IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received December 24, 2018, accepted January 20, 2019, date of publication February 4, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896784

Lossless Data Compression for Communication Systems Based on Optical Frequency Discriminator

GUANGJUN SHAN
School of Information Engineering, East China University of Technology, Nanchang 330013, China
School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: acosmos@126.com

ABSTRACT We propose a novel lossless data compression scheme for a high-speed communication system. Previous works have usually addressed the data compression issue within an identical plane by pure mathematical processing. Here, we convert the binary information into a novel sparse expression form in which the information is divided into two parts: time sequence information and difference value information, which are sent to the destination by two distinct means of transmission. Furthermore, by a deliberately designed symbol called the semantic wavelength, we can select a better compression scheme among candidates. It is expected that our scheme can reach 13.8% data compression ratio.

INDEX TERMS Lossless data compression, compressive sensing, Huffman coding, semantic wavelength, optical frequency Discriminator.

I. INTRODUCTION

The traditional lossless data compression (LDC) schemes can be categorized into three board tactics, data redundancy tactics such as run-length encoding (RLE) [1], [2], probability statistics tactics (entropy coding tactics) such as Shannon-Fano coding [3] and Huffman coding [4], and dictionary-based tactics such as Lempel Ziv coding [5].

Run-length encoding is probably the simplest method of compression. The general idea behind this method is to replace consecutive repeating occurrences of a symbol by one occurrence of the symbol followed by the number of occurrences. For example, AAABBBCCCDDDD can be compressed into RLE format as A3B3C3D4. The method works more efficiently with two symbols (for example 0 and 1) if one symbol is much more frequent than the other. However, its defects are also obvious, if there are no successive repeating values in the data, RLE merely contributes to a very low compression ratio. Usually Run-length encoding can be used in conjunction with other methods to play a positive role in data compression [6], [7].

Shannon-Fano coding ranks symbols in the order of occurrence frequencies from high to low, then divides into two sets whose sum of probabilities are as close as possible to being equal. Huffman coding uses the logic of Shannon-Fano coding, but ranks the symbols in the order of occurrence frequencies from low to high, and assigns shorter codes to symbols that occur more frequently and longer codes to those that occur less frequently. Consequently, Huffman code outperforms Shannon-Fano coding. An alternative application scenario is that the Huffman coding is not directly used to compress the data information but is used as an optimization method to assign the index information bits for subcarrier index modulation orthogonal frequency division multiplexing (SIM-OFDM) [8]. As another form of entropy coding, arithmetic coding [9] also follows the same tactics in which the frequently seen symbols are encoded with fewer bits than rarely seen symbols, it performs slightly more efficiently than Huffman coding, but the implementation is more difficult. Variable length coding (VLC) tactics is widely used in entropy coding schemes, not only Huffman VLC coding [10], but also other coding schemes such as the Unary coding [11], Golomb–Rice (GR) coding [12], [13], Exponential Golomb coding [14], [15], and Context-based Adaptive Binary Arithmetic Coding (CABAC) [16]. Huffman VLC, Unary coding, GR coding, and Exp-GR coding are simple and regular. However, their compression rates are unsatisfactory. CABAC has a superior compression efficiency,

The associate editor coordinating the review of this manuscript and approving it for publication was Zhao Zhang.

but unfortunately, its complexity is too large and the hardware architecture is difficult to implement due to the high irregularity. Moreover, assuming that the occurrences of symbols appear with equal probability, entropy coding methods lose their advantage.

Lempel Ziv (LZ) coding is an example of a category of algorithms called dictionary-based coding. The concept is to create a dictionary of strings used during the communication session. If both the sender and the receiver have a copy of the dictionary, then previously-encountered strings can be substituted by their index in the dictionary to reduce the amount of information transmitted. The LZ algorithm was first proposed in 1977, multiple variations were subsequently generated, such as LZ78 [17], and Lempel-Ziv-Welch (LZW) [18]. Popular compressors such as zip and Unix's compression software are based on the LZW method. In the cases of human languages and images, the LZW algorithm can achieve a compression ratio from 85% to 25%, depending on the statistical characteristics of compressive objectives. However, in terms of communication systems, an optical communication system (for instance), would have a seemingly random arrival of binary digits. It is unlikely that there would exist certain sematic regularities like in human languages, or image characteristics, which can be extracted for compression. Consequently, the effectiveness of category of dictionary-based coding algorithms for communication systems is questionable. On the other hand, as the string table of LZW is dynamical, the algorithm has to add every new string it sees to the string table during the process of compression, therefore the compression speed is too slow to catch up the bit rate in a high-speed communication system. So far, we have not found any report stating that the dictionary-based coding algorithms might be applicable to a communication system.

Besides the above-mentioned LDC schemes, compressive sensing (CS) [19]–[21] is a new methodology of data compression which relies on two assumptions regarding signal's characteristics: sparsity and incoherence. In principle, any signal $f \in \mathbb{R}^n$ has the sparsity when expressed in the proper representation basis Ψ. By the sensing basis Φ, there exists $y = \Psi\Phi f$, where $y \in \mathbb{R}^m$, $m \ll n$, $\Psi\Phi$ is a $m \times n$ sensing matrix, and y is the sparse representation of f under $\Psi\Phi$ [21]. Candès *et al.* [19] proved that when $\Psi\Phi$ meets the RIP (restricted isometry property) conditions, signal f can be recovered from y with an extremely high probability, thus it is questionable whether CS technology is applicable to high-speed communication systems without any data loss. Furthermore, finding a proper representation basis and sensing matrix is not an easy job, in principle, it is a NP-hard issue and results in extensive time-consumption [22], which means CS is not practical for LDC in real-time communication systems.

In this paper, we propose an innovative LDC scheme, which uses an optical frequency discriminator to divide the original bit stream into sparse expressions. At the same time, a semantic symbol, which consists of an individual wavelength, guarantees the practicability of the scheme as we can switch back to conventional transmission methods once our LDC scheme encounters trouble through this symbol's instruction. This scheme is suitable for high-speed communication systems, especially optical communication systems.

The remainder of the paper is organized as follows: in section II, we illustrate how our scheme transforms the bit streams into the sparse representations. In section III, we show the deliberately designed optical device – optical frequency discriminator, which transmits the information in a completely different way, and this is the first time such a device proposed for information transmission. In section IV, we introduce the importance of semantic wavelength. Finally, the performance of our scheme is discussed and analyzed in section V.

II. TIME SEQUENCES INFORMATION (TSI) AND DIFFERENTIAL VALUE INFORMATION (DVI)

One of the fundamental requirements underlying CS is that the signal must be sparse under a certain representation basis. However, not all signals have a sparse representation. Unlike image signals or other signals in which certain regularity is implicated, communication systems usually transmit binary digital 0 and 1, which is random and time-variant, thus it is difficult to find a sparse representation basis for a time-variant random signal.

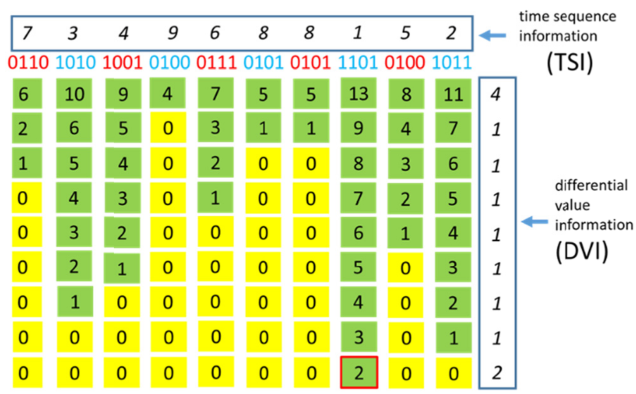

FIGURE 1. Segmentation of bit stream.

To address this issue, we transform the binary stream of communication system into two parts, time sequences information (TSI) and differential value information (DVI). Fig.1 gives an illustrative example in which a bit string which consists of 40 bits. Correspondingly, we may state that the bit string length (BSL) is equal to 40.

Assume the 40 bits string is given by

$$0110101010010100011101010101110101001011$$

We may segment this by 4 bits bundling, that is,

$$0110(6),\ 1010(10),\ 1001(9),\ 0100(4),\ 0111(7),\ 0101(5),\ 0101(5),\ 1101(13),\ 0100(8),\ 1011(11)$$

where the digits included within the brackets are the decimal equivalents of the binary counterparts. It is worth mentioning that the representation of decimal is not necessary in term of machine processing, this is just for the sake of reader's understanding. Instead of the original binary string, our scheme transmits the transformed TSI and DVI from source (***S***) to destination (***D***), where ***D*** is able to reconstruct the original binary string by the rules pre-specified by us, which can indeed simplify the data representation without any information loss.

To obtain the TSI, we sort the segmentation blocks (SB) in descending (or ascending) order. In the case of Fig.1, the values of SB (VSB) are

6, 10, 9, 4, 7, 5, 5, 13, 8, 11

Correspondingly, the TSI is

7, 3, 4, 9, 6, 8, 8, 1, 5, 2

It is clear that the TSI is the information that indicates the descending order of SB values. We note that the TSI has two identical elements 8 because there are two identical values 5 in SB. In principle, the more identical elements TSI has, the better compression ratio can be achieved.

To obtain the DVI, we note the minimum non-zero digital from the row consisting of VSBs, and subtract this digit from every digits of this row to produce the new row. Repeating this step until we get such a row in which all digits are zero except a single non-zero number, or multiple non-zero numbers which are equal each other. Once a certain digit is reduced to zero, it is not further reduced and remains zero whilst other rows are created. We note these differential values (the minuend) of each row, which make up the DVI, the non-zero digital of the last row will be directly recorded into the DVI, such as the red-circled 2 in Fig.1. **It is clear the DVI consists of the minimum non-zero digital(s) of each row**. In Fig. 1, the DVI is

4, 1, 1, 1, 1, 1, 1, 1, 2

Once the ***D*** receives TSI and DVI, it will be able to reconstruct the entire original binary string. The TSI includes the order information that DVI can use to complete the data reconstruction step by step.

For example, in Fig.1,

TSI: 7, 3, 4, 9, 6, 8, 8, 1, 5, 2
DVI:4, 1, 1, 1, 1, 1, 1, 1, 2

The elements of the DVI should be noted from the bottom up in Fig.1, or from left to right in Fig.2. The TSI implies the ninth (the lowest bottom in Fig.1) element of the DVI, which is 2, should be placed at eighth position (starting from left-hand side) of the ninth row as shown in Fig.1, then the eighth element of the DVI, which is 1, should be placed at tenth position of eighth row. Concurrently, the existing non-zero element 2 will be summed with 1, to provide the 3 in the eighth row. Similarly, in the seventh row, we should fill seventh element of the DVI, which is 1, to the second position in the seventh row, and existing non-zero elements 1 and 3

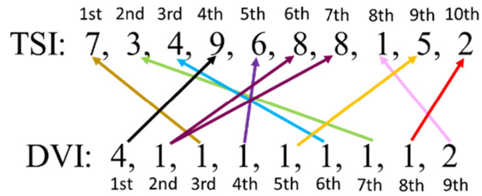

FIGURE 2. Association between TSI and DVI.

will be summed with 1, becoming 2 and 4 in the seventh row. Doing so step by step we will create the entire original binary string in the first row as shown in the Fig.1:

$$6, 10, 9, 4, 7, 5, 5, 13, 8, 11 \rightarrow 0110101010010100011101010101110101001011$$

Although we can reconstruct the original bit string by TSI and DVI, the compression effect has still not been disclosed so far. It is very difficult to express the TSI and DVI within an identical plane (manner in which the data is modulated, transmitted, detected, and recovered by the Nyquist-Shannon principles) while achieving the efficient data compression. Finally, we are aware of that we should use the two incoherent manners to transmit the TSI and DVI for efficient data compression. For this purpose, we design the optical frequency discriminator (OFD) in order to convey the TSI from ***S*** to ***D*** by a different manner.

III. OPTICAL FEQUENCY DISCRIMINATOR (OFD)

The DVI will be transmitted from ***S*** to ***D*** by conventional manner, e.g. fiber channel, which receives the information by sampling (analog signal) or by intensity detection (digital signal). However, it will be difficult to achieve the purpose of data compression if the TSI is also transmitted by the above-mentioned manner. We illustrate as such:

We consider the number of possible values that the TSI could take. This is a permutation and combination problem. In the case of Fig.1, there are 10 SB blocks, we define the total possible values of TSI to be ***R***, since the value of TSI is allowed to occur repeatedly if two or more SB blocks have the same value, thus

$$\boldsymbol{R} = 10^{10} = 10000000000, \quad \log_2 R = 33.2$$

Consequently, 34 bits are required to represent the TSI. This is clearly not an economic method of conveying the TSI. We consider the adoption of an alternative method to the permutation and combination method, for example, assigning bits to represent each element of TSI in a straight way. Taking

$$(7, 3, 4, 9, 6, 8, 8, 1, 5, 2)$$

as an example, at least 30 bits will be required, as we have not taken the delimiter between elements into account. Thus, we can conclude that it is very difficult to achieve the purpose of data compression by conventional processes.

To address this issue, we design an optical device named optical frequency discriminator (OFD), which is placed at the ***D*** side.

DMUX: DMUS means the optical demultiplexer, its function is to receive a beam consisting of multiple wavelengths from a fiber and separate it into individual wavelength components (λ_m, $m = 1 \ldots n$) as they are coupled into the fiber in ***S*** side.

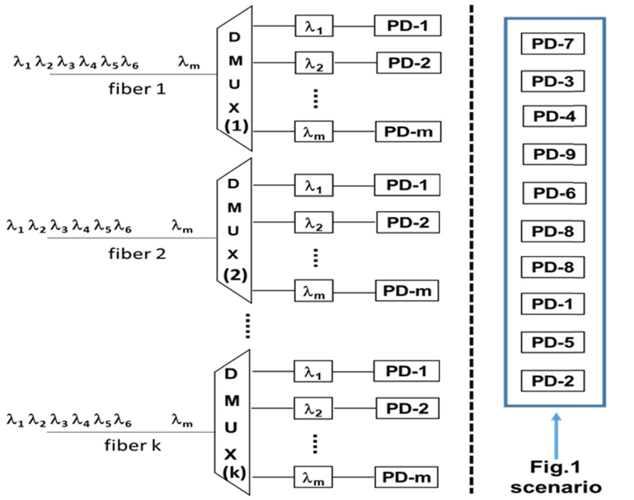

FIGURE 3. Optical frequency discriminator (OFD).

The left half of Fig.3 shows the structure of the OFD, which consists of a series of DMUX. Behind each DMUX, different wavelengths of lightwaves will be refracted to different spatial location and detected by photodiodes (PDs). In short, **the function of the OFD is to convert the time sequence information to the spatial location information.** In the example of Fig.1, $k = 10$ (where k is determined by the number of segmentation blocks), but $m = 9$ (where m is the number of different elements included in the TSI, $0 < m \leq k$). Consequently, we are required to employ 10 DMUXs and 9 wavelengths. In general, as m gets smaller and smaller than k, the effect of data compression will improve. In the ***S*** side, we send the 7th, 3rd, 4th, 9th, 6th, 8th, 8th, 1st, 5th, 2nd wavelengths into the fiber k labeled from 1 to 10 correspondingly. In the ***D*** side, DMUXs will refract the wavelengths to different ports connecting with PDs. The right half of Fig.3 shows the result in which only

PD-7, PD-3, PD-4, PD-9, PD-6, PD-8, PD-8, PD-1, PD-5, PD-2

will produce the photocurrents thereby forming the specific spatial arrangement. In principle, this mechanism can work for any length of bit string and bits bundling, e.g. 250 bits string with 5 bits bundling. Thus, we complete the transmission of TSI by OFD.

IV. SEMANTIC WAVELENGTH

An important fundamental of communication systems is that there is a common rule (or set, thereof) which is recognized and obeyed by ***S*** and ***D***. Sometimes we may name this rule *certain prior knowledge*.

We investigate the issue of transmission of DVI. In the example of Fig. 1, the DVI is

$$4, 1, 1, 1, 1, 1, 1, 1, 2$$

A noticeable fact is that any element of DVI would not be equal to zero, thus we can replace the bit 0 for numerical 1, and bit 1 for numerical 2. The maximum numerical element of the DVI is 4, thus we assign two bits "11" to represent 4. We consider the type of rule which should be formulated for the transmission of the DVI so that to assure that ***D*** can exactly distinguish each element and the boundary between two elements. In general, we have to assign two bits for each element of the DVI due to the bottleneck effect, such as "00" for 1, "01" for 2, so that ***D*** can accurately identify each element without any confusion. Otherwise, we have to introduce a delimiter to indicate the boundary between two elements, obviously this is not an efficient approach. However, assigning two bits in the DVI is also not an efficient approach as, except 4, other elements such as 1 and 2 can actually be represented only by one bit, 0 or 1.

Another noticeable fact is that with the increase of bit string length (BSL), the element of DVI will be no more than 2 with a considerably high probability. This is due to the fact that the VSBs will most likely be closer each other as the BSL increases, furthermore some of them will inevitably be equal. In most cases, we can only assign one bit to each element of DVI. However, although the element of DVI will be limited to 1 or 2 with a high-probability, that does not mean it is a deterministic event. Therefore, when formulating the rule, we cannot assign only one bit for each element of DVI if there is no other counterplan preparing for the occurrence of VSBs are far away from each other.

As a counterplan addressing this problem, we introduce a concept of semantic wavelength as shown in Fig.4.

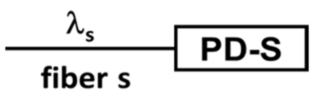

FIGURE 4. Semantic Wavelength, where λ_s is an individual lightwave being specified as the semantic wavelength, and PD-S is a photodiode of semantic wavelength.

The semantic wavelength implies a semantic symbol, this symbol is represented in the form of photocurrent producing from PD-S. Thus, depending on whether PD-S outputs electrical current or not, we will implement the compressive scheme or conventional uncompressed scheme correspondingly. In most cases, the element of DVI will not exceed 2, and we only assign one bit to each element of DVI, we need not use the semantic wavelength and the PD-S does not product any electrical current. IF ***S*** perceives a certain element of DVI which exceeds 2, the lightwave of semantic wavelength will be coupled into the fiber s at ***S*** side, and ***S*** will switch to the conventional scheme for data transmission. When PD-S detects the lightwave and produces the electrical current, $\boldsymbol{D}$ can know the rule has been changed at $\boldsymbol{S}$ side, then it will recover the coming data by the conventional scheme.

V. DISCUSSION AND ANALYSIS

Our scheme segments the bit string into bit bundlings, which consist of a uniform number of bits. By subtracting the minimum element in each row step by step, we finally covert the original bit stream into two parts, TSI and DVI. Furthermore, we arrange a semantic symbol in order to switch back to the conventional transmission in the rare case where elements of DVI differ greatly in value. The TSI will be conveyed by the OFD device from $\boldsymbol{S}$ to $\boldsymbol{D}$, DVI will be transmitted in the form of binary bit stream but in which binary "0" represents for numerical 1, and binary "1" represents for numerical 2. This set of data transformation rules consists of the *prior knowledge*, which is recognized and obeyed by both $\boldsymbol{S}$ and $\boldsymbol{D}$ jointly, and $\boldsymbol{D}$ reconstructs the original bit string from the received TSI and DVI by the inverse transformation. Comparing with uncompressed bit stream, TSI and DVI consume less resource.

We must highlight that the most important aspect is the design idea of OFD. Besides optical carrier, other forms of carriers such as by wireless, or by copper are also applicable to this design idea. **This is the first time such a scheme has been proposed for information transmission. Unlike the conventional scheme in which the information is modulated and transmitted by an individual channel, frequency discriminator uses the order relations of channels to convey information.**

Although OFD has an addition cost with optical fiber resource, in this scheme, it still can save the resources relative to the conventional scheme in which the whole information is transmits within a fiber. For example, as analyzed in section 3 for the example of Fig.1, the conventional scheme would cost around 40 bits. Using the OFD, assuming we complete the light intensity detection within one-bit time, it only costs 10 bits. And we can further reduce the bit consumption if we increase the number of bit bundling (NBB), for example, 8 bit bundling per 40 bits, thus it will only cost 5 bits. Of course, on the other hand, this may increase the probability that elements of DVI are more than 2. Thus, it is a trade-off between NBB and BSL.

It is possible that someone may illustrate the example of WDM (Wavelength Division Multiplexing) to interrogate the resource utilization, as an optical fiber can accommodate multiple wavelengths. We propose the reciprocal approach to resolve this issue. We know that in optical communication systems, a transmitted bit of 0 does not mean that the lightwave does not exist, but exists at a lesser intensity to that of a transmitted bit of 1. A reciprocal approach is that we cut off the lightwave so that the corresponding photodiode will not output the photocurrent at all. Different from the OFD design scheme stated in section III where the output of photocurrent is used to convey the TSI, the reciprocal approach conveys the TSI by without output of photocurrent thus other wavelengths of this fiber still can be used for data transmission.

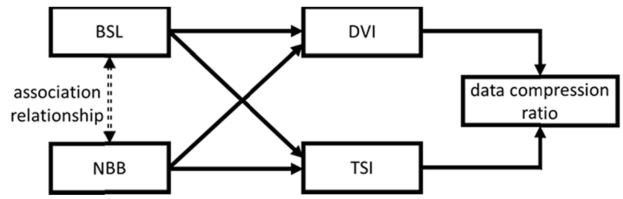

FIGURE 5. Trade-off between NBB and BSL.

Fig.5 illustrates the relationship of BSL and NBB and how that affects the compression rate. It is clear the NBB cannot be arbitrarily increased and set as it is constrained by BSL. Usually we hope $\frac{BSL}{NBB}$ is an integer. When the BSL is fixed, the increase of NBB would reduce the employed lightwaves, thus we can achieve a higher compression. However, the DVIs will be increased with the increase of NBB, and eventually exceed 2, which will lead to a more frequent switch to the conventional scheme.

If we fix the NBB, and increase the BSL, the likelihood of equal or aggregated VSBs will increase, thus we would only be required to assign one bit to each element of the DVI and we would be able to decrease the number of elements of DVI. Both of above aspects are useful for improving the compression ratio. However, increasing the BSL will result in more lightwaves being used (part of lightwaves have the same wavelengths), which leads to the decrease of compression rate.

In addition, we cannot increase the BSL without any restriction as it will increase the complexity of the OFD. Similarly, we cannot arbitrarily increase the NBB without any restriction because the number of wavelengths that can be accommodated in a fiber is limited. Denoting the number of wavelengths as N_W, where $N_W = 2^{NBB}$, we would normally restrict N_W to $N_W \leq 320$,therefore we have the restriction, $NBB \leq 8$.

We now consider the data compression ratio. We denote k to be the number of segmentation blocks, and thus, k = BSL/NBB. Furthermore, k also represents the number of DMUX used in the OFD, as the number of segmentation blocks and number of DMUS are equal. We denote the number of lightwaves as N_L, where multiple lightwaves may have the same wavelength. We denote the number of elements of DVI as N_D, where N_D is a statistical average given in Table 1.

We denote $\rho = \frac{\text{BSL}}{NBB*2^{NBB}}$ which can affect the values of elements of DVI. When $\rho \geq 3$, values of elements of DVI will be likely equal to 1 or 2 with a considerably high probability, thus the semantic wavelength is rarely activated.

Define the data compression ratio as $\gamma = \frac{data\ lenght\ after\ compression}{orginal\ data\ length}$, therefore $\gamma = \frac{N_L+N_D}{BSL}$, because N_L indicates the number of lightwaves which is employed for the transmission of TSI that will be completed within one-bit time. Thus it can be understood there are N_L bits used for the transmission of TSI. In most case, each element of DVI can be represented by one-bit, thus N_D indicates how many bits used for transmission of DVI, it is an estimated value in the Table 1, and leaves large space. Considering that we estimate

TABLE 1. Comparison of compression ratio.

NBB	BSL	k	N_L	N_W	N_D	ρ	γ
2	40	20	21	5	3	5	60%
4	40	10	11	17	15	0.625	78.1%
4	192	48	49	17	15	3	33.3%
4	256	64	65	17	15	4	31.2%
4	320	80	81	17	15	5	30%
4	384	96	97	17	15	6	29.2%
6	1152	192	193	65	63	3	22.2%
6	1536	256	257	65	63	4	20.8%
8	6144	768	769	257	255	3	16.7%
8	8192	1024	1025	257	255	4	15.7%
8	20480	2560	2561	257	255	10	13.8%

the N_D based on the assumption that all DVIs are 1, but in practice, part of DVIs will be equal to 2, that means actually N_D can be further cut down. Thus, $N_L + N_D$ bits can fairly represent the compressed data length.

We note that, N_L= k + 1, as we must take the semantic wavelength into account.

Once the NBB and BSL are fixed, all others parameters and γ are decided. When NBB=2, and BSL=40, we have a fairly simple OFD structure where our scheme can achieve a 60% compression ratio. In the example of Fig.1, γ becomes worse at only 78.1%(where γ will not follow the calculation formula $\frac{N_L+N_D}{BSL}$). This is due to the fact that $\rho = 0.625$ is quite small, which means that the semantic wavelength will often be used and the system will have to switch back and forth between the compression scheme and the uncompressed conventional scheme. At $\rho = 3$, γ rapidly improves to 33.3%, which means that the data is compressed to one-third. However, when we increase ρ to 4, 5, and 6, γ is only slightly improved. As shown in Table 1, an increase of ρ to 6, only results an improvement of 4.1%, from 33.3% to 29.2%, for γ, while OFD must use more DMUXs. If we let $\rho \geq 3$, with an increase of NBB, then γ will rise steadily. Assuming that NBB=8, and BSL=20480, the γ can reach 13.8%. However, in this case, the OFD would have to employ 2560 DMUXs and a great deal of PDs. Unless the OFD is highly integrated, we would not suggest the adoption of such an intensive configuration. In terms of computational complexity, our scheme is only proportional to N_D, this is quite rapid in comparison to the CS technique.

Table 2 gives the comparison of compression ratio between our scheme (NBB=4, BSL=192) and other traditional methods mentioned in the introduction. Four types of data, which are taken from a doc file, a jpg file, a pseudo random sequence (PRS) generator, and a rar file, are used as the benchmarks for comparison. Since the RLE coding is only efficient when the data consists of many consecutively repeating symbols, it is unnecessary to investigate its efficiency in terms of the jpg, PRS and rar. The efficiency of Huffman

TABLE 2. Compression ratio comparison of different schemes.

	doc	jpg	pseudo random sequence (PRS)	rar
RLE coding	53.2%			
Huffman coding	47.1%	93.7%	95.7%	103.1%
LZ coding	40.9%	97.4%	80.7%	100%
Our scheme	29.2%	33.9%	33.3%	35.5%

coding is highly related to the probability distribution of symbols. When the probability distribution of symbols is quite uneven, Huffman coding works well with 47.1% for doc data. However, its efficiency drops as the symbol occurrence probability approaches the uniform distribution. We note the values of 95.7% for PRS, and 103.1% for rar which is even larger than the original data size. LZ coding performs better than Huffman coding in terms of PRS data because it uses the dictionary-based coding mechanism. However, the randomness of symbol occurrence probability of PRS leads to the unsatisfactory compression rate of only 80.7% for PRS. Our scheme not only outperforms the other three coding methods in the aspect of compression rate, 29.2% for doc, 33.9% for jpg. 33.3% for PRS, and 35.5% for rar, but also is less sensitive to the statistical characteristics of the data. This is because we use the data transformation and OFD hardware to divide the original bit stream into sparse expressions, which makes our scheme independent of symbol distribution probability or symbol dictionary.

Given that an OFD device consists of k parallel fibers and m wavelengths per fiber, if do not take WDM effect into account, the total possibilities it can represent is m^k. While for a binary system with k fibers, the total possibilities it can represent is 2^k, thus the symbol ratio between the two systems is k(binary system): k $\log_2 m$(OFD), namely,

$$1(\text{binary system}): \log_2 m(\text{OFD}) \tag{1}$$

That means the transmission capacity of OFD is $\log_2 m$ times that of binary system.

If we take the WDM effect into account, the total possibilities the binary system can represent is $2^{m \bullet k}$. The symbol ratio between the two systems becomes $\log_2 2^{\mathrm{m} \bullet \mathrm{k}}$(binary system): k $\log_2 m$(OFD), namely,

$$\text{m}(\text{binary system}): \log_2 m(\text{OFD}) \tag{2}$$

That means the information transmission capacity of WDM binary system will outperform the OFD. However, it is noteworthy that OFD provides an alternative representation of information transmission, which is based on the permutation and combination of wavelength and spatial position, rather than the conventional intensity detection method that has been widely using in binary coding transmission systems. Thus, the proposed OFD has heuristic meaning

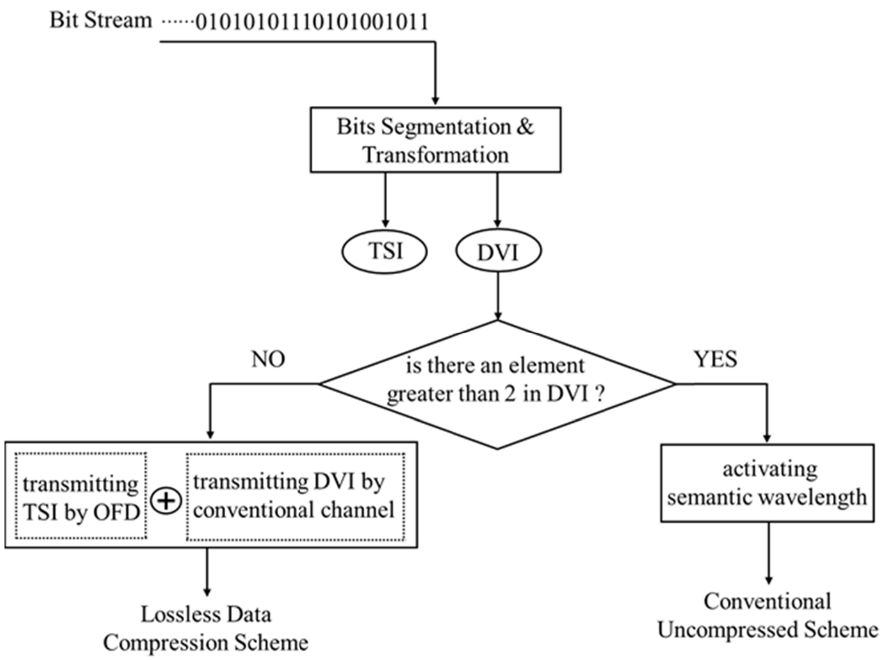

FIGURE 6. Flow chart of the main process steps regarding the proposed scheme.

for information transmission, especially for those scenarios where the carrier frequency cannot be applied to modulation.

In the context of binary, ternary (such as three levels pulse amplitude modulation (PAM)), decimal and hexadecimal systems, the symbol could be either one of values of 2, 3, 10, 16, thus their symbols' degrees of freedom (DOF) are 2, 3, 10, 16, respectively. Depending on the respective representations, the entropy values of the above four systems can be represented as:

$$H_2(X) = \sum_{i=1}^{n} p_i \log_2 p_i^{-1} \tag{3}$$

$$H_3(X) = \sum_{i=1}^{n} p_i \log_3 p_i^{-1} \tag{4}$$

$$H_{10}(X) = \sum_{i=1}^{n} p_i \log_{10} p_i^{-1} \tag{5}$$

$$H_{16}(X) = \sum_{i=1}^{n} p_i \log_{16} p_i^{-1} \tag{6}$$

For the above-mentioned OFD device, if we treat the k PDs' outputs as a complete symbol, then the DOF of OFD's symbol is m^k, denoting m^k as d, then the entropy of OFD is:

$$H_d(X) = \sum_{i=1}^{n} p_i \log_d p_i^{-1} \tag{7}$$

The following conclusion can be derived:

If the DOF of a symbol is d, then its information transmission capacity will be $\log_2$ d times that of binary counterpart.

By the permutation and combination of frequency and its spatial position, OFD greatly extends the DOF of symbol so that it exhibits the stronger transmission capacity than ordinary binary fiber transmission. In particular, by the reciprocal approach, the proposed scheme can not only keep the WDM function effective, but also realize $\frac{1}{\log_2 m}$ compression ratio in terms of TSI part according to the equation (1). For example, assume m = 16, then NBB=4, roughly speaking, TSI part can obtain 25% compression ratio relative to conventional binary transmission method. When BSL=192, considering the occupied resource by DVI part, 33.3% compression rate as Table 2 shows is a well-founded result. But on the other hand, as Fig.5 shows, we could not arbitrarily increase the value NBB because there is a tradeoff between BSL and NBB, if we increase NBB regardless of BSL, the compression effect of DVI part will worsen.

Finally, it worth mentioning that the semantic wavelength can actually represent three symbols, "0" (low level), "1" (high level) and "none" (without output), thus it can be used to indicate the position of those elements of DVI which exceed 2 by combinations of "none", "0" and "1". This means that semantic wavelengths can play a more significant role and we would not have to frequently switch back and forth between the two schemes. Developing more functions of semantic wavelength can greatly simplify the structure of OFD as we need not increase the BSL while keeping the compression ratio in a reasonable range (even if certain elements of the DVIs exceed 2). However, as the specific scheme description would involve a great deal of trivial details, we omit from this paper for brevity. We would study this issue of semantic wavelength in future work.

Fig.6 illustrates the main steps described in section II, III and IV, reader can refer to this flow chart to review the proposed LDC scheme.

VI. CONCLUSION

In this paper, we proposed a hardware-based data compression scheme, which can achieve a fair compression ratio depending on the structure of the OFD. Moreover, a hardware device – OFD which can sparsely represent TSI, was proposed in this paper. As far as we know, this is the first time that such a device has been proposed. It is expected the well-designed semantic wavelength can play a more important role in the next step. In comparison with traditional LDC schemes such as data redundancy, probability statistics and dictionary-based tactics, our scheme demonstrated the universality and practicality for the high-speed communication systems. Relative to the CS technique which is hard to realize the lossless data compression, our scheme not only achieved the lossless data compression, also the low computational complexity.

ACKNOWLEDGEMENT

The author would like to thank Dr Alan Walker of the University of the West of Scotland (UWS) for his help in revising the grammar and format of this paper. He would also like to acknowledge funding support from EU Smartlink project.

REFERENCES

[1] A. H. Robinson and C. Cherry, "Results of a prototype television bandwidth compression scheme," *Proc. IEEE*, vol. 55, no. 3, pp. 356–364, Mar. 1967.

[2] C. M. Agulhari, I. S. Bonatti, and P. L. D. Peres, "An Adaptive Run Length Encoding method for the compression of electrocardiograms," *Med. Eng. Phys.*, vol. 35, no. 2, pp. 145–153, Feb. 2013.

[3] R. Fano, "The transmission of information," Res. Lab. Electron., Cambridge, MA, USA, Tech. Rep. 65, Mar. 1949.

[4] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. Inst. Radio Eng.*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[5] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. IT-23, no. 3, pp. 337–343, May 1977.

[6] B. Dai *et al.*, "Data compression for time-stretch imaging based on differential detection and run-length encoding," *J. Lightw. Technol.*, vol. 35, no. 23, pp. 5098–5104, Dec. 1, 2017.

[7] A. L. Liaghati, W. D. Pan, and Z. Jiang, "Biased run-length coding of bilevel classification label maps of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 10, pp. 4580–4588, Oct. 2017.

[8] M. I. Kadir, H. Zhang, S. Chen, and L. Hanzo, "Entropy coding aided adaptive subcarrier-index modulated OFDM," *IEEE Access*, vol. 6, pp. 7739–7752, 2018.

[9] P. G. Howard and J. S. Vitter, "Arithmetic coding for data compression," *Proc. IEEE*, vol. 82, no. 6, pp. 857–865, Jun. 1994.

[10] S. Li, H. Yin, X. Fang, and H. Lu, "Lossless image compression algorithm and hardware architecture for bandwidth reduction of external memory," *IET Image Process.*, vol. 11, no. 6, pp. 379–388, 2017.

[11] S. Kak, "Generalized Unary Coding," *Circuits, Syst. Signal Process.*, vol. 35, no. 4, pp. 1419–1426, Apr. 2016.

[12] S. Golomb, "On the survival of sequence information in filters (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 2, pp. 310–312, Mar. 1972.

[13] R. F. Rice, "Some practical universal noiseless coding techniques, part 3, module PSl14,K+," Jet Propuls. Lab., Pasadena, CA, USA, Tech. Rep. JPL-91, 1991, vol. 3, pp. 1509–1520.

[14] R. Sugiura, Y. Kamamoto, N. Harada, and T. Moriya, "Optimal Golomb-Rice code extension for lossless coding of low-entropy exponentially distributed sources," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 3153–3161, Apr. 2018.

[15] G. Gong, T. Helleseth, and P. V. Kumar, "Solomon W. Golomb—Mathematician, engineer, and pioneer," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2844–2857, Apr. 2018.

[16] J. Heo and Y.-S. Ho, "Improved context-based adaptive binary arithmetic coding over H.264/AVC for lossless depth map coding," *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 835–838, Oct. 2010.

[17] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 5, pp. 530–536, Sep. 1978.

[18] T. A. Welch, "A technique for high-performance data compression," *Computer*, vol. 17, no. 6, pp. 8–19, Jun. 1984.

[19] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[22] Z. Qin, J. Fan, Y. Liu, Y. Gao, and G. Y. Li, "Sparse representation for wireless communications: A compressive sensing approach," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 40–58, May 2018.

GUANGJUN SHAN received the B.S. degree in optoelectronics engineering, the M.S. degree in electronics and information engineering, and the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China. His research interests include the optical and wireless communication systems, and information and communication theory.