Object recognition in visual sensor networks based on compression and transmission of binary local features

A. Canclini, L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
P.za Leonardo da Vinci 32, Milano, Italy
lastname@elet.polimi.it

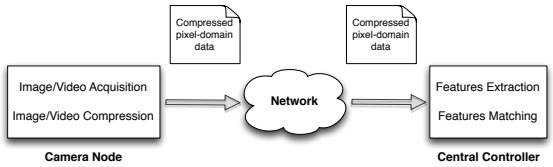
Abstract

This demo features a multi-hop Visual Sensor Network, capable of recognizing objects using 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 alternate analyze-then-compress (ATC) paradigm, the camera node extracts and compresses local binary visual features from the acquired images and transmits them to the central controller, where they are used to perform object recognition. We show that, in a bandwidth constrained scenario, the latter paradigm allows to reach higher application framerates, still ensuring excellent recognition results.

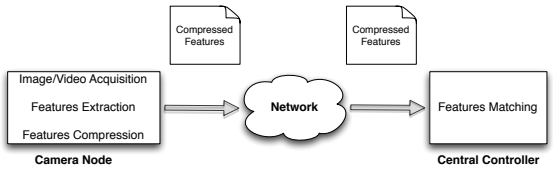
1. Introduction and motivations

Visual Sensor Networks (VSNs), composed of several inexpensive wireless camera nodes, are used to acquire images or video from the environment, which may be further processed to perform a broad range of complex visual analysis tasks, such as object recognition, event detection, localization and tracking, etc. The acquired visual information is typically delivered to a remote destination through low-power wireless transmissions (e.g., IEEE 802.15.4), and multiple visual or generic sensor nodes may be involved in the process of routing this information to its final destination. 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. We focus on a fundamental application of Visual Sensor Networks: the detection and recognition of objects. This is typically accomplished by extracting global or local distinctive features from the acquired images, and by matching these features against a database of pre-stored and labeled features to finally perform object recognition. Such visual task can be practically implemented in different ways in the VSN, depending on *where* in the network the task of feature extraction is performed. Standard implementations rely on a local compression (JPEG or similar) of the acquired images at the camera sensor, which are then delivered through the wireless sensor network to a central controller that extracts the features and performs object recognition. This operating paradigm, sketched in Figure 1(a), is referred to as compress-then-analyze (CTA). The bitstream flowing in the network includes the compressed and lossy pixel-domain representation of the acquired image. As a consequence, depending on the amount of compression, the accuracy of object recognition might be significantly impaired. Moreover, several works in the past demonstrated that multi-hop image transmission in VSNs results in high latency and low application frame rates, due to the struggling between bandwidth availability and requirements [1]. It is important to notice, however, that most visual analysis tasks (e.g. object recognition) can be carried out based on a succinct representation of the image, which entails both global and local features, while it disregards the underlying pixel-level representation. For these reasons, an alternate operating paradigm would be preferrable, in which the visual features are extracted, possibly (lossless) compressed, at the camera sensor, and then delivered to the final destination. This paradigm is referred to as analyze-then-compress (ATC), and it is illustrated in Figure 1(b).

In this demo we present an efficient implementation of the ATC paradigm on a visual sensor network. Scientific details are provided in Section 2, while Section 3 describes the hardware and the user interface. Some conclusions are drawn in Section 4.



(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

2 Scientific and 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 enables higher frame rates with respect to CTA, as the visual information transmitted through the network is much more compact. It is important to notice, however, that the frame rate depends not only on the amount of data to be transmitted, but also on the computational complexity of the tasks performed by the camera node. In fact, while for CTA the computational effort is limited to the compression (e.g., JPEG) of the acquired images, the complexity of the feature extraction algorithm at the base of ATC may be crucial. In the last few years, efficient algorithms for extracting compact binary features have been proposed [2, 3], which outperform, in terms of complexity, the traditional ones (e.g. SIFT [4], SURF [5]) at the cost of a slight worse matching accuracy.

With this demonstrator, we propose two main novelties with respect to the state-of-the-art solutions. First of all, we developed an optimized version of the Binary Robust Invariant Scalable Keypoints (BRISK [3]) detection/description algorithm, with particular focus on ARM architectures, which provide a specific SIMD instruction set (NEON). The optimization of the computation of the BRISK features using the NEON instructions results in an average speed-up of a factor 1.4 of the original BRISK implementation. This makes it possible to extract approximately 50 BRISK descriptors from VGA (640×480 pixels) images in less then $50\,\mathrm{ms}$ onto a $720\,\mathrm{MHz}$ ARM CPU. Furthermore, we implemented a lossless entropy-coding scheme for compressing the extracted BRISK features [6], which achieves a coding gain around 20% for 64-bits descriptors. As a consequence, 50 BRISK descriptors will generate a bitstream constituted of approximately 2.5 kbits. As a matter of comparison, a poor JPEG compression (quality factor Q=20) of a VGA image produced a bitstream of about 10 kbits.

3 Implementation and use

With reference to Figure 2, the demo is composed of 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.
- **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.

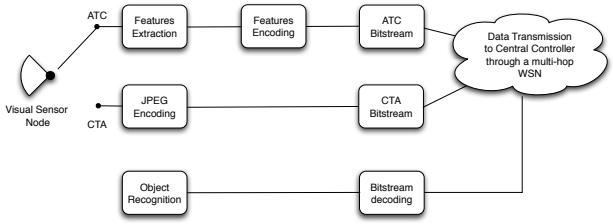


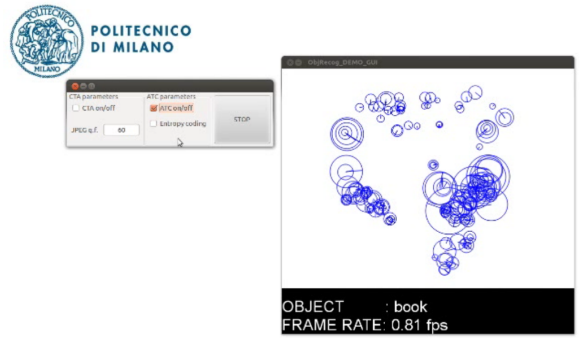
Figure 2. Visual sensor node operational modes. Switching between ATC and CTA is remotely controlled by the user. The compressed multimedia data from the ATC or CTA paradigm is sent to a central controller where it is decoded and used to perform object recognition.

The visual sensor node The camera (visual) sensor node is able to run both visual paradigms. Namely, as far as the CTA paradigm is concerned, the visual sensor node implements the standard JPEG compression algorithm. In the ATC case, the optimized BRISK algorithm discussed in Section 2 is implemented and used to extract local visual features. Moreover, the sensor node offers the possibility to encode the extracted features following the approach mentioned in Section 2.

The central controller The central controller implements a graphical user interface which allows to visualize the result of object recognition, and provides a highly interactive remote controller of the visual sensor node. The user

can switch on the fly between the two operating paradigms (CTA and ATC). 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

(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.

select different detection thresholds and the maximum number of features to be trasnmitted. Moreover, the user interface provides a switch for enabling/disabling the entropy-coding of the descriptors. 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. 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 two paradigms. As an additional feature, when the ATC modality 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 [7]: 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.

4. Conclusions and future developments

The proposed demo demonstrates 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. In the next months, we will investigate the possibility of extending this work including a set of cooperating processors nodes (i.e., multiple ARM BeagleBones), in order to distribute the computational load needed for extracting the visual features and thus obtaining a further increase of the frame rate.

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

References

[1] S. Paniga, L. Borsani, A. Redondi, M. Tagliasacchi, and M. Cesana, "Experimental evaluation of a video streaming system for wireless multimedia sensor networks," in *Med-Hoc-Net*, 2011, pp. 165–170.

[2] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011, pp. 2564–2571.

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

[4] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999, pp. 1150–1157.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[6] 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*, 2013, pp. 900–903.

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