Data-coding Approaches for Organizing Omni-ambient Data

Seishi Takamura

Abstract

Various information sensors are currently deployed to generate data such as images, video, audio, and temperature. The amount of such multi-modal data is rapidly increasing compared to the development of information and communication technology (such as storage, transmission, and processing technology). This means a considerable amount of important Internet-of-Things data has to be abandoned since such data cannot be stored, transmitted, or processed. In this article, I describe our approaches for fully using the huge amount of multi-modal data.

Keywords: video coding, multi-modal signal handling, IoT data handling

1. Introduction

An increasing amount of data is becoming available with advances in technology and the expansion of the Internet of Things (IoT). The potential advantages of having such data available are described in this section.

1.1 Growth in IoT data

It has been reported [1] that the growth in storage devices is expected to increase 10 fold per decade, which is estimated to reach 100 zettabytes (ZB) (1 ZB = 10^{21} bytes) by 2030 and 1 yottabyte (YB) (1 YB = 10^{24} bytes) by 2040. However, the growth in information generated by sensors is expected to increase 40 fold per decade, which is estimated to reach 1 YB by 2030 and 40 YB by 2040 (**Fig. 1**).

Reflecting this rapid growth in IoT data generation and awareness of issues in their processing, a number of international standardization projects, such as Big Media [2], Internet of Media Things [3, 4], Network-Based Media Processing [5, 6], and Network Distributed Media Coding [7], have recently been initiated.

1.2 New opportunities via large-scale multi-modal data

Multi-modal IoT sensors are expected to be deployed around the world to obtain data of the entire earth (**Fig. 2**), which will make various applications possible. For example, in agriculture:

- Monitoring field crops based on super-wide-area video analysis
- Optimizing the timing and amount of fertilizer, water, and agrichemicals
- Maximizing crops and quality of the harvest based on precipitation, temperature, and moisture data.

Weather forecasting, disaster prevention, smart cities, surveillance/security, intelligent transport systems, logistics, infrastructure maintenance/inspection, tourism, etc., may benefit from the data of the entire earth.

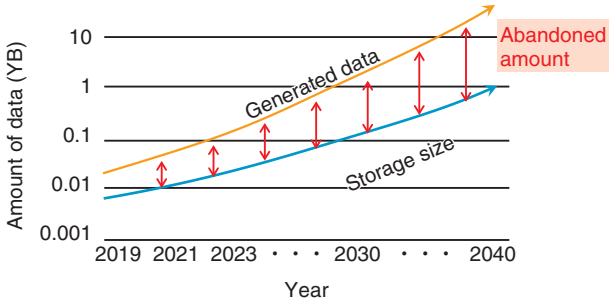

Fig. 1. Growth in generated data and storage size.

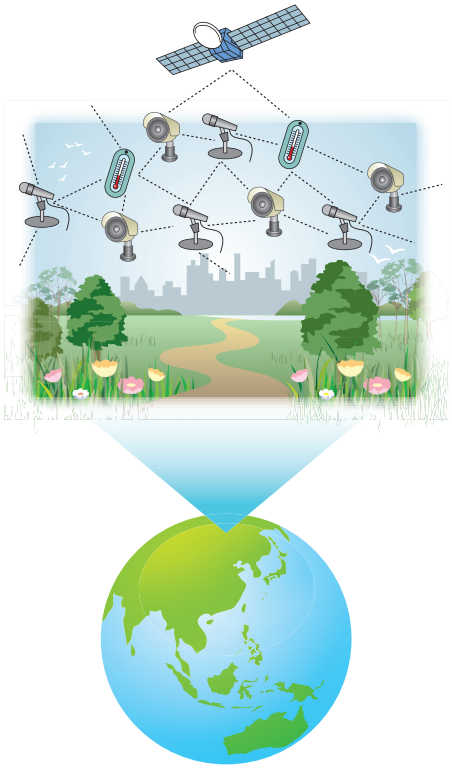

Fig. 2. Multi-modal sensor network over the entire earth.

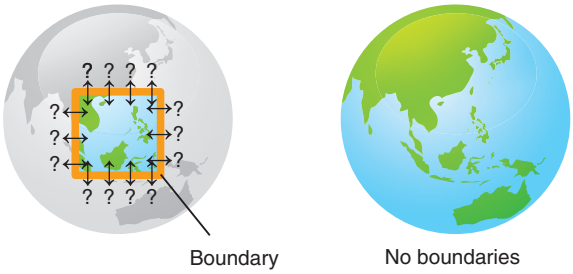

Fig. 3. Partial-earth sensing limitation (left) and entire-earth sensing advantage (right).

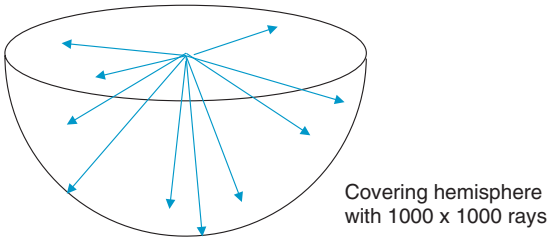

Fig. 4. Light-field-density image covering one hemisphere.

2. Omni-ambient data: data that cover the entire earth

Hereafter, the multi-modal data that cover the entire earth are referred to as omni-ambient data. Let us estimate the data size of omni-ambient data.

2.1 Number of sensors

If we were to observe the entire earth, we would have to deploy multi-modal sensors at every 10-m mesh point at, say, 10 m above the earth's surface. Since the surface area of the earth is 5.1×10^8 km^2, the number of sensors to cover the earth (S) would be 5.1×10^{12}. The rationale of covering the entire earth with sensors is as follows. If only a part of the earth is sensed, there will always be boundaries, which may cause uncertainty in data flux/interaction across them. If the entire world is covered, there would not be any boundaries; hence, no uncertainty would arise (**Fig. 3**).

2.2 Visual data

A hemisphere should be covered by about 1000 × 1000 light rays (**Fig. 4**) to capture the earth's light field. This resolution is about half the full high definition (HD) resolution (1920 × 1080). With this configuration, one ray covers a solid angle of 6.3×10^{-6} sr (steradian). If we assume that video is captured at 30 frames per second and each pixel has 8-bit red (R), green (G), and blue (B) information, the total amount of raw (uncompressed) video data from a single camera becomes 90 Mbit/s, but these data can be compressed with an existing video coding scheme (such as MPEG-H* or High Efficiency Video Coding (HEVC) [8]) to 1/350 its size, which is V (the total amount of compressed video data from a single camera) = 257 kbit/s. The total visual amount of omni-ambient data is S × V = 1.41 Ebit/s (1 exabyte (EB) = 10^{18} bytes), which is 41 YB per year. This amount is equivalent to all data that will be generated by 2040 (Fig. 1).

2.3 Audio data

Compact disc quality single-channel audio is assumed for capturing audio, which is 44.1 kHz × 16

* MPEG-H: International standards for video and audio compression developed by the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) Moving Picture Experts Group (MPEG).

bits = 88 kbit/s. Let it be compressed with a conventional audio coding scheme (such as Advanced Audio Coding) to 1/20 its size, which is 4.4 kbit/s. The total audio amount of omni-ambient data is S × A = 45 Pbit/s (1 petabyte (PB) = 10^{15} bytes), which is 1.4 YB per year.

2.4 Importance of visual data

The amount of audio data is about 30 times less than that of light data in omni-ambient data. The amounts of other data, such as depth, temperature, and moisture, may be comparable or even less. Therefore, the majority of the data is visual data. This is analogous to the visual data in Internet protocol (IP) traffic. It has been reported that mobile video traffic accounted for 59% of all worldwide mobile data traffic in 2017 and will be 79% by 2022 [9]. IP video traffic accounted for 75% of all worldwide IP traffic in 2017 and will be 82% by 2022 [10].

3. Challenges with organizing omni-ambient data

Some of the challenges with and opportunities for organizing omni-ambient data are discussed in this section. In addition to the example techniques described below (3.1–3.4), there should be many techniques to enable such organizing, such as pattern recognition, non-visual signal compression, large-scale archiving, ultrafast database construction, distributed computing, broadband IoT connection, and communication security.

3.1 Further compression via multi-modal synergic coding

Suppose there are two random variables X and Y. We then have the following equation

$$I(X; Y) = H(X) + H(Y) - H(X, Y),$$

where I(X; Y) is the mutual information of X and Y, H(X) and H(Y) are the marginal entropies of X and Y, and H(X, Y) is the joint entropy of X and Y. Since I(X; Y) is non-negative, the above equation can be rewritten as

$$H(X, Y) \geqq H(X) + H(Y).$$

This means that in terms of compression, it is always better to compress two (or even more) sources together. This encourages us to abandon conventional single-modal data coding for multi-modal data coding. One of the possibilities of efficient multi-modal data compression is depicted in **Fig. 5**. Conventionally, depth data and two-dimensional (2D) video data are encoded independently (sizes A and B, respectively, in figure). Since there is a non-negative relationship between depth and the captured image, 3D-shape-aware collaborative video coding may have less compressed data (size B'), which is smaller than B. This applies to not only two but also more than two modalities.

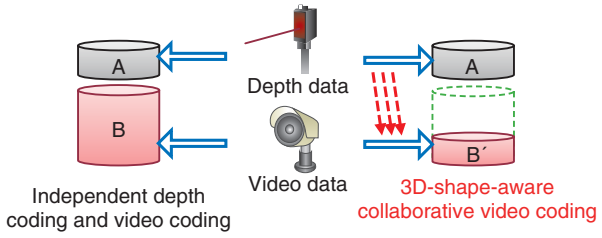

Fig. 5. Example of efficient multi-modal data compression.

3.2 Removing noise via real-entity mining

Sensed data are not always collected under ideal conditions, i.e., acquired data are deemed to contain noise, which is unpredictable and uncompressible by nature. Therefore, from the coding, storing, processing, and transmitting points of view, noise should be removed. Conventionally, acquired pixel values are targeted for encoding. However, the original objects should be behind the acquired pixel values. Therefore, being reminded of the existence of original objects (real-entity-oriented approach) may work better in signal processing, data compression, etc., than not taking care of it (observed-signal-oriented approach) (**Fig. 6**).

One such example is still-camera video coding. By processing such a video sequence and obtaining a real-entity image of the background, the video can be further compressed. Compared to the state-of-the-art video coding standard H.265/HEVC (reference software HEVC test model (HM)16.4) [8], bit-rate savings of 32.40% on average and 56.92% at maximum in terms of the Bjøntegaard Delta rate (BD rate) [11] were observed. It was also observed that the decoded video contains less camera noise than the original, which means the decoded video is even subjectively better than the original (**Fig. 7**). It also provides 21.17% faster encoding [12].

Noise is inevasible in any type of sensed data and it is uncompressible by nature. Therefore, noise reduction from data is crucial for compression efficiency. By tracking noisy rigid objects and temporally aligning

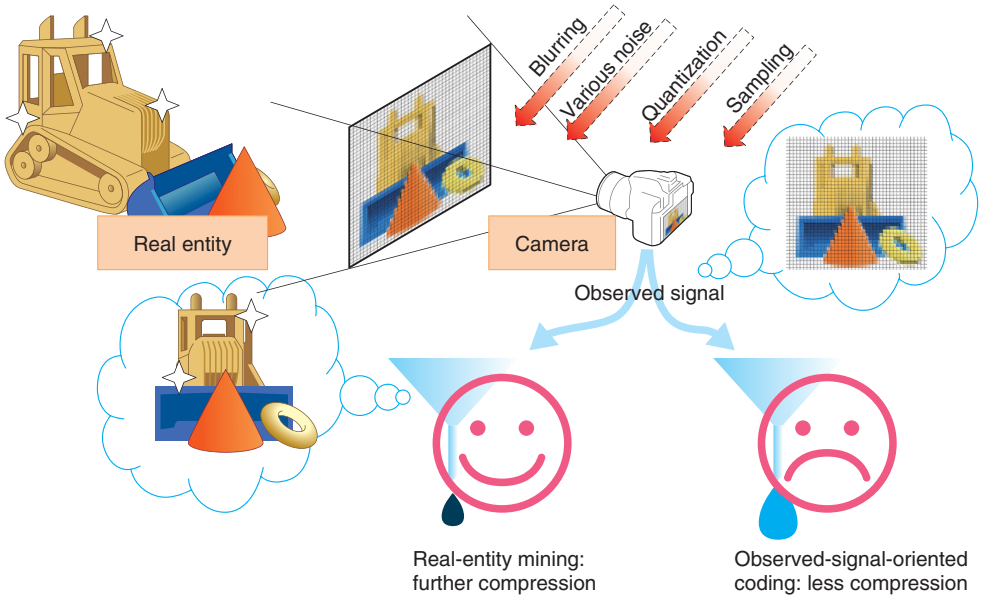

Fig. 6. Observed-signal-oriented coding vs. real-entity mining.

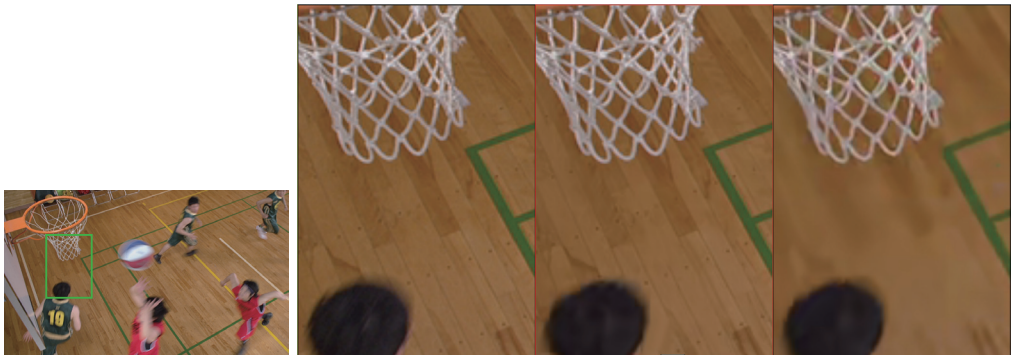

Fig. 7. Still-camera video-coding results based on real-entity background [12]. (a) Original image. Green box is magnified area, (b) original (noisy), (c) real-entity background based coding (Y-PSNR (peak signal-to-noise ratio) = 36.99 dB) 1/330 compression, and (d) H.265/HEVC (HM16.4) (Y-PSNR = 33.60 dB) 1/330 compression.

them and filtering out camera noise, coding efficiency greatly improves. The filtered image can be considered as the real entity of the rigid object and will be used as a reference frame for input video coding. In our experiments, this processing worked well for both objective and subjective metrics (**Fig. 8**). There was a 19–47% increase in the BD rate against H.265/HEVC (HM16.6) and 13–30% against preliminary Versatile Video Coding (VVC) [14] (JEM5.0). In terms of subjective video quality, the decoded video looked even better than the original video while only using 141–229 times fewer bits than JEM5.0 [13].

Another example is water-bottom video coding. Video content through the water surface is generally quite difficult to encode efficiently because of random movement and nonlinear deformation of objects seen through the moving water surface. By generating one additional frame from the input video sequence, which represents the real-entity image of bottom objects (**Fig. 9**), and additionally encoding the

Fig. 8. Rigid-object video-coding results based on real-entity mining approach [13]. (top-left) Former Versatile Video Coding (VVC); experimental model JEM5, rate = 1,857,505 bytes (noisy, similar to original), (top-right) real-entity mining based coding with super-low-rate mode, rate = 13,169 bytes (no noise and crisp), (bottom-left) H.265/HEVC (HM16), rate = 15,061 bytes (distorted), (bottom-right) JEM5, rate = 13,617 bytes (distorted).

(a) (b) (c) (d)

Fig. 9. Real-entity extraction example [15]. (a) Original water-bottom video frame under moving water surface (skewed), (b) temporal median filter result (blurred), (c) real-entity mining based (crisp and not skewed), (d) original water-bottom image under still water surface (ground truth).

frame and storing it as a long-term reference frame, a BD-rate reduction of 12–42% compared to the next-generation standard VVC under development (reference software VVC Test Model (VTM)1.1) [14] and 13–48% compared to VTM4.0, were achieved [15].

3.3 Indexing via machine-to-machine (M2M)-oriented image coding

Adding annotations to the obtained signal is essential for organizing omni-ambient data, and of course it should be done using machines (not by humans). Therefore, information coding that maximizes subjective quality may be less important than coding that maximizes machine recognition. One such M2M-oriented image-coding approach is that by Suzuki et al. [16]. The importance of this concept is reflected in the recent initiation of a new standardization project called Video Coding for Machines [17].

3.4 Reducing amount of original source data via compressed sensing

The above-estimated naïve data amount could become burdensome for initial-stage transmission. Sometimes it may be necessary to reduce the data rate at the sensor and (in return) restore the data by additional data processing. To achieve this, a compressed sensing technique [18] will be applied.

4. Conclusion

An overview was given of the rapidly increasing amount of data generated by ubiquitously deployed multi-modal devices, standardization trends to cope with such data, and possible applications by fully using the data that cover the entire world, i.e., omni-ambient data. The physical amount of such data was evaluated, and it was noted that visual data are dominant; therefore, efficient compression is crucial. Then the possibility of such compression by using mutual information among multi-modal signals was

discussed. How real-entity mining would help reduce noise, enable further compression, and improve subjective video quality was also discussed. We will continue investigating and tackling the challenges and taking advantage of the opportunities of this research, expanding the potential for more applications.

References

[1] H. Muraoka, "Yottabyte-scale Massive Data and Its Impact—Breaking Through a Wall of Ever-increasing Amounts of Information," Presentation material at Tohoku Forum for Creativity Symposium, pp. 11–12, Aug. 2017 (in Japanese), https://www.tfc.tohoku.ac.jp/wp-content/uploads/2017/08/2017EPP_03_Hiroaki_Muraoka.pdf

[2] MPEG Exploration Part 21, Big Media, https://mpeg.chiariglione.org/standards/exploration/big-media

[3] ISO/IEC 23093: "Internet of Media Things (IoMT)."

[4] ISO/IEC JTC 1/SC 29/WG 11, N 18910, "Technology under Consideration for IoMT," Oct. 2019.

[5] ISO/IEC 23090-8: "Network-Based Media Processing (NBMP)."

[6] ISO/IEC JTC 1/SC 29/WG 11, N 18848, "Technologies under Consideration for NBMP."

[7] MPEG Exploration Part 25, Network Distributed Media Coding, https://mpeg.chiariglione.org/standards/exploration/network-distributed-media-coding

[8] ISO/IEC 23008-2:2015: "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High efficiency video coding," 2015.

[9] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper, updated on Feb. 2019. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html

[10] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, updated on Feb. 2019. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[11] G. Bjøntegaard, "Calculation of Average PSNR Differences between RD-curves," ITU-T VCEG-M33, Apr. 2001.

[12] S. Takamura and A. Shimizu, "Simple and Efficient H.265/HEVC Coding of Fixed Camera Videos," Proc. of the 23rd IEEE International Conference on Image Processing (ICIP 2016), TP-L1.3, pp. 804–808, Phoenix, AZ, USA, Sept. 2016.

[13] S. Takamura and A. Shimizu, "Efficient Video Coding Using Rigid Object Tracking," Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017, FP-04.5, Kuala Lumpur, Malaysia, Dec. 2017.

[14] ISO/IEC 23090-3 MPEG-I Versatile Video Coding.

[15] S. Takamura and A. Shimizu, "Water-bottom Video Coding Based on Coding-oriented Reference Frame Generation," Proc. of 2019 IEEE International Conference on Visual Communications and Image Processing (VCIP 2019), Sydney, Australia, Dec. 2019.

[16] S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, "Image Pre-Transformation for Recognition-Aware Image Compression," Proc. of the 26th IEEE International Conference on Image Processing (ICIP 2019), TP.PG.7, Taipei, Taiwan, Sept. 2019.

[17] ISO/IEC JTC 1/SC 29/WG 11, N18772, "Draft Evaluation Framework for Video Coding for Machines," Oct. 2019.

[18] D. L. Donoho, "Compressed Sensing," IEEE Trans. Inf. Theory, Vol. 52, No. 4, pp. 1289–1306, Apr. 2006.

Seishi Takamura

Senior Distinguished Engineer, Signal Modeling Technology Group, Universe Data Handling Laboratory, NTT Media Intelligence Laboratories.

He received a B.E., M.E., and Ph.D. from the Department of Electronic Engineering, Faculty of Engineering, the University of Tokyo, in 1991, 1993, and 1996. His current research interests include efficient video coding and ultrahigh-quality video processing. He has fulfilled various duties in the research and academic community in current and prior roles, including serving as associate editor of the Institute of Electrical and Electronics Engineers (IEEE) Transactions on Circuits and Systems for Video Technology (2006–2014), editor-in-chief of the Institute of Image Information and Television Engineers (ITE), executive committee member of the IEEE Region 10 and Japan Council, and director-general of ITE affairs. He has also served as chair of ISO/IEC Joint Technical Committee (JTC) 1/Subcommittee (SC) 29 Japan National Body, Japan head of delegation of ISO/IEC JTC 1/SC 29, and as an international steering committee member of the Picture Coding Symposium. From 2005 to 2006, he was a visiting scientist at Stanford University, CA, USA.

He has received 51 academic awards including ITE Niwa-Takayanagi Awards (Best Paper in 2002, Achievement in 2017), the Information Processing Society of Japan (IPSJ) Nagao Special Researcher Award in 2006, Picture Coding Symposium of Japan (PCSJ) Frontier Awards in 2004, 2008, 2015, and 2018, the ITE Fujio Frontier Award in 2014, and the Telecommunications Advancement Foundation (TAF) Telecom System Technology Awards in 2004, 2008, and in 2015 with highest honors, the Institute of Electronics, Information and Communication Engineers (IEICE) 100-Year Memorial Best Paper Award in 2017, the Kenjiro Takayanagi Achievement Award in 2019, and Industrial Standardization Merit Award from Ministry of Economy, Trade and Industry of Japan in 2019.

He is an IEEE Fellow, a senior member of IEICE and IPSJ, and a member of Japan Mensa, the Society for Information Display, the Asia-Pacific Signal and Information Processing Association, and ITE.