A Scalable and Flexible Repository for Big Sensor Data

Dongeun Lee, Jaesik Choi, *Member, IEEE,* and Heonshik Shin

***Abstract*—Data generation rates from sensors are rapidly increasing, reaching a limit such that storage expansion cannot keep up with the data growth. We propose a new big data archiving scheme with an optimized lossy coding to handle huge volume of sensor data. Our scheme leverages spatial and temporal correlations inherent in typical sensor data. These correlations, along with quality adjustable nature of sensor data, enable us to compress a massive amount of sensor data without compromising their distinctive attributes. A data aging aspect of sensor data also offers an option to apply scalable quality management while stored. In order to maximize the benefits of storage efficiency, we derive an optimal storage configuration for the data aging scenario. Experiments show outstanding compression ratios of our scheme and the optimality of storage configuration that minimizes system-wide distortion of sensor data under a given storage space.**

***Index Terms*—Quality-adjustable sensor data, storage management, big data archiving, data compression, distributed file systems, wireless sensor networks.**

I. Introduction

DATA generation rates from sensors have increased dramatically, fostering the widespread research of *big sensor data* [1], [2]. As various types of sensors are being deployed, information generated by these sensors are also rapidly increasing [3], [4]. This massive data flow generated by sensor devices now comprises a notable portion of big data and intensifies depending on applications such as large-scale scientific experiments [5]–[7].

While data storage capacities keep increasing with reduced cost, this faster data generation rate now leads to a paradox that increasing storage capacity cannot catch up with the rate of information explosion. It is reported that almost half of information created and transmitted cannot be stored now and this mismatch between available storage and information creation will become more serious [7], [8].

From the perspective of an information repository, this mismatch necessitates the development of a new big data archiving technique that facilitates scalable and flexible usage of the repository. We now propose a quality-adjustable archiving scheme for massive sensor data. Our scheme thoroughly exploits both spatial and temporal correlations inherent in sensor data collections, and generates a digested set of sensor data keeping fidelity under control, which is demonstrated as outstanding compression efficiency with data fidelity corresponding to orders of sensor accuracies. In addition, a concept of data aging is embodied in the quality-adjustable feature of our scheme with multiple fidelity levels: older sensor data are not as representative as recent data and can be represented with less precision [9].

To our best knowledge, there have been no in-depth studies on efficient data archiving techniques that fully exploit spatio-temporal correlation of huge sensor data set. In distributed environments such as wireless sensor networks (WSNs), a few approaches have utilized partial correlation to reduce traffic and storage usage inside the networks themselves [1], [2], [10]–[15]. Although these approaches have achieved their objectives in distributed environments, efficient archiving techniques are still necessary if sensor data are eventually to be stored in central storage.

A massive amount of data from various sensors should be archived in a cost-effective manner such that the system-wide distortion is minimized under a given storage space. In order to solve this issue, we propose new analytical models that closely reflect characteristics of our archiving scheme and eventually an optimal storage configuration problem. Since this optimization problem is convex, we can analytically solve it and obtain optimal parameters. Experimental results demonstrate that our optimal storage configuration effectively minimizes system-wide distortion under a given storage space. The system-wide distortion can otherwise increase drastically, which is translated into inefficient expenditure of storage space.

The rest of this paper is organized as follows. Section 2 explains key characteristics of sensor data and how they can be exploited for storage efficiency using our scheme. Section 3 overviews how our archiving scheme works. In Section 4, we derive analytical models that explain the relationship between controllable quality parameters and rate-distortion, which leads to the rate allocation and storage configuration problems in Section 5. Section 6 exhibits its performance compared to various schemes and also illustrates the importance of the optimal storage configuration. Section 7 reviews related works, followed by concluding remarks in Section 8.

II. Motivation

Our quality-adjustable archiving scheme benefits from the use of lossy coding that exploits three key characteristics of sensor data. The entire storage space can be efficiently utilized through the judicious use of the lossy coding over numerous sensor data blocks with different types.

D. Lee and J. Choi are with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan 689-798, Korea. E-mail: eundong@unist.ac.kr, jaesik@unist.ac.kr

H. Shin is with the Department of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea. E-mail: shinhs@snu.ac.kr

Manuscript received xxxxx xx, xxxx; revised xxxxxxxxx xx, xxxx.

A. Three Key Characteristics of Sensor Data

In many applications, individual sensor data may not require either bit-level accuracy or intactness due to several reasons: (i) each sensor node is equipped with inexpensive and imprecise sensors that only guarantee moderate level of sensing accuracy, (ii) sensor nodes are densely deployed and they periodically capture data that are highly correlated in spatio-temporal domain, which makes storing all of data unnecessary, (iii) we are usually interested in overall trend of sensor data, thus we can tolerate a certain amount of distortion and approximate results are sufficient most of the time [16]–[18]. This property is called the *quality adjustability* in this paper.

Data aging, where data fidelity is gradually decreased, is common practice when handling various kinds of time series data [9], [19]–[21]. Sensor data fidelity can also be gradually decreased as time goes by. Since fresh data are important (e.g., frequently accessed) and should retain high fidelity, aged data could be regarded less important and only find their use in offering a digest of historical trend in sensor readings. Therefore it is sufficient to store key features of sensor data in most sensor applications especially for long-term storage [10], [11], [22].

Because sensors usually capture physical phenomenon such as environmental data, their data are highly correlated in nature within spatial and temporal domain [22]: spatially and temporally close data samples are more correlated than distant counterparts. (Here the degree of correlation is measured by autocorrelation function: one-dimensional in temporal domain and two-dimensional in spatial domain [23].) In particular, the temporal correlation tends to be stronger than the spatial correlation since the sensing frequency of a particular sensor node is in general high enough to surpass the spatial closeness among deployed sensor nodes. This *spatio-temporal correlation*, along with the quality adjustability, allows sensor data to be represented in a compact form.

B. Combating Shortage of Storage Space

The quality adjustability of sensor data and its trade-off between data fidelity and compression ratio provides us with many options of encoding. Among these numerous operating points, we have to select the best possible way of encoding data that yields the maximum fidelity (the minimum distortion) under a given storage space, i.e., the optimal storage configuration.

In other words, we want to solve an optimization problem that requires analytical models, which are unknown. In general, we are not exactly aware of the compressed data size and fidelity prior to an actual encoding that vary depending on a data set. For this reason, we build new analytical models in Section IV. Our models are close enough to reflect operating points of our archiving scheme, which can be adapted to multiple sensor data types using different model parameters.

Given analytical models, their model parameters are determined when an enough number of data samples for each type of sensor data is gathered. We assume a stationarity of data for each type without loss of generality, which can be applicable to most sensor data. (For instance, the dynamic range of temperature data does not change over time.) Therefore, we need to determine model parameters only once for each sensor data type in the *model training* process shown in Fig. 1.

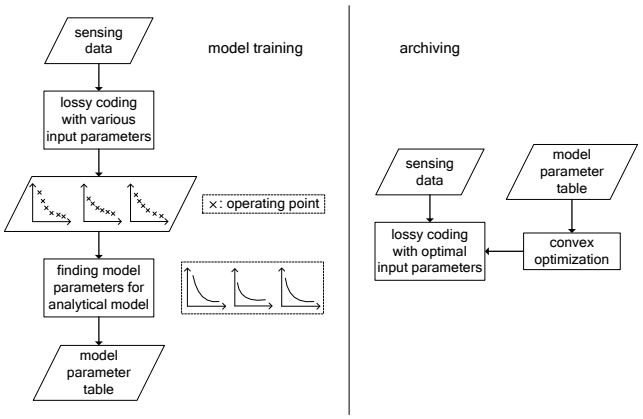


Fig. 1. Flowchart for optimal storage configuration. Each sensor data type has its own model parameter set.

Using these model parameters, we can perform a convex optimization that yields the minimum system-wide distortion with a given storage space, whose solution is presented in Section V. (Our analytical models are convex by virtue of the trade-off relationship between data fidelity and compression ratio. The sum of these convex functions is also convex.) The solution of the convex optimization provides optimal input parameters, with which the *archiving* process shown in Fig. 1 is executed. This way, the entire storage space is efficiently utilized.

III. Overview of Our Archiving Scheme

A. Quality Management Module: Lossy Coding

The characteristics of sensor data described in Section II-A allow us to compress the entire data set into a smaller form with a reasonable loss in fidelity. Fig. 2 illustrates the block diagram of our quality management module, which is designed to work with conventional distributed file system.

Massive data from various sensors are first collected and filtered through the spatio-temporal decorrelation module. Specifically, a sensor value can be predicted by similar values captured by other sensors in close proximity (spatial correlation), or by previous and next similar values captured by that sensor (temporal correlation),[1] whichever is stronger than the other, depending on each data instance. If we take a differential between target and predicted values, we ideally obtain a decorrelated value that is close to zero, which means the redundancy in input data is removed.

In reality, these differentiated sensor values still have a fair amount of correlation inside. Therefore the resulting output in turn undergoes the two-dimensional discrete cosine transform (DCT) for further signal decorrelation and energy compaction. The DCT is an approximation of Karhunen-Loève transform

[1]Predicting a sensor value using similar values in temporal proximity is shown in Fig. 4.

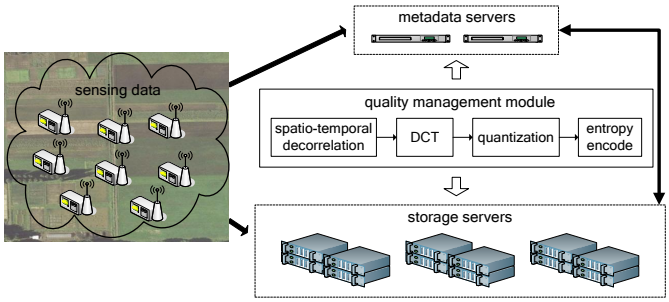


Fig. 2. Quality management module working with conventional distributed file system. Black arrows represent data and control flows of file system. Lossy coding runs on storage servers under control of metadata servers.

that is optimal in reducing the dimensionality of feature space [24].

After the two-dimensional DCT, transformed data are subject to the quantization process that sacrifices precision of data in order to represent them in a compact form, which irrevocably maps a large set of values onto a smaller set. The quantization module controls sensor data fidelity, which can be adjusted through a quantization parameter (QP). The QP determines how much we compress data at the cost of decreased data fidelity. Finally, the entropy encode module compactly produces an encoded data block [23].

It should be noted that this lossy coding part of our scheme is analogous to modern image and video encoding schemes. Specifically, video encoding schemes typically involve computation-intensive operations such as the motion estimation and compensation (ME/MC) [23]. On the contrary, the spatio-temporal decorrelation module shown in Fig. 2 does not involve ME/MC. Rather, previous and next collections of sensor values in the collocated positions are always used for the temporal decorrelation, avoiding complex motion search. Video data and sensor data share similarities in the sense that they both have the spatio-temporal correlation inside. However, sensor data generally do not have the motion of data clusters between consecutive collections of sensor values.

Since storage servers are usually not involved in computation-intensive tasks, they can run the quality management module in online or offline, depending on applications. In fact, the entire chain of processes in Fig. 2 can be further optimized to speed up if we employ single instruction, multiple data (SIMD) instructions that most modern CPUs support [25]–[27].[2]

B. Quality Management Module: Temporal Quality Adjustment

In our scheme, multiple temporal levels are supported with a fixed QP. These multiple temporal levels can be utilized as supplementary layers that are gradually discarded as time elapses to incorporate data aging concept.

Fig. 3 illustrates how incoming sensor data input is handled and archived with our scalable archiving scheme. The quality

[2]The performance optimization of our scheme is beyond the scope of this paper.

Fig. 3. Sensor data flow with our quailty-adjustable archiving scheme. Quality management module adjusts temporal quality through the course of sensor data aging. Number of clusters can vary depending on applications.

management module first compresses raw sensor data block with a selected QP, which is then stored on the highest fidelity cluster, i.e., the cluster 4 in Fig. 3. When a certain amount of time passes, the quality management module discards the top layer and shift the data block to the next cluster. This process continues until the data block finally reaches the cluster 0, where the data block is archived for a long time.

C. Storage Space Optimization

In Section II-B, we described the efficient usage of storage space by determining optimal input parameters to the lossy coding part of our archiving scheme presented in Section III-A. At any given time, the system has numerous sensor data blocks with different types, each of them belonging to one of clusters according to Fig. 3. In this case, the optimization problem would be stated as "minimize system-wide distortion with total rate budget (given storage space)," which is formulated in (15), Section V.

Our solution to this optimization problem is presented as a relationship equation shown in (17). Once model parameters are determined for each sensor data type in the model training process of Fig. 1, the solution (17) can be easily calculated in the archiving process of Fig. 1, yielding the optimal input parameters to lossy coder. These input parameters for each sensor data type can be steadily used, or occasionally recalculated depending on the availability of storage space.

IV. Quality-Adjustable Archiving

We now focus on the quality-adjustability of our archiving scheme. We derive analytical models that reflect the effect of adjusting data fidelity on both rate and distortion aspects. Since our quality management module shown in Fig. 2 is analogous to general video coding schemes, we partially adopt modeling approaches practiced in video coding literature on the one hand [28], [29]. On the other hand, we adopt another tack since these approaches are limited to model every aspect of our scheme; they are designed to model video data. We show our models are close to actual results, while general models in video coding fail to follow actual results. Our models subsequently enable us to develop the optimal storage configuration strategy in the next section.

A. Data Fidelity Model: Rate

While the size of data can be controlled by adjusting QP at the quantization process in Fig. 2, it can also be controlled by adjusting the granularity in temporal domain, which is equivalent to the temporal quality adjustment. Fig. 4 shows the temporal coding structure of our spatio-temporal decorrelation

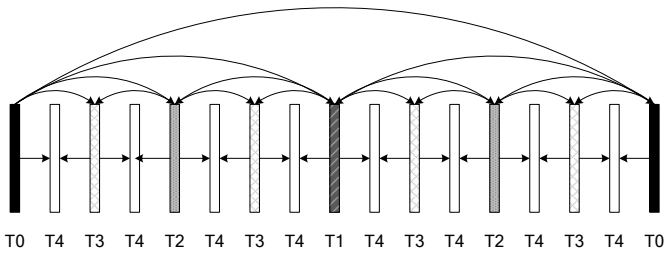


Fig. 4. Temporal coding and prediction structure of our spatio-temporal decorrelation module. Each increasing temporal level coincides with including collections of sensor data labeled the corresponding level, i.e., Ti.

module. There are total five temporal levels shown in Fig. 4, where each increasing temporal level corresponds to a double of frequency at which collections of sensor data at certain time instance are included in coded data set. Here, the range of temporal levels can be extended or reduced depending on applications, which entails modification of the temporal coding and prediction structure. (Without loss of generality, we use the structure shown in Fig. 4 throughout the paper.)

As an example, the temporal level 3 will include collections of sensor data labeled T0, T1, T2, and T3 in Fig. 4. And the highest temporal level 4 shall contain all of data sampled in line with temporal dimension.

Fig. 4 also displays the temporal prediction structure shown by arrows, which exploits strong temporal correlation. Since the prediction of a certain level only involves the lower temporal level collections, adjusting temporal granularity is made possible.

It is quite intuitive to reckon that the size of compressed data block R is reduced by half as the temporal level decreases by one step. (In general, the rate R denotes the number of bits per symbol [30]. We extend its notion to represent a data block size which is a basic unit in our study.) However, due to the temporal prediction structure shown in Fig. 4, the amount of reduction becomes less than half per one temporal level decrease. We can model this relation as

$$R = \alpha(\Delta) \cdot \exp(\beta(\Delta) \cdot T), \tag{1}$$

where $\alpha(\Delta)$ and $\beta(\Delta)$ are model parameters dependent on the quantization step size Δ, and $T \in \{0, 1, 2, 3, 4\}$ denotes the temporal level.

In (1), two model parameters $\alpha(\Delta)$ and $\beta(\Delta)$ have to be estimated from real data based on the quantization step size. Since the quantization step size is directly related to a degree to which a data block is compressed, R is inversely proportional to Δ, which should be reflected on the model parameters given by

$$\alpha(\Delta) = a_\alpha \exp(b_\alpha \Delta) + c_\alpha \exp(d_\alpha \Delta), \tag{2}$$

$$\beta(\Delta) = a_\beta \exp(b_\beta \Delta) + c_\beta, \tag{3}$$

where a_α, b_α, c_α, and d_α are data-dependent constants supplementary to $\alpha(\Delta)$ in (1), and similarly, a_β, b_β, and c_β are constants for $\beta(\Delta)$ in (1). It should be noted that in (2) a mixture of two exponential functions is used to model long-tail shape of $\alpha(\Delta)$ in (1).

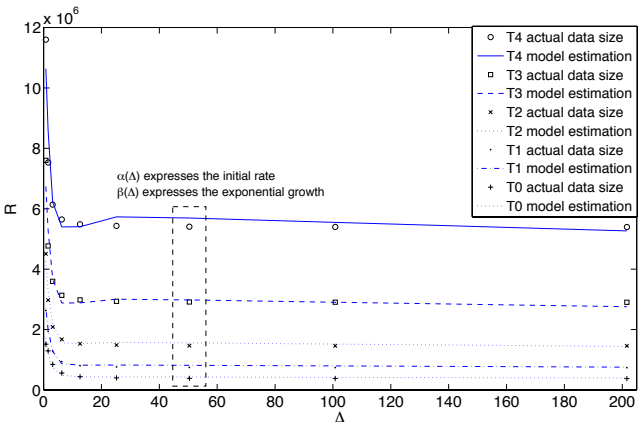


Fig. 5. Sizes of compressed data blocks R as functions of quantization step sizes Δ for different temporal levels T estimated by (1).

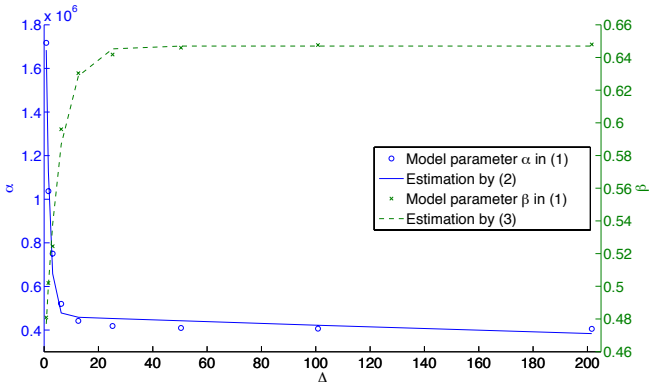


Fig. 6. Two model parameters in (1) as functions of quantization step sizes Δ estimated by (2) and (3).

Combining (2) and (3) with (1), we can represent the total rate as a function of both the quantization step and the temporal level. The resulting model function is plotted in Fig. 5, where five lines represent each temporal level and actual data points are also plotted for comparison. We can confirm the model effectively follows the varying size of actual sensor data.

In Fig. 5, (1) in terms of T is explained along the vertical axis. Fig. 6 shows two model parameters in (1) and their model estimations by (2) and (3).

B. Data Fidelity Model: Distortion

In addition to the rate modeling discussed above, we can estimate the distortion of data due to the quantization as well. Here we represent the distortion in terms of mean squared error (MSE) measure. As more quantization is applied at the quantization process, data fidelity is more decreased, i.e., increased distortion, which is reflected by

$$D_{\text{quant}} = a_{\text{quant}} \cdot \exp(b_{\text{quant}} \cdot QP) + c_{\text{quant}}, \tag{4}$$

where a_{quant}, b_{quant}, and c_{quant} are data-dependent constants. It should be noted that (4) is a function of QP, whose relationship with the quantization step size Δ is expressed

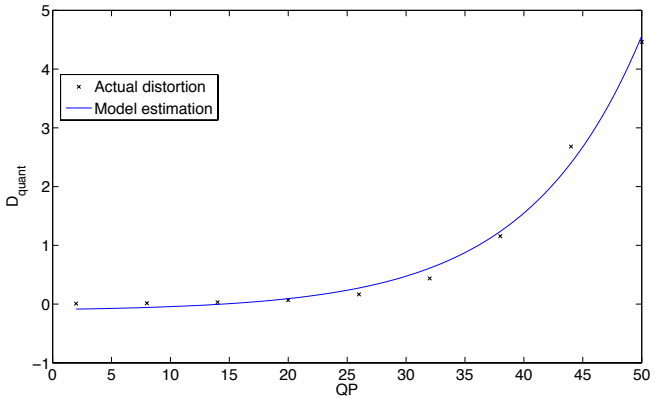


Fig. 7. Distortion curve as a function of QP estimated by (4).

by $\Delta = 0.625 \cdot 2^{QP/6}$ [29].[3] Fig. 7 shows actual distortion points and their approximation using (4).

Although (4) effectively models the distortion caused by quantization, the source of distortion is not limited to the quantization. As the temporal level T varies, the amount of sampled data along temporal dimension varies as well, which causes another distortion. Recalling the temporal coding structure shown in Fig. 4, as T decreases by one step, half of data are excluded from data set. This leads to the condition that omitted data should be estimated using previous data samples. As a result, the total distortion increases as T decreases.

In order to incorporate the temporal distortion into total distortion, we assume that temporal distortion is measured by the mismatch between actual data samples and omitted data samples that are replaced by previous data samples. Although the combination of these two different types of distortion seems tightly coupled, they can be separated as proved in Theorem 2. We first prove the error summation property in the following lemma.

Lemma 1: The total error of the quality management module can be expressed by sum of the quantization error and the temporal omission error.

Proof: In order to better understand how errors are introduced in our archiving scheme, we can model its operating scenario in Fig. 8 concerning errors. In Fig. 8, e_{L} denotes the quantization error and e_{T} the temporal omission error. As shown in Fig. 8, these two errors are from two different distortion sources and are independent.[4] In this scenario, the total error e_{total} is written as

$$e_{\mathrm{total}} = \widehat{x} - x = (\widehat{x} - \hat{x}) + (\hat{x} - x) = e_{\mathrm{T}} + e_{\mathrm{L}}, \qquad (5)$$

where x, $\hat{x}$, and $\widehat{x}$ denote raw sensor data, quantized data, and temporally omitted data, respectively. ■

Using this result, we are now ready to prove the separation property.

[3] QP can cover more range than Δ does. Using QP instead of Δ here is just a matter of better fitting using constants.

[4] In fact, the quantization process affects the quality of data that are subsequently used for the estimation of omitted data in temporal dimension. However, we have empirically found the independence can be assumed in most cases without loss of generality.

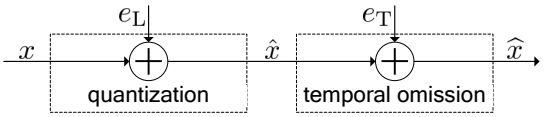


Fig. 8. Block diagram showing errors in quality management module.

Theorem 2 (Separation Property): The joint distortion D_{total} caused by the quantization from lossy coding and the omission of data samples along temporal dimension is separable and can be expressed by sum of both distortions.

Proof: Without loss of generality, we assume an arbitrary probability density function (pdf) of temporal omission error between actual data samples and reconstructed data samples, in which missing samples are covered by previous existing data samples. This pdf is denoted by $f_{E_{\mathrm{T}}}(e_{\mathrm{T}})$, where random variable E_{T} represents the temporal omission error.

It is well known that the pdf of quantization error from lossy coding is approximately uniform as follows [23]:

$$f_{E_{\mathrm{L}}}(e_{\mathrm{L}}) = \begin{cases} \frac{1}{\Delta} & -\frac{\Delta}{2} \le e_{\mathrm{L}} \le +\frac{\Delta}{2} \\ 0 & \text{otherwise} \end{cases}, \qquad (6)$$

where E_{L} is a random variable that denotes the quantization error.

We can express D_{total} using joint distribution:

$$D_{\mathrm{total}} = \iint_{e_{\mathrm{T}} e_{\mathrm{L}}} f_{E_{\mathrm{T}} E_{\mathrm{L}}}(e_{\mathrm{T}}, e_{\mathrm{L}}) \cdot e_{\mathrm{total}}^2 \, \mathrm{d}e_{\mathrm{L}} \, \mathrm{d}e_{\mathrm{T}}. \qquad (7)$$

Using the result in Lemma 1, we have

$$\begin{aligned} D_{\mathrm{total}} &= \iint_{e_{\mathrm{T}} e_{\mathrm{L}}} f_{E_{\mathrm{T}}}(e_{\mathrm{T}}) f_{E_{\mathrm{L}}}(e_{\mathrm{L}}) \cdot (e_{\mathrm{T}} + e_{\mathrm{L}})^2 \, \mathrm{d}e_{\mathrm{L}} \, \mathrm{d}e_{\mathrm{T}} \\ &= \int_{-\infty}^{\infty} f_{E_{\mathrm{T}}}(e_{\mathrm{T}}) \frac{1}{\Delta} \int_{-\Delta/2}^{+\Delta/2} (e_{\mathrm{T}} + e_{\mathrm{L}})^2 \, \mathrm{d}e_{\mathrm{L}} \, \mathrm{d}e_{\mathrm{T}}, \end{aligned} \qquad (8)$$

which continues in

$$\begin{aligned} D_{\mathrm{total}} &= \int_{-\infty}^{\infty} f_{E_{\mathrm{T}}}(e_{\mathrm{T}}) \left(e_{\mathrm{T}}^2 + \frac{\Delta^2}{12} \right) \mathrm{d}e_{\mathrm{T}} \\ &\approx \int_{-\infty}^{\infty} f_{E_{\mathrm{T}}}(e_{\mathrm{T}}) \cdot e_{\mathrm{T}}^2 \, \mathrm{d}e_{\mathrm{T}} + \frac{\Delta^2}{\beta}, \end{aligned} \qquad (9)$$

where β is a denominator which is 12 for a small Δ and larger than 12 in case of a larger Δ compared to the signal variance. This is because when the quantization step size Δ becomes large, quantization errors can no longer be treated as uniformly distributed [28].

In the right-hand side of (9), the first term is a distortion from the temporal omission error and the second term is a distortion from the quantization error. ■

In the above theorem, we assume an arbitrary pdf $f_{E_{\mathrm{T}}}(e_{\mathrm{T}})$ that illustrates the distribution of the temporal omission error. An empirical finding of this distribution is given in Appendix. Using Theorem 2, the total distortion is simply expressed by summing distortions from two different sources, which shortly will be proved as a useful property for modeling distortion.

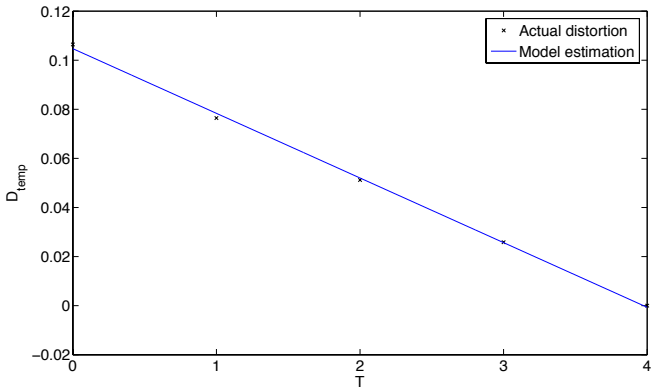


Fig. 9. Temporal distortion as a function of T estimated by (10).

Now we turn to the problem of estimating the temporal distortion model. Specifically, we have empirically found that the temporal distortion D_{temp} is a linear function of the temporal level T, which is given by

$$D_{\text{temp}} = a_{\text{temp}} \cdot T + b_{\text{temp}}, \tag{10}$$

where a_{temp} and b_{temp} are constants. The accuracy of (10) can also be verified by Fig. 9. The linearity of temporal distortion is attributed to the temporal coding structure shown in Fig. 4 where each increasing temporal level corresponds to a double of sensor data collections. (If we represented the temporal distortion as a function of the number of sensor data collection instead, we would have a convex function.)

Thanks to the separation property proven in Theorem 2, we can combine both distortions in (4) and (10) to yield the joint distortion D_{total} as follows:

$$\begin{aligned} D_{\text{total}}(QP,T) &= D_{\text{quant}} + D_{\text{temp}} \\ &= a_{\text{quant}} \exp(b_{\text{quant}} QP) \\ &\quad + a_{\text{temp}} T + a_{\text{total}}, \end{aligned} \tag{11}$$

where c_{quant} in (4) and b_{temp} in (10) are absorbed into one constant.

C. QP-Rate-Distortion Model

We now discuss the accuracy of our analytical model. Thus far, we have discussed the relationship between QP, temporal level, distortion, and rate, i.e., compressed data size. If we express the relationship without temporal level, we obtain the results shown in Fig. 10a, where the temporal change is implied in the variation of the rate, given a particular QP. The actual QP-Rate-Distortion surface graph is also shown in Fig. 10b for comparison. In Fig. 10, we can identify our model estimation is close to the actual result, which was confirmed for two other types of data as well.

It is difficult to model our quality-adjustable archiving scheme using general rate-distortion models. For instance, a well-established modeling of rate and distortion for DCT-based video encoder is [28], [31]

$$D(\Delta) = \frac{\Delta^2}{\beta}, \qquad R(\Delta) = \frac{1}{2}\log_2\left(\frac{\epsilon^2 \beta \sigma_x^2}{\Delta^2}\right), \tag{12}$$

where β is identical to β in (9) that is 12 for a small Δ and larger than 12 in case of a larger Δ compared to the variance of the source σ_x^2, and ϵ^2 is dependent on a source distribution [28].

In (12), β needs to be empirically adjusted to account for a wider range of Δ. However modeling our scheme with (12) yields discouraging results as shown in Fig. 11. In Fig. 11a, β was adjusted according to the actual distortion, which leads to the result identical to the actual distortion curve. On the contrary, the rate modeling of (12) with obtained β is very far from the actual rate, as shown in Fig. 11b. Furthermore, (12) has no provision for data fidelity control over temporal dimension, in contrast to our analytical model. Thus it is imperative that an accurate analytical model is used in order to derive the optimal storage configuration strategy.

V. Optimal Rate Allocation

A. Rate Allocation Strategy

Using the analytical model derived in Section IV, our next concern is how to find the minimum distortion with a given specific rate R_0. We first consider an optimal rate allocation problem of single sensor data block, which can be formulated as follows:

$$\begin{array}{ll} \min\limits_{\{QP,T\}} & D_{\text{total}}(QP,T) \\ \text{s.t.} & R(QP,T) \le R_0 \end{array}, \tag{13}$$

where $D_{\text{total}}(QP,T)$ and $R(QP,T)$ is the distortion and the rate function derived in (11) and (1), respectively.

Fig. 12 shows the surface graph of $D_{\text{total}}(QP,T)$ derived in (11), where 10 contour plots, which are isolines of rate, are drawn together over the surface to reveal contours of same rate over varying distortion. In Fig. 12, we can see that the minimum distortion is obtained along the boundary of QP and T. Specifically, when there is available rate, it has to be first spent on reducing QP, and only after arriving at the minimum QP can the rate be spent on increasing the temporal level.

This allocation strategy can also be explained by deriving the gradient of the distortion function, which is given by

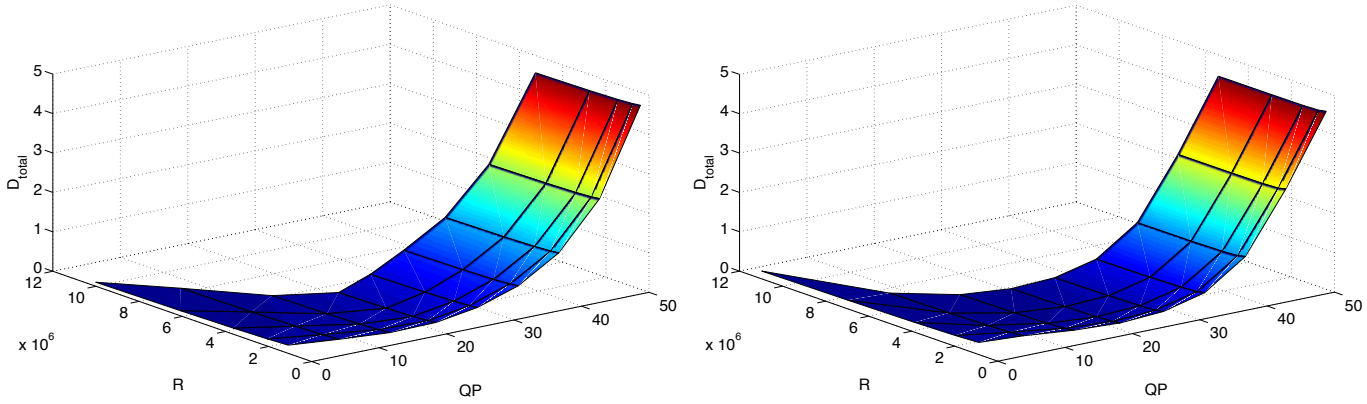
$$\nabla D_{\text{total}}(QP,T) = (a_{\text{quant}} b_{\text{quant}} \mathrm{e}^{b_{\text{quant}} QP}, a_{\text{temp}}). \tag{14}$$

In (14), the magnitude of a_{temp} is much smaller than that of the QP component of the gradient, which means it is more advantageous to adjust QP than temporal level in order to reach the minimum distortion quickly.

B. Optimal Storage Configuration

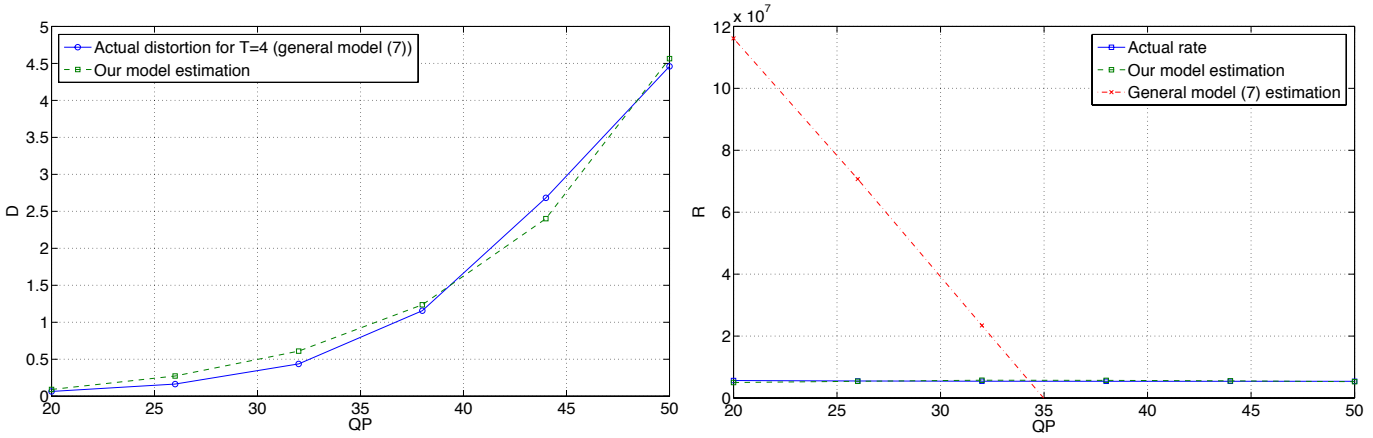
We can furthermore extend the rate allocation problem of single sensor data block to accommodate more general case of storage configuration problem where multiple data blocks have to be stored efficiently. As explained in Section III-B, our scheme supports five supplementary layers that facilitates graceful degradation of data quality. Specifically in Fig. 3, the temporal level T is gradually decreased as a data block ages.

Considering total storage efficiency, we are interested in how to allocate storage to each fidelity cluster and how to determine QP of each data block. Since each data block occupies less storage space in lower fidelity clusters than



(a) model surface (b) actual surface

Fig. 10. QP-Rate-Distortion surfaces of ambient temperature data set. Temporal change is implied in the variation of rate. Other sensor data types show similar surfaces.



(a) distortion curves comparison

(b) rate curves comparison (rate modeling of (12) yields negative values after $QP = 35$)

Fig. 11. Comparison of our model and general model with actual rate-distortion curves.

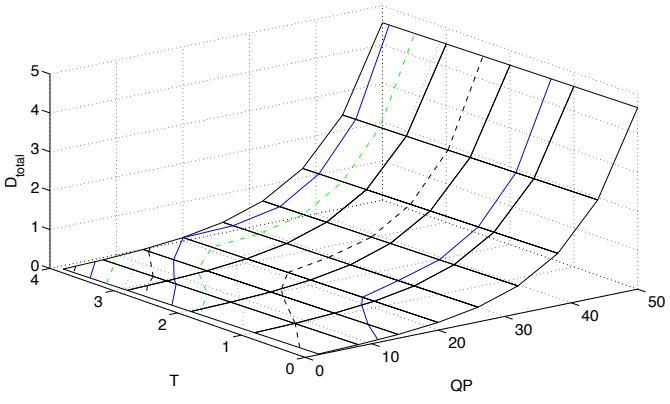


Fig. 12. Isolines of rate over distortion surface of ambient temperature data set. Minimum distortion is obtained along the boundary of QP and T.

higher fidelity clusters, lower fidelity clusters can hold more data blocks given the same capacity. Besides, it is more natural to retain lower fidelity data longer than higher fidelity data. Assuming single sensor data type, the optimal storage configuration problem can then be formulated as follows:

$$\begin{aligned} \min_{\{QP_i, R_j\}} \quad & \sum_{j=0}^{4} \phi_j \sum_{i=1}^{N} D_{\text{total}}(QP_i, j) \\ \text{s.t.} \quad & \phi_j \sum_{i=1}^{N} R(QP_i, j) \le R_j \\ & \sum_{j=0}^{4} R_j \le R_{\text{total}} \\ & \phi_0 \gg \phi_1 > \phi_2 > \phi_3 > \phi_4 = 1 \end{aligned} \quad , \qquad (15)$$

where QP_i denotes QP of each data block, N is the number of data blocks in the cluster 4, and ϕ_j is a natural number denoting the proportion of data block numbers with respect to N. This equation describes a storage configuration at a certain instant where data blocks in lower fidelity clusters inherited QP's from data blocks in higher fidelity clusters. When the total rate budget R_{total} is given, the optimal storage configuration should yield the overall minimum distortion.

The solution to (15) is an equal QP for each data block such that $\sum_{j=0}^{4} R_j \le R_{\text{total}}$, which no longer constrains ϕ_j to be

a natural number: ϕ_j could be any positive rational number not less than 1. Hence the relationship between R_j's is given by

$$\frac{R_j}{R_i} = \frac{\phi_j}{\phi_i} \cdot \exp(\beta(\Delta) \cdot (j-i)) \quad (j \geq i). \tag{16}$$

Note that N and ϕ_j are system parameters that can be appropriately adjusted according to the target duration of retaining sensor data for each cluster.

The same result applies to a case when there are multiple sensor data types: an equal QP for each data block between the same type. However different sensor data types imply different model parameters, which leads to different QP's for different data types. In particular, the relationship between two different sensor data types using QP_{A} and QP_{B} is represented as follows:

$$\frac{\sum_{j=0}^{4} \phi_j D'_{\mathrm{total_A}}(QP_{\mathrm{A}}, j)}{\sum_{j=0}^{4} \phi_j R'_{\mathrm{A}}(QP_{\mathrm{A}}, j)} = \frac{\sum_{j=0}^{4} \phi_j D'_{\mathrm{total_B}}(QP_{\mathrm{B}}, j)}{\sum_{j=0}^{4} \phi_j R'_{\mathrm{B}}(QP_{\mathrm{B}}, j)}, \tag{17}$$

where we used separate distortion and rate function for each QP. In (17), the ratio of the weighted sum of distortion slopes for each temporal level to the weighted sum of rate slopes for each temporal level is fixed. This result is another case of constant slope optimization [32], [33]: we obtain same marginal return for an extra rate spent on either sensor data type.

Utilizing the results, the optimal storage configuration strategy is first to determine proper QP's for each sensor data type in proportion to available storage, and then to encode sensing data input with the maximum temporal level. As time elapses, aged data blocks are shifted to next lower clusters till they reach to the cluster 0. The gradually decreasing fidelity of sensor data with this scheme results in an efficient management of storage space.

VI. Experimental Results

A. Compression Efficiency

In order to suggest the efficiency of our scheme, we compared the compression ratios of popular lossless and lossy coding methods with our scalable data archiving scheme. We used data sets downloaded from the Sensorscope website, which has various WSN deployment scenarios that are mostly environmental data samples [34]. The results convince us that our scheme is a viable solution for archiving huge amount of sensor data.

1) Comparison with Lossless Coding: Lossless coding is ideal for applications that cannot tolerate any difference between the original and reconstructed data. Popular lossless coding schemes that are used in experiments for comparison with our scheme are as follows: *gzip*, based on the combination of LZ77 and Huffman coding [35]; *bzip2*, based on the combination of Burrows-Wheeler transform, move-to-front transform, and Huffman coding [36]; *PPMd*, an optimized implementation of prediction by partial matching (PPM) algorithm [37]; Lempel-Ziv-Markov chain algorithm (LZMA), used in *7-Zip* [38]. These state-of-the-art schemes work well with text and data files.

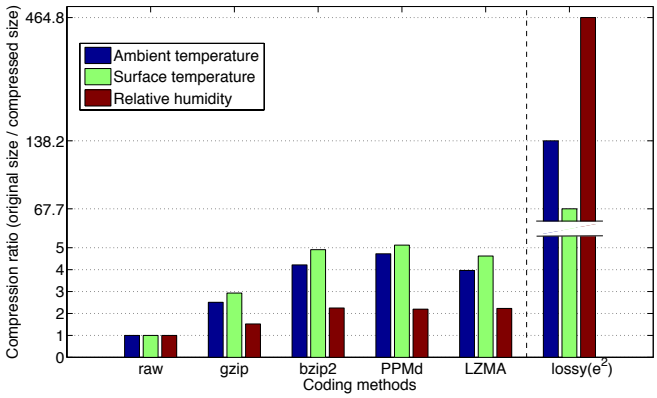


Fig. 13. Compression ratios of our archiving scheme compared with various lossless coding methods. Our scheme allows distortion up to the sensing accuracy.

TABLE I
Sensor Accuracy and Type for Three Data Types [39]

Data Type	Accuracy	Sensor Type
Ambient Temperature (AT)	$\pm 0.3°\mathrm{C}$	Sensirion SHT75
Surface Temperature (ST)	$\pm 0.3°\mathrm{C}$	
Relative Humidity (RH)	$\pm 2\%$	

Fig. 13 shows the compression ratios of various schemes that are expressed by the original raw data size divided by the compressed size. Although the compressed size can be as small as how much we allow distortion, it might be unfair to directly compare lossy coding with lossless coding in terms of coding efficiency. Hence we set out a reference point for distortion, which is the sensing accuracy explained in Section II-A. Table I shows sensor types and their accuracies that correspond to the sensor error margin e. Despite an impressive result shown in Fig. 13, total distortion incurred is comparable to the order of sensor error margin e^2 in terms of MSE distortion measure.

2) Comparison with Lossy Coding of Partial Correlation: In many applications such as the sensor data archiving, we can relax the requirement of a reconstruction to be identical to the original. Lossy coding promises much higher compression ratios than the lossless coding does at the cost of decreased data fidelity. One can adjust the data fidelity depending on a desired quality of the reconstructed data: our archiving scheme accomplishes this through the quantization and the temporal quality adjustment. The lossy coding has been conventionally employed to compress multimedia data such as image and video. We adopt the lossy coding for the sensor data archiving thanks to the quality adjustability of sensor data.

The quality adjustability and the utilization of both spatial and temporal correlations culminate in outstanding compression efficiency as shown in Fig. 14, where our scheme contrasts with wavelet coding methods with partial correlation [10]–[12]. Wavelet coding is another popular lossy coding method apart from DCT-based coding: it is well known that the performance of wavelet-based and DCT-based codings is almost the same [40]. The compression ratio shown in Fig. 14

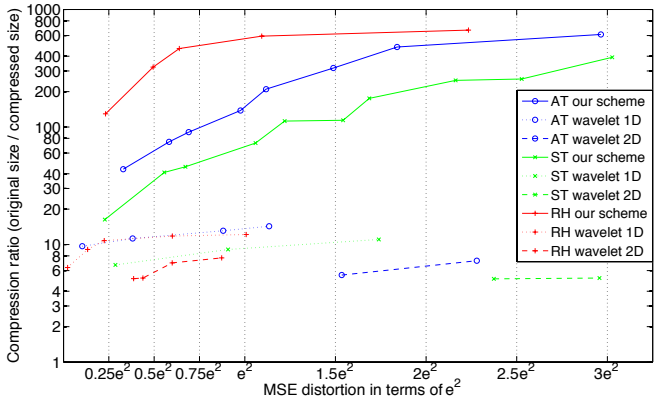


Fig. 14. Log-scale compression ratios of our archiving scheme compared with wavelet-based methods using partial correlations at various data fidelities in terms of the sensing accuracy. Acronyms are explained in Table I.

juxtaposes a consequence of restricting the use of correlation to either the spatial dimension or the temporal dimension: the wavelet 1D only exploits temporal correlation for signal compaction, whereas the wavelet 2D only exploits spatial correlation for signal compaction. After signal compaction, both methods apply threshold, quantization and entropy encode processes for the lossy compaction of signal. Between both wavelet-based methods, the wavelet 1D shows better results than the wavelet 2D, thanks to the stronger correlation in the temporal domain than the spatial domain.

B. Storage Efficiency

Although the solutions to (15) are optimal in analytical sense, we further want to show their optimality for selecting actual operating points of our archiving scheme. Given N, ϕ_j, and R_{total}, we first find the optimal QP's for each sensor data type using our analytical model, then actual operating points corresponding to the optimal QP's are selected to give overall distortion. We compare this system-wide distortion with other selection criteria: (i) uniform selection of arbitrary QP's even in the same sensor types; (ii) equal QP's for the same sensor types, but ignoring their relationship in (17).

Experimental results are shown in Table II, where all of three storage configuration strategies occupy the same storage space. However they exhibit dramatic differences in terms of the system-wide distortion: the *uniform QP selection strategy* is the worst as expected, the *equal QP for the same sensor types strategy* shows better result, but neither of two strategies is comparable to *our optimal configuration strategy*. In other words, we spend the same amount of storage space for poorer overall data fidelity, which is equivalent to maintaining the same quality of data blocks while spending more amount of storage space.

Table II also shows varying distortion ratios depending on the transient duration that denotes how long the transient clusters, i.e., cluster 1, 2, and 3, hold data blocks with respect to the duration of the cluster 4.[5] In particular, parameters for each transient duration are as follows: (i) $\phi_0 = 10$, $\phi_1 = 2.5$, $\phi_2 = 2$, $\phi_3 = 1.5$, $\phi_4 = 1$ (short); (ii) $\phi_0 = 10$, $\phi_1 = 4$, $\phi_2 = 3$, $\phi_3 = 2$, $\phi_4 = 1$ (medium); (iii) $\phi_0 = 10$, $\phi_1 = 7$, $\phi_2 = 5$, $\phi_3 = 3$, $\phi_4 = 1$ (long). In Table II, we can identify that both distortion ratios increase as the transient duration increases.

[5]Obviously, there is no principle of deciding that a given duration is long or short. This is a relative measure depending on the system parameters N and ϕ_j.

TABLE II
DISTORTION RATIOS OF THREE STRATEGIES NORMALIZED BY OUR STRATEGY ($N = 10$)

Optimal	Uniform QP	Equal QP	Transient Duration
1	8.04	5.38	short
1	8.39	5.59	medium
1	8.91	5.91	long

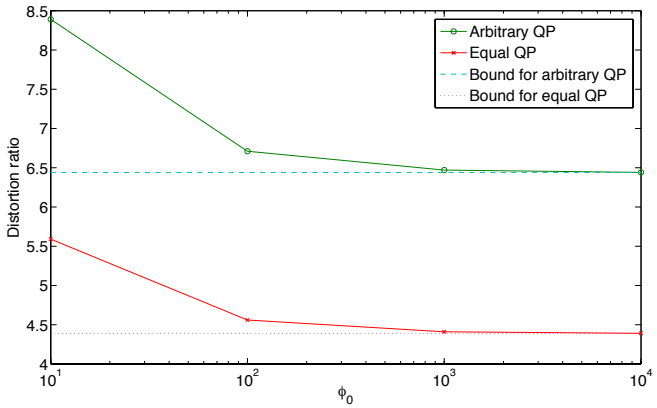


Fig. 15. Distortion ratios with with varying sizes of the cluster 0. The ratio of our optimal configuration strategy, 1, is not shown.

We are also interested in the change of distortion ratios as the number of archival data blocks in the cluster 0 increases. As long as these blocks are to be permanently archived (or at least archived for a long time), the proportion ϕ_0 will keep increasing, representing an increasing portion of the cluster 0. Fig. 15 shows how distortion ratios change with respect to increasing ϕ_0. As ϕ_0 increases, archival data blocks in the cluster 0 will dominate the overall distortion. Thus distortion ratios will be eventually bounded by results solely taking account of the cluster 0, which is also shown in Fig. 15. Although Fig. 15 shows the tendency of decreasing distortion ratios as ϕ_0 increases, we can conclude that inevitable differences exist between the optimal and suboptimal configuration strategies.

Since the results in Table II and Fig. 15 are distortion ratios normalized by our optimal distortion, cumulative distortion will increase as N increases to practical values for storage configuration. This result clearly shows the importance of the optimal storage configuration that has to be derived from proper analytical models; otherwise storage space would be wasted.

VII. CONCLUSION

We have proposed a new archiving scheme for big sensor data that leverages the three key characteristics of typical sensor data such as quality adjustability, data aging, and spatio-temporal correlation. Our lossy coding scheme allows a

significant saving of storage space without compromising key features of sensor data. In addition, quality of sensor data is gracefully degraded over time to relinquish the storage space.

When numerous data blocks from various sensor types are stored, storage should be optimally managed to maximize the space saving for given data fidelity. To this end, we derived analytical models that reflect the characteristics of our lossy coding scheme and solved the optimal storage configuration problem using these models. Experimental results showed significant savings of storage space and the optimality of our storage configuration strategy.

APPENDIX
DISTRIBUTION OF TEMPORAL OMISSION ERROR

We can model the distribution of the temporal omission error using the mixture of the *Dirac delta function* and *Laplacian distribution*, which is an example of *zero-inflated model* [41].

Let p denotes an inflation term that indicates point mass at zero, then the rest of probability mass $(1-p)$ can be represented using the pdf of Laplacian. This zero-inflated Laplacian distribution is given by

$$f_{E_{\mathrm{T}}}(e_{\mathrm{T}}) = \begin{cases} p \cdot \delta(e_{\mathrm{T}}) & e_{\mathrm{T}} = 0 \\ (1-p) \cdot \frac{\lambda}{2} \exp(-\lambda |e_{\mathrm{T}}|) & e_{\mathrm{T}} \neq 0 \end{cases}, \quad \text{(A.18)}$$

where λ is the shape parameter of Laplacian distribution. We can identify that (A.18) follows the actual distributions properly in Fig. 16 where $f_{E_{\mathrm{T}}}(e_{\mathrm{T}})$ was drawn over the histogram of error between actual and omitted data samples.

Since the mean of $f_{E_{\mathrm{T}}}(e_{\mathrm{T}})$ is zero, its variance $\sigma^2_{E_{\mathrm{T}}}$ is equivalent to the distortion from the temporal omission error. The range of errors between actual and omitted data samples that are replaced by previous data samples is widened as more data samples are dropped along the temporal dimension, which equates to decreasing p in (A.18).

REFERENCES

[1] D. Lee and J. Choi, "Low complexity sensing for big spatio-temporal data," in *Proc. Int'l Conf. Big Data (BigData '14)*, 2014, pp. 323–328.

[2] ——, "Learning compressive sensing models for big spatio-temporal data," in *Proc. Intl. Conf. Data Min. (SDM '15)*, 2015.

[3] A. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a service and big data," in *Proc. Int'l Conf. Adv. Cloud Comput. (ACC '12)*, 2012, pp. 21–29.

[4] C. Perera, A. Zaslavsky, C. H. Liu, M. Compton, P. Christen, and D. Georgakopoulos, "Sensor search techniques for sensing as a service architecture for the internet of things," *IEEE Sens. J.*, vol. 14, no. 2, pp. 406–420, Feb. 2014.

[5] J. F. Roddick, E. Hoel, M. J. Egenhofer, D. Papadias, and B. Salzberg, "Spatial, temporal and spatio-temporal databases - hot issues and directions for PhD research," *SIGMOD Rec.*, vol. 33, no. 2, pp. 126–131, Jun. 2004.

[6] P. Ranganathan, "From microprocessors to nanostores: Rethinking data-centric systems," *IEEE Computer*, vol. 44, no. 1, pp. 39–48, Jan. 2011.

[7] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva, "The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011," White Paper, IDC, Mar. 2008.

[8] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, Apr. 2011.

[9] T. Palpanas, M. Vlachos, E. Keogh, D. Gunopulos, and W. Truppel, "Online amnesic approximation of streaming time series," in *Proc. Int'l Conf. Data Eng. (ICDE '04)*, 2004, pp. 339–349.

[10] D. Ganesan, D. Estrin, and J. Heidemann, "Dimensions: Why do we need a new data handling architecture for sensor networks?" *SIGCOMM Comput. Commun. Rev.*, vol. 33, no. 1, pp. 143–148, Jan. 2003.

[11] D. Ganesan, B. Greenstein, D. Estrin, J. Heidemann, and R. Govindan, "Multiresolution storage and search in sensor networks," *Trans. Storage*, vol. 1, no. 3, pp. 277–315, Aug. 2005.

[12] Y.-C. Wang, Y.-Y. Hsieh, and Y.-C. Tseng, "Multiresolution spatial and temporal coding in a wireless sensor network for long-term monitoring applications," *IEEE Trans. Comput.*, vol. 58, no. 6, pp. 827–838, Jun. 2009.

[13] C. M. Sadler and M. Martonosi, "Data compression algorithms for energy-constrained devices in delay tolerant networks," in *Proc. Int'l Conf. Embedded Networked Sensor Syst. (SenSys '06)*, 2006, pp. 265–278.

[14] K. C. Barr and K. Asanović, "Energy-aware lossless data compression," *ACM Trans. Comput. Syst.*, vol. 24, no. 3, pp. 250–291, Aug. 2006.

[15] F. Marcelloni and M. Vecchio, "Enabling energy-efficient and lossy-aware data compression in wireless sensor networks by multi-objective evolutionary optimization," *Inf. Sci.*, vol. 180, no. 10, pp. 1924–1941, May 2010.

[16] T. Srisooksai, K. Keamarungsi, P. Lamsrichan, and K. Araki, "Practical data compression in wireless sensor networks: A survey," *J. Netw. Comput. Appl.*, vol. 35, no. 1, pp. 37–59, Jan. 2012.

[17] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger, "Architecture support for disciplined approximate programming," in *Proc. Int'l Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS '12)*, 2012, pp. 301–312.

[18] A. Sampson, J. Nelson, K. Strauss, and L. Ceze, "Approximate storage in solid-state memories," in *Proc. Int'l Symp. Microarchitecture (MICRO '46)*, 2013, pp. 25–36.

[19] E. Cohen and H. Kaplan, "Aging through cascaded caches: Performance issues in the distribution of web content," in *Proc. SIGCOMM '01*, 2001, pp. 41–53.

[20] M. Palmer, "Seven principles of effective RFID data management," Progress Software, 2005.

[21] H. Cao, O. Wolfson, and G. Trajcevski, "Spatio-temporal data reduction with deterministic error bounds," *VLDB J.*, vol. 15, no. 3, pp. 211–228, Sep. 2006.

[22] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Comput. Netw.*, vol. 45, no. 3, pp. 245–259, Jun. 2004.

[23] K. Sayood, *Introduction to Data Compression*, 4th ed. Morgan Kaufmann, 2012.

[24] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974.

[25] S. K. Raman, V. Pentkovski, and J. Keshava, "Implementing streaming SIMD extensions on the Pentium III processor," *IEEE Micro*, vol. 20, pp. 47–57, Jul./Aug. 2000.

[26] Y.-K. Chen, E. Q. Li, X. Zhou, and S. Ge, "Implementation of H.264 encoder and decoder on personal computers," *J. Vis. Commun. Image Represent.*, vol. 17, no. 2, pp. 509–532, Apr. 2006.

[27] M. Wien, R. Cazoulat, A. Graffunder, A. Hutter, and P. Amon, "Real-time system for adaptive video streaming based on SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1227–1237, Sep. 2007.

[28] H.-M. Hang and J.-J. Chen, "Source model for transform video coder and its application. I. Fundamental theory," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287–298, Apr. 1997.

[29] *Coding of Audiovisual Objects - Part 10: Advanced Video Coding*, ISO/IEC 14496-10 and ITU-T Recommendation H.264, 2003.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[31] M. Dai, D. Loguinov, and H. M. Radha, "Rate-distortion analysis and quality control in scalable Internet streaming," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1135–1146, Dec. 2006.

[32] H. Everett III, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Oper. Res.*, vol. 11, no. 3, pp. 399–417, May/Jun. 1963.

[33] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.

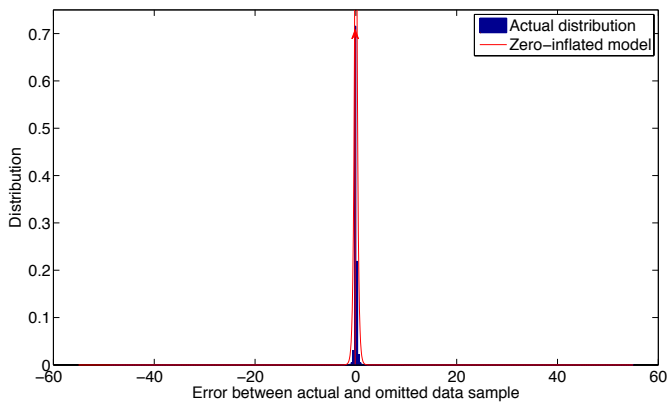
[34] Sensorscope: Sensor networks for environmental monitoring. [Online]. Available: http://lcav.epfl.ch/op/edit/sensorscope-en

[35] gzip. [Online]. Available: http://www.gzip.org

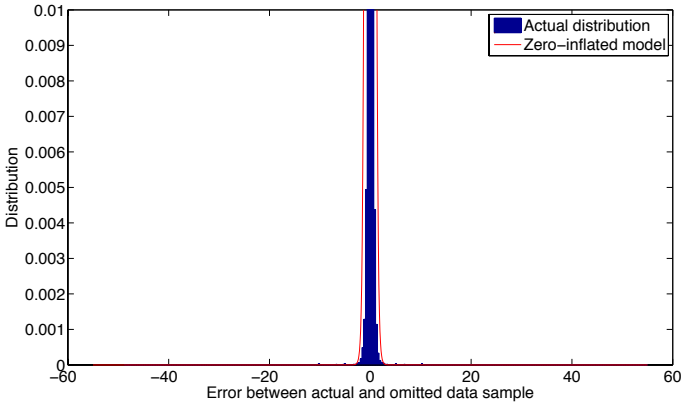
[36] bzip2. [Online]. Available: http://www.bzip.org

[37] A. Moffat, "Implementing the PPM data compression scheme," *IEEE Trans. Commun.*, vol. 38, no. 11, pp. 1917–1921, Nov. 1990.

[38] 7-zip. [Online]. Available: http://www.7-zip.org



(a) distribution showing point mass at 0 (red arrow)



(b) magnified distribution showing details

Fig. 16. Distribution of error fitted with zero-inflated Laplacian distribution.

[39] Sensirion. [Online]. Available: http://www.sensirion.com/en/home/

[40] Z. Xiong, K. Ramchandran, M. T. Orchard, and Y.-Q. Zhang, "A comparative study of DCT-and wavelet-based image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 5, pp. 692–695, Aug. 1999.

[41] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, Feb. 1992.