IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An entropy evaluation algorithm to improve transmission efficiency of compressed data in pervasive healthcare mobile sensor networks

GIACOMO CAPIZZI[1,2], SALVATORE COCO[2], GRAZIA LO SCIUTO[2], CHRISTIAN NAPOLI[3], WALDEMAR HOŁUBOWSKI[1]

[1]Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska, 23, 44-100 Gliwice, Poland (e-mail: w.holubowski@polsl.pl)
[2]Dpt. of Electrical, Electronics and Informatics Engineering University of Catania, Italy (e-mail: gcapizzi@diees.unict.it, glosciuto@dii.unict.it)
[3]Department of Computer, Control, and Management Engineering Sapienza University of Rome, Italy

Corresponding author: Giacomo Capizzi (e-mail: e-mail: gcapizzi@diees.unict.it).

ABSTRACT Data transmission is the most critical operation for mobile sensors networks in term of energy waste. Particularly in pervasive healthcare sensors network it is paramount to preserve the quality of service also by means of energy saving policies. Communication and data transmission are among the most critical operation for such devises in term of energy waste. In this paper we present a novel approach to increase battery life-span by means of shorter transmission due to data compression. On the other hand, since this latter operation has a non-neglectable energy cost, we developed a compression efficiency estimator based on the evaluation of the absolute and relative entropy. Such algorithm provides us with a fast mean for the evaluation of data compressibility. Since mobile wireless sensor networks are prone to battery discharge-related problems, such an evaluation can be used to improve the electrical efficiency of data communication. In facts the developed technique, due to its independence from the string or file length, is extremely robust both for small and big data files, as well as to evaluate whether or not to compress data before transmission. Since the proposed solution provides a quantitative analysis of the source's entropy and the related statistics, it has been implemented as a preprocessing step before transmission. A dynamic threshold defines whether or not to invoke a compression subroutine. Such a subroutine should be expected to greatly reduce the transmission length. On the other hand a data compression algorithm should be used only when the energy gain of the reduced transmission time is presumably greater than the energy used to run the compression software. In this paper we developed an automatic evaluation system in order to optimize the data transmission in mobile sensor networks, by compressing data only when this action is presumed to be energetically efficient. We tested the proposed algorithm by using the Canterbury Corpus as well as standard pictorial data as benchmark test. The implemented system has been proven to be time-inexpensive with respect to a compression algorithm. Finally the computational complexity of the proposed approach is virtually neglectable with respect to the compression and transmission routines themselves.

INDEX TERMS wireless sensor networks; data compression; entropy; quality of service; energy saving; quality prediction; differential information entropy.

I. INTRODUCTION

THE micro-electro-mechanical systems (MEMS) technology has encountered a tremendous evolution in the last decades [1]–[3]. The reached integration level permits us to develop sensors embedding small computational devices with fully functional storage and communication capabilities. Such hardware systems are generally constructed in order to perform some measurements and transmit the collected data as digital signals. A multiplicity of sensors, deployed in a collaborative strategy for data gathering, is called sensors network. Moreover, if a sensor is mounted on a mobile device, it is possible to rearrange their position during time, or ran-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

domly disperse sensors and reposition them in a successive moment (e.g. due to temporary environmental limitations or hazards, or in surveillance operations, etc...) [4], [5].

In this latter fashion, a mobile wireless sensor network (MWSN) is a sensor network constituted by mobile nodes that communicates through a radio signal. A large number of MWSNs have been developed for pervasive healthcare systems [6]: some of them are devoted to continuous monitoring of elderlies, children, chronically ill or impaired people, as well as patients affected by cognitive disorders, such as Alzheimer syndrome; other kind of sensors networks are in development for healthcare oriented environmental monitoring, movement tracking, fall detection, live analysis of human body stats and physiological parameters, etc... Pervasive healthcare mobile sensor networks are capable to join different data coming from different sources gathering a more complete understanding of a diagnostic context, therefore such sensors networks provide for advanced monitoring solutions. Such solutions are extremely valuable due to their improved ability to recognize unusual patterns due to the more complete reference basis (e.g. in the case of body area networks, BAN, or personal area networks, PAN, etc...). Among the many BAN applications of MWSNs, uttermost importance has been gained by ECG-monitoring related solutions [7]–[9] on the other hand for such applications it is paramount to work under guaranteed quality of services conditions also in terms of system autonomy and battery life-cycle [10]. In facts, while remote control and monitoring is one of the main advantages of MWSN healthcare systems, energy efficient sensors are often critical [11], [12]. On the other hand communication interfaces such as wifi and bluetooth, while mandatory components of communicating networks, fail to provide support for energy efficient systems [13]. Due to their nature such sensors must be powered by means of electrical batteries, on the other hand their operation cycle is limited due to the unavoidable power exhaustion during time. It follows that, while in a battery-powered sensor it is paramount to enforce every possible energy-saving policies, in MWSN the data transmission events constitute critical operations that tampers with battery life. In pervasive healthcare sensors networks, the amount of autonomy time between charging cycles makes the difference between a usable technology and a non feasible approach. I.e., while MWSN can be extremely useful in preventing cardiac pathologies or to enforce preemptive alert systems for health operators, it would be pointless to develop such a technology if the resulting device should be put offline, recharging, each few hours. Data compression constitutes a possible solution for energy efficient sensors' data transmission, on the other hand that preventive measure should be carefully evaluated. In facts, while compressed data requires a shorter communication time, and consequently reduces the amount of energy wasted in data transmission, the compression algorithm itself will require a certain amount of energy to be executed. Therefore, as possible trade-off, it would be agreeable to transmit compressed data only when such

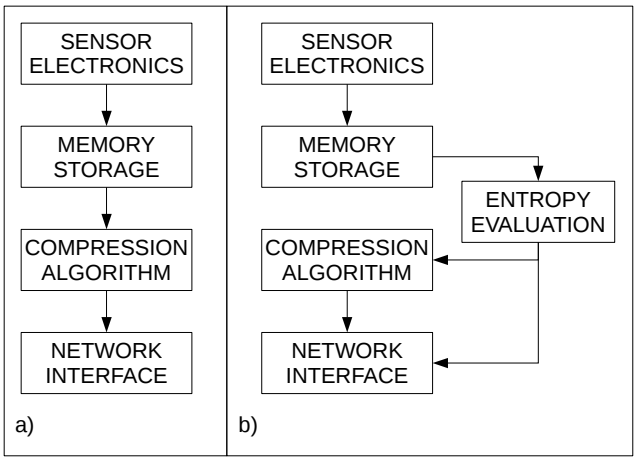

FIGURE 1. The interaction of the proposed entropy evaluation system with the other hardware and software components is shown in the right panel (b), with respect to the traditional interactions design shown in the left panel (a).

operation greatly reduces the transmission time. It follows that, for mobile senors networks communicating by means of wireless signals, data should be compressed only after a positive estimation of the compression efficiency of the data compression algorithm (see Figure 1).

The general problem is easy to state. Given many senders and receivers and a channel transition matrix that describes the effects of the interference and the noise in the network, decide whether or not the sources can be transmitted over the channel [14]. On the other hand a compression algorithm does not represent an optimal solution in all conditions, although there are many optimized software system opportunely designed to achieve the best performances for specific data formats (e.g. [15], [16].) While a large number of algorithms are devoted to data compression for specific applications (see Figure 2), the optimum is generally achieved only by few specific compression algorithm. On the other hand, such an optimality on regards the compressed data size with respect to its original size. Unfortunately, the design and development of an optimal compression system in terms of battery-savings on the field of MWSN would become a strongly data-dependent task and would require a significant effort, both on theoretical and practical side. Moreover the energy efficiency of a compression algorithm would strongly depend on the accuracy of the related model. Hence such a model should have to be meticulously calibrated basing on the structural and semantical topology of the data to compress. Moreover both the complexity of the algorithms and the overall computational effort are strongly influenced by the admissible error in the process. The main issues preliminarily examined when applying data compression are efficiency aspects such as the total compression ratio, and the computing resources (time and memory) required, especially for space communications; other important issues are sensitivity to errors and adaptability to different data types.

Data compression algorithms can be roughly classified

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

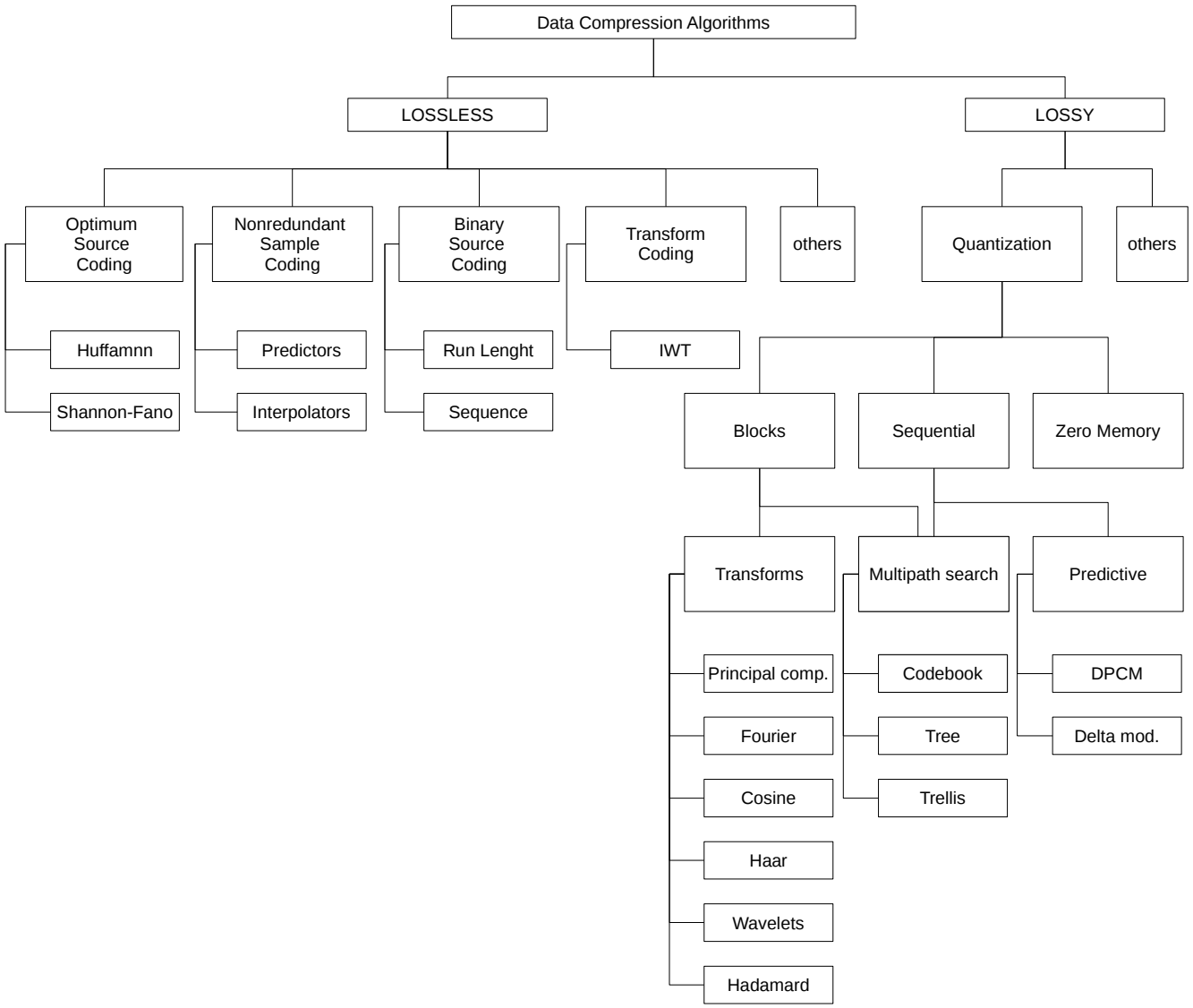

FIGURE 2. A non exhaustive classification of existing compression algorithms.

in two categories: lossy (or non invertible) or lossless (or invertible). While the algorithms that falls in the first of such categories are generally capable of a greater result in terms of compressed data size, this kind of algorithms are unable to fully reconstruct the original data, suffering therefore of an unavoidable information loss, hence they ends by reducing the informational entropy, then definitively neglecting an hopefully small portion of the original data. On the contrary lossless compression systems reduces the transmitted or stored data size by reducing information redundancy from the source, therefore preserving the informational entropy, and allowing the integral reconstruction of the original data.

While lossy compression is in generally suitable for a wide range of applications, on the field of sensors and sensor networks such data requires to be perfectly reconstructed, therefore, often, only lossless compression techniques are applicable. It follows that, in lossless compression, an a priori estimate of the source statistics is highly desirable since it allows us to estimate the maximum theoretically-achievable compression ratio. Such a knowledge becomes helpful to improve the energy efficiency of communicating sensors network. In facts, estimating the data compressibility it is also possible to decide whether or not that procedure would be convenient (e.g. a low compression ratio would not reduce the communication time enough to justify the amount of electrical power spent for the compression itself).

The key quantity which gives us useful information about such items is the entropy content associated with a given data source [17]–[19]. The better the stochastic approximation, the better the compression. First-order entropy compressors do not exploit the internal correlation of source sequences (taken into account by higher-order entropies) unlike the more advanced compression schemes, which obtain a significant gain [20], [21]. ln this paper we present some algorithms

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

which give an evaluation of the source statistics and compute the related absolute and character-relative entropy up to a preassigned order N. The algorithms presented are robust and can deal with data files of arbitrary length. They have been extensively tested with different kinds of files from the Canterbury Corpus. In all cases the observed computing time is typically one order lower than that required to actually compress the files with the best data compressors on the market.

II. ENTROPY BASED COMPRESSIBILITY ASSESSMENT

Elementary calculus then shows that the expected description length must be greater than or equal to the entropy, the first main result. Then Shannon's simple construction shows that the expected description length can achieve this bound asymptotically for repeated descriptions. This establishes the entropy as a natural measure of efficient description length.

An open problem in the field of compressibility theory is about reckoning the deviation from maximum compressibility of the effective compression when using a selected algorithm on the data. In [22] Shannon has devised that the lowest entropy value of an ascii file occurs to be 1.3 bit/digit by using an human being to solve the compression task, but also restricting the related alphabet to 30 different symbols (26 letter from the English alphabet and 4 punctuation symbols). Expert linguists have been able to compress up to 10^8 consecutive digits, while software algorithms can compress 4 to 6 characters long strings. The reason for such a difference lies on the human knowledge of grammar rules, syntax, semantics and the topic-related personal experience. These latter makes the human linguist able to naturally infer or predict portion of the information therefore cumulating the informational entropy of a text in few significant portions, and so naturally implementing a compression procedure ab initio, such a compression capability is unfortunately unquantifiable and actually inimitable by a software algorithm. The best compression algorithms actually developed could achieve 0.88 bit/digit at their best performances, although such algorithms are benchmarked using an alphabet of 256 different symbols (give or take 32 control characters). Effectively the real performances obtained by a compression algorithm depends on the intrinsic compressibility of a file which can be evaluated by characterizing the related informational entropy. In order to evaluate the informational entropy of a file, and consequently its intrinsic compressibility, first order statistics does not suffice, therefore we need to consider larger order statistics. In this context it is mandatory to distinguish between the absolute entropy and the relative entropy of a digit [20].

A. N-TH ORDER ABSOLUTE ENTROPY

Consider an ergodic source emitting sequences of symbols of length L. The number of all possible subsequences of length $N(N \leq L)$ is (L-N+1), therefore if the i-th subsequence S_i occurs M_i times, its relative frequency $f(S_i)$ is:

$$f(S_i) = \frac{M_i}{L - N + 1} \tag{1}$$

By using the interpretation of probability as a relative frequency, we have:

$$P(S_i) = f(S_i) \tag{2}$$

N-th order absolute entropy, $H_{\mathrm{a}}(N)$, is defined as:

$$H_{\mathrm{a}}(N) = \sum_{i=1}^{L-N+1} H_{S_i}(N) \tag{3}$$

where

$$H_{S_i}(N) = P(S_i) \log_2 P(S_i) \tag{4}$$

is the contribution of the generic subsequence S_i to $H_{\mathrm{a}}(N)$. A practical example of absolute N-th order entropy estimation can be given with the text string

ABRACADABRA

and supposing to compute the 2nd order absolute entropy. The subsequences to take into account are constituted by all the pairs of character contained on the string. Such pairs represents all the possible outcome using an alphabet of $256^2 = 65536$ alternatives. To each subsequences can be associated a probability (see Table 1) considering that the string is composed by a total number of 10 possible subsequences of 2 characters (since $L = 11$). Using (3) it follows that

$$H_{\mathrm{a}}(2) = -\left[4 \cdot \frac{1}{10} \cdot \log_2\left(\frac{1}{10}\right) + 3 \cdot \frac{2}{10} \cdot \log_2\left(\frac{2}{10}\right)\right] = \\ = 2.7217 \text{ bits/digit} \tag{5}$$

TABLE 1. Subsequence probability for the string 'ABRACADABRA'.

S_i	AB	AC	AD	BR	CA	DA	RA
$P(S_i)$	0.2	0.1	0.1	0.2	0.1	0.1	0.2

TABLE 2. Occurrences of the subsequences for the string 'ABRACADABRA'.

W_h	A	B	C	D	R
F_{W_h}	5	2	1	1	2

B. N-TH ORDER RELATIVE ENTROPY

The N-th order character-relative entropy for the same sequence of L characters is computed by considering first all the N-order contexts within the sequence (an N-order context is any subsequence of length $N-1$). The entropy associated with the occurrence of the k-th character after the h-th context constitutes the elementary contribution to the N-th order character-relative entropy. The sum of all these contributions gives the total N-th order character-relative entropy:

$$H_r(N) = -\sum_h \frac{F_{W_h}}{L - N + 1} \sum_k \frac{R_{h,k}}{F_{W_h}} \log_2\left(\frac{R_{h,k}}{F_{W_h}}\right) \tag{6}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

TABLE 3. Further occurrences of characters for the string 'ABRACADABRA'

h	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C	D	D	D	D	D	R	R	R	R	R
k	A	B	C	D	R	A	B	C	D	R	A	B	C	D	R	A	B	C	D	R	A	B	C	D	R
$R_{h,k}$	0	2	1	1	0	0	0	0	0	2	1	0	0	0	0	1	0	0	0	0	2	0	0	0	0

where F_{W_h} is the total number of occurrences of the subsequence W_h within the sequence of length L, $R_{h,k}$ is the further occurrences of the subsequence k after the considered one, and $(L-N+1)$ is the number of the occurrences of the k-th character after the subsequence W_h. The above quantity is a useful indicator to establish which of the different subsequences exhibiting the same first-order entropy can be further compressed. Let use the same practical example to compute the first order relative entropy on the string

ABRACADABRA

in this case it follows that $L-N+1=11$, and given the related $R_{h,k}$ (see Tables 2 and 3), from (6) it follows that

$$H_r(1) = -\left\{\frac{5}{11}\cdot\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)\right]\right\} = \\ = 0.6625 \text{ bits/digit} \quad (7)$$

Consider now the following two sequences having the same first-order entropy

ABCDACBDBACDABACBCDC
AAAAABBBBBCCCCCDDDDD

It is evident that the second sequence can easily be compressed whereas the first cannot.

III. THE IMPLEMENTED SOLUTION

In this paper we present an efficient algorithm to compute the N-th order absolute and relative entropy of a string. This latter will be then used to determine when to compress data for transmission in a mobile wireless sensor network. In order to calculate Nth-order absolute and character-relative entropy, existing algorithms are generally articulated into three separate steps. Before calculating entropies, all the strings contained in the sequence are lexicographically ordered and a couple of suitable counters are assigned to each string. These two steps are time-consuming when higher values of N are involved. In the algorithms presented, the above two phases are performed simultaneously; ordering and counter assignment are done in a single step. The algorithm is thus composed of just two steps: first a suitable data structure is constructed and then entropy computations are performed. The data structure used is a modified suffix tree by which the source file is efficiently scanned and an implicit ordering of substrings is simultaneously performed more rapidly than classical ordering algorithms using a modified suffix-tree [23].

A. THE MODIFIED SUFFIX-THREE

For a generic string composed by L digits, each node of the modified suffix-three can represent:

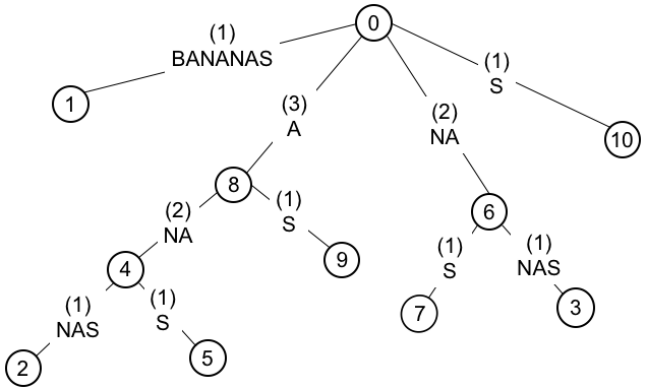

FIGURE 3. The modified suffix three for the string 'BANANAS'.

- a prefix: a substring composed by the first characters of a string
- a suffix: a substring composed by the last characters of a string
- an explicit node: a node with 2 or more children
- an implicit node: a node made by collapsing edges with only one child
- a leaf node: a node without children

Therefore the modified suffix-three can be populated inserting each character form the beginning to the end of the given string by operating three kinds of update:

1) an explicit node update
2) an implicit node update
3) an edge split

Each edge of the suffix-three contains a string, this latter is not entirely stored, in fact, in order to improve the memory occupancy of the algorithm, due the implicit invariance of the sting inherent its suffix-three, we only stored the indexes of the first (`first_char_index`) and last (`last_char_index`) digit as parameters of a class (`Edge`). All the edges are then organized in an hash table. Similarly we store the first and last index for each suffix, along with the origin node index (`origin_node`) which represents the node from which originates the edge containing the suffix at hand (see Table 4).

B. ENTROPY EVALUATION

When evaluating each individual contribution to entropy, it is sufficient to visit the modified suffix-tree structure, avoiding a new complete scan of the file; a sensible reduction in the total computing time is thus achieved. In the modified suffix-tree algorithm, counters are introduced at each branch of the tree. Every counter takes into account how many strings, beginning with the string in the branch considered, there are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

TABLE 4. UML-like table showing several parameters (-) and methods (+) of `Edge` and `Suffix` classes.

Edge	Suffix
-first_char_index:*int* -last_char_index:*int* -start_node:*int* -end_node:*int*	-first_char_index:*int* -last_char_index:*int* -origin_node:*int*
+SplitEdge(s:*Suffix*):*int*	+Explicit():*int* +Implicit():*int*

TABLE 5. Computed subsequence occurrences for the string 'BANANAS' when $N = 3$

W_h	ANA	AS	BAN	NAN	NAS	S
F_{W_h}	2	*	1	1	1	*
* ignored since shorter than N digits						

in the complete sequence of length L. This number is the context frequency and is equal to F_{W_h}.

$R_{h,k}$ is found by considering all the first characters of the strings contained in the subtrees departing from the node considered. Moreover, a complete scanning of the tree does not necessarily have to be performed: if we want to compute, for example, fifth-order entropies, we only have to scan five levels of the tree (in the worst case), because these levels contain all the information about the statistics of 5 character strings. An example of a modified suffix tree for the string BANANAS is shown in Figure 3: it is possible to verify that the edges with only one child have been collapsed. If we want to compute 2nd order entropies we only have to visit the nodes labeled 0,1,8,6,10; all the other nodes do not need to be visited at all; the time saving obtained with this approach is significant, especially in the case of very long sequences (over 10^8 symbols). By using the presented modified suffix-three the computation of entropy for an assigned order N is quite straightforward; all the contributions to the entropy are obtained by visiting only the tree levels from the root to the levels representing N-length o subsequences. All the other tree levels are ignored, thus achieving efficiency and speed in the evaluation. In addition, no preliminary ordering is required. Let consider again the string

BANANAS

and let suppose to compute the 3rd order absolute and relative entropy for such a string. In order to obtain the occurrences of a substring it suffices to count the repetition numbers on the suffixes list. Each time a new substring of different length is found, then we will have yet counted all the occurrences of the previous substring. Therefore it will be possible to compute its entropic contribute immediately. For the string BANANAS the possible substring occurrences are shown in Table 5. Therefore once computed the contributions:

$$H_{\mathrm{a}}(3)[\mathrm{ANA}] = \left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)\right]_{\mathrm{ANA}} \tag{8}$$

$$H_{\mathrm{a}}(3)[\mathrm{BAN}] = \left[\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right)\right]_{\mathrm{BAN}} \tag{9}$$

$$H_{\mathrm{a}}(3)[\mathrm{NAN}] = \left[\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right)\right]_{\mathrm{NAN}} \tag{10}$$

$$H_{\mathrm{a}}(3)[\mathrm{NAS}] = \left[\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right)\right]_{\mathrm{NAS}} \tag{11}$$

it immediately follows that

$$\begin{aligned} H_{\mathrm{a}}(3)[\mathrm{BANANAS}] &= \\ &= -\left[\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right)+\frac{3}{5}\cdot\log_2\left(\frac{1}{5}\right)\right]_{\mathrm{BANANAS}} = \\ &= 0.8643 \text{ bits/digit} \end{aligned} \tag{12}$$

The 3rd order relative entropy computation is a little more difficult. In order to reckon the 3rd order entropy ($N = 3$) we need to take into account substrings of 2 digits ($N - 1$). Therefore the algorithms will execute the following steps:

1) scan the list considering the first two digits for each suffix
2) compare the found substring with the previous suffix
3) count the digits position-related occurrences for each substring
4) proceed to the next substring

After all the suffix occurrences have been computed, the related counts are given as input for a statistical routine. This latter routine determines the occurrence probability for each substring in order to compute the relative entropy as in (6).

C. COMPRESSION EFFICIENCY ESTIMATION

Once the relative entropy have been evaluated at different orders, the obtained values are considered to estimate the possible compression efficiency. It must be pointed out that it is not possible to precisely estimate an a priori compression cost in terms of consumed power due to the many aleatory variables that should be considered otherwise. On the other hand, trough empirical evaluation, it is possible to establish for each given device an entropy descent related threshold, which could eventually be demanded to the hardware constructor. Such a threshold must be related to the slope of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

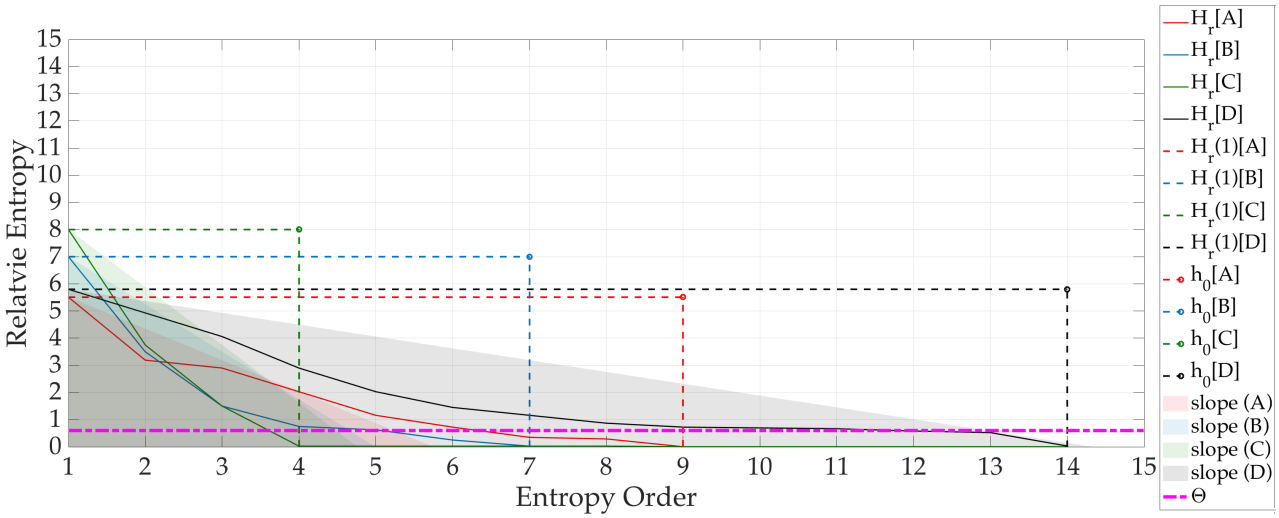

FIGURE 4. The figure shows an order-wise comparison of the relative entropy $H_r(h)$ with respect to the h_0 order, the maximum relative entropy $H_r(1)$ and the computed slope proportional to $H_r(2) - H_r(1)$. The superposition shows the behavior of different files extracted from the *Canterbury Corpus*: [A] an image (*lena.bmp*), [B] an object code for VAX (*obj1*), [C] the first milion digits for π (*pi.txt*), and [D] an english text (*alice29.txt*).

TABLE 6. The used Canterbury corpus (and the image lena.bmp, added for a more representative test).

Corpus collection	Filename	Type	Content	Size
Canterbury	alice29.txt	ascii	English text	152089 B
Canterbury	asyoulik.txt	ascii	Shakespeare	125179 B
Canterbury	cp.html	html	HTML source	24603 B
Canterbury	fields.c	ascii	C source	11150 B
Canterbury	grammar.lsp	list	LISP source	3721 B
Canterbury	kennedy.xls	excel	Excel Spreadsheet	1029744 B
Canterbury	lcet10.txt	ascii	Technical writing	426754 B
Canterbury	plrabn12.txt	ascii	Poetry	481861 B
Canterbury	ptt5	fax	CCITT test set	513216 B
Canterbury	sum	bin	SPARC Executable	38240 B
Canterbury	xargs.1	man	GNU manual page	4227 B
Artificial	a.txt	ascii	The letter 'a'	1 B
Artificial	aaa.txt	ascii	The letter 'a' 100000 times	100000 B
Artificial	alphabet.txt	ascii	alphabet repetitions	100000 B
Artificial	random.txt	ascii	pseudorandom alphanumeric	100000 B
Large	E.coli	ascii	Genome of the E. Coli bacterium	4638690 B
Large	bible.txt	ascii	The Bible of King James	4047392 B
Large	world192.txt	ascii	The CIA world fact book	2473400 B
Miscellaneous	pi.txt	ascii	The first million digits of pi	1000000 B
Calgary	bib	ascii	Latex Bibliography file	111261 B
Calgary	book1	ebook	Fiction book	768771 B
Calgary	book2	ebook	Non-fiction book (troff format)	610856 B
Calgary	geo	raw	Geophysical data	102400 B
Calgary	news	ascii	USENET batch file	377109 B
Calgary	obj1	obj	Object code for VAX	21504 B
Calgary	obj2	obj	Object code for Apple Mac	246814 B
Calgary	paper1	ascii	Technical paper	53161 B
Calgary	paper2	ascii	Technical paper	82199 B
Calgary	pic	gif	Black and white fax picture	513216 B
Calgary	progc	ascii	Source code in C	39611 B
Calgary	progl	ascii	Source code in LISP	71646 B
Calgary	progp	ascii	Source code in PASCAL	49379 B
Calgary	trans	ascii	Transcript of terminal session	93695 B
	lena.bmp	bmp	Bitmap picture	263200 B

relative entropy value with respect to its order, as well as the maximum non-zero entropy order (h_0) defined as

$$h_0 = \min_h \{h : H_{\mathbf{r}}(h) < \epsilon\} \tag{13}$$

where ϵ is a number close to 0 (i.e. 10^{-8}), used to avoid machine's related fluctuations.

Since an high slope for the relative entropy, as well as a small maximum non-zero order suggest a low compressibil-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

ity ratio, and since, on the contrary, an high first order relative entropy value will suggest an high compressibility, it follows that we can define

$$\chi = \frac{H_{\mathbf{r}}(1) \cdot h_0}{H_{\mathbf{r}}(1) - H_{\mathbf{r}}(2)} \tag{14}$$

as an evaluation parameter directly proportional to the compressibility of the data at hand (see Figure 4). In this fashion, given an empirically determined threshold θ, it will follows that data will be compressed only if $\chi > \theta$.

Since the hardware configuration of a device could tamper with the battery lifespan, as well as any implementation and usage choice adopted by the constructor or the user, the said threshold θ must be determined on field and could differ for different devices. It follows that, in general, θ should be provided as data-sheet parameter by the vendor or the implementor of a specific protocol involving such a device. On the other hand, θ could be experimentally determined by measuring in controlled conditions, or in laboratory environment, the maximum battery life-span as a function $T(\vartheta)$, where ϑ represents a threshold candidate. In this manner it is possible to devise an optimal threshold θ so that

$$\theta : T(\theta) = \max_{\vartheta} \{T(\vartheta)\} \tag{15}$$

In the following application for testing purposes the threshold Θ has been defined as approximately 1% of the average χ.

IV. APPLICATION AND TESTING

As common practice in literature, the algorithms implemented in this paper has been extensively tested using the Canterbury Corpus: a set of standard files used to test almost all lossless compression algorithms.

A. THE CANTERBURY CORPUS

The Canterbury Corpus [24] is constituted of several collection of files that are commonly used as benchmark in order to evaluate the performances of compression algorithms on different kinds of file types (such as text files, books, technical papers, source code, object files, raw data, images, etc...). The Canterbury Corpus has been devised as an upgrade of the Calgary Corpus [25]. The purpose of the Canterbury Corpus was to provide researchers with a set of files that could be representative of information that an user would like to compress, as well as provide testing means to gather sufficient statistical data for both an analytical and empirical study of the compression performances of an algorithm. The overall Canterbury Corpus is composed by the following five collections:

- The Canterbury Collection
- The Artificial Collection
- The Large Collection
- The Miscellaneous Collection
- The Calgary Collection

While the Canterbury Collection constitutes the main focus of the corpus, the Calgary Collection has been included mainly for historic reasons, as well as the Large Collection has been included to provide a testing ground for algorithms that are specifically designed, or best performing, for large files. Moreover the corpus also contains the Artificial Collection providing a set of files that should tamper with the standard performances of a compression algorithm due to their intrinsic nature (due to the absence of repetition or due to a large amount of repetitions). This latter Collection, then, is unsuitable for performances characterization, while it is useful to detect outliers. Finally, the Miscellaneous Collection actually contains only a file with the first million digits of π. In Table 6 we report a list of the files constituting the Canterbury Corpus, and that we used to test our algorithm, along with the commonly used image `lena.bmp`.

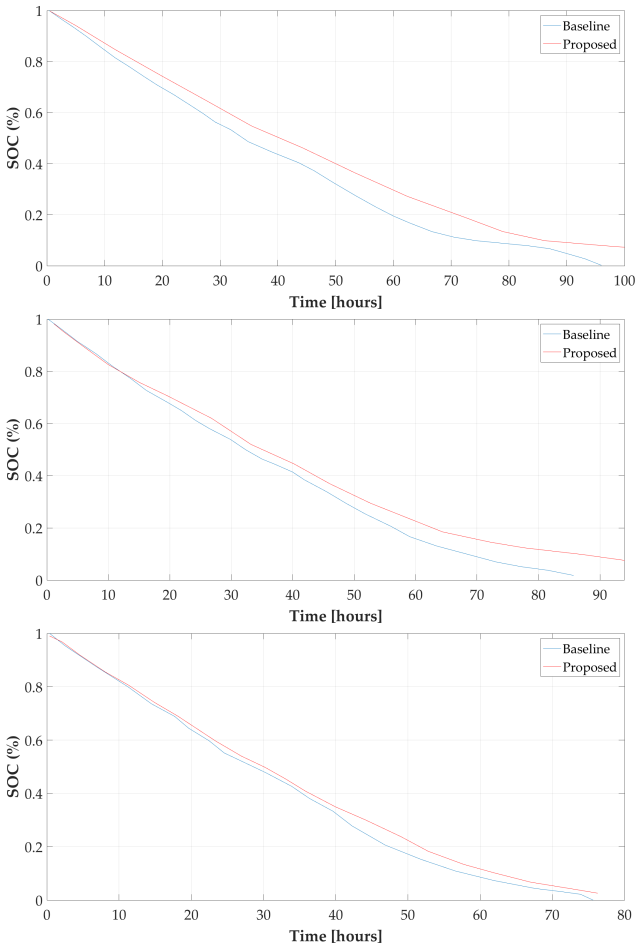

FIGURE 6. Three examples among the many experimental results, collected in different environmental conditions, of battery status of charge (SOC) during time for a communicating sensor (baseline) when the proposed system is implemented (proposed) with a threshold of $\theta = 0.5$.

B. ENTROPY EVALUATION

The various evaluations have been performed by using the files of the Calgary Corpus and Canterbury Corpus that contain different kinds of data. The results of this investigation are summarized in Figure 5 where the absolute and relative entropies of four data files are shown for increasing values

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2962771, IEEE Access

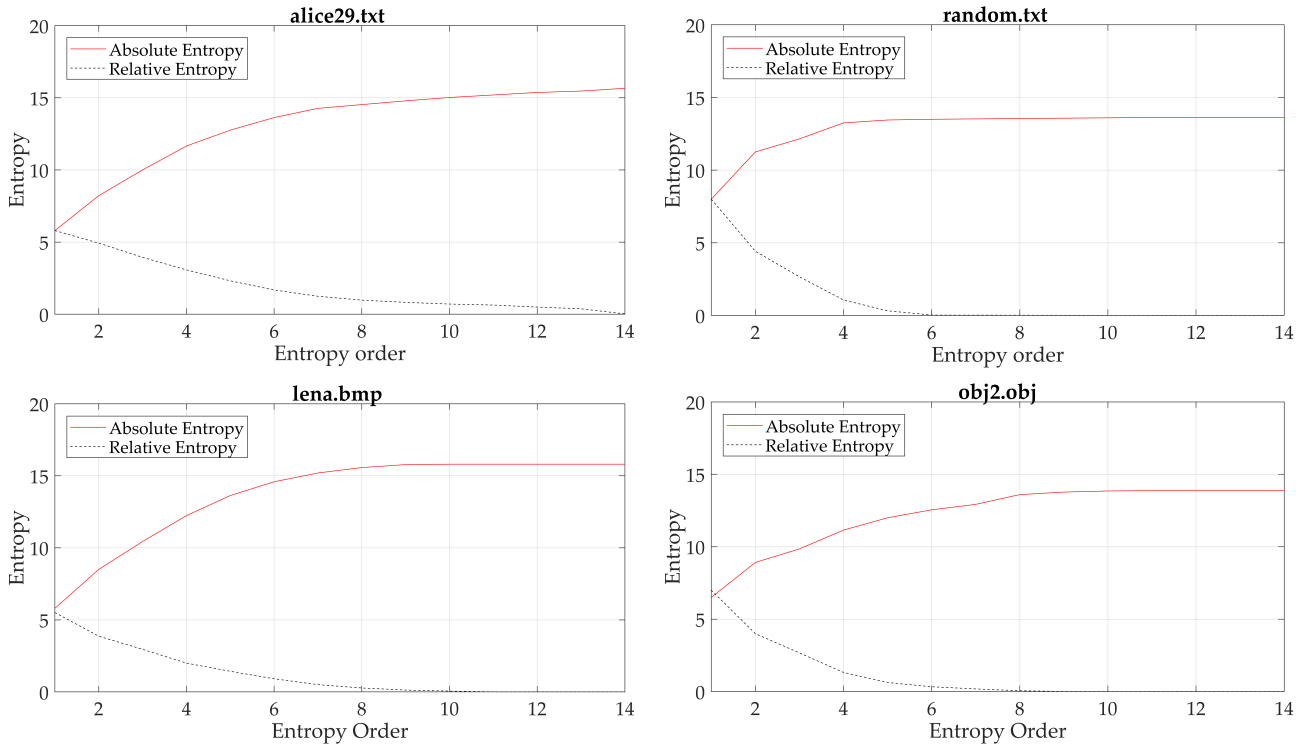

FIGURE 5. Absolute and relative entropy for several of the files used for testing (see Table 6).

of the entropy order N. It is possible to notice that the said entropy values are strongly affected by the analyzed data types. As a matter of fact we observe that the shape of the curves strictly depends upon the kind of data processed; more precisely, shapes tend to be smoother for compressible files while they become sharper for incompressible data. In particular for pseudo-random tiles, character-relative entropy values always fall exactly to zero within the first five orders. It is worth noticing that the behavior of the two quantities is specular with respect to the value assumed for $N = 1$. The character-relative entropy tends to a null value as N increases whereas the absolute entropy reaches an asymptotic value which depends on the nature of the source. Both absolute and character-relative entropy approximately reach their asymptotic values for the same order N. This allows us to consider only one of the two quantities to get an estimate of the entropy content of the source.

C. EXPERIMENTAL RESULTS

The experiments have been conducted by using a Zigbee hardware architecture (Libelium Comunicaciones Distribuidas, Zaragoza, Spain) designed as ultra low power technology due to the extremely small operation current. The architecture, yet know for its use and versatility in mobile sensor networks [26], is provided with 10 sensor boards and 16 radio technologies for short, medium and long range communication. During the experiments (see Figure 6 the board has been tested using the Wi-Fi interface to communicate with a radio-base station at 40 m distance in different kind

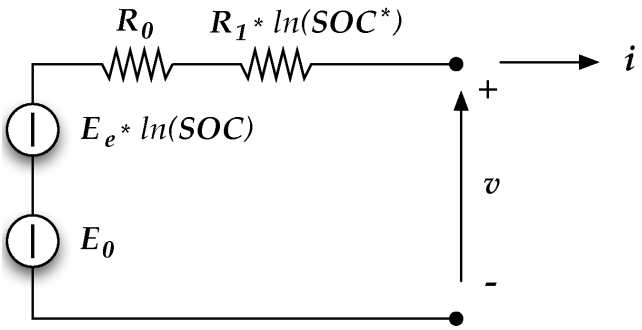

FIGURE 7. Equivalent network of lead-acid battery.

of environments (office building, open field, wood, buildings construction facilities, soccer fields, etc...). During the tuning phase we defined a threshold of $\theta = 0.5$ (approximately 1% of the average χ). Within this configuration, the results show an average the battery-life increment of about 11.8% due to the reduced amount of energy used for data transfer.

The calculation of the average battery life in the test bed scenario used for the validation of the proposed methodology was made by using the experimental apparatus used by one of the authors in [27]. In fact, as shown in the in the previously cited paper, the energy management of the batteries should be based on the state of charge (SOC) checking. The basic equations that relate the SOC to the discharge current and voltage at the battery terminals are the listed in the following

while the equivalent electrical network is shown in fig. 7.

$$\begin{aligned} \frac{dq}{dt} &= i \\ SOC &= 1 - \frac{q}{C_0} \\ v &= E(SOC) - R(SOC^*) * i \end{aligned} \tag{16}$$

where SOC^* it's a fictitious SOC that depends on the effective SOC value, current discharge rate and depth of discharge.

$$\begin{aligned} E(SOC) = &\quad E_0 + E_e * ln(SOC) \\ R(SOC^*) = &\quad R_0 + R_1 * ln(SOC^*) \end{aligned} \tag{17}$$

The energy supplied by the battery to the load is related to itself rated capacity C_t, expressed in Wh, minus a factor accounting the energy lost due the irreversibility of the electrochemical discharge phenomenon.

$$\int vidt = C_t - E(irr) \tag{18}$$

with a little algebra yields the equation.

$$\begin{aligned} \int vidt = &\int E_0 idt - \int R_0 i^2 dt \\ &- \left| \int E_e ln(SOC) idt \right| - \left| \int R_1 ln(SOC^*) i^2 dt \right| \end{aligned} \tag{19}$$

where the integral is taken over the selected discharge time.

For the calculation of the average battery life, in this paper, we used the equations (16) and (17). The calculus of the parameters E_0, E_e, R_0, R_1 and the relationship between the fictitious (SOC^*) and the true SOC have been carried out by using the neural network described in [27], trained with the experimental results, collected in different environmental conditions when the proposed system is implemented with a threshold of $\theta = 0.5$.

V. CONCLUSIONS

Data prediction techniques are often used in sensor networks to mitigate the sensors energy consumption, avoiding unnecessary data transmissions, and extending the network life cycle.

In this work we developed a new approach to increase the energy data trasmission efficiency in pervasive healthcare sensor networks. In the presented approach the sensors battery life has been extended by means of a shorter communication time due to data compression. On the other hand the evaluation of data compressibility has been a paramount asset to avoid energy waste due to inefficient or inappropriate data compression. This evaluations have been performed by means of a novel algorithm for the evaluation of absolute and relative N-th order entropies that allowed an ad-hoc decision system to preliminarily estimate whether or not the reachable compression ratio would justify the amount of energy spent for the data compression itself. The computational cost of this operation is about one order of magnitude lower than a compression operation itself. Therefore entropy computation can be advantageously executed before compressing data, thus avoiding uncertain results.

It can be seen from the experimental results that our scheme can efficiently decrease redundant transmissions while improving the prediction precision. By this means, the energy of sensor nodes is also saved and the fault tolerance is improved. Then the implemented procedure allows an efficient management of data compression for communicating mobile wireless sensor networks, which can be of uttermost importance for pervasive healthcare systems.

REFERENCES

[1] G. Hu and W. Liu, "Nano/micro-electro mechanical systems: a patent view," Journal of Nanoparticle Research, vol. 17, no. 12, p. 465, 2015.

[2] S. Acciarito, G. C. Cardarilli, A. Cristini, L. Di Nunzio, R. Fazzolari, G. M. Khanal, M. Re, and G. Susi, "Hardware design of lif with latency neuron model with memristive stdp synapses," Integration, vol. 59, pp. 81–89, 2017.

[3] G. Khanal, S. Acciarito, G. Cardarilli, A. Chakraborty, L. Nunzio, R. Fazzolari, A. Cristini, M. Re, and G. Susi, "Synaptic behaviour in zno–rgo composites thin film memristor," Electronics Letters, vol. 53, no. 5, pp. 296–298, 2017.

[4] S. Guo, Y. Yang, and C. Wang, "Dagcm: A concurrent data uploading framework for mobile data gathering in wireless sensor networks," IEEE Transactions on Mobile Computing, vol. 15, no. 3, pp. 610–626, 2016.

[5] X. Fu, Y. Yang, W. Li, and G. Fortino, "Topology upgrading method for energy balance in scale-free wireless sensor networks," in Networking, Sensing and Control (ICNSC), 2017 IEEE 14th International Conference on. IEEE, 2017, pp. 192–197.

[6] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," Computer networks, vol. 54, no. 15, pp. 2688–2710, 2010.

[7] M. G. Ball, B. Qela, and S. Wesolkowski, "A review of the use of computational intelligence in the design of military surveillance networks," in Recent Advances in Computational Intelligence in Defense and Security. Springer, 2016, pp. 663–693.

[8] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," IEEE sensors journal, vol. 15, no. 3, pp. 1321–1330, 2015.

[9] F. Beritelli, G. Capizzi, G. L. Sciuto, C. Napoli, and F. Scaglione, "Automatic heart activity diagnosis based on gram polynomials and probabilistic neural networks," Biomedical Engineering Letters, vol. 8, no. 1, pp. 77–85, 2018.

[10] C.-C. Lai, R.-G. Lee, C.-C. Hsiao, H.-S. Liu, and C.-C. Chen, "A h-qos-demand personalized home physiological monitoring system over a wireless multi-hop relay network for mobile home healthcare applications," Journal of Network and Computer Applications, vol. 32, no. 6, pp. 1229–1241, 2009.

[11] B. Liu, Z. Yan, and C. W. Chen, "Medium access control for wireless body area networks with qos provisioning and energy efficient design," IEEE transactions on mobile computing, vol. 16, no. 2, pp. 422–434, 2017.

[12] O. Omeni, A. C. W. Wong, A. J. Burdett, and C. Toumazou, "Energy efficient medium access protocol for wireless medical body area sensor networks," IEEE Transactions on biomedical circuits and systems, vol. 2, no. 4, pp. 251–259, 2008.

[13] D. Bhatia, L. Estevez, and S. Rao, "Energy efficient contextual sensing for elderly care," in Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE. IEEE, 2007, pp. 4052–4055.

[14] K. Sayood, Introduction to data compression. Elsevier, 2005.

[15] M. Wozniak, C. Napoli, E. Tramontana, G. Capizzi, G. L. Sciuto, R. K. Nowicki, and J. T. Starczewski, "A multiscale image compressor with rbfnn and discrete wavelet decomposition," in 2015 International Joint Conference on Neural Networks (IJCNN), July 2015, pp. 1–7.

[16] Y. Zheng, C. Qi, and G. Wang, "A new image pre-processing for improved performance of entropy coding," in 2010 Chinese Conference on Pattern Recognition (CCPR), 2010.

[17] K. Fukunaga, Introduction to statistical pattern recognition. Elsevier, 2013.

[18] K. J. Balakrishnan and N. A. Touba, "Relating entropy theory to test data compression," in Test Symposium, 2004. ETS 2004. Proceedings. Ninth IEEE European, 2004, pp. 94–99.
[19] S. Wegenkittl, "Entropy estimators and serial tests for ergodic chains," IEEE Transactions on Information Theory, vol. 47, no. 6, pp. 2480–2489, 2001.
[20] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd edition. Wiley, 2006.
[21] S. Dhawan, "A review of image compression and comparison of its algorithms," International Journal of Electronics & Communication Technology, IJECT, vol. 2, no. 1, pp. 22–26, 2011.
[22] C. E. Shannon, "Prediction and entropy of printed english," Bell Labs Technical Journal, vol. 30, no. 1, pp. 50–64, 1951.
[23] U. Manber and G. Myers, "Suffix arrays: a new method for on-line string searches," siam Journal on Computing, vol. 22, no. 5, pp. 935–948, 1993.
[24] R. Arnold and T. Bell, "A corpus for the evaluation of lossless compression algorithms," in Data Compression Conference, 1997. DCC'97. Proceedings. IEEE, 1997, pp. 201–210.
[25] T. C. Bell, J. G. Cleary, and I. H. Witten, Text compression. Prentice Hall Englewood Cliffs, 1990, vol. 348.
[26] H. J. Lee, S. H. Lee, K.-S. Ha, H. C. Jang, W.-Y. Chung, J. Y. Kim, Y.-S. Chang, and D. H. Yoo, "Ubiquitous healthcare service using zigbee and mobile phone for elderly patients," International journal of medical informatics, vol. 78, no. 3, pp. 193–198, 2009.
[27] G. Capizzi, F. Bonanno, and G. M. Tina, "Recurrent neural network-based modeling and simulation of lead-acid batteries charge–discharge," IEEE Transactions on Energy Conversion, vol. 26, no. 2, pp. 435–443, 2011.

GIACOMO CAPIZZI received the Laurea degree (summa cum laude) in electronic engineering from the University of Catania, Catania, Italy, in 1993, and the Ph.D. degree in electronic and computer engineering from the University of Reggio Calabria, Reggio Calabria, Italy, in 2000. From 1993 to 1996, he was a Cabling Systems Designer with Itel (a small company of telecommunications). From 2000 to 2002, he was a Lecturer with the Department of Electrical, Electronics and System Engineering, University of Catania, where he joined as an Assistant Professor in 2002. From 2015 to 2016, 2016 to 2017, and from 2017 to 2018, he was an Invited Professor with the Silesian University of Technology to teach in a master course on the topic algorithms and paradigms for pattern recognition (course that currently teaches). His research interests include wavelet theory, neural networks, statistical pattern recognition, Bayesian networks, theory and design of linear and nonlinear digital/analog filters, integrated generation systems, renewable energy sources, and battery storage modeling and simulation.

SALVATORE COCO Salvatore Coco is a Full Professor of Electrotechnics at Catania University. His main scientific interests are in the Finite Element computation of Electromagnetic fields, in innovative circuits and algorithms for signal processing and in the application of neural networks to prediction problems. He is the author of over 150 papers in these fields. Professor Coco is a member of AEI and a founding member of the International Compumag Society.

GRAZIA LO SCIUTO received the Ph.D. degree in applied electronics from the University of Rome Tre in 2016. In 2015, she received the scholarship on the optical calculations for large-scale organic photovoltaic with ENEA-BGU Joint Laboratory, Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. Her research interests include electronic devices, semiconducting polymers, organic materials, novel devices for photovoltaic, and neural networks applied to complex systems, such as renewable energy, signal processing, pattern recognition, and biometrics.

CHRISTIAN NAPOLI is Associate Professor with the Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, since 2019, as well as Scientific Director and co-founder (2015) of the International School of Advanced and Applied Computing (ISAAC). He received the B.Sc. degree in Physics from the Department of Physics and Astronomy, University of Catania, in 2010, where he also got the M.Sc. degree in Astrophysics in 2012 and the Ph.D. in Computer Science in 2016 at the Department of Mathematics and Computer Science. He has been Research Associate with the Department of Mathematics and Computer Science, University of Catania, from 2018 to 2019, while, previously, Research Fellow and Adjunct Professor with the same department from 2015 to 2018. He has been several time Invited Professor at the Silesian University of Technology, Visiting Academic at the New York University. His teaching activity focused on Artificial Intelligence, Neural Networks, Machine Learning, Computing Systems, Computer Architectures, Distributed Systems, and High Performance Computing. His current research interests include neural networks, artificial intelligence, computational models, and high performance computing.

WALDEMAR HOŁUBOWSKI (M'85) received M.S degree from Faculty of Mathematics and Physics of Silesian University of Technology, Gliwice, Poland. In 1991 and 2008 he received PhD and DSc in mathematics from Sankt Petersburg State University, Russia. His research interests include theory of groups and Lie algebras, matrix theory and their applications in engineering. He was visiting researcher at University of Manitoba, Canada and Euler Institute at Sankt Petersburg, Russia. He is currently a Head of the Faculty of Applied Mathematics at the Silesian University of Technology.