Demo: Enabling Image Analysis Tasks in Visual Sensor Networks

L. Baroffio, A. Canclini, M. Cesana, A. Redondi, M. Tagliasacchi
DEIB - Politecnico di Milano
Milano, Italy

G. Dán, E. Eriksson, V. Fodor
KTH - Royal Institute of Technology
Stockholm, Sweden

J. Ascenso, P. Monteiro
Instituto de Telecomunicações
Lisbon, Portugal

ABSTRACT

This demo showcases some of the results obtained by the GreenEyes project, whose main objective is to enable visual analysis on resource-constrained multimedia sensor networks. The demo features a multi-hop visual sensor network operated by BeagleBones Linux computers with IEEE 802.15.4 communication capabilities, and capable of recognizing and tracking objects according to two different visual paradigms. In the traditional compress-then-analyze (CTA) paradigm, JPEG compressed images are transmitted through the network from a camera node to a central controller, where the analysis takes place. In the alternative analyze-then-compress (ATC) paradigm, the camera node extracts and compresses local binary visual features from the acquired images (either locally or in a distributed fashion) and transmits them to the central controller, where they are used to perform object recognition/tracking. We show that, in a bandwidth constrained scenario, the latter paradigm allows to reach better results in terms of application frame rates, still ensuring excellent analysis performance.

Keywords

Binary Local Visual Features, Visual Sensor Networks, ARM, Object Recognition, Object Tracking

1. INTRODUCTION

The integration of low-power wireless networking technologies such as IEEE 802.15.4-enabled transceivers with inexpensive camera hardware, has enabled the development of the so-called visual sensor networks (VSNs). Due to their flexibility and low-cost, VSNs have attracted the interest of researchers worldwide in the last few years, and are expected to play a major role in the evolution of the Internet-of-Things (IoT) paradigm with applications such as video surveillance, object and face recognition, object tracking and many others. Such visual tasks are typically accomplished through the extraction and analysis of global and local features from the pixel domain content: thus, they can be implemented in different ways in the VSN, depending on *where* in the network the task of feature extraction is performed. The traditional *compress-then-analyze* (CTA) approach relies on a local compression (JPEG / H.264) of the acquired images or image sequences at the camera sensor, which are then delivered through the wireless sensor network to a central controller that extracts the features and performs visual analysis. This operating paradigm, sketched in Figure 1(a), is referred to as compress-then-analyze. The bitstream flowing in the network includes the compressed and possibly lossy pixel-domain representation of the acquired image. As a consequence, depending on the amount of compression, the accuracy of the final analysis task might be significantly impaired. Moreover, it is widely known that multi-hop image transmission in low-bandwidth VSNs results in high latency and low application frame rates, due to the struggling between bandwidth availability and requirements. Moreover, when only the result of the visual analysis matters, transmitting image or video data retaining a pixel-level representation is inefficient in terms of the computational and network resources used. For all these reasons, the GreenEyes project considers an alternative approach where the bitstream flowing in the visual sensor network is transformed by some sort of local processing which extracts and encodes visual features, rather than compressing and transmitting a representation of the sensed images in the pixel domain. We call this approach *analyze-then-compress* (ATC) (Figure 1(b)). In this approach, image features are extracted by visual sensor nodes, encoded, and then delivered to the final destination(s) in order to enable higher level visual analysis tasks. In this demo we showcase an efficient implementation of the ATC paradigm on a real visual sensor network testbed, and we demonstrate its benefits compared to the traditional CTA paradigm in a bandwidth-limited scenario. We also show several key results of the GreenEyes project such as binary features encoding and distributed features extraction among neighboring nodes. The rest of the paper is organized as follow: scientific details are provided in Section 2, while Section 3 describes the used hardware and the user interface. Some conclusions are drawn in Section 4.

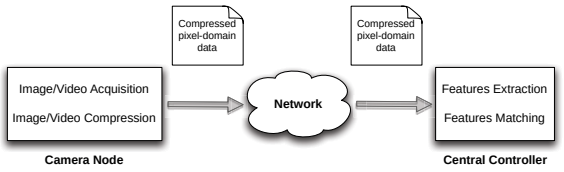
ICDSC '14, November 04 - 07 2014, Venezia Mestre, Italy

http://dx.doi.org/10.1145/2659021.2669477

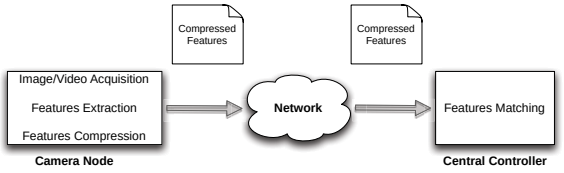
2. TECHNICAL DESCRIPTION

The use of the ATC approach constitutes a novel paradigm shift in the field of VSNs. With the proposed demonstrator, we aim at showing that ATC is indeed a viable option to enable higher frame rates compared to CTA in the case of a bandwidth limited scenario. We demonstrate several novel solutions proposed by the GreenEyes project:

1. **Energy-efficient features extraction:** While for CTA the computational effort is limited to the compression (e.g., JPEG) of the acquired images, the complexity of the feature extraction algorithms at the base of ATC may be critical. In the last few years, efficient algorithms for extracting compact binary fea-



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



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

Figure 1: The two different approaches to perform object recognition in visual sensor networks

tures have been proposed [3], in order to decrease the overall computational complexity. In this demo we leverage a recent work of ours aimed at optimizing the Binary Robust Invariant Scalable Keypoints (BRISK) features extraction algorithm for ARM-based platform. The BRISKOLA (BRISK Optimized for Low-Power ARM Architecture)[4] features extraction algorithm, which uses the ARM specific SIMD instruction set, named NEON and allows to obtain average speed-ups of 1.5 with respect to the original BRISK implementation. This makes it possible to extract approximately 50 BRISK descriptors from VGA (640×480 pixels) images in less then 50 ms on a 720 MHz ARM CPU.

2. **Lossless coding for binary features:** Furthermore, we also showcase a lossless entropy-coding scheme for compressing the extracted BRISK features [5], which achieves a coding gain around 20% for 64-bits descriptors. As a consequence, 50 BRISK descriptors will generate a bitstream constituted of approx. 2.5 kbits. For comparison, a poor JPEG compression (quality factor Q=20) of a VGA image results in about 10 kbits of data.

3. **Lossy keypoints location coding:** Location information of keypoints can be used to check geometric consistency of the descriptors and improve the retrieval performance in terms of accuracy. Since keypoints tend to cluster around particular structures of the image, it is possible to exploit this fact using a spatial grid based quantization and arithmetic coding technique as proposed by the MPEG compressed descriptor visual search (CDVS) framework.

4. **Distributed features extraction:** The camera node may also leverage the presence of several neighboring nodes to distribute the task of feature extraction, in order to reduce the overall processing time through offloading. In this demo we also showcase for the first time a practical way to achieve minimum processing time for features extraction in a distributed fashion, giving birth to a *distributed-analyze-then-compress* (DATC) paradigm. In the proposed solution the camera node uses prediction to optimally allocate slices of the image to cooperating nodes for feature extraction [2]. Each cooperator performs the features extraction algorithm in parallel, thus minimizing the overall processing time. The implemented solution is also able to automatically tune the number of cooperators to use and the dimension of each slice based on real-time network conditions and based on the image content.

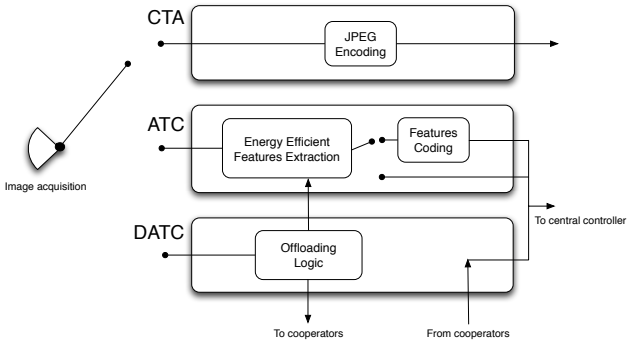


Figure 2: Visual sensor node operational modes. Switching between CTA, ATC or DATC is remotely controlled by the user. The compressed multimedia data from the ATC/DATC or CTA paradigm is sent to a central controller where it is decoded and used to perform object recognition/tracking.

3. IMPLEMENTATION

With reference to Figure 2, the demonstration is built on the following equipment:

- **Visual sensor node:** a battery-operated 720MHz ARM BeagleBone Linux computer which is geared with a Logitech USB camera to capture still images; the visual sensor node is also attached to a IEEE 802.15.4-compliant sensor node (TelosB platform or similar) to remotely transfer the visual content through low-power wireless links.

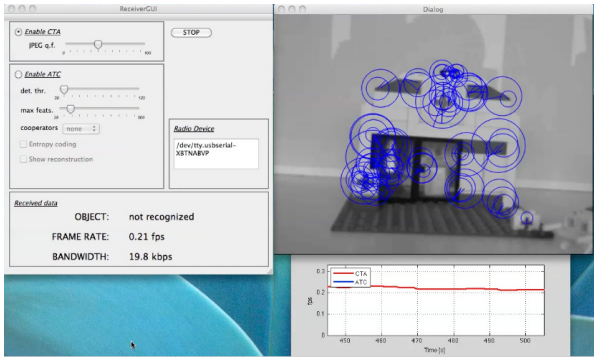
- **Cooperator nodes:** several battery-operated BeagleBone Linux computers similar to the visual sensor node but without sight capabilities. This type of nodes is used to implement the DATC paradigm.

- **Network infrastructure:** a network of battery-operated IEEE 802.15.4-compliant TelosB sensor nodes which is used to route the visual information to a central controller.

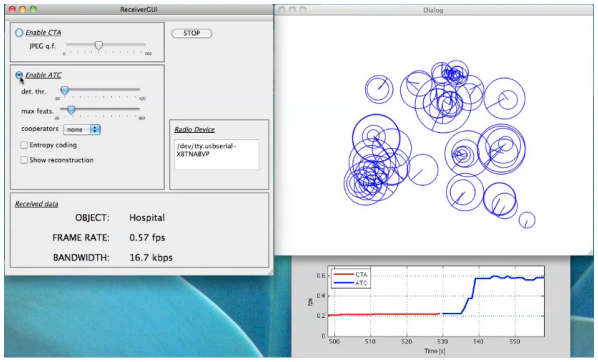
- **Central controller:** a laptop with IEEE 802.15.4 communication capabilities to receive the multimedia content transferred by the visual sensor node and to perform visual analysis.

The visual sensor node: The camera (visual) sensor node is able to to operate following the CTA, ATC and the DATC paradigms. For CTA, the visual sensor node implements the standard JPEG compression algorithm and transmits the compressed pixel-domain data. In the ATC case, the BRISKOLA features extraction algorithm is used to extract binary local visual features. Moreover, the sensor node offers the possibility to encode the extracted features following the approach mentioned in Section 2. In the DATC case, the camera node splits the acquired image into N vertical slices, where N is the number of cooperators used. Each slice is then transmitted and assigned for processing (keypoint detection and descriptor extraction) to a cooperating node, where features extraction is performed. Finally, features are transmitted back to the camera node. Note that N may be smaller than the total number of neighboring nodes present in the network, as the DATC controller will automatically select the optimal number of cooperators based on the actual network conditions.

The central controller: The data received at the central controller (either in CTA or ATC/DATC mode) is decoded and used to perform two different visual analysis tasks: object recognition and



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



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

Figure 3: Graphical user interface of the demonstration. (a) CTA paradigm: a JPEG compressed image is transmitted to the controller for recognition. (b) ATC paradigm: only a set of local visual features are remotely transmitted, ensuring recognition at higher frame rates.

tracking. The central controller implements a graphical user interface which provides a highly interactive remote controller of the visual sensor node. The user can switch on the fly between the operating paradigms (CTA, ATC or DATC) and for each paradigm different parameters may be changed. As far as the CTA case is concerned, when an image is received by the controller, it is displayed along with the positions of the detected keypoints (Figure 3(a)). The user can select the JPEG quality factor in order to control the size of the bitstream generated by the camera node. As for the ATC case, the keypoints associated to the received features are displayed (Figure 3(b)). The user can select different detection thresholds and the maximum number of features to be transmitted. Moreover, the user interface provides a switch for enabling/disabling entropy-coding of features descriptors. In case DATC is activated, the user can select manually how many cooperators to use for offloading the features extraction task, or trigger an automatic selection of the best number of cooperators to use. The received features, in the ATC case, or the features extracted from the received JPEG image, in the CTA case, are matched against a database of labeled features, so that object recognition can be performed. In particular, the demo experiment will showcase a classical object recognition task, with the central controller being able to recognize the type of object which is seen by the visual sensor node. The result of the recognition is displayed on the user interface. In addition, either in CTA or ATC mode, a recognized object can be tracked along time using a tracking-by-detection algorithm, in which the target object is detected frame by frame, i.e. the object location is estimated in every frame independently. This type of approach is suitable for the VSN scenario where only a small amount of descriptors are transmitted, objects may be occluded or disappear from the camera view and thus only a part of the target object is characterized. In this case, the user interface will display a bounding box that defines the object size and position even if no pixel based representation is transmitted (ATC mode). Moreover, the demonstrator estimates and displays the current frame rate, i.e., the maximum number of images which can be processed per unit time, under the different paradigms. As an additional feature, when the ATC paradigm is active, the graphical user interface offers the possibility of reconstructing an approximation of the image captured by the camera node starting from the knowledge of the visual features. In particular, we followed the approach presented in [1]: in a first stage the received features are matched against the features of the database; the image is then reconstructed as a composition of the image patches, extracted from the database, which exhibit the highest matching scores.

Application scenario: The demonstration scenario is composed of a LEGO model of a city containing several different objects. The camera node is mounted on a toy car which is able to move freely inside the city, in order to recognize and track the different objects.

4. CONCLUSIONS

The proposed demo showcases that, in the context of VSNs characterized by a limited transmission bandwidth, the ATC paradigm outperforms the traditional CTA paradigm in terms of the achieved frame rate, for the same performance in terms of visual analysis tasks. Moreover, leveraging the presence of neighboring nodes for distributing the task of features extraction may lead to notable performance improvements. Future work will focus on aspects related to multiple cameras in the VSN, such as inter-view features encoding. We also plan to extend the comparison between CTA and ATC to the case where temporal correlation between acquired images is exploited.

Video demo: A detailed video describing the demonstrator is available at `www.greeneyesproject.eu`

5. 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.

6. REFERENCES

[1] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *ICPR*, pages 935–938, 2012.

[2] E. Eriksson, G. Dán, and V. Fodor. Real-time distributed visual feature extraction from video in sensor networks. In *DCOSS*, pages 152–161, 2014.

[3] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555, 2011.

[4] A. Redondi, L. Baroffio, M. C. A. Canclini and, and M. Tagliasacchi. Briskola: Brisk optimized for low power arm architectures. In *IEEE International Conference on Image Processing 2014*, 2014.

[5] A. Redondi, L. Baroffio, M. C. J. Ascenso and, and M. Tagliasacchi. Rate-accuracy optimization of binary descriptors. In *IEEE International Conference on Image Processing 2013*, pages 900–903, 2013.