be transmitted, their quantization rate and the accuracy of the analysis task. Moreover, we formulated the resource allocation problem in a visual network considering energy and bandwidth constraints. Experimental simulations show the benefits in terms of network lifetime that can be obtained by using the proposed model.

Parameter	Description	Value
P_{tx}	Transmission Power	71.20 mW
P_{rx}	Reception Power	71.20 mW
R	Channel rate	250 kbps
E	Energy budget	10kJ
E_{acq}	Cost of image acquisition	20 mJ
E_{det}	Cost of key-points detection	45mJ
E_{desc}	Cost of one descriptor computation	0.7 mJ

Table 1. Simulation parameters

6. REFERENCES

[1] E. Culurciello, Joon Hyuk Park, and A. Savvides, "Address-event video streaming over wireless sensor networks," in *Circuits and Systems. ISCAS 2007. IEEE International Symposium on*, may 2007, pp. 849 –852.

[2] S. Paniga, L. Borsani, A. Redondi, M. Tagliasacchi, and M. Cesana, "Experimental evaluation of a video streaming system for wireless multimedia sensor networks," in *Ad Hoc Networking Workshop (Med-Hoc-Net), The 10th IFIP Annual Mediterranean*, june 2011, pp. 165 –170.

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

[4] Tinne Tuytelaars and Krystian Mikolajczyk, "Local invariant feature detectors: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[6] S. A. Nene, S.K. Nayar, and H. Murase, "Columbia object image library (coil-100),," *Technical Report*, , no. CUCS-006-96, feb. 1996.

[7] H. Shao, T. Svoboda, and L. Van Gool, "Zubud-zurich buildings database for image based recognition," *Technical report, Swiss Federal Institute of Technology*, , no. 260, 2003.

[8] M. Brown and D.G. Lowe, "Unsupervised 3d object recognition and reconstruction in unordered datasets," in *3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on*, june 2005, pp. 56 – 63.

[9] Yifeng He, I. Lee, and Ling Guan, "Distributed algorithms for network lifetime maximization in wireless visual sensor networks," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 5, pp. 704 –718, may 2009.

[10] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, April 2011.

[11] Christopher Evans, "Notes on the opensurf library," Tech. Rep. CSTR-09-001, University of Bristol, January 2009.

[12] L. Nachman, J. Huang, J. Shahabdeen, R. Adler, and R. Kling, "Imote2: Serious computation at the edge," in *Wireless Communications and Mobile Computing Conference, 2008. IWCMC '08. International*, aug. 2008, pp. 1118 –1123.

[13] P. Drews, R. de Bern, and A. de Melo, "Analyizing and exploring feature detectors in images," *IEEE International Conference on Industrial Informatics (INDIN)*, pp. 305–310, 2011.