Performance Evaluation of Object Recognition Tasks in Visual Sensor Networks

Luca Baroffio, Matteo Cesana, Alessandro Redondi, Marco Tagliasacchi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Piazza Leonardo da Vinci, 32 - 20133 Milano - Italy
Email: {name.surname}@polimi.it

***Abstract*—We analyze the performance of TCP and UDP transport protocols when applied to image retrieval or object recognition in wireless visual sensor networks (VSN). We focus on two different paradigms for image analysis, namely compress-then-analyze (CTA) and analyze-then-compress (ATC). The former entails the transmission of JPEG encoded images from camera nodes to a server, where the analysis takes place. The latter consists in extracting and compressing local visual features on board camera nodes, before transmission to a remote location. The presented analysis is useful to assess the best coupling between application and transport layers under delay and accuracy constraints for different networking conditions.**

I. Introduction

Visual Sensor Networks (VSNs) are composed of multimedia sensor nodes equipped with inexpensive camera hardware to acquire images or sequences of images, which may be further processed to perform digitalization and compression. The acquired visual information is then delivered to a final destination through low-power wireless transmissions. Multiple visual or generic sensor nodes may be involved in the process of routing the 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 [1], [2], and are expected to play a major role in the evolution of the Internet-of-Things (IoT) paradigm [3] as they can support a broad range of complex visual analysis tasks, such as object recognition, event detection, localization and tracking, etc.

We focus on a fundamental task in computer vision: the detection and recognition of objects. In this task, multimedia sensors (cameras) acquire images which are then processed to extract global or local distinctive features; the features extracted from an image are then matched against a database of pre-stored 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. Broadly speaking, two distinct paradigms can be categorized: *compress-then-analyze* (CTA) and *analyze-then-compress* (ATC) [4], [5].

In CTA, the acquired images are locally compressed at the multimedia sensor using standard techniques like JPEG and delivered through the wireless sensor network to a central controller which performs object recognition. The bitstream flowing in the network thus includes the compressed version of a pixel-representation of the acquired image. Image analysis (object recognition) performed at the central controller is thus based on a compressed and lossy representation of the original image, which might significantly impair the recognition accuracy [6]–[8].

Alternatively, in ATC image features are extracted locally at the multimedia sensor, processed (compressed), and then delivered to the final destination(s) in order to enable higher level visual analysis tasks. The key tenet is that most visual analysis tasks 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.

The performance of the object recognition task under the two aforementioned paradigms obviously depends on the wireless communication network used to deliver the information remotely and the combination of communication protocols used at different layers to support such delivery. In this work, we evaluate the impact of different combinations of communication protocols on the performance of object recognition under the ATC and CTA paradigms when the communication network provides a lossy communication channel. The relevant metrics used to assess the performance include the overall accuracy of the object recognition task, and the time-to-recognition which influences the maximum frame rate which can be supported. Protocol-wise, the reference lower transmission layers are based on low-power WiFi at the physical layer, and standard IEEE 802.11 Distributed Coordination Function at the MAC layer. At the transport layer, the performance when using both UDP and TCP is compared. The evaluation is carried experimentally over a testbed which implements the full pipeline for the ATC and CTA paradigms and allows to determine the *best* combination of visual paradigm (CTA or ATC) and communication protocol (UDP or TCP) in specific network conditions.

The paper is organized as follows: Section II reviews related work in the field and further highlights the main novel contribution of the present work. In Section III the main building blocks to implement visual analysis tasks for object recognition are described. The reference scenario and the related testbed implementation are thoroughly discussed in Section IV. Performance evaluation is presented and commented in Section V, whereas Section VI conveys our concluding remarks and future research directions.

978-0-9836283-9-2

II. Related Work

The matter of designing efficient wireless networks for supporting multimedia transmission has been largely debated in the literature. The work in the field can be broadly classified depending on the specific multimedia information that flows in the wireless network. A broad span of the literature in the field focuses on video delivery. In [9] the authors propose an optimization framework to maximize Peak Signal to Noise Ratio (PSNR) in cooperative wireless networks; namely, centralized and distributed PSNR-optimal strategies are proposed to jointly control the video encoding rate, the selection of relaying nodes and the allocated power level to perform wireless transmissions. See [10] for survey on the topic.

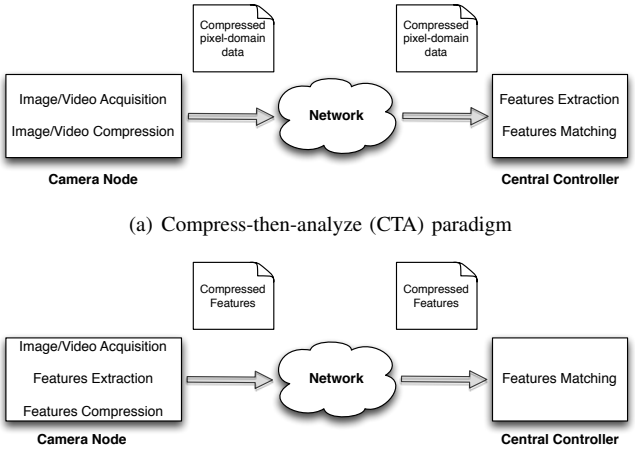
Specific work for video delivery in VSNs is carried out in [11] where the authors introduce an optimization framework to jointly optimize the coding rate and the routes in wireless sensor networks where correlated visual sensors operate under distributed source coding. A similar contribution and networking scenario is considered in [12] and [13], in which power control is also considered in the optimization problem formulation. In [14], an adaptive retransmission scheme for MPEG-encoded video is proposed in the field of Bluetooth-based sensor networks.

A good deal of work is available also on improving the quality of still image transmission by playing with network-related or source-related parameters. Different measures for the quality of the image transfer process are discussed in [15].

In [16], Wu *et al.* propose a method to improve the perceived PSNR of images transmitted in wireless sensor networks by playing with two ingredients: path diversity and Forward Error Correction (FEC) codes. The proposed scheme is evaluated for wavelet-based image compression. A similar reference scenario and solution concept is considered also in [17]. Always in this field, Lecuire *et al.* propose in [18] a in-network packet discarding technique to reduce the energy consumption of the image transmission process, while still matching quality requirements on the perceived PSNR. The main idea is to adapt the discarding rule for packets on the relevance of the visual content they are transporting. Again, image are compressed using wavelets.

An architecture for supporting reliable image transmission over wireless sensor network is presented in [19] and successively refined in [20]. The architecture includes optimized algorithms for wavelet compression on the camera nodes implemented on FPGA and optimized communication protocols to enforce reliability in the information delivery. Costa *et al.* evaluate in [21] the impact of packet segmentation when transferring DWT-based still images over lossy wireless channels. The evaluation framework only accounts for energy consumption neglecting the achieved quality of the image transfer process.

In the field of image retrieval, there has been recent work addressing the problem of feature extraction from lossy representation of images [22]. Chao *et al.* in [23] target the optimization of JPEG compression standard to improve feature detection performance when applying state-of-the art scale-based detectors.



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

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

Fig. 1: The two different approaches to enable image analysis in visual sensor networks

Common characteristics of the aforementioned work are that (i) it focuses on generic video or still-image transmission, and/or (ii) it proposes modifications or new protocols/solutions either at the multimedia source (new source coding algorithms) or in the communication networks (MAC, routing, etc.). Differently, we focus on the specific visual task of automatic image retrieval and object recognition based on feature extraction algorithms, and we evaluate the impact of standard communication paradigms/configurations on the performance of two different paradigms to implement the reference visual task. Specifically, to the best of our knowledge, this is the first time that object recognition is analyzed from the point of view of an unreliable communication channel.

III. Background on Object Recognition

The reference working mode for the two paradigms is illustrated in Figure 1. Regardless of the specific paradigm, object recognition builds on the concept of local visual feature, which can be qualitatively defined as a concise, yet efficient, representation of the underlying pixel domain content of an image. Extracting visual features from an image is accomplished with the aid of two main components: the detector, which identifies salient keypoints in the image like corners or blob-like structures, and the descriptor which takes as input the detected keypoint and a patch of pixels in its surroundings and returns a compact representation of the keypoint itself.

In this section, we provide a brief overview of the two algorithms used in this paper for detecting/describing keypoints, namely SIFT [24] for the CTA paradigm, and BRISK [25] for ATC.

A. SIFT in CTA

In the compress-then-analyze case, the query images acquired by camera nodes are first compressed with JPEG at different rates by varying the JPEG Quality Factor QF from 1 to 100. The images are then sent to a central controller, which extracts local features from (compressed) query images and matches them against the features extracted from (uncompressed) images in the reference database. SIFT features are extracted from compressed query images at the central controller. SIFT features are considered as the gold standard in visual analysis, as they typically achieve state-of-the-art performance in most applications.

The SIFT detector is composed by three main stages, namely (i) scale-space extrema detection, (ii) keypoint localization and (iii) orientation assignment. In the first step, keypoints are detected at all locations and scales in the image, being each scale a resized and smoothed version of the original image. Detecting keypoints at different scales allows the visual task to be robust to scale transformation of the reference image. Then, at each candidate location, a detailed model is fit to determine accurately the location and scale of each keypoint. Finally, an orientation is assigned to the keypoint, based on local image gradient direction.

The detection step provides each keypoint with a location, scale and orientation. These parameters define a local 2D coordinate system in which to describe the local patch, therefore achieving invariance to these image transformations. The number of keypoints provided as output by the detector obviously depends on the visual content of the reference image. Now, a descriptor has to be computed for the local image patch in a highly distinctive way, and possibly with invariance with respect to other transformations, such as illumination changes. First, image gradient magnitudes and orientations are sampled around the keypoint: all the coordinates are rotated using the keypoint orientation to achieve rotation invariance. A Gaussian weighting function is used to weigh the magnitude of each sample point, in order to avoid sudden changes in the descriptor with small changes in the position of the window. Then, the keypoint descriptor is built by creating orientation histogram over 4x4 sample regions. The descriptor is formed from a vector containing the values of all the orientation histograms entries. The standard SIFT descriptor is built from a 4x4 array of histograms with 8 orientation bins in each, thus the descriptor is formed by a 4x4x8 = 128 element feature vector for each keypoint.

B. BRISK in ATC

In the analyze-then-compress (ATC) case, visual features are (i) extracted from uncompressed query images directly at the camera node, (ii) sent to the central controller where (iii) they are matched with the features extracted from the reference database of images. In this work, the camera node runs the Binary Robust Invariant Scalable Keypoints (BRISK) detector and descriptor. The choice of BRISK is motivated by the fact that it constitutes a valid alternative to state-of-the-art methods such as SIFT, providing similar performance with significantly lower computational complexity, thus being particularly suitable for running on camera sensor nodes [26] [27].

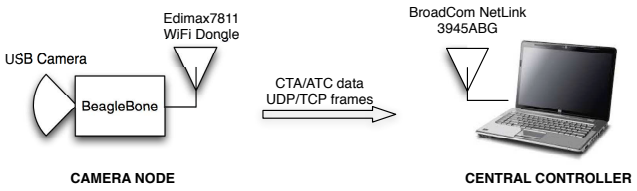


Fig. 2: Experimental setup

BRISK implements a multi-scale corner detector. A candidate point p of the image, (with intensity I_p) is classified as a corner if n contiguous pixels in the Bresenham circle of radius 3 around p are all brighter than I_p+t, or all darker than I_p-t, with t a predefined threshold. In BRISK, the detection phase is applied not only in the original image plane, but also in the scale-space; that is, the image is progressively half-sampled to originate octaves; intra-octaves are further created by downsampling the lower octave by a factor of 1.5 and keypoints are searched with the aforementioned procedures in all the octaves and intra-octaves. Upon keypoint identification, BRISK assigns a *binary descriptor* to each keypoint, which is a string of bits obtained from pairwise comparisons between smoothed intensities of pixels within the region of interest of such a keypoint. The final dimension of BRISK descriptor is 512 bits. The dimension of the BRISK descriptor is further reduced by resorting to state-of-the art compression techniques. In this work, we leverage the compression approach proposed in [28], which allows to scale down the descriptor dimension to 360 bits on average.

IV. Experimental Framework

The pipeline for object recognition in Figure 1 is realized with the following hardware/software building blocks; the camera node is a BeagleBone platform running Linux (Ubuntu 12.04) which is geared with a WiFi Edimax 7811Un, set to operate according to the IEEE 802.11g standard extension with a nominal bit rate of 54 Mbps. The central controller is a PC based on an Intel Core Duo CPU (2.40 GHz) CPU, with 4 GB of RAM and geared with a Broadcom Netlink 3945ABG WiFi Network Interface card (NIC). Both the Edimax dongle on the Beagleboard and the NIC at the central controller are configured to operate in *ad hoc mode*. Static IP addresses are assigned to both WiFi interfaces. To provide vision capabilities to the camera node, a Logitech C170 USB camera is attached to the BeagleBone. An illustration of the testbed is provided in Figure 2.

The BeagleBone camera node acquires an image and runs OpenCV 2.4.6 libraries to implement JPEG compression for the CTA paradigm, and the BRISK detector/descriptor algorithms provided by the authors[1] for the ATC case. The bitstream generated out of JPEG compression (CTA) or BRISK

[1] http://www.asl.ethz.ch/people/lestefan/personal/BRISK

detector/descriptor phases (ATC) is then encapsulated into UDP or TCP segments which are finally delivered to the central controller. The Network Emulator (NetEm)[2] software is used to emulate uncorrelated channel losses over the wireless channel, by discarding a specific percentage of IP frames at the transmitter.

In the case of ATC, TCP/UDP segments are decoded in order to reconstruct the encoded BRISK local features. Those are then matched against the ones of each image in the central database. In the case of CTA, SIFT local features are extracted from the received JPEG-compressed image. Such features are then matched against the local database. In this last case, the OpenCV v.2.4.6 implementation of SIFT is used. The matching process is based on Euclidean distance between SIFT features and Hamming distance between BRISK features in the case of CTA and ATC, respectively. Furthermore, matches are refined resorting to the ratio test algorithm and RANSAC is applied to ensure geometric consistency.

The Mean of Average Precision (MAP) measure is used to assess the quality of object recognition. Given an image query q, the Average Precision (AP) is defined as:

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

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

$$MAP = \frac{\sum_{q=1}^{Q} AP_q}{Q} \quad (2)$$

The performance evaluation is carried out using the Zurich Building Database (ZuBUD)[3] which contains 1005 color images of 201 buildings of the city of Zurich. Each building has five VGA images (640x480), taken at random arbitrary view points under different seasons, weather conditions and by two different cameras. A separate archive containing 115 images (at a resolution of 320x240) of the same buildings (with different imaging conditions) is available as query dataset.

V. Performance Evaluation

This section reports the experimental performance evaluation of the two paradigms (ATC and CTA) under different conditions as far as the application and the network set up is concerned. From the communication network perspective, the degrees of freedom for comparison include the use of the lightweight but unreliable UDP against the reliable but "heavier" TCP over lossy wireless channels.

[2]http://www.linuxfoundation.org/collaborate/workgroups/networking/netem
[3]http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html

The evaluation is carried out having in mind two main performance measures: the accuracy of the object recognition task and the maximum frame rate of the object recognition task. The former parameter is measured using the Mean Average Precision (see Section IV); the latter parameter can be formally defined as the inverse of the visual task completion time:

$$t_c = t_{\text{acq}} + t_{\text{proc}} + t_{\text{tx}} + t_{\text{matching}},$$

where t_{acq} is the time needed to acquire one image, t_{proc} is the processing time (JPEG compression in CTA, feature extraction in ATC), t_{tx} is the transmission time to deliver the required bitstream to the sink (JPEG images in CTA and BRISK descriptors in ATC), and t_{matching} is the time required to match the features against the reference database. The acquisition time is fixed and thus equal in the two used paradigms. The delay component related to the matching procedure is obviously comparable in the CTA and ATC cases; moreover, as the matching is implemented at the central controller, which generally features powerful processing hardware, the t_{matching} component can be assumed to be small with respect to the other components. Referring to the processing component to the delay, t_{proc}, recent studies show that the processing time for performing feature extraction and description (ATC) on embedded hardware is comparable to the processing time required to run JPEG compression [29], [30]. Building on these results and observations, in the following analysis we focus on the impact of the communication network onto the visual analysis paradigms, thus we analyze the t_{tx} component to the delay. In Sections V-A and V-B the focus is on CTA and ATC respectively, whereas the comparison between the two paradigms is discussed in V-C.

A. Performance under CTA

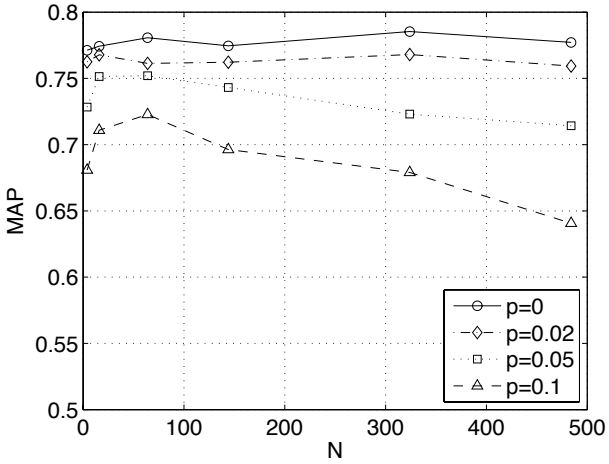
In case CTA is used in combination with UDP over lossy channels, the loss of one single packet out of the packet stream coming from the compressed JPEG image makes impossible the decoding of the entire image, which, in turn, totally impairs the accuracy of the object recognition task. Thus, in order to use the lightweight UDP to support CTA, we have considered the case where the original image is not encoded in one single JPEG block, but rather JPEG compression is performed over independent slices of pixels of the original image. The main rationale for this approach is to reduce the impact of packet losses over the accuracy of the object recognition task. Defining N the number of slices the original image is divided in, one would expect that the accuracy of the object recognition task increases with N for a given value of JPEG Quality Factor (QF) and packet error rate (p).

Figure 3(a) shows the behavior of MAP when varying the parameter N when QF=30 and UDP is used. If one slice is used, MAP goes to zero when the channel is lossy. All the curves show a global maximum which is generated by the trade-off between two distinct effects: for low values of N, the accuracy is first increasing in N; in this regime, even if some slices get lost due to packet errors on the channel,

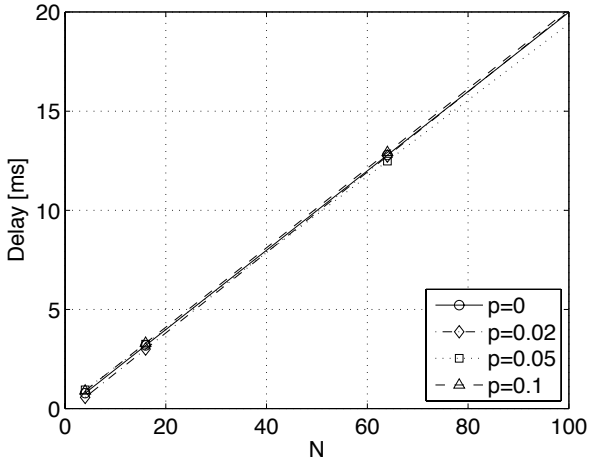
still the information collected at the sink allows to perform the object recognition task with reasonably high accuracy. As N keeps increasing, the original image is sliced and encoded in smaller and smaller pieces; as packet losses on the channel turn into slice losses at the sink, the reconstructed image is characterized by several missing patches (see Figure 5). The absence of such patches impairs the quality of the keypoint detection algorithm, which is cheated in detecting artificial keypoints around each missing patch. This in turn, impairs the image matching procedure and the accuracy of the object recognition task. To countermeasure this artifact, two different techniques are also tested: in the first one, error concealment techniques (e.g., based on inpainting) are applied on the received image to interpolate the missing pixels; in the second one, the detector is disabled in the surrounding area of the black patches, thus preventing the detection of unwanted keypoints. In this last case, the keypoints detected along the perimeter to the black patches of the reconstructed image are not considered in the matching process. Figure 4 shows the very same results of Fig. 3(a) when these two techniques are used for a specific packet error rate and QF ($p = 0.1$ and $QF = 30$). The curve labelled as *Plain* refers to the original scheme, whereas the two curves labeled as *EC* and *Skip* refer respectively to the case in which error concealment and keypoint skipping techniques are applied to the received image. As clear form the figure, even if the adoption of these technique smoothes down the accuracy loss when N increases, still the MAP versus N curves show a global maximum in N. Very similar results have been obtained for different combinations of QF and p parameters.

It is also worth analyzing the impact of the slicing procedure on the latency of the object recognition task. Figure 3(b) shows the transmission delay as a function of parameter N. As expected, the delay increases almost linearly with N; the reason being that the slicing procedure reduces the overall coding efficiency of JPEG, that is, the produced bitstream is larger when the original image is first sliced and then encoded with respect to the case in which one single slice corresponding to the original image is used. A larger encoded bitstream also leads to larger packet segmentation overhead at the lower (IP and MAC) layers.

Figure 6 summarizes the MAP/delay performance for the CTA paradigm when UDP and TCP are used. The curves referring to UDP are obtained by using the value of N which maximizes the MAP at different PER (see Figure 3(a)) and the error concealment technique based on inpainting, whereas when TCP is used, the slicing is not applied as TCP is inherently reliable. In both cases (UDP and TCP) the MAP first increases when increasing the QF of JPEG compression (which is proportional to the transmission delay); then, under TCP the MAP levels off to a value close to 80%, whereas when using UDP on lossy channels the MAP/delay function has a monotonic non-increasing behavior for large QF. This is due to the fact that as the QF increases, the size of the bitstream to be transmitted also increases and so does the required number of UDP segments and IP packets to deliver one single slice of



(a) MAP vs N



(b) Delay vs N

Fig. 3: Impact of the number of slices N over MAP and delay; JPEG compressed image with QF=30, UDP transport protocols under different packet error rate values (p)

the JPEG image. Thus, for a given packet error rate, as the size of the bitstream increases, also the slice error rate increases. In other words, bigger slices are more likely to get lost on the channel for a given PER. This is why the MAP/delay curves tend to bend downwards when UDP is used. Finally, the comparison between the two transport paradigms clearly shows that TCP has superior performance than UDP when implementing object recognition applications under the CTA paradigm.

B. Performance under ATC

Figures 7-9 summarize the performance evaluation of the ATC paradigm. Figure 7 analyzes the impact of the number of keypoints extracted from the images in case of ideal channel and UDP transport layer. Qualitatively, increasing the number of keypoints per image, that is, decreasing the BRISK threshold parameter t, increases on one side the accuracy of

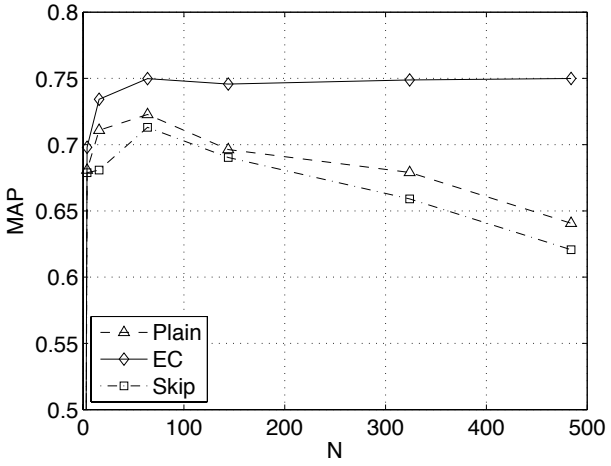
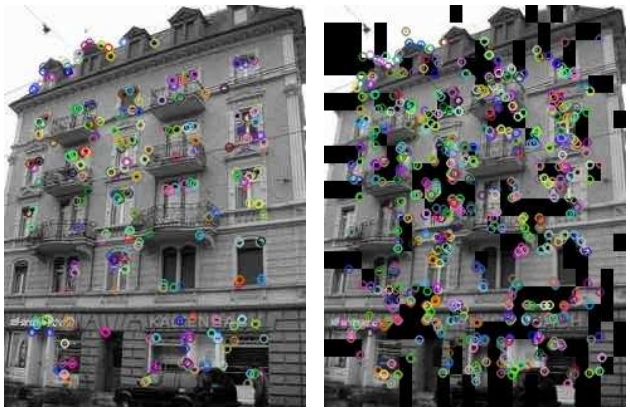


Fig. 4: Impact of the number of slices N over MAP; JPEG compressed image with $QF = 30$, UDP transport protocol under different techniques to cope with errors in the received image. Packet error rate $p = 0.1$. *Plain* refers to the original scheme, *EC* refers to the case in which error concealment techniques are applied to the received image, and *Skip* refers to the case in which the keypoints detected along the perimeter to the black patches of the reconstructed image are not considered in the matching process.



(a) N=1 (b) N=64

Fig. 5: Artificial keypoints generated by image slicing procedure under lossy channels (p=0.05); JPEG compressed image with $QF = 30$, UDP transport protocol.

the visual task (Fig. 7(a)), but on the other side also increases the transmission delay (Fig. 7(b)). A similar behavior can be observed also over lossy channels and/or when using TCP in place of UDP.

Figure 8 shows the impact of the number of keypoints which are encapsulated into one UDP segment. Increasing the number of keypoints per UDP/TCP segment reduces the transmission overhead (and thus the transmission delay), but on the other side may lower the accuracy over lossy channel for a given packet error rate p. The loss of one single packet

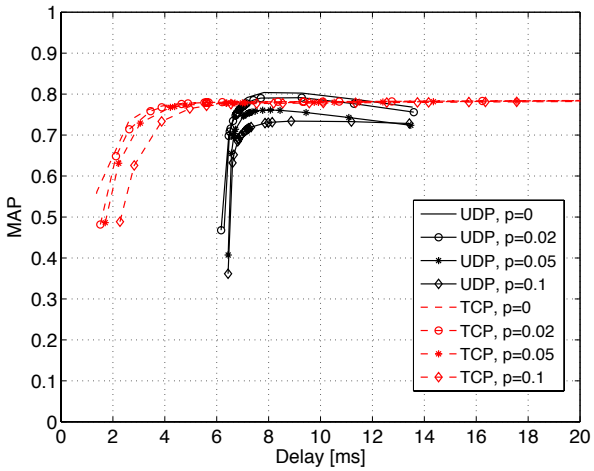


Fig. 6: MAP vs Delay curves for the CTA paradigm under different channel conditions (packet error rate, p) in case when TCP and UDP are used at the transport layer.

leads to the loss of an increasing number of keypoints.

Figure 9 finally shows the MAP-delay curves when ATC is used on top of UDP and TCP for different values of packet error rate. Expectedly, a higher packet error rate leads to lower accuracy for a given delay. In general, UDP shows superior performance than TCP when coupled with the ATC paradigm.

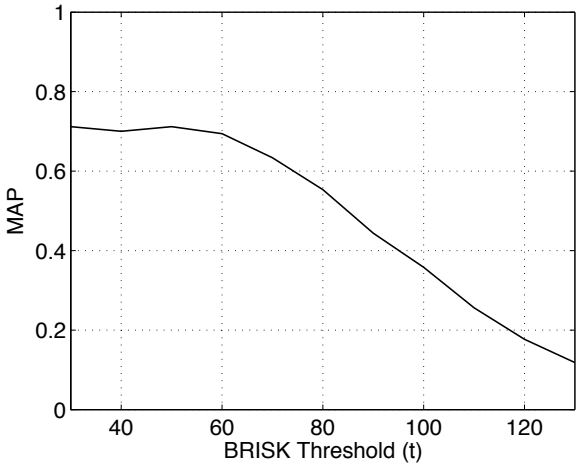
C. CTA vs ATC comparison and discussion

Figure 10 summarizes the performance comparison between ATC and CTA paradigms for different values of packet error rate, p. One first observation from the figures is that CTA always provides the highest maximum accuracy; in case of ideal channel ($p = 0$) the accuracy loss of ATC is around 10%. This is due to the fact that CTA leverages SIFT as detector/descriptor algorithm which is known the provide better performance than BRISK.

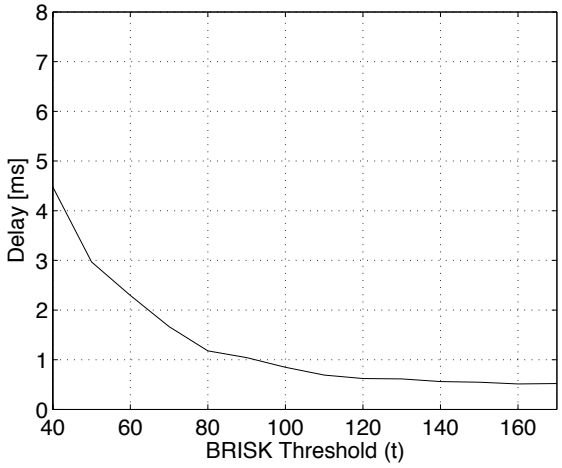
Conversely, when the required delay is low, the only possibility is to use ATC, as the dimension of the bitstream generated by CTA is lower bounded by the JPEG encoding process. The best combination of visual paradigm/transport protocol is given by the upper envelope of the curves reported in Figure 10; that is, there exist a reference delay value below which the best MAP/delay performance is achieved using ATC/UDP and above which the best paradigm/protocol combination is CTA/TCP.

VI. Conclusions

In this work, we considered two alternative paradigms to implement visual task of object recognition in wireless sensor networks; in the Compress-Then-Analyze paradigm, images are collected at the camera sensor node, compressed through standard techniques as JPEG and sent remotely to a sink where the visual processing task is carried out. Conversely, in the Analyze-Then-Compress paradigm, part of the visual

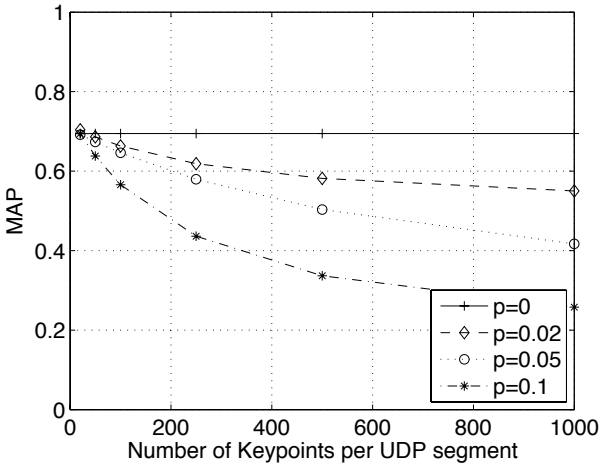


(a) Impact on MAP

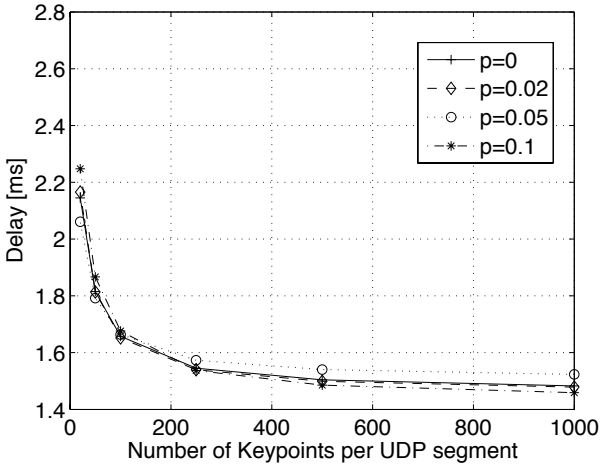


(b) Impact on Delay

Fig. 7: Impact of the number of descriptors collected and sent out under ATC paradigm; UDP transport protocols under ideal channel condition ($p = 0$)



(a) Impact on MAP



(b) Impact on Delay

Fig. 8: Impact of the number of keypoints carried by one UDP segment under ATC paradigm for different channel conditions.

task is accomplished directly at the camera node, and only more succinct information is delivered for further analysis at the sink node.

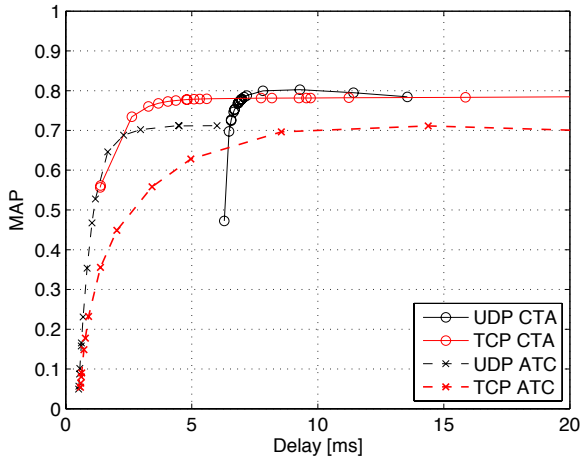
We have implemented a testbed to evaluate the performance of the object recognition visual task when adopting the two aforementioned paradigms in terms of recognition accuracy and recognition delay; the interplay has been further studied between each visual analysis paradigm and the specific transport layer protocol used to deliver the required information from the camera node to the sink. The main results coming from this analysis can be summarized as follows:

1) the accuracy of the visual task under CTA is extremely sensitive to wireless channel impairments, thus a reliable transport protocol is needed (TCP-like);
2) conversely, the ATC paradigm is robust to wireless channel impairments, thus a lightweight, responsive, best effort transport layer protocol is preferable;
3) overall, when the reference visual task requires high frame rate, ATC+UDP is the best combination, whereas if the visual task has loose constraints on the required frame rate, the combination CTA+TCP is the one which guarantees higher accuracy.

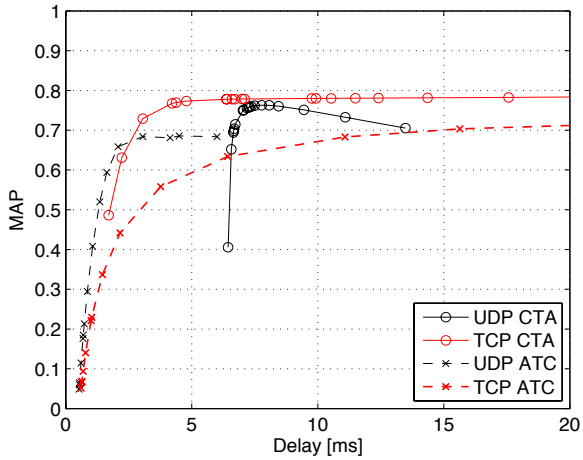
As a future work, we plan to investigate on the energy performance of the considered protocols/paradigms combinations.

ACKNOWLEDGEMENTS

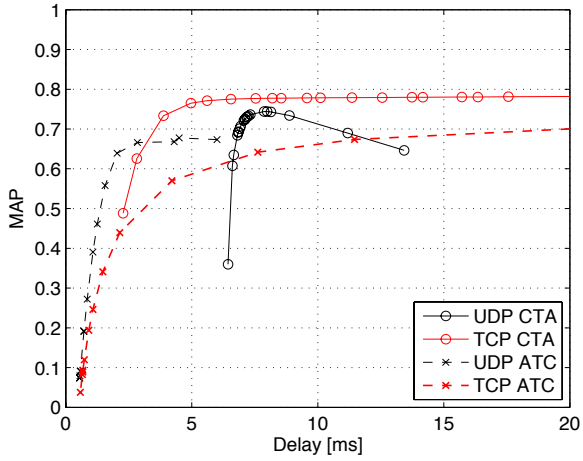
This work has been carried out within project GreenEyes with the support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 296676.



(a) p=0



(b) p=0.05



(c) p=0.1

Fig. 10: Comparison of the best MAP vs Delay curves for the two paradigms ATC and CTA.

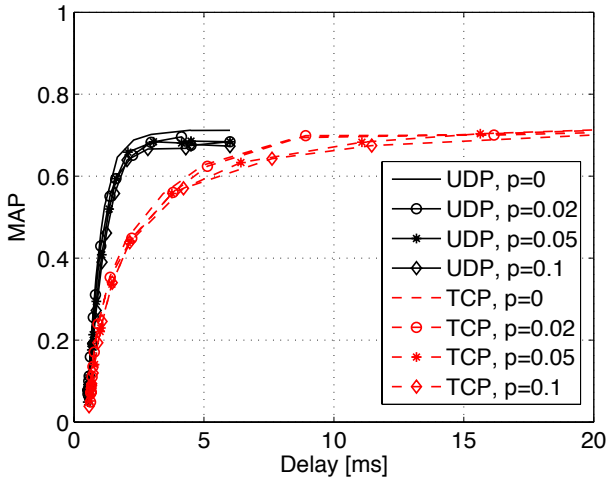


Fig. 9: MAP vs Delay curves for the ATC paradigm under different channel conditions (packet error rate, p) in case TCP and UDP are used at the transport layer.

REFERENCES

[1] S. Soro and W. Heinzelman, "A Survey of Visual Sensor Networks," *Advances in Multimedia*, vol. 2009, pp. 1–22, 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/640386

[2] T. Melodia and I. Akyildiz, "Research challenges for wireless multimedia sensor networks," in *Distributed Video Sensor Networks*, B. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, and D. Terzopoulos, Eds. Springer London, 2011, pp. 233–246. [Online]. Available: http://dx.doi.org/10.1007/978-0-85729-127-1_16

[3] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787 – 2805, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128610001568

[4] A. Redondi, L. Baroffio, M. Cesana, and M. Tagliasacchi, "Compress-then-analyze vs. analyse-then-compress: Two paradigms for image analysis in visual sensor networks," in *IEEE International Workshop on Multimedia Signal Processing*, 2013.

[5] A. Canclini, L. Baroffio, M. Cesana, A. Redondi, and M. Tagliasacchi, "Comparison of two paradigms for image analysis in visual sensor networks," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '13. New York, NY, USA: ACM, 2013, pp. 62:1–62:2. [Online]. Available: http://doi.acm.org/10.1145/2517351.2517378

[6] A. Zabala and X. Pons, "Impact of lossy compression on mapping crop areas from remote sensing," *International Journal of Remote Sensing*, vol. 34, no. 8, pp. 2796–2813, 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01431161.2012.750772

[7] A. D. Bagdanov, M. Bertini, A. D. Bimbo, and L. Seidenari, "Adaptive video compression for video surveillance applications," in *ISM*, 2011, pp. 190–197.

[8] G. Gualdi, A. Prati, and R. Cucchiara, "Video streaming for mobile video surveillance," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1142–1154, 2008.

[9] Z. Guan, T. Melodia, and D. Yuan, "Jointly Optimal Rate Control and Relay Selection for Cooperative Wireless Video Streaming," *IEEE/ACM Transactions on Networking*, vol. 21, no. 4, pp. 1173–1186, August 2013.

[10] Y. He and L. Guan, "Optimal resource allocation for video communication over distributed systems," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 1414–1423.

[11] J. Zou, H. Xiong, C. Li, R. Zhang, and Z. He, "Lifetime and distortion optimization with joint source/channel rate adaptation and network coding-based error control in wireless video sensor networks," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 3, pp. 1182–1194, 2011.

[12] B. Peng, J. Zou, C. Tan, and M. Wang, "Network lifetime optimization in wireless video sensor networks," in *Wireless Mobile and Computing*

(CCWMC 2009), IET International Communication Conference on, 2009, pp. 172–175.

[13] Y. Chen, X. Hu, H. Yang, and L. Ge, "Power control routing algorithm for maximizing lifetime in wireless sensor networks," in *Advances in Mechanical and Electronic Engineering*, ser. Lecture Notes in Electrical Engineering, D. Jin and S. Lin, Eds. Springer Berlin Heidelberg, 2013, vol. 178, pp. 129–136. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31528-2_22

[14] R. Kapoor, M. Cesana, and M. Gerla, "Link layer support for streaming mpeg video over wireless links," in *Computer Communications and Networks, 2003. ICCCN 2003. Proceedings. The 12th International Conference on*, 2003, pp. 477–482.

[15] P. Sarisaray-Boluk, "Performance comparisons of the image quality evaluation techniques in wireless multimedia sensor networks," *Wirel. Netw.*, vol. 19, no. 4, pp. 443–460, May 2013. [Online]. Available: http://dx.doi.org/10.1007/s11276-012-0477-5

[16] H. Wu and A. A. Abouzeid, "Error resilient image transport in wireless sensor networks," *Computer Networks*, vol. 50, no. 15, pp. 2873 – 2887, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128605003877

[17] P. Boluk, S. Baydere, and A. Harmanci, "Robust image transmission over wireless sensor networks," *Mobile Networks and Applications*, vol. 16, no. 2, pp. 149–170, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11036-010-0282-2

[18] V. Lecuire, C. Duran-Faundez, and N. Krommenacker, "Energy-efficient transmission of wavelet-based images in wireless sensor networks," *EURASIP Journal on Image and Video Processing*, vol. 2007, no. 1, p. 047345, 2007. [Online]. Available: http://jivp.eurasipjournals.com/content/2007/1/047345

[19] S. Aziz and D. M. Pham, "Energy efficient image transmission in wireless multimedia sensor networks," *Communications Letters, IEEE*, vol. 17, no. 6, pp. 1084–1087, 2013.

[20] D. M. Pham and S. M. Aziz, "Object extraction scheme and protocol for energy efficient image communication over wireless sensor networks," *Computer Networks*, vol. 57, no. 15, pp. 2949 – 2960, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128613002144

[21] D. Costa, L. Guedes, F. Vasques, and P. Portugal, "Effect of frame size on energy consumption in wireless image sensor networks," in *Imaging Systems and Techniques (IST), 2012 IEEE International Conference on*, 2012, pp. 239–244.

[22] J. Chao, A. Al-Nuaimi, G. Schroth, and E. Steinbach, "Performance comparison of various feature detector-descriptor combinations for content-based image retrieval with JPEG-encoded query images," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Pula, Sardinia, Italy, Oct 2013.

[23] J. Chao, C. Hu, and E. Steinbach, "On the design of a novel jpeg quantization table for improved feature detection performance," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2013.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, Nov. 2004.

[25] 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.

[26] J. Heinly, E. Dunn, and J.-M. Frahm, "Comparative evaluation of binary features," in *Computer Vision ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, pp. 759–773. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33709-3_54

[27] A. Canclini, R. Cilla, A. Redondi, J. Ascenso, M. Cesana, and M. Tagliasacchi, "Evaluation of visual feature detectors and descriptors for low-complexity devices," in *Digital Signal Processing Conference*, 2013.

[28] 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.

[29] M. L. Kaddachi, A. Soudani, V. Lecuire, L. Makkaoui, J.-M. Moureaux, and K. Torki, "Design and performance analysis of a zonal dct-based image encoder for wireless camera sensor networks," *Microelectronics Journal*, vol. 43, no. 11, pp. 809 – 817, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026269212001449

[30] J.-S. Park, H.-E. Kim, and L.-S. Kim, "A 182 mw 94.3 f/s in full hd pattern-matching based image recognition accelerator for an embedded vision system in 0.13- μm cmos technology," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 5, pp. 832–845, 2013.